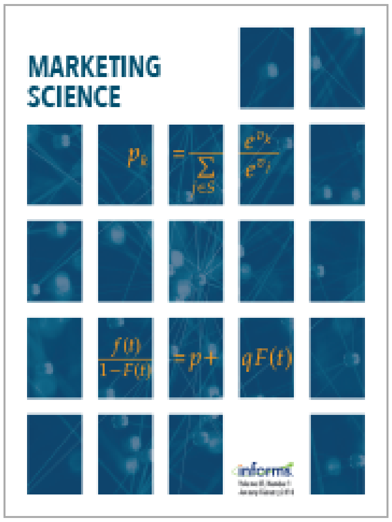


Marketing Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Models of Sequential Evaluation in Best-Worst Choice Tasks

Tatiana Dyachenko, Rebecca Walker Reczek, Greg M. Allenby

To cite this article:
Tatiana Dyachenko, Rebecca Walker Reczek, Greg M. Allenby (2014) Models of Sequential Evaluation in Best-Worst Choice Tasks. Marketing Science 33(6):828-848. https://doi.org/10.1287/mksc.2014.0870

MARKETING SCIENCE
Vol. 33, No. 6, November–December 2014, pp. 828–848
ISSN 0732-2399 (print) | ISSN 1526-548X (online)

http://dx.doi.org/10.1287/mksc.2014.0870

Models of Sequential Evaluation in Best-Worst Choice Tasks

Tatiana Dyachenko
McDonough School of Business, Georgetown University, Washington, DC 20057,
tld40@georgetown.edu

Rebecca Walker Reczek, Greg M. Allenby
Fisher College of Business, Ohio State University, Columbus, Ohio 43210
{reczek.3@osu.edu, allenby.1@osu.edu}

We examine the nature of best-worst data for modeling consumer preferences and predicting their choices. We show that contrary to the assumption of widely used models, the best and worst responses do not originate from the same data generating process. We propose a sequential evaluation model and show that people are likely to engage in a two-step evaluation process and are more likely to select the worst alternative first before selecting the best. We find that later choices have systematically larger coefficients as compared to earlier choices. We also find the presence of an elicitation effect that leads to larger coefficients when respondents are asked to select the worst alternative, meaning that respondents are surer about what they like least than what they like most. Finally, we investigate global inference retrieval in choice tasks, which can be represented by the central limit theorem and normally distributed errors, versus episodic retrieval represented by extreme value errors. We find that both specifications of the error term are plausible and advise using the proposed sequential logit model for practical reasons. We apply our model to data from a national survey investigating the concerns associated with hair care. We find that accounting for the sequential evaluation in the best-worst tasks and the presence of the scaling effects leads to different managerial implications compared to the results from currently used models.

Keywords: Bayesian estimation; choice modeling; mathematical psychology
History: Received: June 1, 2012; accepted: May 30, 2014; Preyas Desai served as the editor-in-chief and Jean-Pierre Dubé served as associate editor for this article. Published online in *Articles in Advance* September 8, 2014.

1. Introduction

Experimental choice tasks are commonly used in marketing research to get respondents to discriminate among a set of options. Data generated from choices are often easier for respondents to provide because people make choices every day, and they avoid the tendency people have to evaluate everything as being good or acceptable. However, choice data suffer from being less informative than continuously valued data. At an individual level, respondents also become fatigued after being asked a dozen or more questions about the same construct. A challenge in the use of choice data in marketing is the limited information it reveals to the analyst.

Researchers have dealt with data deficiency at the individual level in a number of ways. The most prevalent way is to employ Bayesian random-effect models that pool information across respondents while retaining the ability to study an individual respondent's choice model coefficients (see Rossi et al. 2005). Although this approach helps shore up data deficiencies at the individual level, it rarely leads to coefficients possessing small levels of uncertainty because of the sparseness of information available at the individual level. An alternative approach is to collect additional data by asking respondents to indicate more than just their top choice. Examples include the use of the exploding logit model of Chapman and Staelin (1982), in which a full ranking of the choice alternatives is obtained, and the best-worst format (Louviere 1991, Finn and Louviere 1992) for data collection, in which the top and bottom choices are collected. According to the Sawtooth Software company, a leading organization offering software for conjoint and best-worst analyses, in 2012, 54% of software users have utilized the best-worst tool, called MaxDiff (Sawtooth Software 2012).

In this paper, we examine the nature of best-worst data for modeling consumer preferences and predicting their choices. We present evidence indicating that the best and worst responses do not originate from the same data generating process. Contrary to existing models that treat the worst data as the minimum and the best as the maximum of a single evaluative process, we find evidence of systematic departures along two dimensions. First, we find evidence indicating the presence of sequential evaluation where respondents are more likely to select the worst alternative, followed by the best, from the remaining choice alternatives. We also find that attribute coefficients are systematically greater in the second decision, indicating greater surety in choice.

The second finding is the presence of an elicitation effect that leads to larger coefficients when asked to select the worst choice alternative. We find that respondents are surer about what they like least than what they like most. This finding interacts with the first finding in that its presence depends on the ability to control for the decision sequence. For the minority of respondents who select the best alternative first, their coefficients for the worst alternative are larger because of both the sequence and the elicitation effects. Conversely, for the majority of respondents who select the worst alternative first, the sequence and elicitation effects tend to partially cancel each other out.

