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

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

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

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

Torrin M. Liddell*, John K. Kruschke

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

ARTICLE INFO

Handling editor: Roger Giner-Sorolla

Keywords:
Ordinal data
Likert
Ordered-probit
Bayesian analysis

ABSTRACT

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

1. Introduction

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

2. Ordinal data and approaches to modeling them

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

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

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

https://doi.org/10.1016/j.jesp.2018.08.009
Received 6 November 2017; Received in revised form 3 August 2018; Accepted 16 August 2018
0022-1031/

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

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

2.1. Ordinal data are routinely analyzed with metric models

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

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

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

2.2. Metric and ordinal models

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

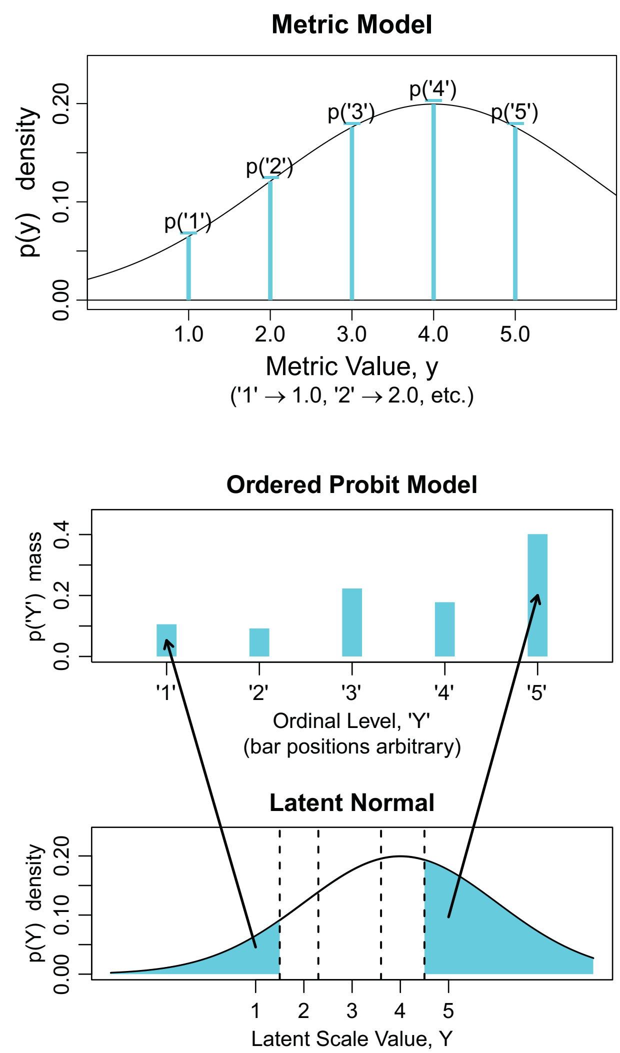

Fig. 1. *Upper panel: Metric model of ordinal data.* Ordinal values are mapped to corresponding metric values on the horizontal axis, with ‘1’→1.0, ‘2’→2.0, and so forth. The probability of each value is the normal density as shown by the heights of the lines. *Lower pair of panels: Ordered-probit model.*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

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