Draft of November 6, 2017

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

Torrin M. Liddell and John K. Kruschke
Department of Psychological and Brain Sciences
Indiana University, Bloomington

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 new evidence that analyzing ordinal data as metric is problematic. We demonstrate that treating ordinal data as metric can yield low correct detection rates, distorted effect size estimates, and greatly inflated false alarm rates. Moreover, we demonstrate that the averaging of multiple ordinal items into a Likert scale does not solve these problems. We provide examples of real data in the contexts of the difference of two groups and simple linear regression. To solve these problems we use an ordered probit model with Bayesian estimation of parameters. The ordered probit model shows appropriate correct detection rates and false alarm rates, and produces accurate effect sizes estimates and response probabilities. Bayesian estimation of this ordinal model is straight forward, yields rich and accurate information, and has no need for auxiliary sampling assumptions. We conclude that ordinal data ought to be analyzed with ordinal models, and that Bayesian estimation is an excellent method for accomplishing that goal.

Keywords: Ordinal data; Likert scales; Bayesian data analysis; false alarm rate; effect size

---

Correspondence can be addressed to Torrin M. Liddell at `torrin.liddell@gmail.com` or to John K. Kruschke at `johnkruschke@gmail.com`, or to either at the Department of Psychological and Brain Sciences, Indiana University, 1101 E. 10th St., Bloomington IN 47405-7007.

NOTE: This table of contents appears only for the benefit of reviewers; it will not appear in the final version.

Contents

Ordinal data and approaches to modeling them **3**
Ordinal data are routinely analyzed with metric models 4
Metric and ordinal models 5
Formal specification of ordered probit model 7
Bayesian analysis for ordered probit model 9
Is treating ordinal data as metric innocuous? 12

Examples of errors when treating ordinal data as metric **13**
Two groups, single item 13
False rejection of null effect (over-estimation of effect) 13
Failure to detect non-zero effect (under-estimation of effect) 14
Monte Carlo simulations demonstrating inflated false alarm rate . . 16
Why the ordinal-as-metric analysis does not reflect the underlying parameters 18
Example real data 19
Case 1: Ordinal-as-metric model indicates a significant difference, Bayesian ordered probit model indicates no credible difference 20
Case 2: Ordinal-as-metric model indicates no significant difference, Bayesian ordered probit model indicates a credible difference 21
Two groups, average of multiple items (Likert "scale") 22
False rejection of null effect (over-estimation of effect) 22
Failure to detect non-zero effect (under-estimation of effect) 24
Summary: Taking average of multiple items does not solve the problem 24
Linear regression 28
Monte Carlo Simulation 28
Example real data 30
Quadratic regression 32

General discussion **34**
Detecting differences in variances 34
Other analysis methods 37
Metric Bayesian Analysis 37
Frequentist ordinal model 37
Conclusions 40

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 mis-estimated effect sizes and reduced correct detection rates (i.e., reduced statistical power), in the contexts of group comparison and linear regression. We also demonstrate greatly inflated false alarm rates (i.e., the rate of detecting an effect where none exists, Type I error) in group comparison, and we identify the conditions that yield these errors. In the context of quadratic trends, we again demonstrate the potential for false alarms and failure to detect trends in the data. We also provide evidence that multi-item Likert scales do not solve these problems, another novel result. In contrast, we show that an ordinal model yields more accurate interpretations. Although frequentist approaches to ordinal models are available for some applications, we instead favor a Bayesian approach because of its flexible applicability and its thorough, exact information.

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., Jamieson, 2004; Carifio and Perla, 2007; Vickers, 1999; Spranca et al., 1991; Clason and Dormody, 1994; Hui and Bateson, 1991; Feldman and Audretsch, 1999). The ubiquity of ordinal data is due in large part to the widespread use of Likert items and Likert scales (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. A Likert *"scale"*, as opposed to a single-item, is an average of related Likert items. For example, one item might ask for a rating of happiness, another item might ask for a rating of satisfaction, and third item might ask for a rating of displeasure (reverse scaled). The average of the three responses constitutes a Likert scale.

Ordinal data do not have metric information. Although the response options might be numerically labelled 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 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.

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 1,063 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, 2016). 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 the vast majority of researchers 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.

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 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 and 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.

Figure 1 illustrates the difference between normal (metric) and thresholded cumulative normal (ordered probit) models. Suppose we have data from a Likert response scale, with possible ordinal values labelled '1', '2', '3', '4', and '5'. According to the metric model, shown in the upper panel of Figure 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-sized intervals between the ordinal responses, and maps the ordinal responses to corresponding metric scale values. The upper panel of Figure 1 shows the normal probability densities at these values. When modeling the data, the analyst estimates the values of mean and standard-deviation parameters (μ and σ) in the normal distribution.

On the other hand, according to the ordered-probit model shown in the lower panel of Figure 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

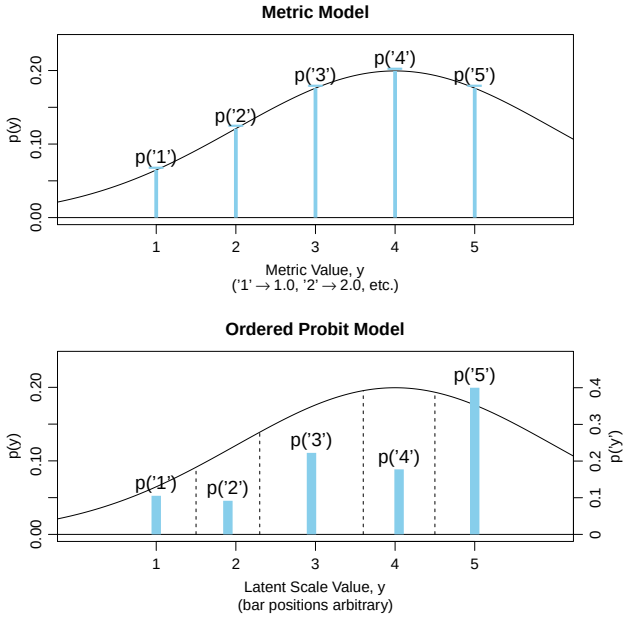

Figure 1. Upper panel: Metric model of ordinal data. Ordinal values are mapped to corresponding metric values on the horizontal axis, with '1'→1.0, '2'→2.0, and so forth. The probability of each value is the normal density as shown by the heights of the lines, with numerical scale indicated on the left vertical axis. *Lower panel: Ordered probit model.* A latent scale on the horizontal axis is divided into subintervals with thresholds marked by dashed lines. The cumulative normal probability in the subintervals is the probability of the ordinal values. The cumulative normal probability within each interval is indicated by the height of the corresponding bar, with numerical scale indicated on the *right* vertical axis.

divided at some thresholds for mapping into discrete responses. The lower panel of Figure 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 lower panel of Figure 1 plots the cumulative-normal probabilities as bar heights with scale on the right vertical axis. 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 threshold parameters. This sort of model is sometimes called an "ordered probit," "ordinal probit," or "thresholded cumulative normal" model (e.g., Winship and Mare, 1984; McKelvey and Zavoina, 1975). This model is used as one of the standard models of ordinal data in both frequentist and Bayesian analysis (e.g., Albert and Chib, 1997; Lynch, 2007; Kruschke, 2015).

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.

Formal specification of ordered probit model

The ordered-probit model assumes that ordinal responses are generated from a latent metric variable. The metric value is normally distributed with mean μ and standard deviation σ. Thresholds θ_j on the metric continuum determine the corresponding ordinal response. The probability of ordinal response k is

$$p(y{=}k|\mu,\sigma,\{\theta_j\}) = \Phi\left((\theta_k-\mu)/\sigma\right) - \Phi\left((\theta_{k-1}-\mu)/\sigma\right) \tag{1}$$

where Φ is the standardized cumulative normal function. Equation 1 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, $k = 1$, the threshold θ_{k-1} is negative infinity, and for the highest level, $k = K$, the threshold θ_K is positive infinity. The probabilities in the lower panel of Figure 1 were computed using Equation 1. To indicate that the y values are distributed according to the probabilities in Equation 1, we use the following notation:

$$y \sim \text{cat}\left(p(y{=}1),...,p(y{=}K)\right) \tag{2}$$

where K is the number of possible ordinal responses. In Equation 2, the symbol "$\sim$" is read "is distributed as", and the term cat(.) indicates a *categorical* probability distribution with category probabilities specified by the components within the parentheses.

When there are multiple groups, each group is modeled with its own mean and standard deviation, denoted by superscript index g, set in square brackets so it is not misinterpreted as a power:

$$p(y{=}k|\mu^{[g]},\sigma^{[g]},\{\theta_k\}) = \Phi\left((\theta_k-\mu^{[g]})/\sigma^{[g]}\right) - \Phi\left((\theta_{k-1}-\mu^{[g]})/\sigma^{[g]}\right) \tag{3}$$

All groups are assumed to use the same thresholds because thresholds represent properties of the response scale, and the response procedure is the same across groups. Therefore the θ values in Equation 3 do not have group superscripts.

The parameter values in this 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 '7' we set $\theta_1 = 1.5$ and $\theta_6 = 6.5$. 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 arbitrarily sets $\sigma = 1$ and $\mu_1 = 0$ (e.g., Winship and Mare, 1984; McKelvey and Zavoina, 1975), 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 and 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.

We also explore applications in which ordinal data are predicted by continuous covariates. For example, we might predict people's ordinal rating of subjective happiness as a function of their annual income. In these situations, the ordinal data are modeled as in Equation 1, but with the underlying mean being a function of the predictor. For example, an analyst could consider a quadratic trend in the latent mean as a function of the predictor, x:

$$\mu = \beta_0 + \beta_1 x + \beta_2 x^2 \tag{4}$$

This model therefore involves the parameters β_0, β_1, β_2, σ, and $\{\theta_k\}$.

Finally, we will also consider applications to Likert *scales*, wherein users compute the arithmetic mean of several related Likert items and treat the resulting average as the dependent measure. Suppose there are several related Likert items indexed by i. As a simplification for purposes of illustration, we assume that there is a single latent variable that underlies all the items simultaneously. (A more elaborate model would use a distinct latent variable for every item, with the latent dimensions strongly correlated across responders. Our simplified model essentially assumes perfect correlation across the latent dimensions.) As before, each group is assumed to be normally distributed on the latent variable, with mean $\mu^{[g]}$ and stan-

dard deviation $\sigma^{[g]}$, but the latent value is mapped to each item by an item-specific linear function, involving intercept $\beta_0^{[i]}$ and slope $\beta_1^{[i]}$. Formally, the probability of ordinal value $y^{[i,g]}$ on item i in group g is specified as:

$$\begin{aligned} & p\left(y^{[i,g]}=k \,\middle|\, \mu^{[g]}, \sigma^{[g]}, \{\theta_j^{[i]}\}, \beta_0^{[i]}, \beta_1^{[i]}\right) \\ & = \Phi\left(\left((\theta_k^{[i]} - \beta_0^{[i]})/\beta_1^{[i]} - \mu^{[g]}\right) \Big/ \sigma^{[g]}\right) - \Phi\left(\left((\theta_{k-1}^{[i]} - \beta_0^{[i]})/\beta_1^{[i]} - \mu^{[g]}\right) \Big/ \sigma^{[g]}\right) \end{aligned} \tag{5}$$

Notice that when $\beta_0^{[i]} = 0$ and $\beta_1^{[i]} = 1$ then Equation 5 reduces to Equation 3. As before, because of indeterminacies in the algebraic formulation, we fix the lowest and highest thresholds of every item, with $\theta_1^{[i]} = 1.5$ and $\theta_{K-1}^{[i]} = K - 0.5$ for all i, where K is the maximum ordinal value for item i. Moreover, we fix the intercept and slope of item 1 at $\beta_0^{[1]} = 0$ and $\beta_1^{[1]} = 1$. Notice in this case of Likert *scales* that the data being modeled by the metric model are the averaged Likert ratings, whereas the data being modeled by the ordered-probit model are the ratings of the individual items, described by a common latent factor.

Bayesian analysis for ordered probit model. There are various goals when modeling data. First and foremost, we would like to estimate the parameter values of the model, because the parameter values are meaningful. For example, if we have two groups of ordinal data, we are interested in the magnitude of the difference of means, $\mu_1 - \mu_2$, because that describes the difference in the central tendencies of the groups. Another goal is to accurately describe the data distribution, which is to say that the data probabilities should be accurately mimicked by the model, because if the data distribution is not well described then the parameter values may be meaningless. For instance, the mean of a bimodal distribution does not indicate where most of the data fall. A third goal is that we may want to make decisions about hypothetical null values of the parameters. For example, we might ask, Is a difference of means non-zero?