We also investigate the presence of global inference versus an episodic memory retrieval process. Episodic memory retrieval is consistent with the use of maximal extreme value theory when respondents are asked to indicate their most preferred option. Similarly, when asked for the least preferred option, we expected to find minimum extreme value theory to fit the data best. Both distributions are derived from selecting a best or worst event from a set of events. Global inference retrieval (Tulving 1972, Carlston 1980) is characterized by an averaging process leading to a normal distribution of errors by the Central Limit Theorem. To our knowledge, this investigation is unique to the literature because assumed error distributions tend to be made either out of convenience or for theoretical considerations and are not based on empirical evidence.

We examine data from a national survey investigating consumers' concerns associated with hair care. The data were collected to study the prevalence of needs among a subset of the U.S. population. Our results caution against the use of dual elicitation schemes such as best-worst analysis unless accompanied by models accommodating sequential and elicitation effects. We find that the presence of these effects leads to different inferences about individual-level concerns compared to the results from currently used models without these effects.

The remainder of this paper is organized as follows. Section 2 reviews models of choice consistent with economic theory where best and worst data are thought to arise from the same preference ordering. Section 3 introduces a sequential model of evaluation and shows that the traditional models are special cases of that model. An empirical application is described in §4. Coefficient estimation and results are presented in §5 where we show that not accounting for the effects of sequential evaluation inflates coefficient estimates. Implications for marketing research and concluding comments are offered in §6.

2. Models of Single Evaluation

Models involving the single evaluation of alternatives in a choice set are based on an economic view of choice that assumes respondents have a predetermined preference ordering for the alternatives (Manski 1977). The error terms in these models represent private information not revealed to the analyst that partially determines preferences. We consider three models of single evaluation used to evaluate best-worst choice tasks:

1. A partial ranking model that uses an exploded logit specification (Chapman and Staelin 1982) based on an extreme value error term.
2. A double censored probit model based on an error term that is normally distributed.
3. The MaxDiff model (Louviere 1991, Finn and Louviere 1992) that assumes respondents perform a pairwise comparison of the alternatives and select a pair that has maximum difference between the items.

The first two models assume that respondents have a latent utility vector z for the choice alternatives and then select the alternative with the highest utility as the most preferred and the one with the lowest utility as the least preferred. When the choice alternatives are not product offerings, the latent vector z can be thought of as the level of feeling, or intensity, associated with an issue, object, or construct. The issue with the highest level of concern is selected as the "most important" or "best describing," and the item with the lowest level would be chosen as the "least important" or "worst describing." Models 1 and 2 only differ in terms of the distribution of the error term representing information not revealed to the analyst.

The model of one best-worst task assuming an extreme value distribution is

$$\begin{aligned} z &= X\beta + \varepsilon, \quad \varepsilon \sim \text{MaxEV}(0, I), \\ y_{\text{best}} &= \sum_{k=1}^{p} k \times I(z_k = \max(z)), \\ y_{\text{worst}} &= \sum_{k=1}^{p} k \times I(z_k = \min(z)), \end{aligned} \tag{1}$$

where "MaxEV" is Extreme Value Type I distribution (Johnson and Kotz 1995), also known as maximum extreme value distribution; β is the vector of importance weights; X is a design matrix for one observation with p alternatives in the set; ε is a vector of error term realization; and y_{best} and y_{worst} are the best and worst responses.

The evaluation of choice probabilities for the best and worst responses in the model with normal distribution of the error term requires the evaluation of the error region R_y corresponding to the latent utilities, z, consistent with the observed best and worst responses

$$\Pr\big(y = (y_{\text{best}}, y_{\text{worst}}) \mid X, \beta, \Sigma\big) \\ = \Pr(z \in R_y \mid X, \beta, \Sigma) = \int_{R_y} \phi(z \mid X, \beta, \Sigma)\, dz. \quad (2)$$

The direct evaluation of the integral can be avoided with Bayesian estimation using data augmentation (Tanner and Wong 1987). The advantage of using data augmentation is especially noticeable when the number of choice alternatives is large and the integration would be numerically intensive. A disadvantage is that bypassing the integral evaluation leads to an inability to evaluate the likelihood, which is needed for model assessment.

