Research Data Reports

FEDERAL RESERVE BANK OF BOSTON™

No. 17-2

Battery Order Effects on Relative Ratings in Likert Scales

Marcin Hitczenko

Abstract:

Likert-scale batteries, sequences of questions with the same ordinal response choices, are often used in surveys to collect information about attitudes on a related set of topics. Analysis of such data often focuses on the study of relative ratings or the likelihood that one item is given a lower (or higher) rating than another item. This work studies how different orderings of the items within a battery and, in particular, the relative location of items affect relative rating distributions. We take advantage of data from the 2012–2014 Survey of Consumer Payment Surveys, in which item order in six Likert-scale batteries is varied among respondents.

We find that ordering effects are real and consistent across years. The most prominent effect relating to relative locations of items is that the farther one item is placed after another item, the more likely that item is to have a lower rating.

JEL Classifications: C83

Keywords: hierarchical models, survey design, ordinal responses

Marcin Hitczenko is a survey methodologist and a member of the Consumer Payments Research Center in the research department of the Federal Reserve Bank of Boston. His e-mail address is marcin.hitczenko@bos.frb.org.

This paper, which may be revised, is available on the web site of the Federal Reserve Bank of Boston at http://www.bostonfed.org/economic/rdr/index.htm.

This version: July 24, 2017

CPRC CONSUMER PAYMENTS Research Center

1 Introduction

Many fields of research, especially those in the social sciences, rely on surveys as a means of collecting data. Indeed, certain types of information, such as attitudes or self-assessments, can only be gathered in this manner. Experience has shown that response patterns are heavily influenced by the questionnaire design, with variations in the instrument often introducing systematic tendencies and introducing a "survey effect" in the distribution of sample statistics (Bradburn, Sudman, and Wansink 2004; Schuman and Presser 1996; Sudman, Bradburn, and Schwarz 1996). Aspects as fundamental as the survey mode (Bowling 2005) to seemingly trivial details of question presentation (Dawes 2008; Schwarz et al. 1991) are known to make a difference. As a result, an entire field of research, survey methodology, has emerged to better understand these aspects of data collection and to establish conventions for consistency.

One of the prominent themes in the survey design literature is that order often matters. While in some contexts there is no evidence that the order of questions in a survey affect responses, (Bradburn and Mason 1964), most analyses find otherwise (McFarland 1981; Sigelman 1981; Schwarz and Hippler 1995). There is also strong evidence that the order of response options influences respondents' tendencies (Knauper 1999; Chan 1991; Schwarz and Hipler 1991).

In this work, we focus on order effects within a very narrow, but common, form of survey question: a battery of Likert-scale questions. A Likert-scale question asks a respondent to select a response from a set of categorical, ordered options (Likert 1932). Likert-scales are useful for categorizing attitudes or feelings when quantitative measures are impossible or impractical to obtain. While the nature of the questions and response options can vary, we use the terminology that the respondent is "rating" an "item," which is often the case. If a Likert-scale question asks respondents to rate one particular item, then a battery of such questions asks for ratings of several different, but presumably related, items on the same scale in one table or screen. An example from the field of consumer payment surveys is given in Figure 1.

Likert-scale batteries allow respondents to efficiently provide ratings for a group of items in the context of one another. For this reason, analyses often center on the relative rating distributions to two items in the battery, or how often one is given a lower rating than the other. Unlike mean ratings or even the distribution of ratings themselves, relative rating distributions provide direct insight into how the population feels toward one item relative to another. Looking at relative ratings avoids issues caused by heterogeneity of responses, in which certain individuals tend to give high ratings while others tend to give low ratings (Tourangeau, Rips, and Rasinki 2000).

As far as we know, there has been little research on the effects of item ordering in Likert-scale batteries. The lone example, Siminski (2008) found significant differences in response patterns under two orderings in the context of a medical survey, but that analysis did not examine relative rating distributions nor did it attempt to quantify effect sizes or trends. In this work, we look to understand the degree of variation in relative rating distributions that is caused by different orderings and whether trends in ordering effects depend on the relative location of two items. This paper is organized as follows. Section 2 defines the notation and research goal. Section 3 introduces the data used in our analysis, and Section 4 does some preliminary data analysis that motivates the model developed in Section 5. Section 6 discusses the implications of the model fits, and a broad summary of findings and potential future work is given in Section 7.

2 Notation

In this section, we define the basic concepts of interest and introduce any relevant notation. Our analysis is limited to Likert-scale questions with five possible responses fielded in batteries of eight items, just as in Figure 1. Considering a collection of respondents, indexed by u, we let $R_i[u]$ represents the rating given to item i by individual u. In the case of a five-point Likert-scale, $R_i[u] = 1, 2, \ldots, 5$ represent the ordered responses. In the example in Figure 1, a rating of 1 corresponds to "very hard to get or set up" and a rating of 5 corresponds to "very easy to get or set up."

The focus of this work is not on individual responses, but rather aggregate trends within the entire sample. Specifically, we are interested in the relative rating distributions for two items, i and j, defined by $\text{Prob}(R_i[u] < R_j[u]), \text{Prob}(R_i[u] = R_j[u])$, and $\text{Prob}(R_i[u] > R_j[u])$ where u is assumed to be drawn at random from the entire population of U.S. consumers. To distinguish patterns under different orderings, we let o represent a particular ordering of items, generally represented as a permutation of the integers one to eight. A battery with eight items has $8! = 40{,}320$ possible orderings. For any two items, i and j, and ordering, o, the relative ratings of interest can be uniquely identified by the quantities

$$p_{ij}(o) = \text{Prob}\left(R_i[u] < R_j[u] \mid \text{item order } o\right) \quad \text{and} \quad q_{ij}(o) = \text{Prob}\left(R_i[u] \leq R_j[u] \mid \text{item order } o\right), \quad (1)$$

along with the fact that $\text{Prob}\left(R_i[u] = R_j[u] \mid \text{item order } o\right) = q_{ij}(o) - p_{ij}(o)$ and $\text{Prob}\left(R_i[u] > R_j[u] \mid \text{item order } o\right) = 1 - q_{ij}(o)$.

Ideally, the quantities in (1) are identical for all o, but based on general findings about order effects, we expect differences. One goal of this work is to gain insight into how $\{p_{ij}(o), q_{ij}(o)\}$ compares with $\{p_{ij}(o'), q_{ij}(o')\}$. Order effects, if present, likely result from a variety of factors, many of which may be specific to the particular set of items

being considered. However, some effects might relate to known, predictable components of any ordering, one of which is the relative location of the two items in question. Relative location takes into account order of appearance as well as distance between two items. Thus, we define $\ell_i(o)$ to be the location in the battery of item i in ordering o, coded as integers from 1 (first in the list) to 8 (last in the list). Then, we define $d_{ij}(o) = \ell_j(o) - \ell_i(o)$, which takes integer values from -7 to 7, with the exception of 0.

3 Data

The data analyzed in this paper come from the Survey of Consumer Payment Choices (SCPC), a survey conducted annually since 2008 by the Consumer Payment Research Center at the Boston Federal Reserve. The survey, generally administered at the end of September and beginning of October, asks respondents about preferences and typical behavior regarding various aspects of household economics, with a particular emphasis on the use of payment instruments. The entire survey is taken online and takes around 30 minutes to complete. In our analysis, we use data from the 2012 – 2014 SCPCs. The administration of the Likert batteries using several different orderings began in 2012, and the survey in all three included years provides responses for the exact same set of batteries. Studying data from several renditions of the survey helps estimate the size of any universal ordering effects. Below, we discuss the SCPC sample and introduce the survey design for the questions of interest.

3.1 Survey of Consumer Payment Choice

SCPC respondents come exclusively from RAND's American Life Panel (ALP). The ALP originated in 2006 with around 1,000 respondents and has been growing since. Recruitment into the ALP has relied on a variety of methodologies, with the general goal of making the panel as representative of the U.S. adult population as possible. More information about the ALP can be found at `http://mmic.rand.org/alp`.