The goals can be approached within frequentist or Bayesian frameworks. The primary concern in a frequentist framework is controlling error rates when making decisions about null values. This is achieved by computing p values and rejecting a null value only when $p < .05$. The primary concern in a Bayesian framework is ascertaining the credibility of candidate parameter values conditional on the observed data. This is achieved by re-allocating probability across parameter space, conditional on the observed data. Both frequentist and Bayesian approaches provide estimates of parameters and measures of uncertainty in those estimates. In a frequentist approach, the estimate is provided by the *maximum likelihood estimate*

(MLE), and the uncertainty is provided by a *confidence interval* (CI). In a Bayesian approach, the estimate is provided by the an entire posterior probability distribution on the joint parameter space, which can be summarized by the mode of each parameters marginal distribution and its *highest density interval* (HDI). A tutorial comparison of frequentist and Bayesian approaches is provided by Kruschke and Liddell (2017b).

We will use Bayesian methods to estimate the parameters of the model. Relative to a frequentist approach, the Bayesian approach has many advantages:

1. Bayesian analysis avoids dealing with p-values. We will not repeat here the many arguments against the use of p-values. For more details see, e.g., Kruschke (2013); Kruschke and Liddell (2017b); Wagenmakers (2007).

2. Instead of a point estimate, Bayesian analysis produces a full posterior distribution over the joint parameter space. In the traditional approach a maximum likelihood estimate (MLE) for the parameters of the model is found. Then, to produce a p value and confidence interval for a given parameter of interest, auxiliary assumptions are needed to create a sampling distribution from counterfactual hypotheses. In many cases (including ordered probit models), the resulting p values and confidence intervals can only approximated for asymptotically large sample size, and typically the results underestimate the p value and width of the confidence interval (e.g., Wickens, 1982). In the Bayesian approach, a full posterior distribution on the joint parameter space is a natural product of the analysis, and it is always accurate regardless of sample size. Because the posterior distribution reveals all parameters simultaneously, many questions of interest can be evaluated. For instance, even if the primary focus is the difference of means between two groups, the difference of standard deviations is easily investigated with no further analysis, merely by querying the joint posterior distribution.

If binary decisions about null values are required, the posterior distribution can be examined to consider the relation of the most credible parameter values to the null value. The 95% highest density interval (HDI) contains parameter values that are more credible (i.e., have higher probability density) than parameter values outside the interval, such that the total probability of points in the interval is 95%. The 95% HDI is compared to a region of practical equivalence (ROPE) around the null value, which defines values considered to be practically equivalent to the null for practical purposes. For example, when applied to effect size in comparison of means, a ROPE from -0.1 to $+0.1$ might be used because it spans only half of a conventionally "small" effect size (Cohen, 1988). If the 95% HDI falls outside of the ROPE, the null value is rejected, because the 95% most credible values are all not practically equivalent to the null value. If the 95% HDI falls entirely within the ROPE, the null value is accepted for practical purposes, because the 95% most cred-

ible values are all practically equivalent to the null value. More information about this decision rule can be found in Kruschke and Liddell (2017b,a), and Kruschke (2011a,b, 2015, Ch. 12).

3. Bayesian analysis as instantiated in modern Bayesian software is easily adapted to variations on the typical ordinal scenario. For example, though the model we present here uses a normal distribution to describe noise on the underlying metric variable, it is easy to use a different distribution (logistic, t distribution, etc.).

This discussion of the general advantages of Bayesian analysis is necessarily brief. For more complete discussions of the benefits of the Bayesian approach, see Kruschke (2015). Importantly, we do not intend to advance the position that Bayesian estimation of the parameters of this model is the only method to solve the problems inherent in the ordinal-as-metric approach. A frequentist implementation of the model presented here could also address some of the problems of treating ordinal data as metric, but we prefer Bayesian estimation for the reasons described above as well as in the works cited here. We discuss issues regarding a frequentist implementation of this model in the general discussion at the end of the article.

In our Bayesian analyses, we use prior distributions that are broad and have minimal influence on the results. The thresholds are given wide normal priors:

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

where the second argument in norm(,) 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 but with a large standard deviation. Although the priors specified in Equation 6 allow inverted thresholds (e.g., $\theta_k < \theta_{k-1}$), the model as implemented in computer programs gives zero probability to data from inverted thresholds and therefore they never actually occur.

The prior on $\mu^{[g]}$ is a normal distribution that is broad on the scale of the data, with a mean set to the midpoint of the ordinal scale ($(K + 1)/2$) and with a standard deviation equal to the number of ordinal categories (K):

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

As mentioned before, the second argument in norm(,) is its standard deviation, not its variance or precision.

The prior on $\sigma^{[g]}$ is a broad uniform distribution, from a thousandth of the number of categories to 1000 times the number of categories:

$$\sigma^{[g]} \sim \text{unif}\left(\frac{K}{1000}, 1000K\right) \tag{8}$$

This extremely broad prior distribution is designed to have minimal influence on the posterior distribution.

The priors on the trend coefficients (Eqn. 4) are also chosen to be broad on the scale of the data. The priors are set to be broad so they have minimal influence on the results. The intercept β_0 has a prior centered at the middle of the ordinal scale, with a large standard deviation (as in Equation 7):

$$\beta_0 \sim \text{norm}\left((K+1)/2, K\right) \tag{9}$$

To make it easy to specify generic priors and to make the MCMC process more efficient, the predictor x is standardized. The prior for the slope β_1 is centered at 0, with a large standard deviation (when x is standardized):

$$\beta_1 \sim \text{norm}\left(0, K\right) \tag{10}$$

The intercept and slope are transformed back to the original scale.

The coefficients for the item response scales in Equation 5 are also given broad priors, with $\beta_0^{[i]} \sim dnorm(\frac{K^{[i]}}{2}, 10)$ and $\beta_1^{[i]} \sim dgamma(1.393, 0.393)$ which has a mode of 1.0 and standard deviation of 3.0.

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, 2013). For more implementation details, see Kruschke (2015).

Is treating ordinal data as metric innocuous?

Aside from not having equal intervals 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 be positioned at equal intervals, are often strongly skewed, heavy-tailed or thin-tailed, or multi-modal.

Thus, treating ordinal data as if they were normally-distributed equal-interval 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. We show that analyzing ordinal data with ordinal models is straightforward, and therefore analysts *should* use ordinal models for ordinal data. Moreover, while the models introduced here are not novel, we do demonstrate the benefits of using Bayesian estimation in the context of these ordinal models, and in the discussion we directly compare the application of frequentist alternatives to the Bayesian methods presented through-

out.

Examples of errors when treating ordinal data as metric

The examples of this section illustrate a variety of 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 with 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 false alarms (Type I errors) when the true parameters are null, failures to detect non-zero effects (Type II errors), and poor descriptions of data probabilities. The examples include single Likert items for two groups, Likert scale averages of multiple items for two groups, linear regression, and quadratic trend analysis.

Two groups, single item

False rejection of null effect (over-estimation of effect). The data were generated by starting with random metric values from an underlying normal distribution, and then assigning the metric values to discrete ordinal values according to thresholds on the metric scale, as was described in Figure 1. In our first example, thresholds were placed at 1.5, 2.5, 3.5, ..., and 6.5 to yield 7 intervals. The normal distribution for group 1 had a mean of 2.5 and a standard deviation of 3.0, while group 2 also had a mean of 2.5 but a standard deviation of 1.5. Because the means of the groups are equal, the true effect size is zero. Data were generated by randomly sampling $N = 301$ from the normal distribution of group 1, and $N = 302$ from the normal distribution of group 2.

Figure 2 shows histograms of the data and the results of the analyses. The upper panel of Figure 2 shows the results when treating the ordinal data as if they were metric. The results from Welch's t test (Welch, 1938), which allows for different variances across groups, are displayed in the title of the plot. As can be seen, the analysis when treating the data as metric indicates a significantly non-zero effect, that is, a false alarm.

The lower panel of Figure 2 shows the results when analyzing the ordinal data with an ordered probit model. The title of the plot indicates the estimated effect size along with the Bayesian credible interval (the 95% highest density interval, HDI). As can be seen, the ordered-probit analysis recovers the true generating effect size very accurately.

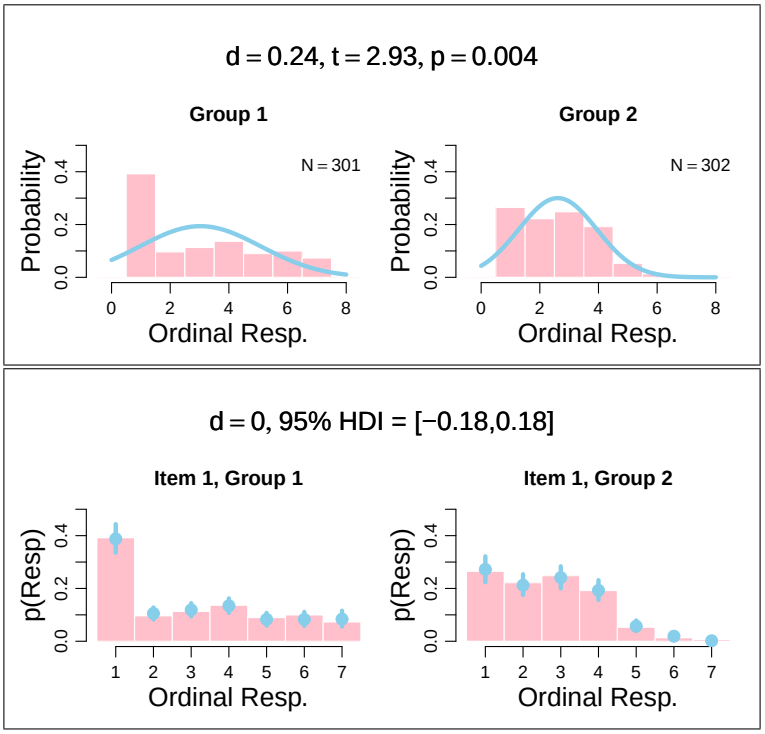

Figure 2. Ordinal data from a single item for two groups, displayed in histograms (same data in upper and lower). The generating parameters had $d = 0.0$. *Upper: Treated as metric.* The title indicates estimated effect size (Cohen's d) and result of Welch's t test. Best fitting normal curves are superimposed on data histogram. *Lower: Treated as ordinal.* The title indicates estimated effect size (Cohen's d) and the 95% highest density interval (HDI) of the effect size. Predicted response probabilities are shown as dots superiposed on data histogram; short vertical bars through the dots indicate 95% HDI of predicted probability.

The upper and lower panels of Figure 2 also show the model predictions superimposed on the data. The metric model finds the best-fitting continuous normal distributions, which are shown in the upper panel of Figure 2. Clearly the data are not well described by normal distributions. The ordered-probit model, on the other hand, predicts discrete data probabilities as shown in the lower panel of Figure 2. Clearly the data are very well described by the ordered-probit model.

Failure to detect non-zero effect (under-estimation of effect). For our next example, data were generated in the same way as the previous example, with thresholds placed at 1.5, 2.5, 3.5, ..., and 6.5 to yield 7 intervals. The normal distribution for group 1 had a mean of 3.0 and a standard deviation of 1.5, while group 2 had a mean of 2.5 and a standard deviation of 3.0. The true effect size is therefore $d = (\mu_1 - \mu_2)/\sqrt{(\sigma_1^2 + \sigma_2^2)/2} = 0.21$.

Figure 3 shows histograms of the data and the results of the analyses. The

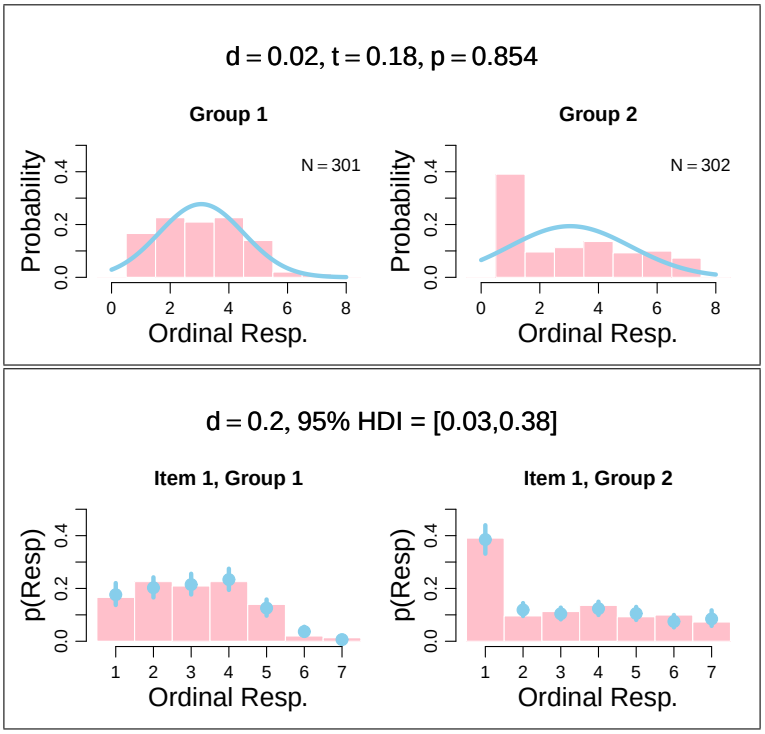

Figure 3. Ordinal data from a single item for two groups, displayed in histograms (same data in upper and lower). The generating parameters had $d = 0.21$. *Upper: Treated as metric.* The title indicates estimated effect size (Cohen's d) and result of Welch's t test. Best fitting normal curves are superimposed on data histogram. *Lower: Treated as ordinal.* The title indicates estimated effect size (Cohen's d) and the 95% highest density interval (HDI) of the effect size. Predicted response probabilities are shown as dots superiposed on data histogram; short vertical bars through the dots indicate 95% HDI of predicted probability.

upper panel of Figure 3 shows the results when treating the ordinal data as if they were metric. The results from Welch's t test (Welch, 1938), which allows for different variances across groups, are displayed in the title of the plot. As can be seen, the analysis when treating the data as metric indicates no significant effect, that is, a failure to detect the true difference between groups.