The evaluation of the likelihood is greatly simplified by assuming an extreme value distribution, leading to the logit model. The choice probability for a particular ordering of choices can be shown equal to (see Chapman and Staelin 1982) the following:

$$\Pr\big(z_{(p)} > z_{(p-1)} > \cdots > z_{(2)} > z_{(1)}\big) = \prod_{i=1}^{p} \frac{\exp(x_i'\beta)}{\sum_{k=i}^{p} \exp(x_k'\beta)}. \quad (3)$$

When only the best and worst responses are known, we observe only a partial ordering of the other items, and the likelihood is defined by marginalizing over all possible permutations of the choice set items not chosen as best or worst

$$\Pr(y_{\text{best}} = i, y_{\text{worst}} = j) \\ = \Pr\big(z_i = \max\{z_k\}, z_j = \min\{z_k\},\ k = 1, \ldots, p;\ i \neq j\big) \\ = \sum_n \Pr\big(z_i > \text{Perm}_n[z_{k,-i,-j}, (p-2)] > z_j\big). \quad (4)$$

In the case of a five-alternative choice task, where the respondent selects i as the best alternative and j as the worst, the probability for one of the six possible unobserved orderings with permutation Perm of the remaining items (a, b, c) for extreme value error is

$$\Pr\big(z_i > \text{Perm}[z_{k,-i,-j}, (p-2)] > z_j\big) \\ = \Pr\big(z_i > z_a > z_b > z_c > z_j\big), \quad (5)$$

which is calculated based on Equation (3).

The MaxDiff model (Louviere 1991, Finn and Louviere 1992) assumes that a respondent evaluates all possible pairs of the presented alternatives by calculating the differences in retrieved importance weights between the attributes and selects the pair with the highest value. The error term is assumed to follow Type I extreme value distribution. For example, for the pair of alternatives i and j, the model is as follows:

$$z_{ij} = (x_i\beta - x_j\beta) + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \text{EV}(0, 1),$$

$$\Pr\big(y_{\text{best}} = i, y_{\text{worst}} = j \mid X, \beta, i \neq j\big) \\ = \Pr\big(z_{ij} > z_{kl},\ k \neq i,\ l \neq j,\ k \neq l;\ k, l = 1, \ldots, p\big) \\ = \Pr\big((x_i\beta - x_j\beta) + \varepsilon_{ij} > (x_k\beta - x_l\beta) + \varepsilon_{kl}; \\ k \neq i,\ l \neq j,\ k \neq l;\ k, l = 1, \ldots, p\big) \\ = \frac{\exp(x_i\beta - x_j\beta)}{\sum_{\{k,\,l\},\,k\neq l} \exp(x_k\beta - x_l\beta)}, \quad (6)$$

where x_i and x_j are the corresponding rows of the design matrix X. However, a choice set with five alternatives, such as the one in our data set, would require evaluation of 20 pairs according to the MaxDiff model. Although the model is similar to the single evaluation logit, we see value in including the MaxDiff model in our model comparison because of its popularity in applied studies.

3. Sequential Evaluation Models

In this section, we present a model of sequential evaluation in best-worst tasks. The model has three main components that are developed based on specific theories from the psychological literature. These components reflect different assumptions for the data generating processes for these tasks than do the models of single evaluation. Under specific conditions, the models of sequential evaluation nest the single evaluation model discussed in the previous section, thus allowing for the more general case.

First, we propose that respondents in best-worst tasks do not engage in single evaluation of the items but make decisions about the two questions at hand, "select-the-best" and "select-the-worst," and do so using a sequential process. The assumption of sequential evaluation reflects possible task simplification that respondents might be performing by solving two small tasks instead of one big single evaluation task and by reducing the set of items to process in the second choice. Thus, our model will include data generating processes for two decisions: selecting the best and selecting the worst alternatives.

Regardless of possible sequence of the decisions—best then worst versus worst then best—there are several ways that two sequential data generating events can be related to each other in one model. A naive approach assumes independence of choices in best-worst tasks, which means that the same item can be both best and worst. In most commercial applications, the data collection software does not allow the selection of items as both best and worst. As a result, the order of the decision sequence matters in the model, with the first decision involving the entire choice set and the second decision involving a choice set with the first choice eliminated.

3.1. Decision Sequence (γ)

We assume that there is a latent decision sequence for a respondent h indexed by a dummy variable θ_h that is equal to 1 if the best choice is selected first, followed by the worst choice. Thus, the general form of the choice probability for the model is

$$\begin{aligned}&\Pr(y_{\text{best}}=i, y_{\text{worst}}=j \mid X, \beta_h, \theta_h)\\&=\theta_h \Pr(y_{\text{best}}=i \mid X, \beta_h)\Pr(y_{\text{worst}}=j \mid X_{-i,}, \beta_h)\\&\quad+(1-\theta_h)\Pr(y_{\text{worst}}=j \mid X, \beta_h)\Pr(y_{\text{best}}=i \mid X_{-j,}, \beta_h). \quad (7)\end{aligned}$$

where subscripts $-i$ and $-j$ indicate the exclusion of the corresponding row from design matrix X. Model (7) can accommodate an observed as well as probabilistically inferred sequence of decisions. Using the observed sequence of choices requires a strong assumption that the observed data for the order of selection are a good representation of the underlying latent processes of sequential evaluation of alternatives. We test for this assumption on the actual data set by comparing performance of the models. For the model with inferred decision sequence for each respondent h, we specify the data generation mechanism as follows:

$$\theta_h \sim \text{Bernoulli}(\gamma), \quad (8)$$

where parameter γ represents the overall probability that the respondents select the best alternative first. The identification of this parameter comes from the property that γ is conditionally independent of the other parameters in the model given the sequence indicator vector $\theta = \{\theta_h\colon h = 1, \ldots, H\}$. Thus, we can integrate out latent parameter θ and make inferences about γ.