A key strategy in SCPC sample selection every year has been to balance the improvement of sample coverage with respect to the target population of U.S. adults and to continue a longitudinal component to the sample. As a result, many individuals in one year are the same as in a previous year. In 2012, there were 3,170 respondents, including almost 1,000 individuals who had recently been added to the ALP as part of a targeted recruiting strategy for lower-income and minority households. In the following two years, there were fewer respondents, 2,082 in 2013 and 1,805 in 2014. There were 1,347 respondents who participated in all three years.

The analysis in this paper is based on treating each respondent with equal weight, rather than weighing to better match population composition or explicitly modeling separate effects for different demographic groups. While the

SCPC sample in any given year does not perfectly match the demographic compositions of the U.S. adult population, with respect to certain common demographics it is generally no farther off than a simple random sample of the target population might be expected to be. Although it is certainly possible that the effect of order on response tendencies varies across various demographic variables, such as age, making this type of assessment is beyond the focus of this work. Instead, we are interested in the average effect for the survey-taking population.

3.2 Assessment of Payment Instrument Batteries

Survey data analyzed in this paper are responses to a series of six Likert batteries relating to the assessment of eight payment instruments with respect to different characteristics. The characteristics to be rated were: acceptance, cost, convenience, security, ease of setting up, and access to payment records. Each characteristic is presented on a separate screen with the instruments listed vertically in a table as shown by the screenshot of the cost assessment question in Figure 1. The wording and definitions relevant to these survey questions, shown in Table 1, were the same for all three years of the SCPC.

Prior to 2012, the order in which the eight payment instruments were presented to the respondent was fixed, but in 2012 the SCPC began randomizing the order in which the items are listed. The eight payment instruments are grouped into three general types of payment instruments: paper, plastic, and online. The top panel in Table 2 lists the eight instruments by type. The randomization of the survey instruments was done by permuting the order of the three general groups of instruments while maintaining the same order within each group to preserve a degree of similarity between instruments of the same group. Therefore, there are six possible orderings for the instruments, as shown in the bottom panel of Table 2. Each year, the instrument orderings are assigned randomly to each respondent (and maintained for all six characteristics for that individual) independently of any ordering observed by that respondent in previous years. In order to accommodate the yearly component, we use the index $t = 1, 2, 3$ to identify data from the years $2012, 2013$, and 2014, respectively. We let $N_t(o)$ be the number of respondents in year t to see the batteries under ordering o. Because of the increased sample size in 2012, $N_1(o)$ is around 500 respondents, while $N_2(o)$ and $N_3(o)$ are closer to 300. Because of random assignment of ordering and sample changes, comparing data within a particular ordering across years generally involves responses from a different set of respondents. No more than 12 percent of respondents in a given year saw the same ordering in a different year.

Implementing this form of randomization is advantageous from the point of view of survey methodology, as it allows us to study patterns under different orderings. Restricting the number of orderings to six options has the benefit of yielding substantial sample sizes for each ordering, making inferences about the particular orderings

possible. However, if we are primarily interested in studying trends with respect to relative locations of the items, a strategy in which orderings were taken at random from the $8! = 40{,}320$ possible options would yield a greater number of meaningful observations. As currently done, certain pairs of instruments are always in the same relative locations: checks always directly follow cash. This means we cannot compare how relative ratings of checks and cash change as their relative locations change. Additionally, by grouping instruments according to similarity, it is possible that we are masking certain forms of variation, as we do not observe larger distances between items of the same payment groups. Perhaps relative ratings for instruments that have similar attributes, such as credit and debit cards, have a different pattern across orderings than relative ratings for pairs of instruments with less in common.

4 Preliminary Data Analysis

We begin with some preliminary data analysis that examines the nature of the variability in sample relative rating frequencies under different item orderings. The extent of the variation, the presence of any trends with respect to the relative location of items, and the extent to which the trends are consistent across years all inform the statistical model for the data. In particular, having responses for identical batteries from a set of largely different respondent sets in subsequent years provides great insight into whether effects are largely fixed for each given battery. Because of the yearly component of the data, we introduce the subscript t to the notation in (1),

$$p_{ijt}(o) = \text{Prob}\left(R_{it}[u] < R_{jt}[u] \mid \text{item order } o\right) \quad \text{and} \quad q_{ijt}(o) = \text{Prob}\left(R_{it}[u] \leq R_{jt}[u] \mid \text{item order } o\right), \tag{2}$$

so that yearly effects can be recognized. Natural estimates of the quantities in (2) are the sample frequencies of observed relative ratings under corresponding orderings. Thus, if for respondent u, we let o_{ut} be the battery ordering observed in year t, then

$$X_{ijt}(o) = \sum_u 1\left[R_{it}[u] < R_{jt}[u] \mid o_{ut} = o\right] \quad \text{and} \quad Y_{ijt}(o) = \sum_u 1\left[R_{it}[u] \leq R_{jt}[u] \mid o_{ut} = o\right],$$

represents the number of individuals who gave a lower rating to item i than to item j under ordering o in year t and the number of individuals who rated item i with an equal or lower rating than item j under ordering o in year t, respectively. Sample estimates of (2) take the form

$$\bar{p}_{ijt}(o) = \frac{X_{ijt}(o)}{N_t(o)} \quad \text{and} \quad \bar{q}_{ijt}(o) = \frac{Y_{ijt}(o)}{N_t(o)}.$$

Perhaps the simplest measure of the order effect comes by contrasting the relative rating frequencies for a given

ordering to the averaged relative frequencies across all observed orderings, given by

$$\bar{p}_{ijt} = \sum_{o=1}^{6} \frac{N_t(o)}{N_t} \bar{p}_{ijt}(o) \quad \text{and} \quad \bar{q}_{ijt} = \sum_{o=1}^{6} \frac{N_t(o)}{N_t} \bar{q}_{ijt}(o).$$

Using weighted estimates rather than simply pooling data helps account for the fact that the relative prevalence of each ordering varies across years. We note that any changes in $\bar{p}_{ijt}$ and $\bar{q}_{ijt}$ across t are of little interest. While these quantities tend to be fairly consistent, some do show statistically significant differences across years. Because our analysis focuses on relative changes from one ordering to another, we do not want to confound such ordering effects with changes in overall attitudes from one year to the next. Therefore, we always estimate baseline frequencies separately for each year.

Deviations, given by $p_{ijt}(o) - p_{ijt}$ and $q_{ijt}(o) - q_{ijt}$, provide a measure of variation in relative ratings under different orderings. Figure 2 shows scatterplots of deviations for $t = 1$ versus corresponding deviations for $t = 2, 3$. As a point of comparison, consider six independent draws from a $\text{Binomial}(300, 0.5)$ distribution. The deviation of the sample frequency from the first draw from the sample frequency among data pooled over all six draws has a standard deviation of 0.026, meaning over 90 percent of such deviations fall between -0.05 and 0.05. If the number of draws is changed to 500, this interval drops to $(-0.04, 0.04)$. Of course, if the likelihood of success moves away from 0.5, these intervals will shorten as well. For a success likelihood of 0.2 and number of draws fixed at 500, the standard deviation is 0.016. Given that values of p_{ijt} and q_{ijt} are not clustered around 0.5, the deviations in Figure 2 suggest greater variation than expected, especially in the case of $\bar{q}_{ijt}(o)$.

Perhaps the most remarkable aspect of Figure 2 is the consistency observed in deviations for different years. If deviations were drive only by sampling variation, the expected correlation would be 0. Instead, we observe correlations of over 0.5 and 0.6 for $p_{ijt}(o)$ and $q_{ijt}(o)$, respectively. The strength of correlation cannot be explained by the fact that some of the same respondents are providing responses within the same ordering. These people constitute no more than 12 percent of the sample and correlations of item ratings from one year to the next among those who participated in at least two years range from 0.25 to 0.6 with a mean of 0.45. Dependence driven by presence of the same individuals in two samples is not great enough to see the correlations in Figure 2. The similarity of the size and direction of deviations across years is evidence of real order effects. In addition, we find a relatively strong correlation of 0.46 between $\bar{p}_{ijt}(o) - \bar{p}_{ijt}$ and $\bar{q}_{ijt}(o) - \bar{q}_{ijt}$.