The lower panel of Figure 3 shows the results when analyzing the ordinal data with an ordered probit model. The title of the plot indicates the estimated effect size along with the Bayesian credible interval (the 95% highest density interval, HDI). As can be seen, the ordered-probit analysis recovers the true generating effect size very accurately.

The upper and lower panels of Figure 3 also show the model predictions superimposed on the data. The metric model finds the best-fitting continuous normal distributions, which are shown in the upper panel of Figure 3. Clearly the data are

not well described by normal distributions. The ordered-probit model, on the other hand, predicts discrete data probabilities as shown in the lower panel of Figure 3. Clearly the data are very well described by the ordered-probit model.

Monte Carlo simulations demonstrating inflated false alarm rate. The above examples demonstrate that cases exist where ordinal-as-metric analyses can yield incorrect conclusions. In this section, we apply the same general methodology to a large scale simulation demonstrating that not only does such potential for error exist, but that the rate of false alarms can be much larger than the rate specified by the nominal alpha level. We do so in the context of the following fictitious example: A school administrator is interested in comparing student satisfaction across two high schools. A Likert-item survey is given to 100 students in each school, with a single response ranging from 1 = "not at all satisfied" to 7 = "very satisfied." We want to find out how different the schools' ratings are. The simulated data were generated as normally distributed underlying metric values, with the mean of both schools at 5.0 on the underlying scale. However, one school had higher variability than the other school, with one standard deviation at $\sigma = 4$ and the other at $\sigma = 2$. These fictitious data have relatively high variability because it creates a clear illustration, but realistic data can exhibit similar distributions. The underlying metric values were transformed into ordinal responses according to six thresholds set at 1.5, 2.5, 3.5, 4.5, 5.5, and 6.5. Thus, if the underlying metric satisfaction was less than 1.5 then the ordinal response was 1, and if the underlying metric satisfaction was between 1.5 and 2.5 then the ordinal response was 2, and so on.

We generated 10,000 fictitious data sets with these parameters, and to each applied a Welch's t test for a difference of means, as well as a Bayesian ordered probit model. For the t test, we tallied the proportion of data sets that produced $p < 0.05$ to determine the false alarm rate, because the true difference of means is known to be zero. The simulation showed an actual false alarm rate of 0.182, far above the 0.05 anticipated by the t test. Thus, we have established that the ordinal-as-metric approach has inflated false alarm rates under the conditions described above.

Using the same generated data, we also assessed the false alarm rate of the Bayesian ordered probit model. To do so, we computed the difference between $\mu^{[1]}$ and $\mu^{[2]}$ at every step in the MCMC chain, which yields a complete posterior distribution on the difference, $\mu^{[1]} - \mu^{[2]}$. We computed the 95% HDI of the difference and compared it to a ROPE around zero. To make our false alarms rates strictly comparable to those computed by NHST, we used a ROPE of zero width. In other words, because NHST rejects a null value when it falls outside the 95% confidence interval, we will reject a null value when it falls outside the 95% HDI. This use of a zero-width ROPE is not recommended in general under this approach to assessing

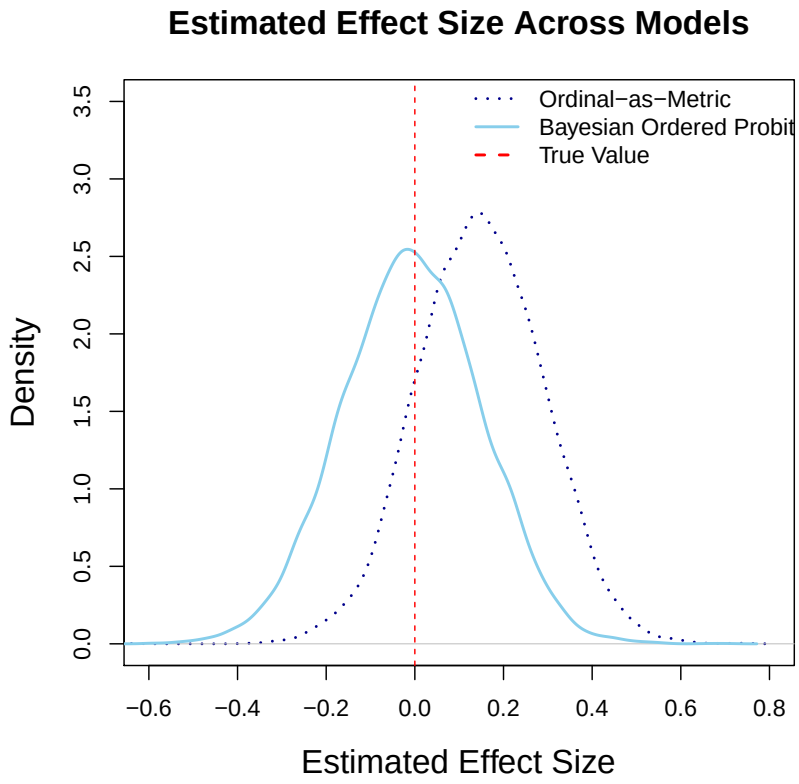

Figure 4. Sampling distributions of the estimated estimated effect size from two models. The generating mean of 0.0 is marked by a vertical dashed line (red). Across the 10,000 simulated data sets, the mean estimated effect size for the ordinal-as-metric model was 0.148, which badly mis-estimates the true generating value of 0.0. By contrast, the estimated effect size from the Bayesian ordered probit model had a mean of −0.007, which is essentially equal to the true generating value of 0.0.

null values, because it implies a null value can never be accepted. We stress that the use of a zero-width ROPE is only for comparison with NHST; in this particular method of assessing null values, a ROPE with non-zero width should be used in usual circumstances. Notice that using a non-zero width ROPE would reduce the false alarm rate.

When the Bayesian ordered probit model was applied to these data we observed a false alarm rate of only 0.047, which is essentially the 5% false alarm rate expected by using a zero-width ROPE and 95% HDI. If a non-zero ROPE were used (as is recommended for real applications) the false alarm rate would be further reduced.

The benefits of the Bayesian approach do not end with reduced false alarm rates: The effect size is also recaptured in an unbiased fashion. Figure 4 shows

a density plot of the sampling distribution of the estimated effect sizes from the two models. Effect size is equal to the difference of means divided by the pooled standard deviation (Cohen, 1988). In the ordinal-as-metric model, the effect size is computed using the MLE of the means and standard deviations (treating the ordinal data as though they were metric). In the Bayesian model, for every simulated sample of data the effect size is computed at every step in the MCMC chain as $(\mu_1 - \mu_2)/\sqrt{(\sigma_1^2 + \sigma_2^2)/2}$ and the mode of the posterior distribution is used as the estimated effect size for that sample. As can be seen in Figure 4, the Bayesian ordered probit model correctly recovers the generating effect size of 0.0, whereas the ordinal-as-metric approach badly mis-estimates the effect size.

Why the ordinal-as-metric analysis does not reflect the underlying parameters. The problems yielded in the ordinal-as-metric approach stem from one of two types of asymmetric ordinal "bin" sizes. In the simulation described in the previous section, this asymmetry results from a central tendency of the underlying metric values that is closer to one end of the ordinal scale or the other. This asymmetry exploits a key loss of information in ordinal responses: Any underlying metric value greater than the highest threshold only produces the highest ordinal response no matter how extreme the underlying metric value is. The analogous censoring occurs on the low end of the ordinal scale. If the mean of the underlying metric data is closer to the highest threshold than to the lowest threshold, then there will be more underlying metric values above the highest threshold than there are metric values below the lowest threshold. Consequently, the arithmetic mean of the ordinal values will be less than the mean of the underlying metric values. In general, an underlying metric distribution that falls asymmetrically on the thresholds will tend to have its mean estimated to be less extreme than it really is. We call this condition a floor/ceiling asymmetry. This floor/ceiling asymmetry causes one of the end point bins to be larger than another in terms of the amount of the distribution that it contains.

The second way for the asymmetric bin size to occur, is for some of the interior bins to be unequally sized in a manner that is asymmetric around the mean of the underlying distribution. For instance, if a there is a 5-point ordinal scale with thresholds at 1.0, 1.5, 2.0, and 4.5 and an underlying metric mean of 3.0, this asymmetry of the thresholds would cause distortions in the mean estimate. By contrast, a 5-point ordinal scale with thresholds at 1.5, 2.0, 4.0, and 4.5 and again an underlying metric mean of 3.0 would not cause distortions of the mean estimate despite the large middle bin, because the distortion is symmetric around the underlying mean.

If either of these conditions are present, the ordinal-as-metric approach will produce a biased estimate of the underlying central tendency. However, this bias will

tend to affect all measured groups equally, if the groups are identical in all ways. That is, this problem will not cause the mis-estimated effect size present in the above simulation on its own; it requires a difference in variance. When the variance is different across the two groups, the estimate of the mean is distorted to different degrees across the two groups. For example, suppose there is a 5-point ordinal response with thresholds at 1.5, 2.5, 3.5, and 4.5, with the mean of the underlying normal distribution set $\mu = 4.0$. When the underlying standard deviation is $\sigma = 2.0$, the expected value of the ordinal responses is 3.67, and when the underlying standard deviation is $\sigma = 4.0$, the expected value of the ordinal responses is 3.38. For both standard deviations, the ordinal mean is less extreme than the true generating mean, but the underestimation is worse for the larger variance. Thus, even though the true generating means are equal, the expected ordinal means are different.

These conditions can cause multiple types of problems when treating ordinal data as if they were metric. These problems include increased false-alarm rates, low correct-detection rates, and distorted effect-size estimates. Why did previous research not find inflated false alarm rates?(e.g., Havlicek and Peterson, 1976; Glass et al., 1972; Hsu and Feldt, 1969; Heeren and D'Agostino, 1987) To our knowledge, previous work has not investigated the combination of floor/ceiling asymmetry or the more general case of unequal bin width and unequal variances across groups. Without the unequal variances across groups, the mean estimates will be distorted equally in the two groups, and will not cause inflation in false alarm rates. However, this is not the case for all analyses and all contexts, and we will show later that in the context of quadratic trend analysis that both Type I and Type II errors can occur even with equal variance across all values of the predictor.

To summarize, unequal bin width relative to the underlying metric distribution causes a distortion in the observed mean. If the underlying variances differ across groups being compared, this distortion will have a different magnitude across the two groups, which will distort the observed difference of means (or lack thereof) between the two groups

Example real data. In this final section on two-group single-item comparisons, we present real data that suggest the type of errors demonstrated in the above simulated examples. The data presented here come from the U.S. Department of Education Educational Longitudinal Study of 2002, available at `http://nces.ed.gov/surveys/els2002/avail_data.asp` (for more details, see Rogers et al., 2004). This is a large data set consisting of a variety of measures relevant to education, from numerous schools across the United States. We used a single Likert item (discussed earlier in our fictitious example) presented to individual students in the context of a larger survey, namely Question 20F of the student questionnaire: "Teachers are

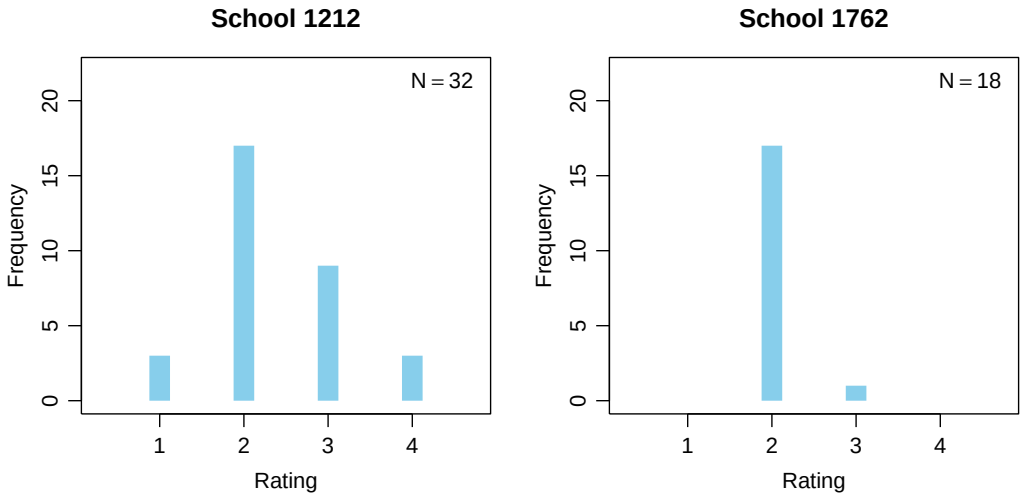

Figure 5. Histograms of the responses to question 20F for schools 1212 and 1762. Notice that school 1762 has much less variation than school 1212. There is a significant difference of means if the ordinal responses are treated as metric values in a t test, but not according to the Bayesian ordered probit model.

interested in the students." The question was rated on a 1 to 4 scale from "Strongly Agree" to "Strongly Disagree."