3.2. Sequential Effect (ψ)

Another important aspect of the sequential evaluation process is the potential difference in decision consistency from the first to the second choice. The idea of sequential scalars was earlier proposed by Ben-Akiva et al. (1992). In their work, the change in scale of parameters is a function of the ranking "depth." Their estimation was performed on the data from tasks where respondents selected the best option three times from subsets that excluded the previously selected alternative. Thus, the underlying data generating mechanism was assumed to be the same in these ranking tasks. Our model, in contrast, is more general and allows for possible differences in the data generating processes in the two best-worst tasks. Moreover, we include the scaling effect ψ as a proportion between the first and second decisions, irrespective of whether the sequence of decision was from best to worst or from worst to best.

$$\begin{aligned}&\Pr(y_{\text{best}}=i, y_{\text{worst}}=j \mid X, \beta_h, \theta_h, \psi_h)\\&\quad=\theta_h \Pr(y_{\text{best}}=i \mid X, \beta_h)\Pr(y_{\text{worst}}=j \mid X_{-i,}, \psi_h\beta_h)\\&\quad\quad+(1-\theta_h)\Pr(y_{\text{worst}}=j \mid X, \beta_h)\\&\quad\quad\cdot\Pr(y_{\text{best}}=i \mid X_{-j,}, \psi_h\beta_h), \quad (9)\end{aligned}$$

where the coefficient vector β_h becomes $\psi_h\beta_h$ in the second decision. If the parameter ψ_h is greater than one, then there is more certainty in the second decision. This improvement in certainty could be expected from the reduction in the set size between the first and second tasks.

3.3. Elicitation Effect (λ)

The assumption of sequential evaluation in best-worst tasks allows incorporation of another effect, which we call an elicitation effect, that reflects the possibility that the "select-the-best" and "select-the-worst" question may lead to different processes

$$\beta_{\text{worst}} \neq \beta_{\text{best}}.$$

Whereas economic theory assumes that preferences are known to the respondents, stable and retrievable during choice tasks, psychological theory suggests this may not be true. Research offers evidence that preferences and importance weights are constructed and are context dependent (Bettman et al. 1998, Slovic 1995, Tversky and Simonson 1993). The psychology literature shows that respondents are likely to engage in selective information search, characterized by paying attention to information that would help confirm a proposition or hypothesis that is created by the context. This process is known as hypothesis testing (Snyder 1981, Snyder and Swann 1978, Hoch and Ha 1986), or confirmatory selectivity bias.

We believe that the two elicitation modes, or the questions, create two frames under which this "preferential evidence gathering" would be expected. These frames may trigger different associations in respondents' memories and thus generate two possibly different memory samples that are congruent with each of the questions, giving rise to different preference parameters β_{best} and β_{worst} in these two tasks.

We consider the unrestricted model where importance weights constructed in the best and worst responses differ from each other.

$$\begin{aligned}&\Pr(y_{\text{best}}=i, y_{\text{worst}}=j \mid X, \beta_{h,\text{best}}, \beta_{h,\text{worst}}, \theta_h, \psi_h)\\&=\theta_h \Pr(y_{\text{best}}=i \mid X, \beta_{h,\text{best}})\Pr(y_{\text{worst}}=j \mid X_{-i,}, \psi_h\beta_{h,\text{worst}})\\&\quad+(1-\theta_h)\Pr(y_{\text{worst}}=j \mid X, \beta_{h,\text{worst}})\\&\quad\cdot\Pr(y_{\text{best}}=i \mid X_{-j,}, \psi_h\beta_{h,\text{best}}). \quad (10)\end{aligned}$$

A more parsimonious model relates the best to the worst coefficients through a scaling parameter:

$$\beta_{\text{worst}} = \lambda\beta_{\text{best}} = \lambda\beta,$$

where $\lambda > 1$ indicates overall, across all items, greater certainty for the worst choice than for the best choice.

Ben-Akiva et al. (1992) showed that the high ranked items, or best choices, are more certain than the lower ranked ones. There is also research that has shown that people are often more sure of what they do not like than what they do like (Meloy and Russo 2004, Lyubomirsky and Lee 1999). Although we investigate the direction of the elicitation effect in best-worst tasks, in general, this parameter can be thought of as an overall adjustment present between two contexts. We note that the sequential model with sequential and elicitation effects can be shown to reduce the MaxDiff model and other models (Marley and Louviere 2005, Marley and Pihlens 2012) for certain values of θ and when $\lambda = \psi = 1$.

Our scaling parameter is conceptually different from the scaling parameters investigated in Ben-Akiva et al. (1992). In their work, respondents sequentially identified their best, second best, and third best alternatives. Our model imposes the change in the size of parameters as the result of change of the procedure: select-the-best versus select-the-worst.

3.4. Distributional Assumptions

The probabilities in (10) are obtained through distributional assumptions for an underlying error term in the model. The current practice is to use type I extreme value distribution for the error term as in Equation (1), which provides a closed-form expression for the probabilities of choice and, thus, is computationally significantly more convenient than the assumption of the normal distribution. We believe that it is important to rationalize the error distribution primarily based on conceptual differences among distributions, not computational convenience. In this paper, we propose one possible way to do so based in terms of semantic versus episodic memory (Tulving 1972).