To look at trends with respect to relative locations, we capitalize on the above finding to combine data from all

three years. For each of the 14 observable distances, $d = -7, \ldots, -1, 1, \ldots, 7$, Figure 3 shows boxplots for

$$\bigcup_{(o,i,j,t)} \{\bar{p}_{ijt}(o) - \bar{p}_{ijt} \mid d_{ij}(o) = d\} \quad \text{and} \quad \bigcup_{(o,i,j,t)} \{\bar{q}_{ijt}(o) - \bar{q}_{ijt} \mid d_{ij}(o) = d\},$$

the former in the top plot and the latter in the bottom plot. Because the deviation for any particular value of d is measured relative to a baseline determined by the six observed orderings observed, combining deviations for different item pairs is not always comparing the same concepts. The average effect over observed values of d may not be the same for any two pairs of items, so the baseline of comparison is not standardized. For example, regarding cash and checks the distance is always $d = 1$, while for cash and online bill payment $d = 1$ is observed among $d = -7, -4, 3, 5, 1, 4$. In the former, expected deviation for any observed ordering is 0, because all orderings have $d = 1$. In the latter, the expected deviation for the ordering with $d = 1$ will depend on the average effects for the observed d, which may not be the same as the effect when $d = 1$.

Despite the concern in comparing deviations across item pairs, Figure 3 suggests a very interesting dynamic in the changes of relative rating distributions as a function of the relative distance. Following the medians of each boxplot, the data suggest that the likelihood of giving one item a lower rating than a second item is largely constant as long as the second item is after the first. Whether the second item is directly after or seven spots after the first item, the first item seems equally likely to have a lower rating than the second. However, as revealed by looking at cases where $d > 0$ in the second plot, the earlier item is progressively less likely to be rated less than or equal to the second item as the distance between them increases. The implication is that items later in the battery tend to get lower ratings than when they are earlier in the list. The sample trends in both plots when $d < 0$ (item i is after item j) largely mirrors the case where $d > 0$, as it should since $\text{Prob}(R_i[u] < R_j[u] \mid d_{ij} = d) = 1 - \text{Prob}(R_j[u] \leq R_i[u] \mid d_{ji} = -d)$.

Based on the results of Figure 3, a parametric model for trends in relative ratings as a function of d should be smooth and non-linear in order to capture the asymmetry between cases where $d > 0$ and those when $d < 0$. Based on the extent of noise around the median trends in the plots, relative location explains only a portion of variation due to order effects. Therefore, our model should allow for additional, order-specific effects independent of the relative location of items.

5 Model

In this section, we delineate a model for the 2012 – 2014 SCPC data that can be used to generate broader inferences about battery order effects in a general context. The model is based around formulating a set of distributions for

$\{p_{ijt}(o), q_{ijt}(o)\}$. Given these, a natural model for the observed data might be

$$\{X_{ijt}(o), Y_{ijt}(o) - X_{ijt}(o), N_t(o) - Y_{ijt}(o)\} \sim \text{Multinomial}\left(N_t(o), \{p_{ijt}(o), q_{ijt}(o) - p_{ijt}(o), 1 - q_{ijt}(o)\}\right). \quad (3)$$

In practice, fitting hierarchical models, in which $p_{ijt}(o)$ and $q_{ijt}(o)$ are themselves random variables, is somewhat onerous in most statistical software. Therefore, we greatly simplify the fitting by adopting the model

$$X_{ijt}(o) \sim \text{Binomial}(N_t(o), p_{ijt}(o)) \quad \text{and} \quad Y_{ijt}(o) \sim \text{Binomial}(N_t(o), q_{ijt}(o)). \quad (4)$$

Expected trends, effectively estimated by averaging over functions of $\bar{p}_{ijt}(o)$ and $\bar{q}_{ijt}(o)$, data statistics that do not violate assumptions of the Multinomial distribution, will yield sensible results (Agresti 2002). Potential problems may arise if one attempts to use our model to generate distributions of ordering effects, which in certain cases may yield implausible Multinomial distributions. This aspect is discussed in further detail below.

Defining $\text{logit}(z) = \log\left[\frac{z}{1-z}\right]$, for all $z \in (0, 1)$, we adopt the following logistic regression model:

$$\text{logit}\left[p_{ijt}(o)\right] \quad = \quad \mu_{ijt} + \lambda_{ijt}(o). \quad (5)$$

By virtue of the fact that

$$\begin{aligned} \text{logit}\left[q_{ijt}(o)\right] &= \log\left[\frac{q_{ijt}(o)}{1 - q_{ijt}(o)}\right] \\ &= \log\left[\frac{1 - p_{jit}(o)}{p_{jit}(o)}\right] \\ &= -\left[\mu_{jit} + \lambda_{jit}(o)\right], \end{aligned} \quad (6)$$

specifying a distribution for $\{\lambda_{ijt}(o)\}$ and $\{\mu_{ijt}\}$ defines the entire data likelihood. The parameters, $\{\mu_{ijt}\}$, define the "base rate" of relative ratings, to which relative rating distributions under different orderings are compared. In estimation, we allow separate effects for μ_{ijt} and μ_{jit}, with little interest in estimating the joint distribution of these variables, and we assume independence between base rates and order effects.

We also assume that the random variables $\lambda_{ijt}(o)$ and $\lambda_{i'j't'}(o)$ are independent as long as it is not true that $i = i', j = j'$ or $i = j', j = i'$. In particular, this means that changes in how item i is rated relative to item j has no bearing on changes to how item i is rated relative to item j'. It is easy to imagine how such an assumption may be wrong. If items j and j' are sufficiently similar, a factor that affects the comparison of i to j might have a similar effect on the comparison of i to j'. In addition, certain items may be more susceptible to external factors influencing their given rating, in which case relative ratings involving that item might be generally more variable

across orderings. However, this level of complexity to the dependence structure is beyond the scope of this work.

We decompose the order effect, $\lambda_{ijt}(o)$, into two components: one dealing with the relative location of the two items and one related to all other sources of variation. We write the function as:

$$\lambda_{ijt}(o) \quad = \quad f_{ijt}\left(d_{ij}(o)\right) + \epsilon_{ijt}(o). \tag{7}$$

Because $d_{ji} = -d_{ij}$, the model in (7) implies that

$$\lambda_{jit}(o) \quad = \quad f_{jit}\left(-d_{ij}(o)\right) + \epsilon_{jit}(o). \tag{8}$$

In particular, the variable, $\epsilon_{ijt}(o)$, serves to distinguish response patterns in two different orderings in which the locations of the two items in question happens to be the same. There are myriad potential causes for these effects, an example of which might be the choice of items that fall between item i and item j.

For our smooth, non-linear function, we choose a cubic form, so that the model is

$$\begin{aligned} f_{ijt}(d) &= \sum_{k=1}^{3} \alpha_k d^k + \alpha_{ijt} d, \\ \alpha_{ijt} &= \beta_{[ij]} + \beta_{ij} + \beta_{[ij]t} + \beta_{ijt} \\ \epsilon_{ijt}(o) &= \epsilon_{[ij]o} + \epsilon_{ijo} + \epsilon_{[ij]to} + \epsilon_{ijto}, \end{aligned} \tag{9}$$

where the subscript $[ij]$ specifies effects that differ only in the sign of the ordered pairs (i, j) and (j, i). Thus, $\beta_{[ij]t} = -\beta_{[ji]t}$. The motivation for this specification comes from the positive correlation observed between deviations $\bar{p}_{ijt}(o)$ and $\bar{q}_{ijt}(o)$, which under the models in (5) and (6) corresponds to positive correlations between $\lambda_{ijt}(o)$ and $-\lambda_{jit}(o)$. In fact, if $\lambda_{ijt}(o) = -\lambda_{jit}(o)$, our model guarantees that $p_{ijt}(o) < q_{ijt}(o)$.

In considering model specifications, we found that models in which varying quadratic and cubic terms were included in addition to the varying linear term were superfluous. Fits resulted in linear combinations of the varying terms as having zero variance, suggesting that there was not enough variation in the distance effects to justify more complex models. Very similar results to those of the adopted model were found by using a piecewise linear function, with varying slopes for different signs of d. However, the simpler use and exposition of the model in (10) led us to its selection.