Using an automated program, we searched a subsection of schools for pairs where the Bayesian ordered probit model and the t test disagreed on their conclusions, and we present a selection of these below. We recognize that the existence of these cases "in the wild" does not imply that any particular data set, or even any sizable proportion of data sets, will be examples of these cases. However, we contend that a researcher cannot know if her ordinal data set is one that is especially problematic for the application of metric methods without first applying a more generally appropriate analysis like the Bayesian analysis presented here. And of course, avoiding the consequences of these more severe cases is just one benefit among many benefits of Bayesian estimation.

Case 1: Ordinal-as-metric model indicates a significant difference, Bayesian ordered probit model indicates no credible difference. We compared the scores of two schools (schools 1212 and 1762) on question 20F. Figure 5 shows histograms summarizing the two data sets. We applied both Welch's t test and the Bayesian ordered probit model. We found a significant difference between groups when using Welch's t test, $t(39.7) = 2.12, p = 0.040$, 95% Confidence Interval from 0.015 to 0.624. Conversely, we found no credible difference of means when applying the Bayesian ordered probit model, mean difference = 0.152, 95% HDI from -0.324 to

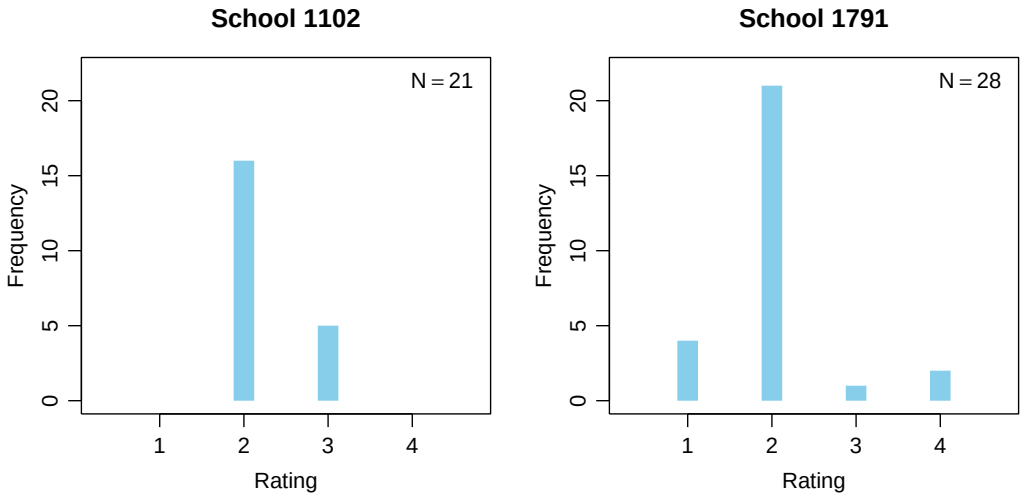

Figure 6. Histograms of the responses to question 20F for schools 1102 and 1791. Notice that school 1102 has much less variation than school 1791.

0.618. This data set is a clear example of the unequal bin width effect: both groups peak in similar locations (approximately 2 in this case) but one group (school 1212) has considerably higher variability than the other. The higher variability is expressed disproportionately in the larger direction due to the censoring effect of the ordinal scale, leading to an apparently higher mean. The Bayesian ordered probit model interprets the apparent difference between group central tendency as a difference not of means, but of standard deviation ($\sigma_1 - \sigma_2 = 0.539$, 95% HDI from 0.199 to 0.917).

Case 2: Ordinal-as-metric model indicates no significant difference, Bayesian ordered probit model indicates a credible difference. We compared the scores of two schools (schools 1102 and 1791) again on question 20F. Figure 6 shows histograms of the data. Just as in the previous case, Welch's t test and the Bayesian ordered probit model were applied. As this case is primarily intended to demonstrate the detection abilities of the Bayesian method, we used a ROPE on estimated effect size from -0.1 to 0.1. Opposite of Case 1, this data set shows no significant difference of means when the data are treated as metric in NHST: $t(45.813) = 1.25$, $p = 0.22$, 95% confidence interval from -0.124 to 0.528 when using Welch's t test. But when using the Bayesian analysis, there is a non-zero effect size: mean effect size = 1.049, 95% HDI from 0.126 to 1.988 (note that the HDI excludes not just zero, but the ROPE from -0.1 to 0.1). In this case, it appears that the asymmetric bin width effect obscures the magnitude of the underlying difference of means by causing an overestimate of the ordinal-as-metric mean for school 1791. On the other hand, the

results of the Bayesian analysis indicate that school 1791 has a lower mean but a higher variance. This difference of standard deviations is estimated by the Bayesian ordered probit model to have a mean difference of $\sigma_1 - \sigma_2 = 0.537$ with 95% HDI from 0.098 to 0.982.

Two groups, average of multiple items (Likert "scale")

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 and call the resulting value a Likert "*scale*". 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 put into standard metric analyses such as the t test.

To generate simulated data for four items, we sampled from a four-dimensional multivariate normal distribution that had correlations of 0.8 for all pairwise combinations of dimensions. The thresholds for every dimension were the same as the previous example, namely 1.5, ..., 6.5. The means and standard deviations were the same on every dimension, each group had its own mean and standard deviation just as in the previous examples.

When analyzing the data with an ordered-probit model, we used the model described back in Equation 5. In other words, the model assumed a single latent dimension instead of four correlated latent dimensions.

False rejection of null effect (over-estimation of effect). The true generating normal distribution for group 1 had a mean of 2.5 and a standard deviation of 3.0, while group 2 also had a mean of 2.5 but a standard deviation of 1.5. Because the means of the groups are equal, the true effect size is zero. Data were generated by randomly sampling $N = 301$ from the multivariate normal distribution of group 1, and $N = 302$ from the multivariate normal distribution of group 2. Because the sample of data is random, the actual difference of sample means will not be exactly the generating effect size.

Figure 7 shows histograms of the data and the results of the analyses. The upper panel of Figure 7 shows the results when treating the ordinal data as if they were metric. The results from Welch's t test (Welch, 1938), which allows for different

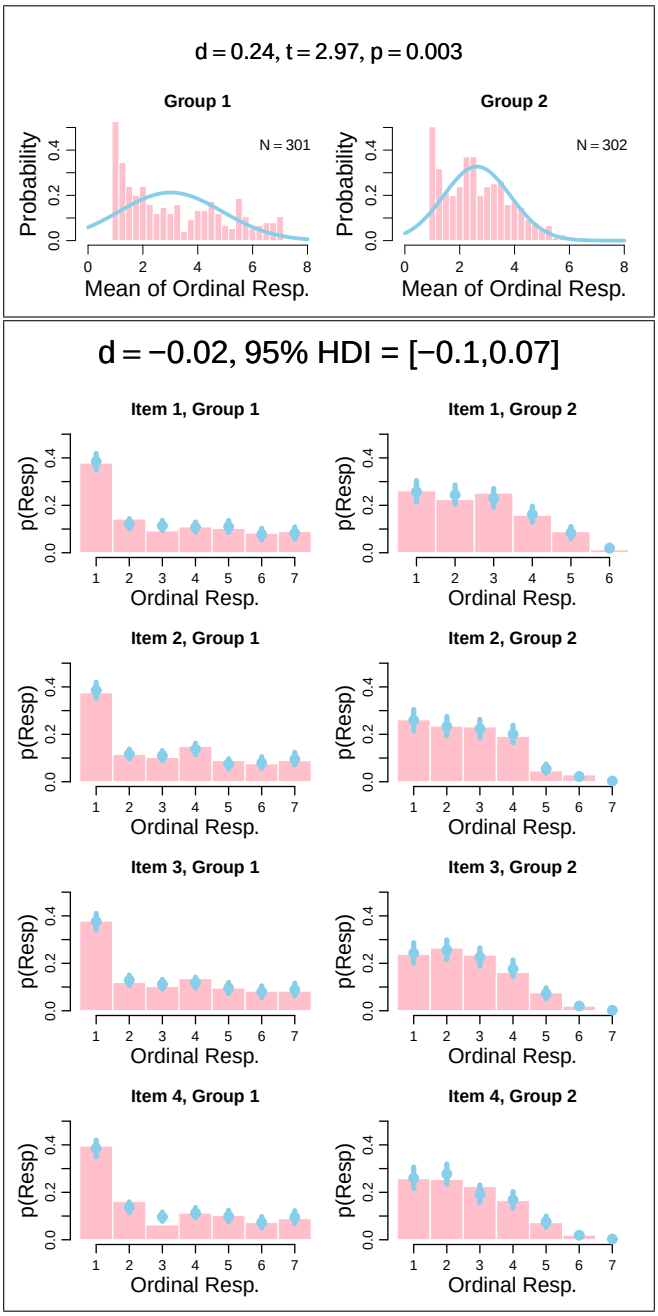

Figure 7. Ordinal data from four items for two groups, displayed in histograms (same data in upper and lower). The generating parameters had $d = 0.0$. *Upper: Treated as metric.* Best fitting normal curves are superimposed on data histogram; title indicates estimated effect size (Cohen's d) and result of Welch's t test. *Lower: Treated as ordinal.* Predicted response probabilities are shown as dots superiposed on data histogram; title indicates estimated effect size (Cohen's d) and the 95% highest density interval (HDI) of the effect size.

variances across groups, are displayed in the title of the plot. As can be seen, the analysis when treating the data as metric indicates a significantly non-zero effect, that is, a false alarm.

The lower panel of Figure 7 shows the results when analyzing the ordinal data with an ordered probit model. The title of the plot indicates the estimated effect size along with the Bayesian credible interval (the 95% highest density interval, HDI). As can be seen, the ordered-probit analysis recovers the true generating effect size very accurately.

The upper and lower panels of Figure 7 also show the model predictions superimposed on the data. The metric model finds the best-fitting continuous normal distributions to the average responses across items, as shown in the upper panel of Figure 7. Clearly the data are not well described by normal distributions. The ordered-probit model, on the other hand, predicts discrete data probabilities for all the items as shown in the lower panel of Figure 7. Clearly the data are very well described by the ordered-probit model.

Failure to detect non-zero effect (under-estimation of effect). The true generating multivariate normal distribution for group 1 had a mean of 3.0 and a standard deviation of 1.5, while group 2 had a mean of 2.5 and a standard deviation of 3.0. The true effect size is therefore $d = (\mu_1 - \mu_2)/\sqrt{(\sigma_1^2 + \sigma_2^2)/2} = 0.21$.

Figure 8 shows histograms of the data and the results of the analyses. The upper panel of Figure 8 shows the results when treating the averaged ordinal data as if they were metric. The results from Welch's t test (Welch, 1938), which allows for different variances across groups, are displayed in the title of the plot. As can be seen, the analysis when treating the data as metric indicates no significant effect, that is, a failure to detect the true difference between groups.

The lower panel of Figure 8 shows the results when analyzing the ordinal data with an ordered probit model. The title of the plot indicates the estimated effect size along with the Bayesian credible interval (the 95% highest density interval, HDI). As can be seen, the ordered-probit analysis recovers the true generating effect size very accurately.

The upper and lower panels of Figure 8 also show the model predictions superimposed on the data. The metric model finds the best-fitting continuous normal distributions, which are shown in the upper panel of Figure 8. Clearly the averaged data are not well described by normal distributions. The ordered-probit model, on the other hand, predicts discrete data probabilities as shown in the lower panel of Figure 8. Clearly the data are very well described by the ordered-probit model.