According to semantic memory theory, respondents retrieve some form of aggregated, or summarized, information to generate responses. The support for this kind of inference retrieval, which we call "global" in the paper, comes from experimental literature. Carlston (1980) found that people generate judgments based at least partially on recalled inferences. (Hintzman 1986, p. 411) states that "the information retrieved from memory reflects the summed content of all activated traces responding in parallel." Although there is a discussion in the literature on what kind of algebraic operations can represent formation of these global inferences, we consider the most simple one—an averaging or summation of information, which would be consistent with the Central Limit Theorem in statistics. The Central Limit Theorem would justify the symmetric normal distribution for the error term because it would represent the averaging or summation over memory samples and association regardless of how the individual samples are generated, i.e., distribution of the individual samples.

On the other hand, if responses are made based on retrieved episodes and events rather than averaged information, then respondents could search their episodic memory for the most representative occasions or samples from memory to answer the question at hand. The modeling assumptions for that mechanism can be described by the compatibility principle (Fitts and Seeger 1953, Fitts and Deininger 1954, Shafir 1993, Shafir et al. 1993, Machin 2006) that states that there exists compatibility between stimulus (i.e., question) and response. We assume that respondents select the maximum value if asked to provide the "best" choice and select the minimum if asked to provide the "worst" choice. This process can be represented by the extreme value theory (Gumbel 2004), which says that the maximum of a sample has a maximum extreme value (MaxEV) distribution and the minimum of a sample has a minimum extreme value (MinEV) distribution. Thus, we model episodic memory retrieval with extreme value errors as MaxEV for the best choice and MinEV for the worst choice (see Appendix A for derivation). The maximum extreme value distribution has a tail to the right, and the minimum extreme value distribution has a tail to the left.

Thus,

$$\Pr(y_{\text{best}} = i \mid X, \beta) = \Pr(x_i\beta + \varepsilon_i > x_k\beta + \varepsilon_k,\ k \neq i) = \frac{\exp(x_i\beta)}{\sum_k \exp(x_k\beta)}, \tag{11}$$

and

$$\Pr(y_{\text{worst}} = j \mid X, \beta) = \Pr(x_j\beta + \varepsilon_j < x_l\beta + \varepsilon_l,\ l \neq j) = \frac{\exp(-x_j\beta)}{\sum_l \exp(-x_l\beta)}, \tag{12}$$

where x is the row vector corresponding to the item's row of a design matrix X in the discrete choice experiment.

For normally distributed errors related to global inference retrieval, the choice probabilities do not have a closed form but can be evaluated using the technique of Bayesian data augmentation.

$$\Pr(y_{\text{best}} \mid X, \beta, \Sigma) = \Pr(z_{\text{best}} \in R_{y_{\text{best}}} \mid X, \beta, \Sigma) = \int_{R_{y_{\text{best}}}} \varphi(z_{\text{best}} \mid X, \beta, \Sigma)\, dz_{\text{best}}, \tag{13}$$

$$\Pr(y_{\text{worst}} \mid X, \beta, \Sigma) = \Pr(z_{\text{worst}} \in R_{y_{\text{worst}}} \mid X, \beta, \Sigma) = \int_{R_{y_{\text{worst}}}} \varphi(z_{\text{worst}} \mid X, \beta, \Sigma)\, dz_{\text{worst}}. \tag{14}$$

To summarize, the modeling relationship between the sequential and elicitation effects described above

Table 1 Summary of the Models

	Parameters			Error term	
Model	Preference parameter	Elicitation scaling	Sequential scaling	EV (asymmetric)	Normal (symmetric)
			Single evaluation models		
Exploded logit	$\beta_{\text{best}} = \beta_{\text{worst}}$	—	—	SL	—
Probit with double censoring	$\beta_{\text{best}} = \beta_{\text{worst}}$	—	—	—	SP
MaxDiff	$\beta_{\text{best}} = \beta_{\text{worst}}$	—	—	SMD	—
			Sequential evaluation models		
Inferred sequence	$\beta_{\text{best}} = \beta_{\text{worst}}$	$\lambda = 1$	$\psi = 1$	SQL1	SQP1
	$\beta_{\text{best}} = \lambda\beta_{\text{worst}}$	λ estimated	$\psi = 1$	SQL2	SQP2
	$\beta_{\text{best}} \neq \beta_{\text{worst}}$	—	$\psi = 1$	SQL3	SQP3
	$\beta_{\text{best}} = \beta_{\text{worst}}$	$\lambda = 1$	ψ estimated	SQL4	SQP4
	$\beta_{\text{best}} = \lambda\beta_{\text{worst}}$	λ estimated	ψ estimated	SQL5	SQP5
	$\beta_{\text{best}} \neq \beta_{\text{worst}}$	—	ψ estimated	SQL6	SQP6
Observed sequence	$\beta_{\text{best}} = \beta_{\text{worst}}$	$\lambda = 1$	$\psi = 1$	SQL1s	SQP1s
	$\beta_{\text{best}} = \lambda\beta_{\text{worst}}$	λ estimated	$\psi = 1$	SQL2s	SQP2s
	$\beta_{\text{best}} \neq \beta_{\text{worst}}$	—	$\psi = 1$	SQL3s	SQP3s
	$\beta_{\text{best}} = \beta_{\text{worst}}$	$\lambda = 1$	ψ estimated	SQL4s	SQP4s
	$\beta_{\text{best}} = \lambda\beta_{\text{worst}}$	λ estimated	ψ estimated	SQL5s	SQP5s
	$\beta_{\text{best}} \neq \beta_{\text{worst}}$	—	ψ estimated	SQL6s	SQP6s