We model $\beta_{\cdot}$ and $\epsilon_{\cdot}$ as independent, Normal random variables with mean zero. Specifically, we take

$$\begin{array}{llll} \beta_{[ij]} \sim \text{Normal}(0, \sigma^2_{s0}) & \beta_{ij} \sim \text{Normal}(0, \sigma^2_{a0}) & \beta_{[ij]t} \sim \text{Normal}(0, \sigma^2_{s1}) & \beta_{ijt} \sim \text{Normal}(0, \sigma^2_{a1}) \\ \epsilon_{[ij]o} \sim \text{Normal}(0, \tau^2_{s0}) & \epsilon_{ijo} \sim \text{Normal}(0, \tau^2_{s0}) & \epsilon_{[ij]to} \sim \text{Normal}(0, \tau^2_{s1}) & \epsilon_{ijto} \sim \text{Normal}(0, \tau^2_{a1}), \end{array}$$

where the naming framework of the variance parameters takes "s" to mean symmetric with respect to (i, j), "a" to mean asymmetric with respect to (i, j), "0" to represent effects that are fixed across years, and "1" to those that vary across years.

The model used for the SCPC data accounts for the longitudinal nature of the data. However, it can easily be consolidated to a more general case, outside the context of repeated surveys. Thus, a hypothetical researcher who is interested in the distribution of ordering effects for a new battery of items is interested in a model of the form

$$\lambda_{ij}(o) = \sum_{k=1}^{3} \alpha_k d^k + \alpha_{ij} d + \epsilon_{ij}(o) \quad ,$$

$$\alpha_{ij}, \alpha_{ji} \sim \text{Normal}\left(0, \sigma_{s0}^2 + \sigma_{s1}^2 + \sigma_{a0}^2 + \sigma_{a1}^2\right) \quad \text{with} \quad \text{Cov}\left(\alpha_{ij}, \alpha_{ji}\right) = -\left(\sigma_{a0}^2 + \sigma_{a1}^2\right), \tag{10}$$

$$\epsilon_{ij}(o), \epsilon_{ji}(o) \sim \text{Normal}\left(0, \tau_{s0}^2 + \tau_{s1}^2 + \tau_{a0}^2 + \tau_{a1}^2\right) \quad \text{with} \quad \text{Cov}\left(\epsilon_{ij}(o), \epsilon_{ji}(o)\right) = -\left(\tau_{a0}^2 + \tau_{a1}^2\right). \tag{11}$$

6 Results

Models are fit using the glmer function in R. Estimates of all model parameters are provided in Table 3. We begin by discussing some implications of the model fits themselves.

The nature of the varying effect terms suggests strong similarity of effects for the same pairs across different years, $\lambda_{ijt}(o)$ and $\lambda_{ijt'}(o)$, as well as heavy dependence between $\lambda_{ijt}(o)$ and $\lambda_{jit}(o)$. For the linear component, in fact, only the time-invariant terms, $\beta_{[ij]t}$ and β_{ijt}, have estimated variances greater than zero. With regard to the order-specific term, the model does find the presence of a year-specific term, $\epsilon_{[ij]to}$, but it is relatively small, accounting for about a third of the variance in the order-specific effects. This result is consistent with empirical similarities in the data across years.

Additionally, the effect sizes are such that the model implies great similarity between $\lambda_{ijt}(o)$ and $-\lambda_{jit}(o)$, by virtue of the fact that the asymmetric terms have relatively small variance compared to the symmetric ones. Therefore, the distribution of $\lambda_{ijt}(o)$ conditional on $-\lambda_{jit}(o)$ is relatively concentrated. The practical implication is that effects related to rating item i less than item j are heavily dependent on effects related to rating item i less than or equal to item j, with strong positive correlations in shifts for any item pair.

One nice aspect of the strong dependence in ordering effects, $\lambda_{ijt}(o)$ and $\lambda_{jit}(o)$, is that simulations based on the fitted model are likely to produce results in which $p_{ijt}(o) \leq q_{ijt}(o)$, a condition that is not explicitly imposed by our estimation methodology. Given (5) and (6), we have that

$$\text{logit}\left[q_{ijt}(o)\right] - \text{logit}\left[p_{ijt}(o)\right] \quad = \quad \left(-\mu_{jit} - \mu_{ijt}\right) + \left(\alpha_{ij} - \alpha_{ji}\right) d + \left(\epsilon_{ijo} - \epsilon_{jio}\right). \tag{12}$$

The first term on the right-hand side of (12), representing difference in base rates, is greater than zero by definition. Therefore, implausible probabilities ensue if the remainder of right-hand side is negative and its absolute value is greater than the difference in base rates. Under our model, this quantity is a Normal random variable with variance greatest when $d = 7$ or $d = -7$, in which case, the standard deviation is 0.075. This means that if $-\mu_{jit} - \mu_{ijt} > 0.225$, it is virtually certain that the model will result in technically sound probabilities. A difference of 0.225 in base rates is the difference between a frequency of 0.5 and 0.55 or the difference between a frequency of 0.2 and 0.24. Therefore, as long as the likelihood of providing the same ranking is larger than 0.05, which is the case in our data, the fact that our model does not meet the technically required criteria should have little practical impact on simulations.

Algorithm 1 details a procedure to simulate data, $X^*_{ijt}(o)$ and $Y^*_{ijt}(o)$. The procedure detailed uses the estimated baseline rates, $\hat{\mu}_{ijt}$, so as to make comparisons of relative ordering effects for particular items in the SCPC survey easier, but it can easily be adapted to generate data for any unobserved batteries. Adaptations to simulate without an annual component involve removing the time-specific loop in (1) and updating distributions to those given in (10 – 11).

Algorithm 1 Simulating Data

for unordered pairs (i, j) **do**
  Draw $\beta_{[ij]} \sim \text{Normal}(0, \sigma^2_{s0})$
  Draw $\beta_{ij}, \beta_{ji} \sim \text{Normal}(0, \sigma^2_{a0})$
  for $o = 1, 2, 3, 4, 5, 6$ **do**
    Draw $\epsilon_{[ij]o} \sim \text{Normal}(0, \tau^2_{s0}$
    Draw $\epsilon_{ijo}, \epsilon_{jio} \sim \text{Normal}(0, \tau^2_{a0})$
    for $t = 1, 2, 3$ **do**
      Draw $\epsilon_{[ij]to} \sim \text{Normal}(0, \tau^2_{s1})$
      Define $\lambda_{ijt}(o)$ and $\lambda_{jit}(o)$, as given in (7), (8), and (10)
      Define $p_{ijt}(o)$ and $q_{ijt}(o)$ using $\hat{\mu}_{ijt}$ and $\hat{\mu}_{jit}$, as given in (5) and (6)
      Draw $\left\{X^*_{ijt}(o), Y^*_{ijt}(o)\right\}$ from $\text{Multinomial}\left(N_t(o), \left\{p^*_{ijt}(o), q^*_{ijt}(o) - p^*_{ijt}(o), 1 - q^*_{ijt}(o)\right\}\right)$
    end for
  end for
end for

6.1 Posterior Checks

Before drawing conclusions and inferences about trends, it is prudent to check that our model adequately captures certain aspects of the data. We do so through posterior predictive checks, in which properties of simulated data are compared with those of the observed data (Gelman and Hill 2007). We are primarily interested in expected trends

with respect to the relative locations, which are defined by the parameters α_k. Therefore, ideal posterior checks involve functions of the data whose distribution is predominantly determined by those parameters as opposed to others, thus isolating the effects of the α_k from other components, most notably the base rates.

A convenient pair of functions for any pair of items, (i, j), is

$$\begin{aligned} S_{ijt}(o, o') &= \text{logit}\left[p_{ijt}(o)\right] - \text{logit}\left[p_{ijt}(o')\right] \\ T_{ijt}(o, o') &= \text{logit}\left[q_{ijt}(o)\right] - \text{logit}\left[q_{ijt}(o')\right] \end{aligned}$$

because their expectations with respect to random draw of (o, o') are

$$\begin{aligned} \text{E}\left[S_{ijt}(o, o')\right] &= \sum_{k=1}^{3} \alpha_k \left[d_{ij}^k(o) - d_{ij}^k(o')\right] \\ \text{E}\left[T_{ijt}(o, o')\right] &= \sum_{k=1}^{3} \alpha_k \left[d_{ji}^k(o') - d_{ji}^k(o)\right]. \end{aligned}$$

In fact, for unique pairs (d_1, d_2), the set defined by

$$U(d_1, d_2) = \left\{S_{ijt}(o, o') \mid d_{ij}(o) = d_1, d_{ij}(o') = d_2\right\} \bigcup \left\{T_{ijt}(o, o') \mid d_{ij}(o) = -d_2, d_{ij}(o') = -d_1\right\}.$$

is a collection of variables with an identical mean that is defined by α_k, d_1, and d_2.