Summary: Taking average of multiple items does not solve the problem. The previous sections illustrated that the same problems with the single-item case can

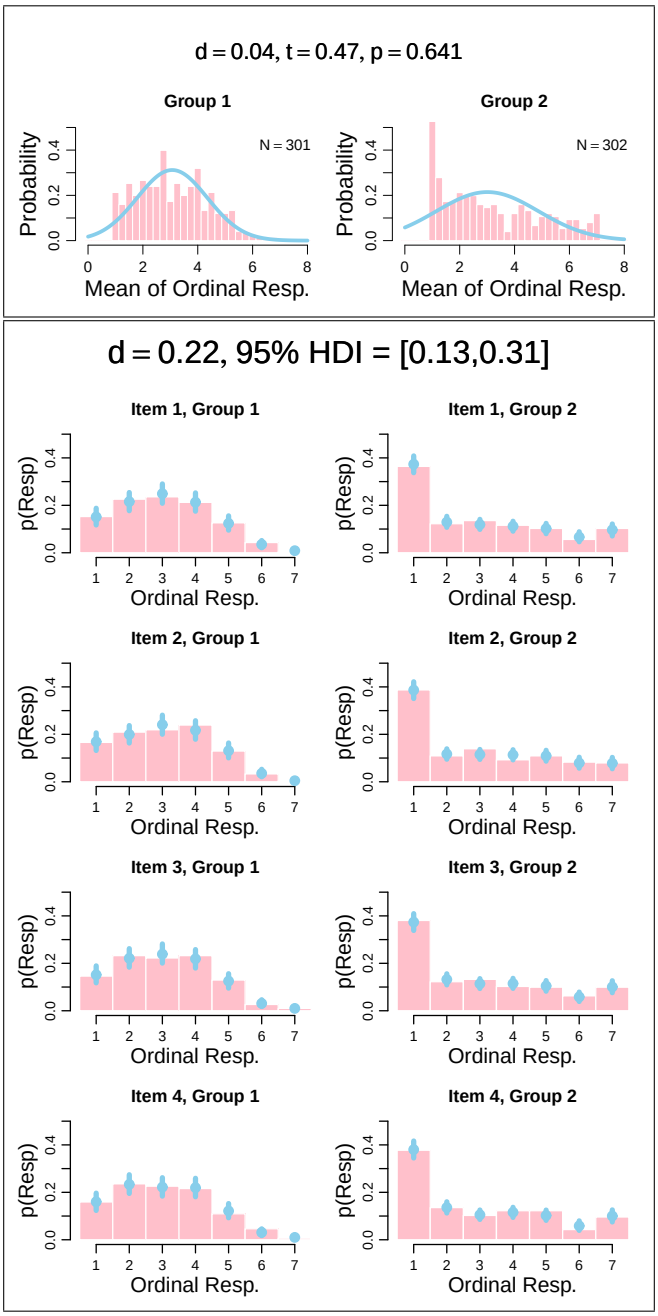

Figure 8. Ordinal data from four items for two groups, displayed in histograms (same data in upper and lower). The generating parameters had $d = 0.21$. *Upper: Treated as metric.* Best fitting normal curves are superimposed on data histogram; title indicates estimated effect size (Cohen's d) and result of Welch's t test. *Lower: Treated as ordinal.* Predicted response probabilities are shown as dots superiposed on data histogram; title indicates estimated effect size (Cohen's d) and the 95% highest density interval (HDI) of the effect size.

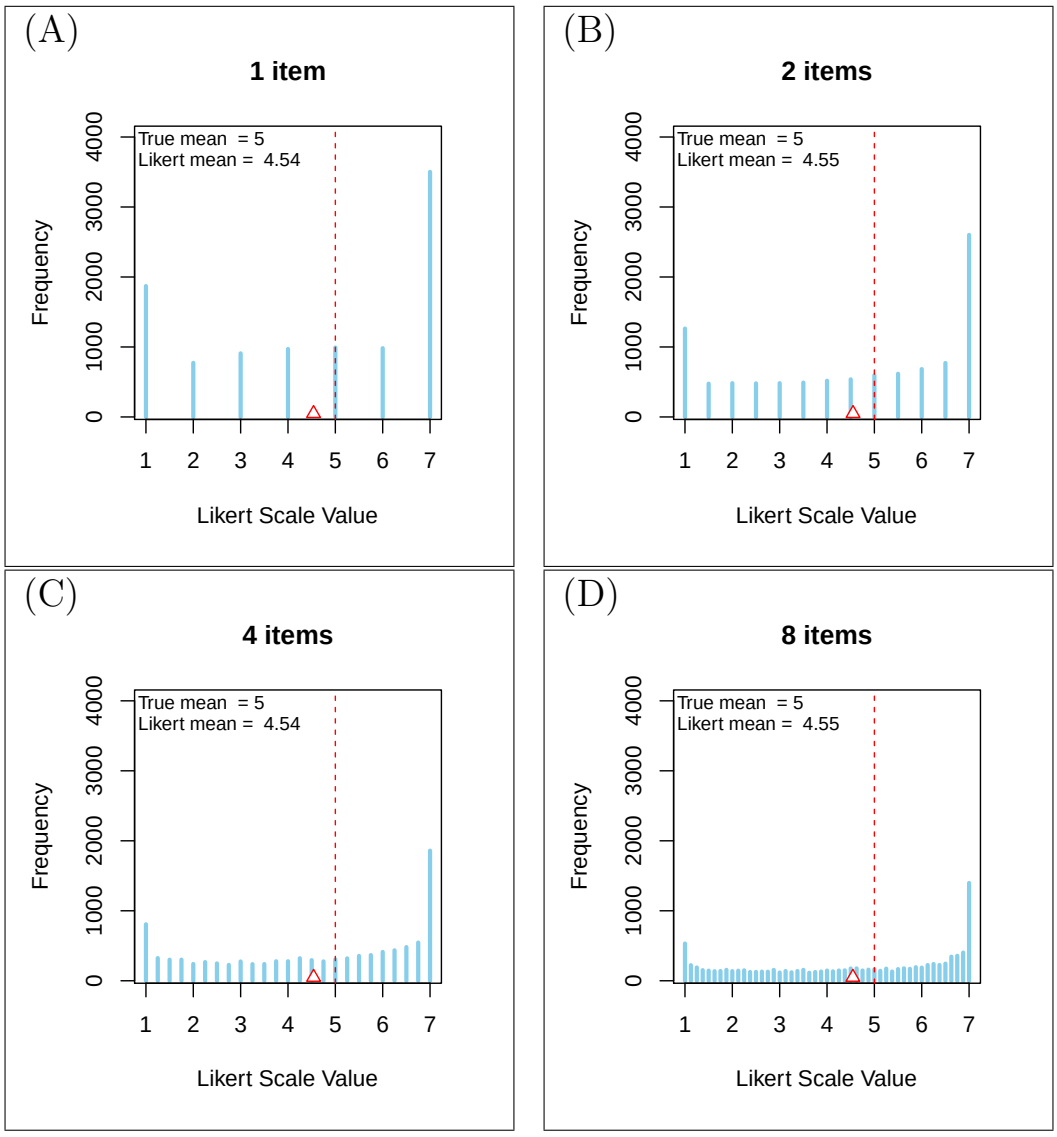

Figure 9. Histograms of Likert scale scores for different numbers of averaged questions or items. For Q items, averaged ordinal scores can occur at discrete levels of Q/Q, $(Q+1)/Q$, ..., KQ/Q. Panel A shows the single-item case ($Q = 1$). Panel B: $Q = 2$. Panel C: $Q = 4$. Panel D: $Q = 8$. For each panel, the observed mean of the ordinal values is annotated at the top of the panel, and also marked on the axis by a triangle. The true generating mean is marked by a vertical dashed line. Notice that despite increasing granularity, the observed mean underestimates the generating mean.

occur in a multi-item Likert scale. This is because computing a mean of Likert items does not prevent asymmetric bin widths. This is especially easy to see in the case of the high and low bins, which are still censored just as are responses for singular Likert items.

We illustrate this point with a simple example. Suppose we are measuring satisfaction with Likert items that use response levels from 1 to 7 (i.e., $K = 7$ in this example). There are Q items targeting the same underlying construct. The averaged ordinal responses can have the values Q/Q, $(Q + 1)/Q$, ..., KQ/Q. We wish to determine the probabilities of each of those possible outcomes. We did so via a Monte Carlo simulation. We generated these data using the same multivariate normal distribution described in the previous section. Thresholds for each dimension were again evenly spaced from 1.5 to 6.5, with pairwise correlations on each dimension being 0.8. We generated data from a single group with a mean of 5 and a standard deviation of 4. For each subject, we created an ordinal-as-metric Likert "scale" score by taking the arithmetic average of all questions. The area of interest in this simulation is what occurs as the number of items or questions Q increases.

Figure 9 shows the frequency histograms from 10,000 simulated subjects. As the number of items Q increases, the number of discrete Likert scale values increases, and the probability of receiving the most extreme scores is attenuated. However, this probability is attenuated roughly equally across extremely high scores and extremely low scores such that the biasing effect of the asymmetric floor and ceiling persists as Q increases. In particular, the expected value of the Likert scale remains steady at about 4.5 (recall that the true mean is 5) as Q increases.

Although the mean of the Likert scale value does not change much as Q increases (in this example), the standard deviation of the Likert scale values does decrease as Q increases. This decrease in standard deviation is caused by fewer of the Likert scale scores falling in the more extreme bins. Unfortunately this increase in precision merely makes the erroneous estimation of the mean *more* pronounced when tested against the true mean.

To illustrate this problem of higher precision in erroneous estimation, we repeated the two-school comparison simulation described in the two-group single-item comparison, including the elaborations necessary to produce Likert scale observations. Each scale was constituted by $Q = 8$ Likert items. The standard deviation of each item was set at 2 for the first school and set at 4 for the second school. The observed false alarm rate produced by this procedure was .201, even greater than when using a single Likert item. The reason for this increase in false alarm rate is the decrease in standard deviation noted in the previous paragraph: a lower standard deviation leads to greater certainty in the erroneous observed mean, leading to a higher false alarm rate.

In summary, Likert scales do not solve any of the problems highlighted here with the ordinal-as-metric approach, and in some cases may even exacerbate them.

Linear regression

In the case of simple linear regression, we are predicting an ordinal variable from a single metric predictor. The model is similar to the two groups case, except instead of a nominal predictor (e.g., group or school) we have a metric predictor (e.g., time spent on homework). We will again see advantages of Bayesian estimation of a ordered probit model relative to a frequentist analysis that treats the ordinal predicted values as metric. For this case, we focus on errors in capturing response probabilities, utilizing a Monte Carlo simulation and a real-data example.

Monte Carlo Simulation. In this simulation, we assessed the abilities of the models to make accurate predictions regarding the probability of each ordinal response. We compared the Bayesian ordered probit model to ordinal-as-metric linear regression.

Our simulation procedure was similar to the single-item two-group case. First, a metric x value was generated from a uniform distribution between 0 and 12. Then, a y value was generated using the x value as a predictor via Equations 4 and 1 with $\beta_0 = -1$, $\beta_1 = 0.8$, and $\sigma = 3.0$. The thresholds were identical to the two-group case (i.e., 1.5, 2.5, 3.5, 4.5, 5.5, and 6.5). In each simulation, 100 $\langle x, y \rangle$ pairs were generated, and both models were applied to the generated data. In total, 10,000 simulated data sets were analyzed.

As described above, our primary interest is not the correct detection rates of the methods. Indeed, for these generating parameter values (with a large non-zero slope), both models had correct detection rates of 100%. Instead, our primary interest is the two models' abilities to accurately describe the data in terms of predicted outcome probabilities.

The predicted response probabilities are much more accurate for the Bayesian ordered probit model than for the ordinal-as-metric model. To see this graphically, consult Figure 10, which plots the category probabilities predicted by each model alongside the true generating probabilities for a single representative simulation. The true generating probabilities are represented by the wide bars. The thinner superimposed bars are model predictions: the lower bar is the ordinal-as-metric prediction and the upper bar is the Bayesian ordered probit prediction. For the ordinal-as-metric model, the predictions represent the normal density as predicted using the MLE parameter values, normalized across the seven responses. In the Bayesian ordered probit case, the bars represent the mean predicted probability in the posterior distribution. Notice that the Bayesian ordered probit predictions more

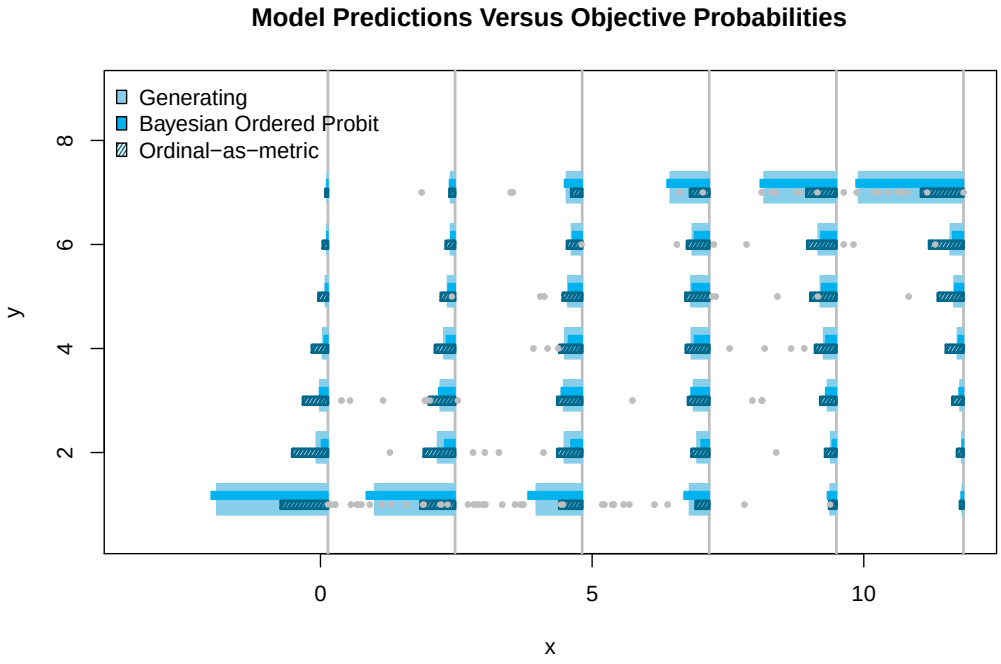

Figure 10. Predicted ordinal probabilities for the ordinal-as-metric model and the Bayesian ordered probit model, along with the generating probabilities. The sideways bars represent the probabilities of each ordinal response at the selected values of x indicated by the grey vertical lines. At each ordinal response there are three bars. The wide bar is the true generating probability. The left (lower) bar is the ordinal-as-metric prediction. The right (upper) bar is the Bayesian prediction. The small grey dots are data values. The Bayesian predictions are much more accurate than the ordinal-as-metric predictions.

closely match the generating probabilities across the range of x-values.