and the error term assumptions can be described as follows. The sequential effect ψ, resulting from changes in choice sets between the first and second decision, and elicitation effect λ, which is derived from confirmation bias literature and indicates the dependence from elicitation mode ("select the best" versus "select the worst"), reflect the changes in the location parameter in these models. The compatibility principle justifies the use of symmetric (normal), right-skewed (MaxEV) and/or left-skewed (MinEV) distributions around the location parameter. Thus, the sequential logit model with sequential and elicitation effects becomes (we omit subscript h here for clarity) the following:

$$\begin{aligned}\Pr(y_{\text{best}} = i, y_{\text{worst}} = j \mid \beta, \lambda, \psi, \theta) &= \theta \frac{\exp(x_i\beta)}{\sum_k \exp(x_k\beta)} \frac{\exp(-x_j\psi\lambda\beta)}{\sum_{l,-i} \exp(-x_l\psi\lambda\beta)} \\ &\quad + (1-\theta)\frac{\exp(-x_j\lambda\beta)}{\sum_l \exp(-x_l\lambda\beta)} \frac{\exp(x_i\psi\beta)}{\sum_{k,-j}\exp(x_k\psi\beta)}. \qquad (15)\end{aligned}$$

The sequential probit model with the same sequential and elicitation scaling is

$$\begin{aligned}\Pr(y_{\text{best}} = i, y_{\text{worst}} = j \mid \beta, \lambda, \psi, \theta) &= \theta \int_{R_{y_{\text{best}}}} \varphi(z_{\text{best}} \mid X, \beta, \Sigma)\, dz_{\text{best}} \\ &\quad \cdot \int_{R_{y_{\text{worst}}}} \varphi(z_{\text{worst}} \mid X_{-i}, \psi, \lambda, \beta, \Sigma)\, dz_{\text{worst}} \\ &\quad + (1-\theta) \int_{R_{y_{\text{worst}}}} \varphi(z_{\text{worst}} \mid X, \lambda, \beta, \Sigma)\, dz_{\text{worst}} \\ &\quad \cdot \int_{R_{y_{\text{best}}}} \varphi(z_{\text{best}} \mid X_{-j}, \psi, \beta, \Sigma)\, dz_{\text{best}}. \qquad (16)\end{aligned}$$

Table 1 summarizes the models and their elements that we use for model comparison. These models include proposed models with inferred sequence of decisions as well as with observed order. We discuss the elements of the estimation for these models in §4.

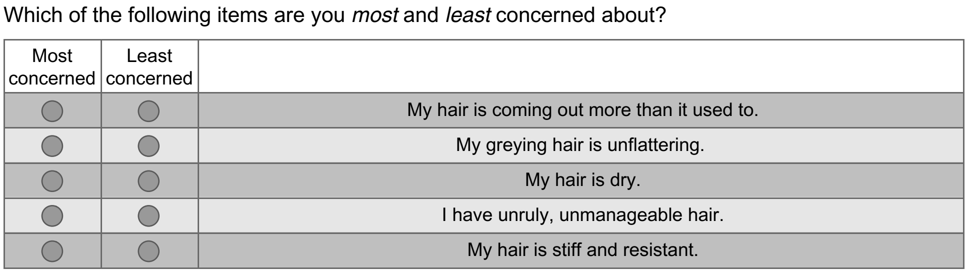
4. Empirical Application

We investigate the performance of the models in Table 1 using data from a national study of hair products related to the care of aging and damaged hair. We elected to emphasize the predictive validity of our results, realizing that a field test such as ours cannot provide the tight controls needed to positively relate the underlying theory to the parameter estimates. That is, we can establish that our estimates are consistent or inconsistent with the proposed theories but cannot prove that they arise from them only, as is the case in laboratory studies. Our analysis does not rule out other explanations for the data, and although we report on process measures like response latencies, we recognize they are insufficient for establishing the internal validity of our measures. The goal of our analysis is to caution against the use of single evaluation models in best-worst choice tasks by demonstrating the presence of various effects.

4.1. Data Description

Respondents recruited into the study were females age 50 years or older screened for inclusion in the survey. All respondents were given 15 best-worst tasks of five items from the list. Three sets of data were collected that differed in the presentation of the best-worst tasks. The first group of respondents ($n = 594$) had to complete the traditional best-worst tasks where both columns were presented at the same time as shown in

Figure 1 Example Screen of the Best-Worst Tasks



Which of the following items are you *most* and *least* concerned about?

