The Impact of Likert or Rating Scale Mis-responding (Differential Responding) on Commonly Used Statistical Methods:

An Overview of the Symposium and the Methodology

Bruno D. Zumbo
University of British Columbia

Charles O. Ochieng
CTB/McGraw-Hill

Paper Presented at the 2003 American Educational Research Association (AERA) annual meeting in Chicago, Illinois

Introduction

- Many of the measures used in our research are ordered responses to questionnaires, surveys, ratings of complex assessment tasks, or experimental tasks in a research study.
- The random variables that characterize these ordered responses are commonly referred to as “ordinal”, “rating scale”, or “Likert” variables -- we will use these terms interchangeably.
- The focus of this symposium centers on two points:
  - the fact that these ordered random variables are not continuous variables but yet are commonly treated as continuous variables when they are used in statistical techniques such as factor analysis, principal components analysis, inter-rater reliability indices, and regression analysis.

Introduction

  - There is an “unstated” assumption that that everyone in your population is responding using the same response process – that is, the same thresholds for a given variable.

- I believe that many researchers have a sense of this assumption. That is, a common question we hear is:
  - With these rating scales what happens to the results of our analyses if some sub-group of respondents are using the scale differently than the rest of the respondents.
- This prototypical question involves: (a) the effect of the number of scale points, and (b) the effect of “differential” or “mis-responding”. Researchers have investigated the former but not the latter.

A Model for the Item Response Process

- Let us see if we can articulate this assumption with a bit of formal detail.
- As Zumbo and Zimmerman (in press) note, there are a variety of conceptualizations of ordered response variables but the most common one in the educational, social, and behavioral sciences characterizes these variables as ordered-categorical observed variables wherein the underlying variable is completely unobserved (i.e., latent). Furthermore, as the normally distributed latent variable increases beyond certain threshold values, the observed variable takes on higher scores, referred to as scale points.
- The debate about this conceptualization, as Zumbo and Zimmerman remind us, goes back to the early 1900s with Pearson-Heron-Yule and is really at the root of the "levels of measurement" arguments in social research methodology.

A Model for the Item Response Process

More formally, for an observed ordered variable *y* it is assumed that there is an underlying (unobserved) continuous variate y^* that

(a) represents the latent variable (e.g., an aptitude, attitude, propensity, or personal characteristic of the respondent) underlying the ordered responses to y, and
(b) is assumed to rest on the real line, hence having a range of $-\infty$ to $+\infty$ and assigning a metric to the otherwise metric-free ordered response variable.

The observed random variable, *y*, is typically characterized as taking on one of *m* possible categories – if we were to use the language of Likert scales, this would be, for example, a 4-point Likert response scale or more generally an *m*-point response or rating scale.

A Model for the Item Response Process

Therefore, if y has m ordered categories, the link between y and y^* is

$$y = j \Leftrightarrow \tau_{j-1} < y^* < \tau_j, \ j = 1,2,3,\ldots,m,$$

where

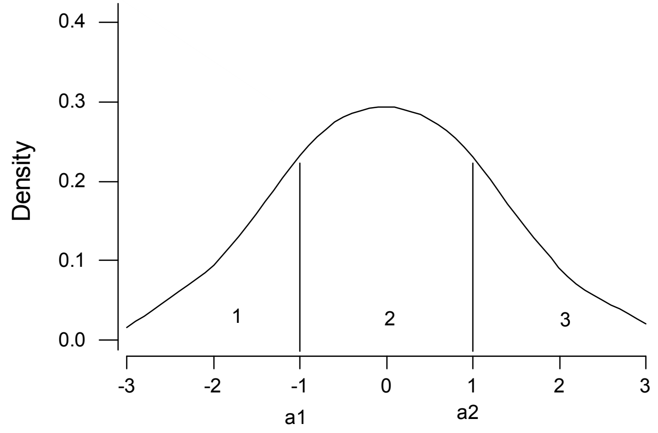
$$-\infty = \tau_0 < \tau_1 < \tau_2 < \tau_3 < \cdots \tau_{m-1} < \tau_m = +\infty,$$

are threshold values. For an m-point response scale, there are m-1 threshold values. It is both conventional and statistically convenient to characterize the random variate y^* as a standard normal (mean zero and unit variance) distribution.

A Model for the Item Response Process

- Given the above description of an ordered response variable, all we know is that in responding to a task or question a person who chooses one category has more of a characteristic than if he/she had chosen a lower category, but we do not know how much more.
- As such, the ordered discrete random variable, *y*, is not a continuous variable although it is implicitly treated as such in commonly used statistical methods in educational and psycho-social research.
- This assumption is further strained when (a) there is reason to believe that the relation between the underlying variate and the observed score may not be linear, and (b) when there may be sub-groups of individuals who are using different response thresholds.

A Model for the Item Response Process



- This is an example of a three point item response format.
- Anyone less than a threshold of –1 will respond with a 1.
- What would happen to our statistical results, however, if a sub-group of respondents were not able to discern the top two thresholds and responded "2" to anything in the top two areas.