We used KL divergence (Kullback and Leibler, 1951) to quantitatively assess the departure of the model predictions from the true probabilities. The KL divergence of one model's predictions from the generating probabilities, for a particular x value, is $\sum_{k\in 1:7} P_{gen}(k) ln \frac{P_{gen}(k)}{P_{mod}(k)}$ where $P_{mod}(k)$ is the probability of response k predicted by the model, and $P_{gen}(k)$ is the generating probability of response k. In each simulation, we calculated the KL divergence of each model for 30 equally spaced x values across the range of x. As in Figure 10, the predicted probabilities of the ordinal-as-metric model were the normalized densities of the respective y response using the MLE parameters, and the predicted probabilities for the Bayesian ordered probit model were the mean category probabilities for the respective y responses in the posterior distribution. Figure 11 shows a density plot of the sampling distribution of the observed KL divergence for the two models across the 10,000 simulations. The distributions of KL divergence values for the two models are drastically differ-

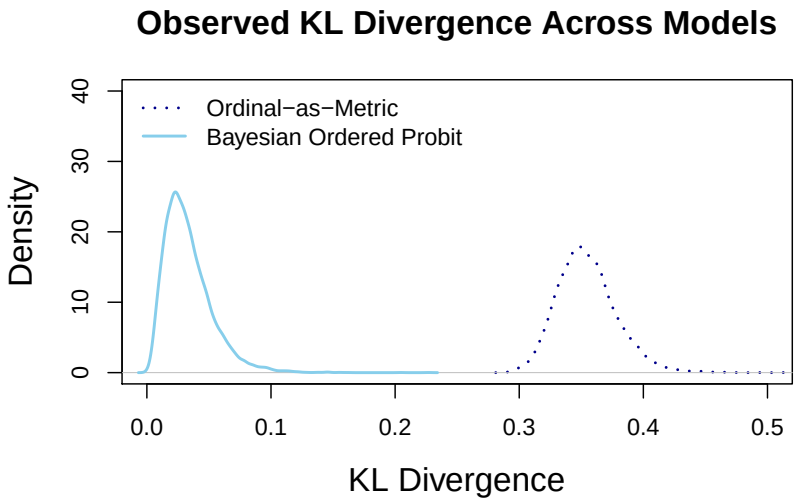

Figure 11. Distributions of the KL divergence of model predictions from generating probabilities. The mean KL divergence for the ordinal-as-metric model is 0.355, whereas the mean KL divergence for the Bayesian ordered probit model is an order of magnitude lower, at 0.033.

ent, with the ordinal-as-metric model having considerably higher divergence.

Example real data. In this section we present exemplary real data wherein the ordinal-as-metric model and the Bayesian ordered probit model disagree on the predicted probabilities of ordinal outcomes at various x values. We again make use of the ELS data set question 20F. However, instead of comparing the scores of two schools on question 20F, we now predict the scores of individual students on question 20F using the response to question 34 part B as the sole predictor. Question 34 was "Overall, about how much time do you spend on homework each week, both in and out of school," and part B was the response to the box labeled "Out of school." Thus we predict (ordinal) student attitudes about teachers from (metric) hours spent on homework out of school. The data presented here come from School 1141.

We applied ordinal-as-metric simple linear regression and the Bayesian ordered probit regression model. The slopes are similarly estimated for the two models: the ordinal-as-metric model yields an MLE slope of -0.037, $p = 0.106$, with 95% CI from -0.082 to 0.008. The Bayesian model yields a modal slope of -0.050 with 95% HDI from -0.110 to 0.014. However, the two models make disparate predictions about the probability of each categorical response at given values of the predictor. Figure 12 illustrates the difference in prediction probabilities across the two models. Unlike in the simulated data, there is no generating probability to compare the predictions of the models against. Instead, we plot the observed proportions of each

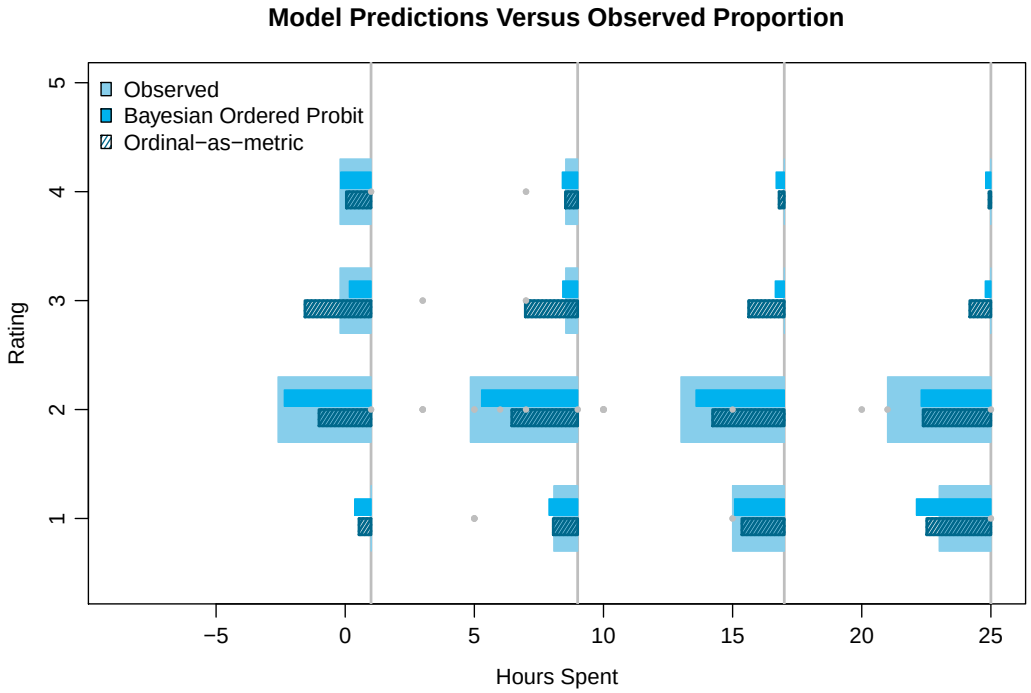

Figure 12. Predicted ordinal probabilities for the ordinal-as-metric model and the Bayesian ordered probit model, along with the observed proportions. The sideways bars represent the probabilities of each ordinal response at the selected values of x indicated by the grey vertical lines. At each ordinal response there are three bars. The wide bar is the observed proportion. The left (lower) bar is the ordinal-as-metric prediction. The right (upper) bar is the Bayesian prediction. The small grey dots are data values. The Bayesian predictions tend to match the observed proportions better than the ordinal-as-metric predictions.

ordinal response in four sub-intervals of the predictor. The x range is split into four equal sections and these sections are used to bin the data. Within each bin, the proportion of responses in each ordinal category is represented by the thicker bar. As before, the overlaid thinner bars represent model predictions of the two models, evaluated at the mid-point of each bin. Notice that the Bayesian ordered probit model is usually closer to the observed proportions than the ordinal-as-metric model, mirroring the results of the simulated data from the previous subsection. To quantify this difference of model predictions, we again computed the KL divergence between each model's predictions and the observed proportions, computed for each of the four x values plotted in Figure 12. For the ordinal responses where no data were observed, KL divergence was treated as zero for both models. The mean KL divergence of the Bayesian ordered probit model was 0.100, whereas the mean KL divergence of the ordinal-as-metric model was 0.246. Clearly, the probabilities predicted by the Bayesian ordered probit model more closely match the observed

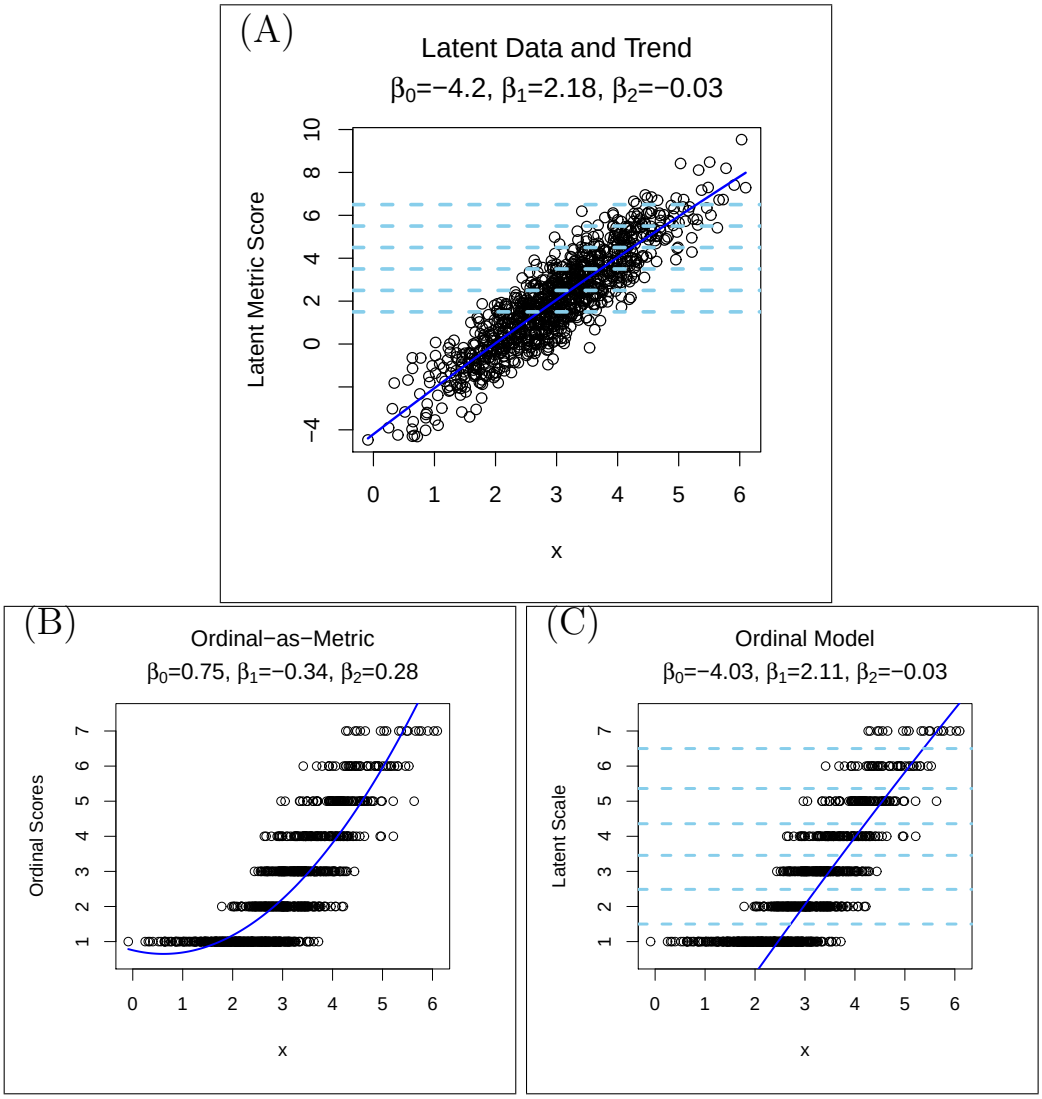

Figure 13. Example data with no underlying quadratic trend, analyzed with ordinal-as-metric quadratic regression and an ordered probit quadratic regression. Panel A shows the underlying latent data, with the thresholds indicated as blue dotted lines, labeled with the maximum likelihood estimates for a linear regression model. Panel B indicates the maximum likelihood estimate trend for the ordinal-as-metric quadratic trend. Note that it estimates a high degree of quadratic curve, in contrast with the true generating values. Panel C shows the maximum likelihood estimate trend for an ordered probit regression model with estimated thresholds indicated by the blue dotted lines, and demonstrates that the true linear trend is recovered.

data.