Most concerned	Least concerned	
○	○	My hair is coming out more than it used to.
○	○	My greying hair is unflattering.
○	○	My hair is dry.
○	○	I have unruly, unmanageable hair.
○	○	My hair is stiff and resistant.

Figure 1. Data for two additional groups of respondents were obtained in which we manipulated the order of responses to be either best-worst ($n = 290$) or worst-best ($n = 299$). This was done by removing one of the columns from the task table in Figure 1 and sequentially providing two tables—one to select the best alternative and the second to select the worst. Table 2 displays a list of items reflecting concerns associated with hair care. We also collected other measures to assist in validating our model's assumptions and inferences. We discuss these data in more detail in §5.

4.1.1. Identification of the Scale Origin. Our proposed models have a problem common to many choice models—the models cannot identify all preference parameters without certain constraints. Traditionally, researchers have to set one of the parameters to zero and interpret the rest of the preference parameters as a difference from the fixed attribute. However, testing models like SQL3, SQP3, SQL6, and SQP6 would not be possible under such restrictions as they require a common scale for parameters β_{best} and β_{worst}. Thus, to facilitate testing of the hypotheses presented by the models in Table 1, it is critical to have a nonarbitrary scale origin, which we accomplish by introducing auxiliary data. Other problems with arbitrary scale origin have been discussed in the literature. Bacon and Lenk (2012) showed that the spread of the distribution of heterogeneity can be either overstated or understated depending on the actual correlational structure among variables. Also, Böckenholt (2004) demonstrated that the actual covariance structure of attributes cannot be recovered from the traditional choice models with relative attribute importance: one cannot infer from the relative judgment data whether any two variables are positively or negatively correlated.

Table 2 Concerns for the Best-Worst Tasks

1. My hair is too oily.
2. My hair is breaking.
3. My hair is stiff and resistant.
4. My hair is coming out more than it used to.
5. My hair is dry.
6. My hair is coarse and frizzy.
7. I have split ends.
8. I have unruly, unmanageable hair.
9. My hair's color is faded and dull.
10. My hair seems finer.
11. My hair is damaged from products, treatments, or sun.
12. My hair has been impacted by stress and hormone fluctuations.
13. My greying hair is unflattering.
14. I am seeing more of my scalp or receding hairline.
15. My hair lacks shine.

We follow the line of work of Bacon and Lenk (2012) and combine the information from the choice decision in the best-worst tasks with the information on the items measured on the absolute scale by asking respondents whether or not they were concerned with specific items. These absolute judgment data are collected on the binary scale and serve as a restriction for the preference parameters β. These data are modeled using a binary logit or probit likelihood for each item y_i, where $y_i = 1$ corresponds to a realization of a latent importance parameter being greater than zero

$$\Pr(y_i = 1) = \Pr(\beta_i + \varepsilon_i > 0). \tag{17}$$

The absolute judgment data allow us to statistically identify parameters for each of the 15 concerns in Table 2 by augmenting the likelihood of the models in Table 1 by the likelihood for the absolute judgment data. The concerns for which absolute judgment data are available are items 1, 2, 3, 4, 11, 12, 13, and 14. Approximately 94% of the respondents indicated they are concerned with at least one of the items.

4.2. Likelihood and Priors

The likelihood used to estimate the model parameters combines a best-worst likelihood from §§2 or 3 ℓ_{BW} with the likelihood for the absolute judgment data ℓ_{absolute}

$$L = \ell_{\text{BW}} \times \ell_{\text{absolute}}.$$

The sequence indicator parameter θ is defined in (8). We allow all other parameters in the models to vary among respondents and include matrix of covariates Z, which consists of an intercept, level of involvement with, level of expertise in, and age variables. For importance parameters β and respondent h we assume

$$\beta_h \sim \text{MVN}(\bar{B}' Z_h, V_\beta),$$

where $\bar{B}$ has dimensions $k \times N_{\text{var}}$, where k is the number of covariates in Z including intercept. The sequential scaling parameter ψ and the elicitation scaling parameter λ are also restricted to the positive domain through reparameterization

$$\psi_h = \exp(\psi_h^*) \quad \text{and} \quad \lambda_h = \exp(\lambda_h^*),$$

and we estimate

$$\ln \psi_h = \psi_h^*, \quad \ln \psi_h \sim N(\bar{\psi}' Z_h, \sigma_\psi^2), \tag{18}$$

$$\ln \lambda_h = \lambda_h^*, \quad \ln \lambda_h \sim N(\bar{\lambda}' Z_h, \sigma_\lambda^2). \tag{19}$$

Parameter estimates are obtained using Bayesian Markov chain Monte Carlo (MCMC) methods (see Rossi et al. 2005) using proper but relatively uninformative priors.