We define $e(d_1, d_2)$ to be the average of the elements in a given collection, $U(d_1, d_2)$. Therefore, $e(d_1, d_2)$ is itself a random variable with expectation uniquely defined by the parameters α_k. The variance of these statistics depends on many variables, including the number of elements in $U(d_1, d_2)$, the size of $N_t(o)$, the base rates associated with item pairs in the collection, and the variance parameters of varying effects in our model.

In generating posterior predictions, we try to keep as many components as possible constant. Thus, we let $X_{ijt}^*(o)$ and $Y_{ijt}^*(o)$ be simulated counts using the estimated base rates $\hat{\mu}_{ijt}$ and the appropriate sample sizes, $N_t(o)$, of the SCPC data. Simulated results can be used to generate $p_{ijt}^*(o), q_{ijt}^*(o)$ and, in turn, $S_{ijt}^*(o, o')$ and $T_{ijt}^*(o, o')$. For each set of simulated data, we can create $U^*(d_1, d_2)$ and calculate its average, $e^*(d_1, d_2)$. Corresponding averages for the observed data are noted as $\bar{e}(d_1, d_2)$.

We simulate 500 independent sets of SCPC data and thus generate 500 sets of $e^*(d_1, d_2)$ for each of the 52 unique pairs of (d_1, d_2). Figure 4 compares the observed statistics $\bar{e}(d_1, d_2)$ to the mean and 95 percent prediction intervals, defined by central-most 475 values, of the simulated data. Appropriate specification of model parameters should lead to general consistency between the observed statistics and their simulated distributions. As expected, when $d_1 = d_2$, the expectation of $e^*(d_1, d_2)$ is zero. Otherwise, as $d_2 - d_1$ increases, the mean increases, because

our model predicts generally larger discrepancies when distances are greater. In addition, the obvious trends within (d_1, d_2) for a fixed value of $d_2 - d_1$ result from the fact that plotted results are organized according to increasing values of d_1 and d_2, and the largest effects are observed for negative values of d.

Only 47 out of 52 (90.4 percent) posterior intervals contain the observed means, although it is important to remember that confidence intervals are not independent of one another, since data for a given pair of items goes into results for $(d_{ij}(o), d_{ij}(o'))$, and $(-d_{ij}(o), -d_{ij}(o'))$. Although the observed patterns mimic the simulated trends, they are sometimes shifted up ($d_2 - d_1 = 11$) or down ($d_2 - d_1 = 10$). This too might be explained by dependence of observations or the inability of our model to properly capture effect variances, which, as noted throughout this paper, likely have complicated structures. Indeed, the use of medians of $U(d_1, d_2)$ rather than means leads to better agreement between observed and simulated results, perhaps indicating that variance in effects is more heterogeneous across pairs than our model allows. Nevertheless, the SCPC statistics seem to generally match the simulated trends, suggesting, at least, that our parametric trend curve, given by α_k, is adequate.

6.2 Implications

Given that our model seems to do an adequate job of capturing mean effects as a function of the relative location of items, we explore implications for survey data. The overall trend determined as a function of the distance between two items, $d_{ij}(o)$, is given by the estimated parameters α_k. It is useful to consider effects relative to a baseline, which we choose to be an ordering in which $d = 1$. Thus, the solid black line in Figure 5 shows the expected multiplicative effect on the odds relative to an ordering in which $d_{ij} = 1$, $\exp\left\{\sum_{k=1}^{3} \alpha_k d^k - \sum_{k=1}^{3} \alpha_k\right\}$. The shape of this trend line closely matches the trend in medians shown in Figure 3, confirming the empirical findings.

Figure 5 also shows the variation in trends due to the varying slope term, which allows pairs of items to have different trend lines across different orders. The variation in these trend lines is substantial. As an example, positive values of α_{ij} tend to increase the relative influence of the linear term and at the same time increase the slope, resulting in more drastic differences in $f_{ij}(d)$ across d with negative values of $f_{ij}(d)$ when $d > 0$ and positive values when $d < 0$. When $\alpha_{ij} < 0$, the quadratic term takes precedence, resulting in an $f_{ij}(d)$ with relatively small, but positive values for all d, although with the largest values when the distance is large ($|d|$ is big). Finally, Figure 5 also shows the relative variation due to the order-specific effect. In particular, we see that for a fixed distance between items i and j, other changes in the ordering change the odds by factors from 0.8 to 1.2.

Multiplicative effects on odds are revealing, but it is hard to take those results and visualize practical effects on estimated relative rating distributions under different orderings. In Figure 6, we use the fitted trend line to show

expected effects on relative rating distributions as the relative locations of the items changes for four examples defined by different base rates. The expected changes under different relative locations are non-trivial. The greatest changes relate to the probability that item i is greater than item j as item j moves down the battery relative to item i. The greater the distance, the more likely that item j is given a higher rating than item i. Comparing extreme cases, in which item i and j switch spots at the very top and bottom of the list, shows differences ranging from 0.05 to 0.09 in the likelihood of giving item i a higher rating than item j.

The changes observed in Figure 6 are certainly large enough to bias comparisons of groups in which data were collected under different orderings. For example, if a researcher were interested in making inferences about relative ratings of two items in one-subpopulation to those in a second sub-population, but responses for the two were collected under different battery orderings, such comparisons would likely be unfair. While the order-specific differences are impossible to predict, changes in reported attitudes based on item distances can be adjusted for. Otherwise, sub-populations that have similar tendencies may be distinguished and marked as different solely due to different relative locations of items in the survey batteries.

7 Discussion

In this paper, we use data from six different question batteries that were asked of survey respondents under different orderings to study how the ordering of items in the battery influenced responses. We found evidence that different orderings have distinct effects on tendencies in relative item ratings. In particular, we see a pattern that depends on the relative locations of the items in the battery. At the very least, researchers who use similar Likert-scale batteries should be aware of these effects when writing questionnaires and analyzing survey data.

Further research into the nature of these effects can take many forms. Using different sets of items to check for robustness, developing more elaborate parametric models and dependence structures for effects under different orderings are perhaps most useful. Simple extensions, such as considering Likert-scales with a greater number of response options or longer batteries would also be interesting. Expanding analysis to include not only relative locations, but other predictable aspects of the battery, such as the actual location of the items, could also help give us insight into the dynamics at hand. More nuanced analyses could potentially distinguish trends between item pairs in which lower ratings are considered positive to those in which lower ratings are negatively associated. In general, psychological explanations for the underlying cognitive processes involved would be interesting and could potentially help in future questionnaire design.

From the viewpoint of a survey methodologist, how best to collect responses for a series of Likert-scale questions

is unclear. One option for modeling is to allow for an ordering effect in the stochastic model used to describe relative response tendencies. The use of randomization of ordering will also be effective at averaging out the influence of ordering effects as long as sample sizes are sufficiently large. Finally, because our finding is that effects are greater when distances between items are greater, it is possible to use shorter Likert batteries. However, it is unclear how the order of the blocked, Likert batteries would affect results. Again, more research is necessary to ascertain the potential benefits of such an approach.

References

Agresti, Alan. 2002. *Categorical Data Analysis*. New York, NY: Wiley, 2nd ed.

Bowling, Ann. 2005. "Mode of Questionnaire Administration Can have Serious Effects on Data Quality." *Journal of Public Health* 27(3): 281–291.

Bradburn, Norman, Seymour Sudman, and Brian Wansink. 2004. *Asking Questions: The Definitive Guide to Questionnaire Design - For Market Research, Political Polls, and Social and Health Questionnaires*. San Francisco, CA: Jossey-Bass.