Quadratic regression

One area with high potential for an ordinal-as-metric analysis to make mistakes is in the detection of quadratic trends. A quadratic trend analysis is one way

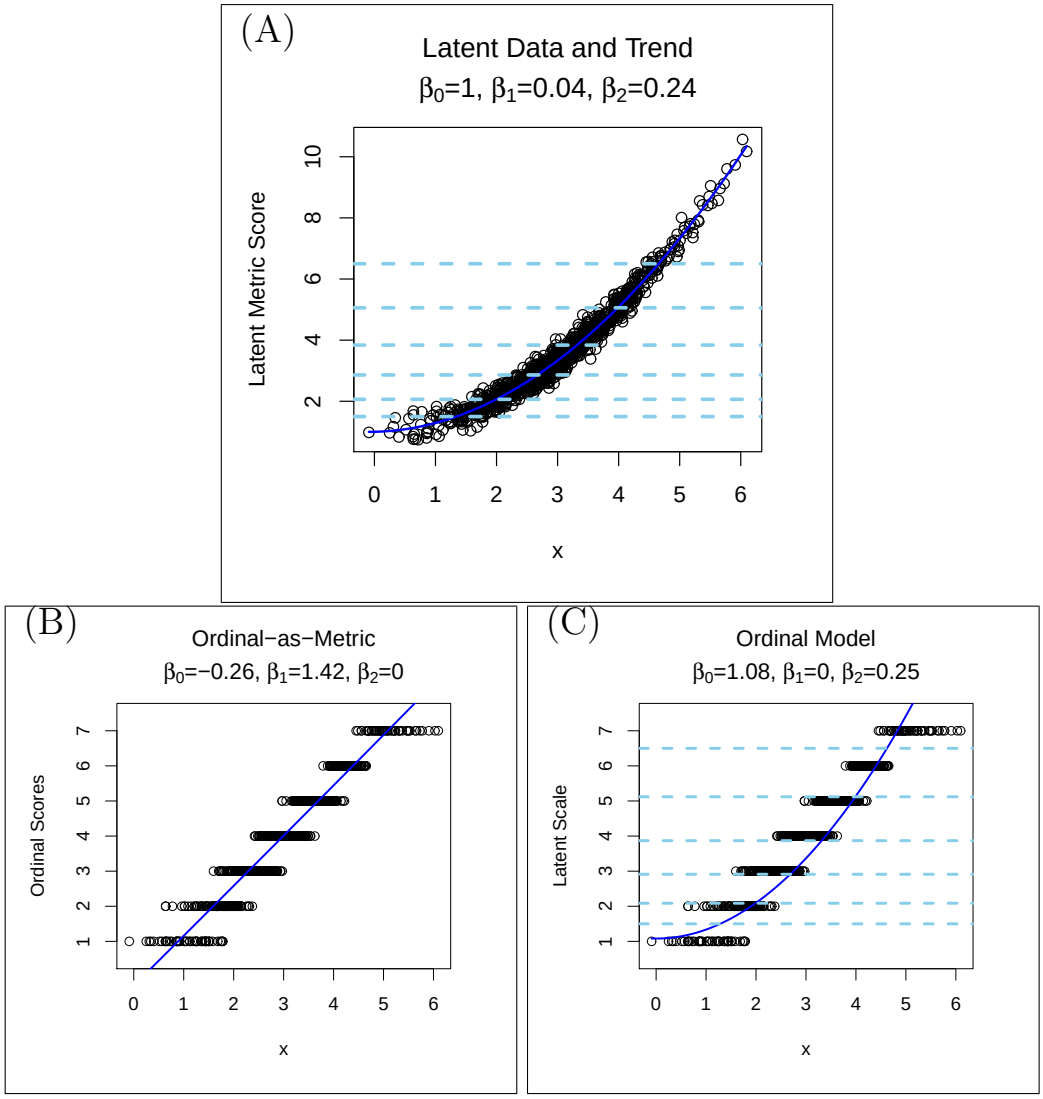

Figure 14. Example data with an existing underlying quadratic trend, analyzed with ordinal-as-metric quadratic regression and an ordered probit quadratic regression. Panel A shows the underlying latent data, with the thresholds indicated as blue dotted lines, labeled with the maximum likelihood estimates for a linear regression model. Panel B indicates the maximum likelihood estimate trend for the ordinal-as-metric quadratic trend. Note that it estimates zero quadratic curvature, in contrast with the true generating values which have a high degree of curvature. Panel C shows the maximum likelihood estimate of trend for an ordered probit regression model with estimated thresholds indicated by the blue dotted lines, and demonstrates that the true quadratic trend is recovered; compare the estimated parameter values in the titles of the panels.

to detect a curvature in the relationship between a predictor x and a predicted y. This is accomplished by utilizing x^2 as a separate predictor with its own estimated predictor weight, encoding the direction and degree of curvature. The simple linear model is extended such that μ is defined as a quadratic function of x as defined in Equation 4.

The unequal bin width of the ordinal scale can produce data that look curved if they are treated as metric, when the underlying latent value is completely linear. See Figure 13 for an example. The large amount of values below the bottom threshold creates an impression of curvature when the data are treated as metric. Conversely, an ordinal model correctly recovers the (lack of) quadratic curvature.

This effect of unequal bin width can also obfuscate an existing quadratic trend. Figure 14 shows an example where the narrower intervals at the lower end of the scale obfuscates the curvature of the underlying metric data, if the ordinal data are treated as metric. Again, the ordinal model correctly recovers the quadratic curvature.

This quadratic example is included to provide another illustration of the distortions that are endemic to treating ordinal data as metric, caused by the unequal bin-width problem identified earlier in the manuscript. The problems outlined here are not intended to be exhaustive; on the contrary, they are intended to convey the severity and pervasiveness of problems with the general practice of ordinal-as-metric analyses.

General discussion

So far we have presented several examples of problems for ordinal-as-metric analyses, all under the general umbrella of the unequal bin-width problem. We've shown that effect size can be mis-estimated and correct detection rates can be decreased in the context of linear regression. When combined with unequal variances across groups, we found these problems and high false alarm rates in group comparison. In the context of quadratic trends, we again demonstrated the potential for false alarms and failure to detect trends in the data. We have also shown that multi-item Likert scales are incapable of providing a solution to the unequal bin width problem.

In this section, we look at other potential ways to model ordinal data and conclude that Bayesian ordinal models are preferable, though any ordinal model will avoid many of the problems described here.

Detecting differences in variances

One of the major problems with ordinal-as-metric analysis identified here is the unequal bin width problem combined with a difference of variances in group com-

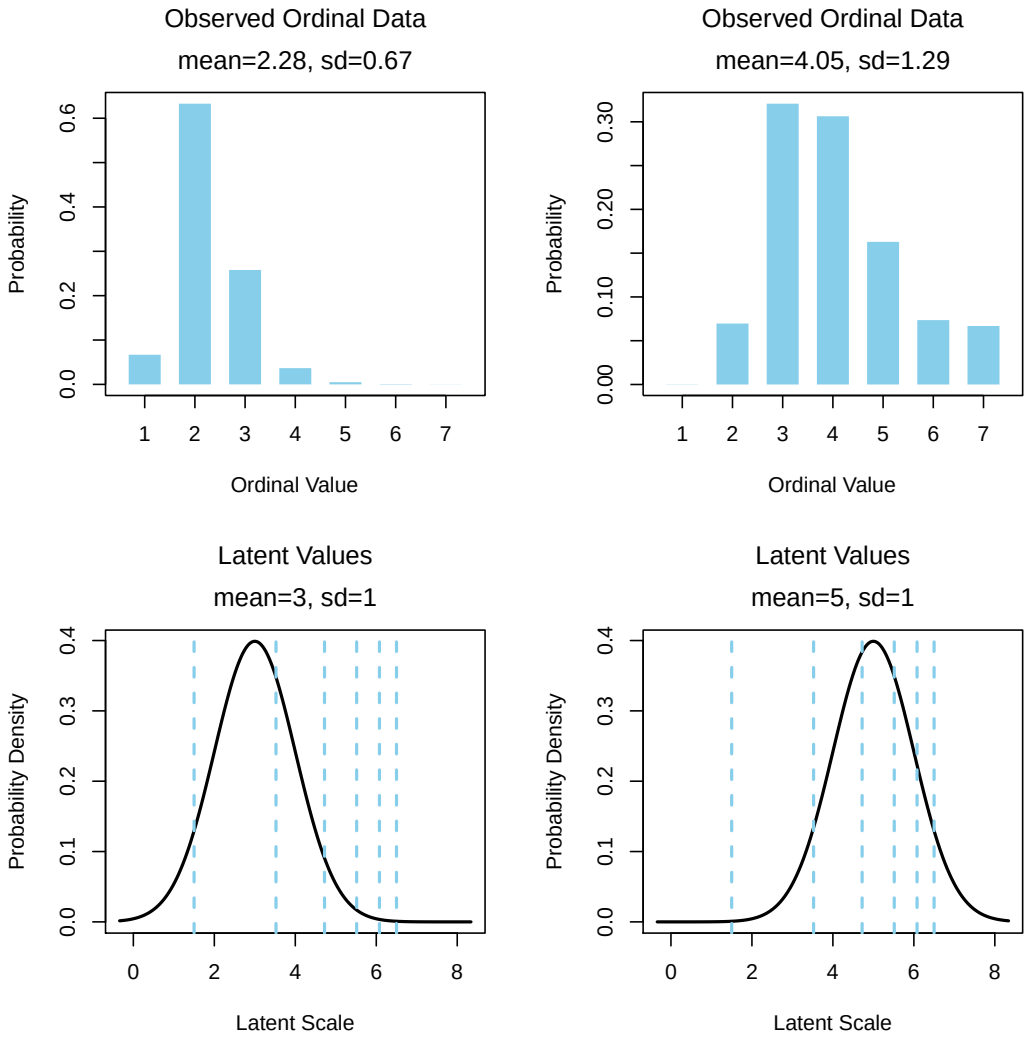

Figure 15. Plots of two distributions with identical latent variance, but different observed ordinal-as-metric variance. The bar plots show the observed ordinal data, labeled with the ordinal-as-metric mean and standard deviations. The density plots show the corresponding latent continuous distributions, with the thresholds indicated as dotted lines.

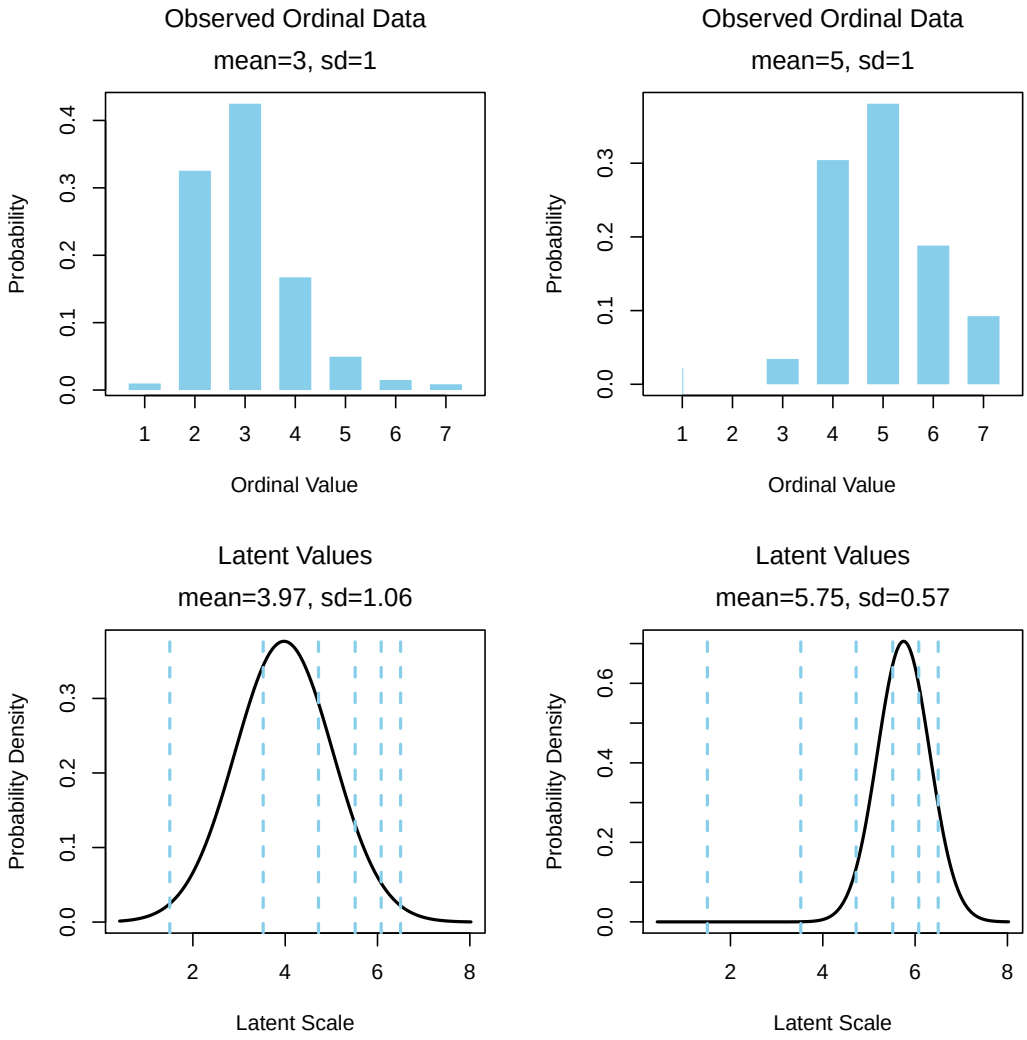

Figure 16. Plots of two distributions with different latent variance, but identical observed ordinal-as-metric variance. The bar plots show the observed ordinal data, labeled with the ordinal-as-metric mean and standard deviations. The density plots show the corresponding latent continuous distributions, with the thresholds indicated as dotted lines.

parison. It might be possible to avoid this problem while maintaining an ordinal-as-metric approach if we attempt to detect differences in variance across groups before beginning an analysis. This might allow the analyst to at least rule out the possibility of high false alarm rates demonstrated here. However, despite initial plausibility, this method is not tenable. This is because the condition is a difference of variances on the *latent metric scale*, which is not necessarily detectable in the observable ordinal data. Figure 15 and Figure 16 demonstrate this visually. Figure 15 shows a case where the variances of the two latent metric distributions are the same, but the variances of the observed ordinal datasets are quite different. Figure 16 shows the opposite case, where the variances of the two latent metric distributions are different, but the variances of the observed ordinal datasets are the same. This demonstrates an important point: often the only method to determine potential problems in an ordinal-as-metric approach is to apply an ordinal model, in which case the results of the ordinal analysis ought to be utilized regardless.