The research focus on the symposium

- In short, then, the matter of concern for this symposium is what happens to the statistical results of analyses like regression, factor analysis, reliability estimates, and intraclass correlations when some sub-group of respondents (or raters) use a different response process.
- We will report on a series of studies that investigate the impact of (a) using ordered response variables and (b) a sub-group of respondents using a different response process on some commonly found research applications:
  - fitting a regression model via ordinary least-squares regression;
  - estimating the reliability of measure or scale;

Mis-responding / Differential Responding

  - determining the dimensionality of a measure via exploratory factor analysis or principal components analysis; and
  - assessing the reliability of rating data – i.e., investigating the inter-rater reliability via an intraclass correlation
- There two fundamental methodological questions we need to address before we go on:
  - When might this “differential” or “mis-responding” happen?
  - Clearly there are many types of mis-responding that could be studied, what type are we investigating?

Mis-responding / Differential Responding

- When might this "differential" or "mis-responding" happen?
  - First, it is important to note that we are still struggling with what term to use to describe this phenomenon. We have chosen the term "mis-responding" because an analogue to what we are proposing occurs in multi-way contingency tables under the rubric of "misclassification". In one sense we are investigating ordinal misclassification or, focusing on the response process, "mis-responding".
  - Another way of looking at the phenomenon, however, is to focus solely on the response process and note that it is not "mis-responding" (because that implies incorrect responding) but rather that a sub-group of individuals are using different response thresholds – akin to DIF.

Mis-responding / Differential Responding

- When might this happen? Well here are two examples of when it might happen.
  - We are investigating responses to a depression scale. It is well documented in the literature that men and women have different response thresholds on some of those items (e.g., the "crying" item)
  - In a 1997 study of the effects of consent form information on self-disclosure, it was shown that men respond differently than women when asked information about the consent form. When asked details on the consent form they just signed, men under-report on depression and women do not.
- Our focus is not on whether you know misresponding is present in your data but rather what happens to your statistical results if misresponding is present.

Mis-responding / Differential Responding

- Clearly there are many types of mis-responding that could be studied, what type are we investigating?
- Our goals today are to highlight the issue of mis-responding, call for a more extensive program of research, and to present some preliminary findings on a variety of statistical methods.
- Therefore, the type of mis-responding we will focus on in this symposium is: there is a sub-group of the population that cannot discern the difference between the top two response categories and hence they always respond with the lower.

Common Research Methodology

- Before describing each of the papers in the symposium, it is appropriate, at this point, to say a few words about the commonalities among the papers:
- The papers report on a series of computer simulation studies.
- There is no widely accepted recommendation on the number of scale points for a rating scale therefore each of the studies will explore rating scale data ranging from 3 to 9 scale points. Studies reporting more than 9 scale points are rarely seen in the research literature.

Common Research Methodology

- Although skewed item response data occur with frequency we have focused our attention on the case of symmetric item response distributions as an initial cut into the problem.
- This symposium will not deal with the matter of hypothesis testing or sample-to-sample variability of the statistics but rather on what happens to statistics such as the eigenvalues and loadings from a factor or principal components analysis, the R-squared from a regression analysis, and the intraclass correlation coefficient from an inter-rater study when we manipulate the: (a) number of scale points and (b) response distribution of the rating scale variable(s).

Common Research Methodology

- Therefore, as in Zumbo and Ochieng (2002) and Ochieng (2001), to side-step the matter of sample-to-sample variability and focus on large-sample impact (a form of bias), 100,000 continuous normally distributed scores will be generated and all comparisons will be made at this population analogue level.
- These normally distributed scores will represent the (typically unobserved) latent scores from which the order responses will be simulated. For each of the finite populations of 100,000, the continuous scores (which represent the unobserved latent variable) will be manipulated to mimic responses on a rating scale.
- In essence, the simulation methodology mimics the process of responding to a rating scale format and then uses the responses as variables in the analyses.

Common Research Methodology

- The objective of the simulation is to compare the statistic (e.g., the factor loadings in factor analysis) produced by analyzing the rating scale data to the same statistic that one would have obtained, with the same data, had they been able to conduct the statistical using the continuous latent variable, rather than the rating scale responses. Zumbo and Zimmerman (in press; 1993) describe the advantages of comparing the "ideal" (in our case the latent continuous variate) to "observed" data to study the impact of scaling.
- Where possible we will also use statistical (regression) modeling to statistically study the outcomes of the simulation study. In some cases, however, the result show little to no variation across the simulation conditions so we use the modeling cautiously. Response surface approach similar to Harwell and Zumbo (1999).

The papers

- #1 Slocum, Ochieng, & Zumbo look at regression and correlation
- #2 Gelin, Beasley, & Zumbo look at estimating coefficient alpha for a scale and examining the dimensionality of the scale
- #3 Rupp, Koh, & Zumbo look at the matter of using a polychoric correlation matrix with exploratory factor analysis (LISREL and EQS)
- #4 Witarsa, Breithaupt, and Zumbo look inter-rater reliability and intraclass correlation.
- #5 I will come back at the end and try and wrap it all up before we turn to the discussants.