Bradburn, Norman M., and William M. Mason. 1964. "The Effect of Question Order on Responses." *Journal of Marketing Research* 1(4): 57–61.

Chan, Jason C. 1991. "Response-Order Effects in Likert-Type Scales." *Educational and Psychological Measurement* 51(3): 531–540.

Dawes, John. 2008. "Do Data Characteristics Change According to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales." *International Journal of Market Research* 50(1): 61–77.

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York, New York: Cambridge University Press.

Knauper, Barbel. 1999. "The Impact of Age and Education on Response Order Effects in Attitude Measurement." *Public Opinion Quarterly* 63(3): 347–370.

Likert, Rensis. 1932. "A Technique for the Measurement of Attitudes." *Archives of Psychology* 140: 1–55.

McFarland, Sam G. 1981. "Effects of Question Order on Survey Responses." *Public Opinion Quarterly* 45(2): 208–215.

Schuman, Howard, and Stanley Presser. 1996. *Question and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. Thousand Oaks, CA: Sage Publications.

Schwarz, Norbert, and Hans-J. Hipler. 1991. "Response Alternatives: The Impact of Their Choice and Presentation Order." In *Measurement Error in Surveys*, eds. Paul P. Biermer, Robert M. Groves, Lars E. Lyberg, Nancy A. Mathiowetz, and Seymour Sudman, 41–56. Wiley.

Schwarz, Norbert, and Hans-J. Hippler. 1995. "Subsequent Questions May Influence Answers to Preceding Questions in Mail Surveys." *The Public Opinion Quarterly* 59(1): 93–97.

Schwarz, Norbert, Barbel Kanuper, Hans-J. Hipler, Elisabeth Noelle-Neumann, and Leslie Clark. 1991. "Numeric Values May Change the Meaning of Scale Labels." *The Public Opinion Quarterly* 55(4): 570–582.

SCPC. Various Years. "Survey of Consumer Payment Choice." `https://www.bostonfed.org/publications/survey-of-consumer-payment-choice.aspx`.

Sigelman, Lee. 1981. "Question-Order Effects on Presidential Popularity." *Public Opinion Quarterly* 45(2): 199–207.

Siminski, Peter M. 2008. "Order Effects in Batteries of Questions." *Quality and Quantity* 42(4): 477–490.

Sudman, Seymour, Norman M. Bradburn, and Norbert Schwarz. 1996. *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco, CA: Jossey-Bass.

Tourangeau, Roger, Lance J Rips, and Kenneth Rasinki. 2000. *The Psychology of Survey Response*. Cambridge, UK: Cambridge University Press.

Table 1: The text for the six Likert-scale batteries used in the 2012 – 2014 SCPCs.

Characteristic	Question Text
Acceptance	Please rate how likely each payment method is to be ACCEPTED for payment by stores, companies, online merchants, and other people or organizations.
Cost	Please rate the COST of using each payment method. Examples: Fees, penalties, postage, interest paid or lost, subscriptions, or materials can raise the cost of a payment method. Cash discounts and rewards (like frequent flyer miles) can lower the cost of a payment method.
Convenience	Please rate the CONVENIENCE of each payment method. Examples: speed, control over payment timing, ease of use, effort to carry, ability to keep or store.
Security	Suppose a payment method has been stolen, misused, or accessed without the owners permission. Please rate the SECURITY of each method against permanent financial loss or unwanted disclosure of personal information.
Ease of Setting Up	Rate the task of getting or setting up each payment method before you can use it. Examples: getting cash at the ATM, length of time to get or set up, paperwork, learning to use or install it, or travel.
Access to Payment Records	Rate the quality of payment records offered by each payment method. Consider both paper and electronic records. Examples: proof of purchase, account balances, spending history, usefulness in correcting errors or dispute resolution, or ease of storage.

Table 2: The three different groups of instruments referenced in the SCPC along with the six different orderings presented at random to respondents. The different orderings reflect different permutations of the three instrument types.

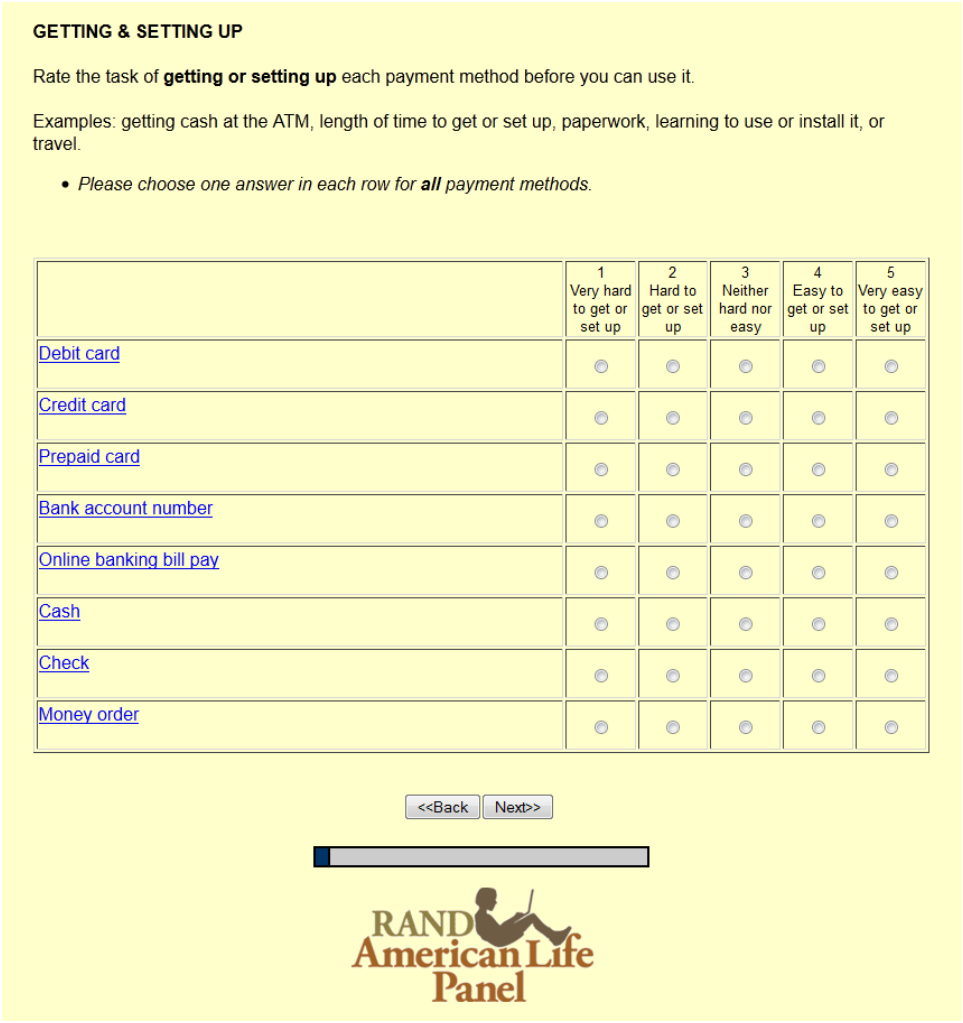
Instrument Group	**Instruments**		
Paper	Cash (C)	Check (Ch)	Money Order (MO)
Plastic	Credit (CC)	Debit (DC)	Prepaid Card (PC)
Online	Bank Acct. # Payments (BA)	Online Bill Payment (OB)	

Order 1	C	Ch	MO	CC	DC	PC	BA	OB
Order 2	C	Ch	MO	BA	OB	CC	DC	PC
Order 3	CC	DC	PC	C	Ch	MO	BA	OB
Order 4	CC	DC	PC	BA	OB	C	Ch	MO
Order 5	BA	OB	C	Ch	MO	CC	DC	PC
Order 6	BA	OB	CC	DC	PC	C	Ch	MO

Table 3: Fitted model parameter estimates ($\times 100$). For non-varying parameters, standard errors are included in parentheses.

	Non-Varying			**Varying Slope**				**Order-Specific Effect**			
Parameter	α_1	α_2	α_3	σ_{s0}	σ_{a0}	σ_{s1}	σ_{a0}	τ_{s0}	τ_{a0}	τ_{s1}	τ_{a1}
Estimate	-1.28 (0.27)	1.299 (0.12)	-0.06 (0.04)	1.43	0.55	0.00	0.00	9.04	3.66	5.38	0.00

Source: Author's calculations.