Other analysis methods

In this paper, we have described several problems with the common method of analyzing ordinal data via frequentist metric analyses. We have also shown that Bayesian ordinal models do not suffer these issues. Though we have focused on the status quo frequentist metric analyses, and the alternative approach of Bayesian ordinal models, these two methods are not the only potential approaches to ordinal data. As discussed previously, the two orthogonal changes we advocate, from frequentist to Bayesian and from metric models to ordinal models, suggest two other potential approaches. Here we briefly discuss these two alternatives, and present arguments for why we believe neither of these methods are adequate when compared to Bayesian ordinal methods.

Metric Bayesian Analysis. It is possible to apply the same or similar metric models that are traditionally utilized in frequentist analyses of ordinal data, and estimate the parameters in a Bayesian fashion. However, any model that does not respect the ordinal character of ordinal data will have the same issues we have discussed in this paper, regardless of how the parameters of the model are estimated. Even if Bayesian analysis provides additional detail and allows further flexibility in the metric models, the parameters from a model that fails to describe the data is meaningless. As such, we do not pursue a numerical demonstration of the inadequacy of this methodology, as it would essentially repeat the problems of frequentist metric models demonstrated throughout this paper.

Frequentist ordinal model. Frequentist ordinal models have more promise; they do not treat ordinal data as though they were metric, and thus avoid many of the

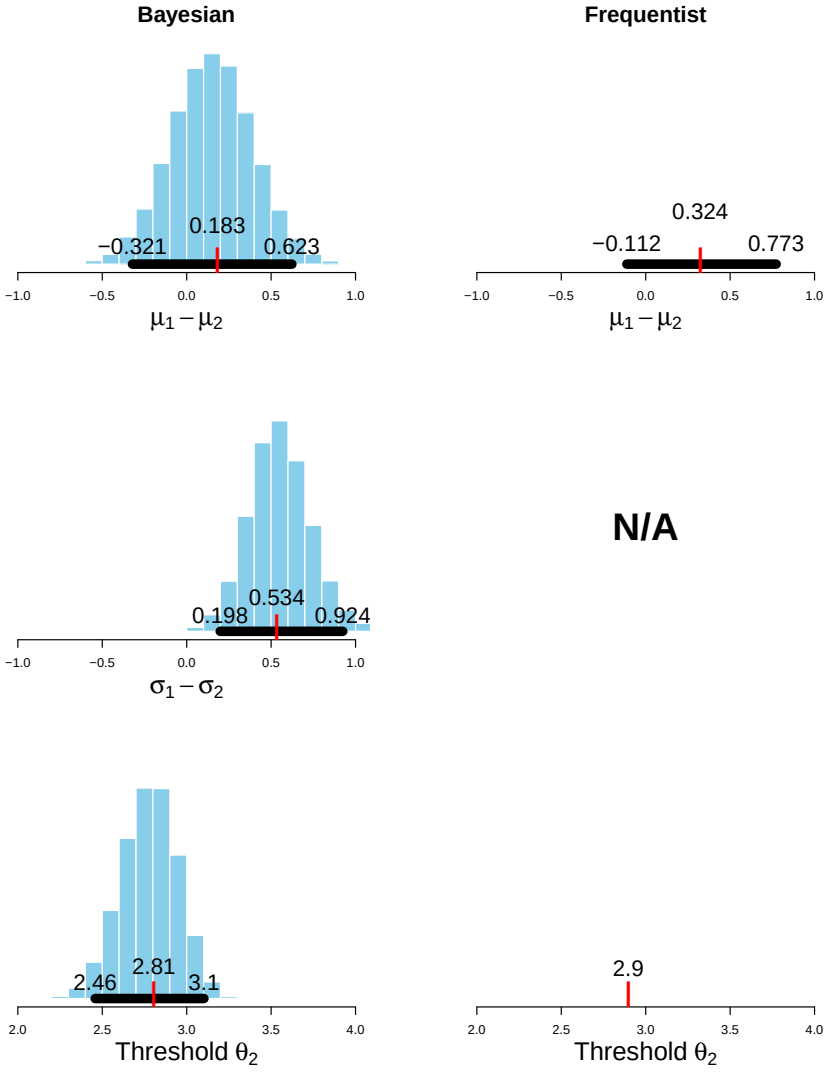

Figure 17. Comparison of frequentist (using the polr function from the MASS package in R, Venables and Ripley, 2002) and Bayesian implementations of the ordered probit model. For all plots, labeled red tick marks indicate modal posterior or MLE values, and the black bars along the axis represent measures of uncertainty (Bayesian HDI or frequentist confidence interval). The first row of plots compares the two approaches on the difference of latent means. Note that the frequentist plot has no distributional information, and also that the estimate of the difference between the two groups is biased upward due to the homogeneous variance assumption. The second row compares the two approaches to difference of variances. Note that due to the homogeneous variance assumption, the frequentist implementation cannot provide an estimate. Finally the last row shows estimates of one of the thresholds. Note that while the Bayesian implementation maintains the full distributional information, the frequentist implementation only has the MLE, with no measure of uncertainty.

issues presented here. However, Bayesian ordinal models have distinct advantages over frequentist approaches.

The typical frequentist analysis for ordinal data is ordered probit regression (and the closely related ordinal logit regression). As described above, the ordered probit model is the same model we presented earlier and estimated the parameters by using Bayesian means. The ordinal logit model is similar, but instead of using the cumulative normal distribution to link the latent continuous values to ordinal responses, the similar-in-shape logistic distribution is used. We focus on the probit model here, but for the issues important to present purposes we can consider the two to be essentially identical.

Given that this class of models is the same in structure as the Bayesian ordinal models presented here, this approach is immune to the problems described earlier. That said, this approach does have two disadvantages when compared to Bayesian ordinal models. The first is that, like a large class of more complicated models, confidence intervals and p-values associated with parameters of the model are dependent on asymptotic assumptions that can cause errors for small sample sizes (e.g. Albert and Chib, 1993; McKelvey and Zavoina, 1975). Moreover, even an approximation of a confidence interval is not available for all parameters of the model (like the category thresholds) in many circumstances. This information is easily available in the posterior distribution produced by a Bayesian analysis.

The second and more serious error for our purposes is that the frequentist models assume equal variances. That is, these models assume that across all levels of the predictors, the variance in the predicted value is equal. Neither our simulated data nor our example real data respect this assumption in the context of group comparison. In the case of the t test, which also traditionally assumes equal variance, there are easily available alternatives that relax the equal variance assumption, and these alternatives were used here. However, with ordinal regression models, versions that relax the equal variance assumption are not usually available in modern software. To test the extent of this effect, we applied ordered probit regression to our fictional two school example (using the polr function from the MASS package in R, Venables and Ripley, 2002). Recall that both schools had means of 5.0 on the underlying metric scale, one school had $\sigma = 4.0$ and the other had $\sigma = 2.0$, with 10,000 samples generated from these values. These conditions led to a false alarm rate of 0.100 for the frequentist ordered probit regression, and a mean estimated effect size of 0.288. Recall that the ordinal-as-metric t test yielded a false alarm rate of 0.182 and a mean estimated effect size of 0.148. Thus the equal variance assumption was in fact more damaging than treating the data as metric when estimating the effect size. Conversely, separate variances are easily implemented in modern Bayesian estimation software. This is reflective of a broadly applicable benefit of Bayesian

analysis: models are easily and flexibly modified to appropriately model the data, whereas frequentist analysis requires a different sampling distribution to be derived, which may not be available in modern software or possibly even the statististical literature, depending on the specific application and model.

Figure 17 demonstrates some of the important distinctions between frequentist and Bayesian implementations of the ordered probit model discussed above. We analyzed the actual data presented in Figure 5, though the actual properties of the data are not important to the distinctions presented in the figure.

Conclusions

We have shown that applying metric models to ordinal data can cause various problems:

1. Despite existing demonstrations in the literature that false alarms rates are not greatly inflated by treating ordinal data with metric models, we have pointed out systematic conditions in which false alarms are indeed greatly inflated. These conditions include latent distributions with unequal variances across groups that are located asymmetrically relative to the response scale. We illustrated these conditions with novel simulations and real data.

2. We have shown in novel simulations that treating ordinal data with metric models can lead to greatly reduced power, that is, reduced ability to correctly reject the null.

3. We have shown that treating ordinal data with metric models can misestimate effect sizes and variances.

4. We have shown that treating ordinal data with metric models can make very poor predictions of response probabilities for the ordinal values.

5. We have shown that averaging several ordinal items (to produce a "Likert scale") does not solve these problems.

To address these problems, we used an ordered probit model with parameters estimated through Bayesian analysis. In various Monte Carlo simulations, we showed that

1. for the Bayesian ordered probit models, the false-alarm rates and correct-detection rates were appropriate in scenarios where the metric models had greatly inflated false alarm rates and reduced correct detection rates,

2. the Bayesian ordered probit estimates of parameters were very accurate, unlike the metric models, and

3. the Bayesian ordered probit predicted probabilities were very accurate, unlike the metric models.

Moreover, we illustrated the differences between analyses with examples from real data. 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 Bayesian estimation is an excellent way to estimate the parameters of such a model.

References

Albert, J. H. and Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*, 88(422):669–679, doi:10.1080/01621459.1993.10476321.

Albert, J. H. and Chib, S. (1997). Bayesian Methods for Cumulative , Sequential and Two-step Ordinal Data Regression Models. *Technical Report, Bowling Green State University.*

Becker, W. E. and Kennedy, P. E. (1992). A Graphical Exposition of the Ordered Probit. *Source: Econometric Theory*, 8(8):127–131, doi:10.1017/S0266466600010781.

Carifio, J. and 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.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Erlbaum, Hillsdale, NJ, 2nd edition.

Denwood, M. J. (2013). runjags: An R package providing interface utilities, distributed computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software*.

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

Glass, G. V., Peckham, P. D., and 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, doi:10.3102/00346543042003237.

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

Heeren, T. and 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. and 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, doi:10.3102/00028312006004515.

Hui, M. and 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):1217–8, doi:10.1111/j.1365-2929.2004.02012.x.

Kruschke, J. K. (2011a). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, 6(3):299.

Kruschke, J. K. (2011b). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Academic Press/Elsevier, Burlington, MA.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology. General*, 142(2):573–603, doi:10.1037/a0029146.

Kruschke, J. K. (2015). *Doing Bayesian data analysis, second edition: A tutorial with R, JAGS, and Stan*. Academic Press/Elsevier, Burlington, MA, 2nd edition.

Kruschke, J. K. and Liddell, T. M. (2017a). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review*, **(**):**–**, doi:10.3758/s13423-017-1272-1.

Kruschke, J. K. and Liddell, T. M. (2017b). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, **(**):**–**, doi:10.3758/s13423-016-1221-4.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, doi:10.1214/aoms/1177729694.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*.

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

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

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*.

Rogers, S. J., Pratt, D. J., Rogers, J. E., Siegel, P. H., and Stutts, E. S. (2004). Education longitudinal study of 2002: Base year data file user's manual.

Spranca, M., Minsk, E., and 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, doi:10.1126/science.103.2684.677.

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

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

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.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, 14(5):779–804, doi:10.3758/BF03194105.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29(3/4):350–362.

Wickens, T. D. (1982). *Models for behavior: Stochastic processes in psychology*. W.H. Freeman & Co., San Francisco.

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