GETTING & SETTING UP

Rate the task of **getting or setting up** each payment method before you can use it.

Examples: getting cash at the ATM, length of time to get or set up, paperwork, learning to use or install it, or travel.

- *Please choose one answer in each row for **all** payment methods.*

	1 Very hard to get or set up	2 Hard to get or set up	3 Neither hard nor easy	4 Easy to get or set up	5 Very easy to get or set up
Debit card	○	○	○	○	○
Credit card	○	○	○	○	○
Prepaid card	○	○	○	○	○
Bank account number	○	○	○	○	○
Online banking bill pay	○	○	○	○	○
Cash	○	○	○	○	○
Check	○	○	○	○	○
Money order	○	○	○	○	○

Figure 1: Example of Likert-scale battery from the Survey of Consumer Payment Choice. Source: Author's calculations.

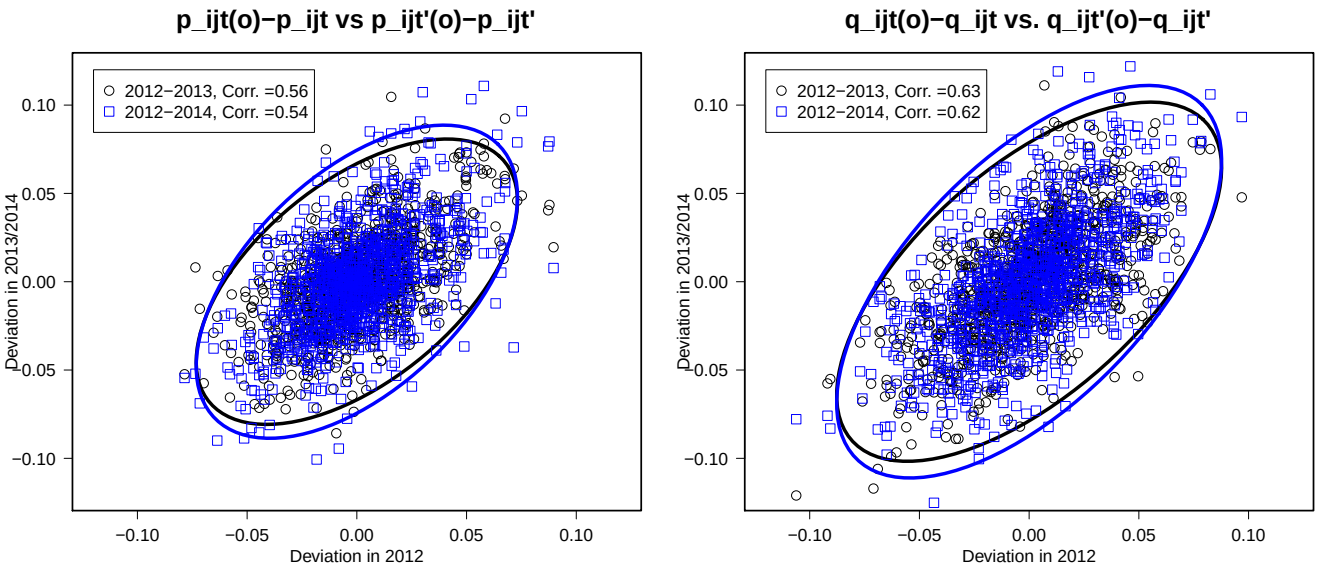


Figure 2: $p_{ijt}(o) - p_{ijt}$ and $q_{ijt}(o) - q_{ijt}$ across $t = 1$ and $t = 2, 3$. Ellipses correspond to 98 percent confidence interval under the assumption of a bivariate Normal distribution.

Source: Author's calculations.

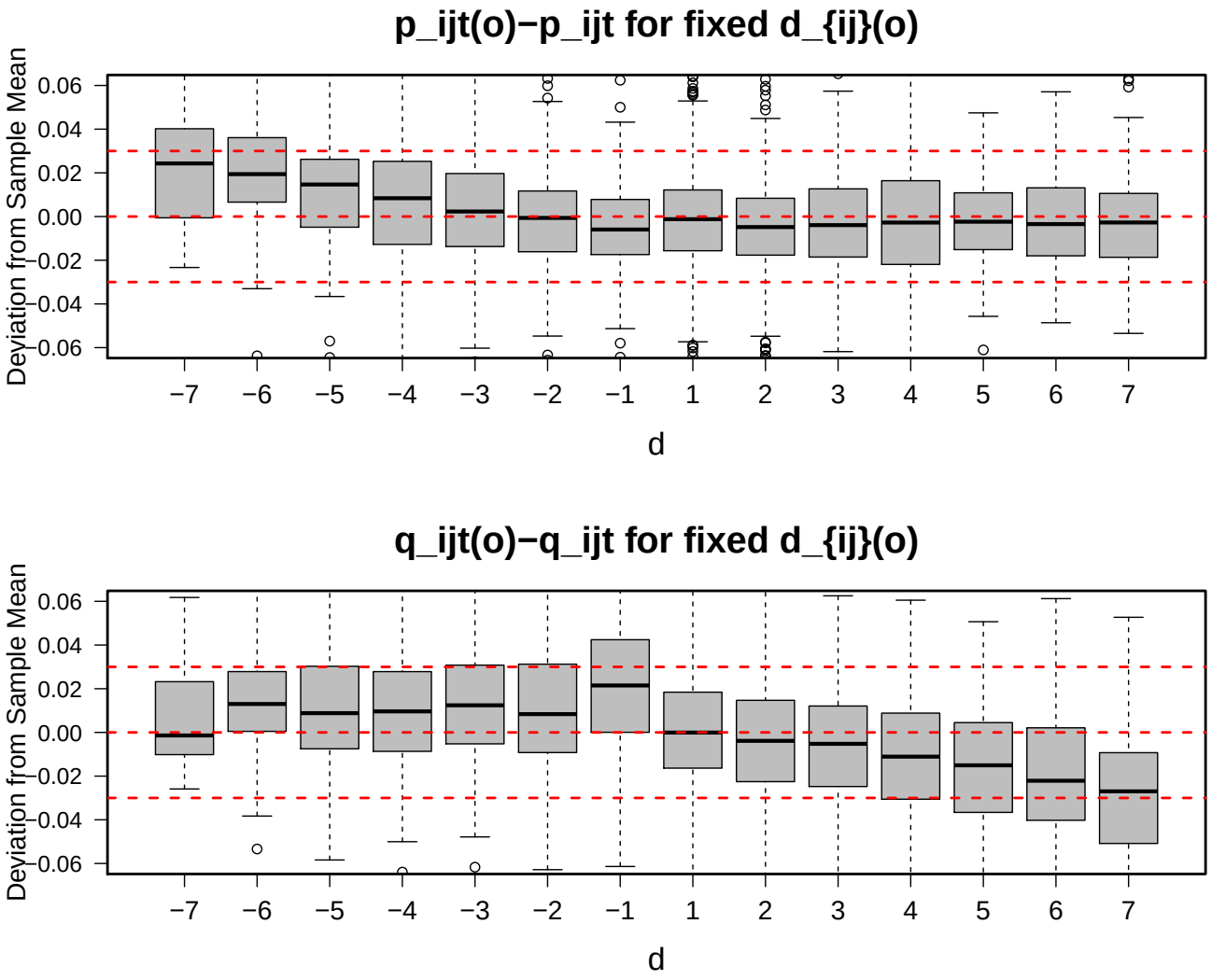


Figure 3: Distribution of deviations for different relative locations, indexed by d.
Source: Author's calculations.

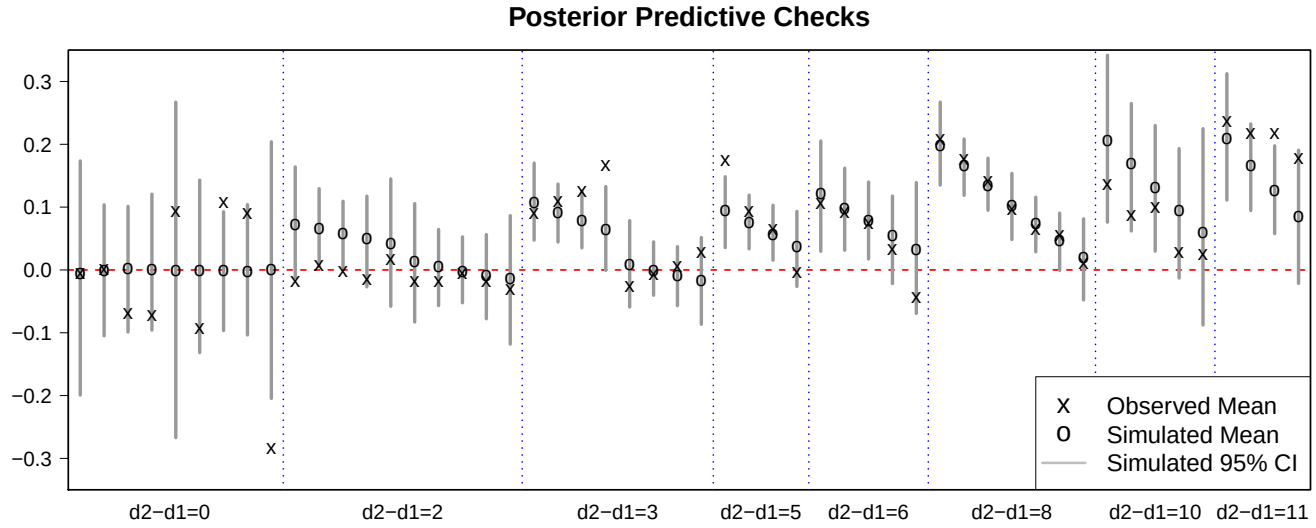


Figure 4: $\bar{e}(d_1, d_2)$ vs. 95 percent posterior intervals for $e^*(d_1, d_2)$.
Source: Author's calculations.

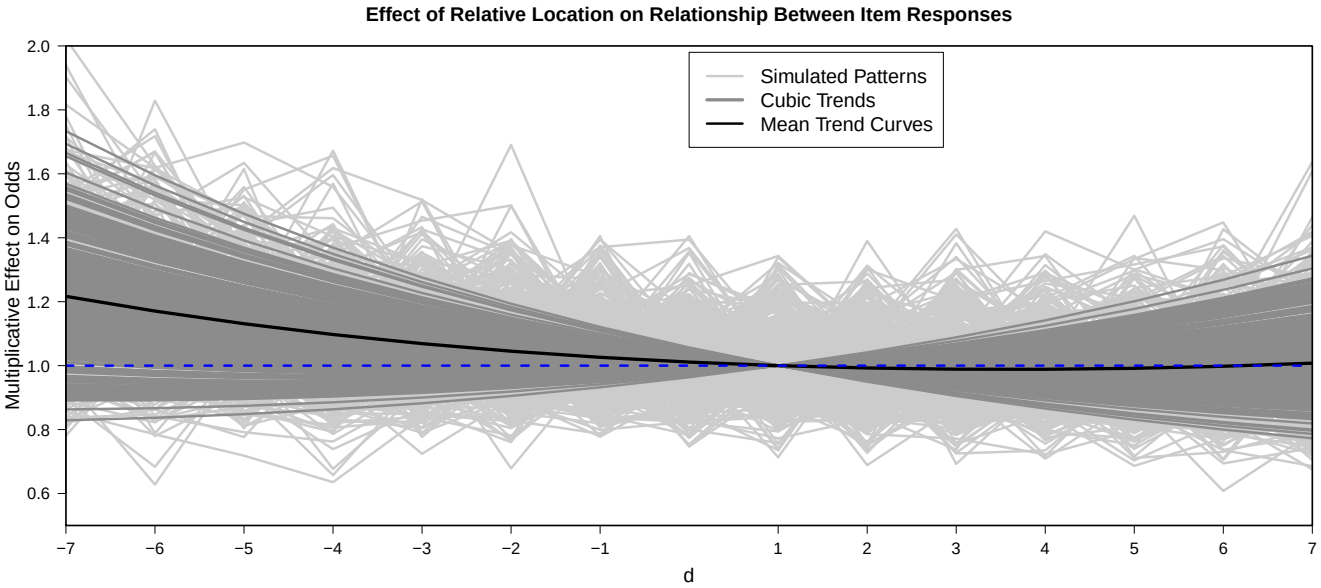


Figure 5: Relative log-odds compared to hypothetical ordering for which $d_{ij}(o) = 1$.
Source: Author's calculations.

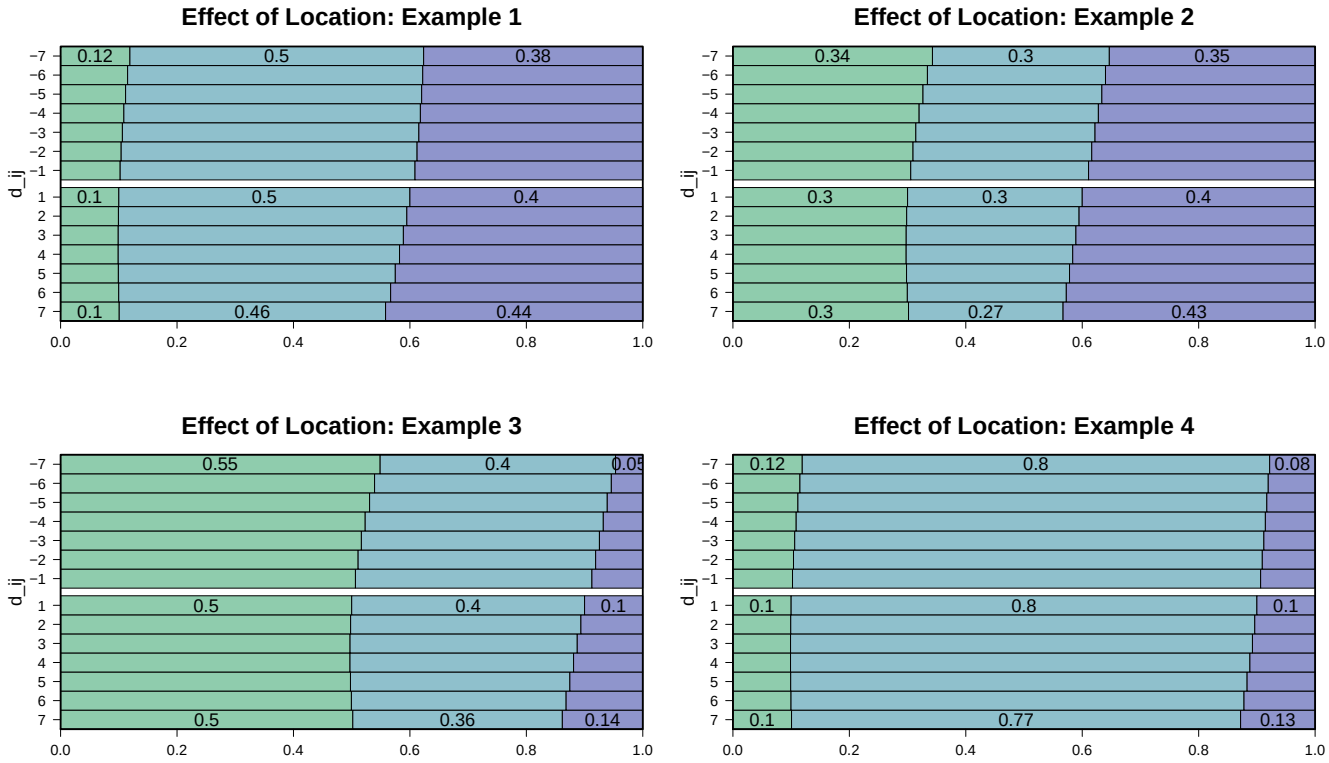


Figure 6: Expected distributions of relative ratings under different relative locations for four examples. The left-most bar corresponds to item i having a rating less than that of item j. The middle bar is the frequency with which item i and item j have the same rating. The right-most bar is the frequency with which item j is rated higher than item i.

Source: Author's calculations.