$$\begin{aligned}
\bar{B} &\sim N(\bar{\bar{B}} = \mathbf{0}, A = 10^{-5} I_k),\\
V_\beta &\sim IW(\nu_\beta = 18, V = 8I),\\
\bar{\psi} &\sim N(\bar{\bar{\psi}} = (0)_k', T_\psi = 0.001 I_k),\\
\sigma_\psi^2 &\sim \text{InvChiSq}(\nu_\psi = 5, q_\psi = 5),\\
\bar{\lambda} &\sim N(\bar{\bar{\lambda}} = (0)_k', T_\lambda = 0.001 I_k),\\
\sigma_\lambda^2 &\sim \text{InvChiSq}(\nu_\lambda = 5, q_\lambda = 5),\\
\gamma &\sim \text{Beta}(\nu_\gamma = 3, q_\gamma = 3).
\end{aligned} \tag{20}$$

Algorithms for estimating the models are provided in Appendix B. A simulation experiment was conducted to understand the identification of the parameters in the models. The results are presented in Appendix E. It confirms that the absolute data (ℓ_{absolute}) statistically identify the parameters β in all models. All other parameters of the models are also well identified.

5. Estimation Results

Parameter estimates are based on the first 13 best-worst tasks, reserving the last two tasks for predictive testing. The in-sample performance of the models is evaluated using the log-marginal density (LMD) estimator proposed by Newton and Raftery (1994) and hit probabilities. Out-of-sample performance is measured by the hit probabilities and hit rates using the same methodology (see Appendix D). The probabilities for the single evaluation probit models are evaluated using the trapezoid rule and for sequential evaluation probit models using the GHK algorithm (Keane 1994, Hajivassiliou et al. 1996). Details are provided in Appendix C.

Table 3 **In-Sample and Predictive Hit Rates for Single Evaluation Models of "Best" and "Worst" Responses**

	In-sample		Predictive	
Estimation	Best	Worst	Best	Worst
Logit with best only responses	0.7523	0.3676	0.6042	0.3840
Logit with worst only responses	0.4141	0.6037	0.4018	0.4727
Logit with best and worst responses (SL)	0.7071	0.6197	0.5861	0.4888
Probit with best only responses	0.7521	0.3683	0.6045	0.3830
Probit with worst only responses	0.4124	0.6004	0.4013	0.4721
Probit with best and worst responses (SP)	0.7192	0.6010	0.5950	0.4841

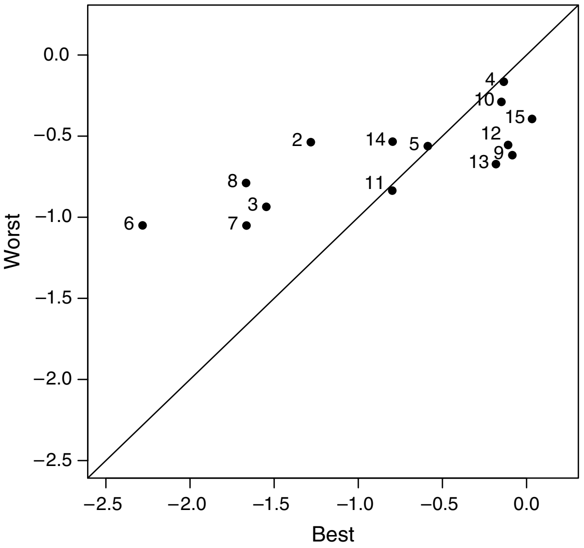
5.1. Estimation Results of Single Evaluation Models

We first examine in-sample and predictive fits for the logit and probit single evaluation models (3), (2), and (6) to motivate the need for greater model complexity. Table 3 displays the results for these models. As mentioned before, models of single evaluation assume that the same data generating mechanism produces the best and worst evaluations, with the difference in responses due to different censoring mechanisms of the same latent evaluation. Thus, the hit rates for "cross" responses (predicting best responses with parameters from the estimation based on worst responses or predicting worst responses with parameters from the best-based estimation) are expected to be similar to the hit rates for own responses (best responses with best estimation or worst responses with worst estimation) when examining in-sample fits or predicting the holdout choice tasks. However, this is not the case.

For both the logit and probit models, the hit rates exhibit a marked reduction when used to predict the cross responses. That is, models calibrated on the best responses predict the worst responses poorly, and models calibrated using the worst responses predict the best responses poorly. It is interesting to note that combining information from the best and worst choices in the models hurts both in-sample and holdout hit rates for best responses (for example, from 0.604 to 0.586 in logit models out of sample), making the fit for worst responses slightly better (for example, from 0.472 to 0.484 for probit models). These findings provide first initial evidence that the best and worst responses are not likely to be generated according to the models of single evaluation.

Additional evidence that the single evaluation assumption might not be appropriate for these tasks is presented in Figure 2, which displays posterior means of the estimated parameters for the logit model (3) estimated using the best and worst data separately. A 45-degree line is added to the figure to help with the comparison of the two estimates, which should be equal if the same data generating mechanism is present. We find that the estimated parameters based

Figure 2 Means of Importance Weights from Best and Worst Only Responses (Single Evaluation Logit Model) for Low Involvement and Low Expertise Group



Note. Numbers represent item numbers from Table 2.

on the worst data are smaller than those based on the best data. The linear association between the two sets of coefficients suggests that the scaling parameters (λ, ψ) might be successful at relating the best and worst response coefficients without resorting to models in which β_{best} and β_{worst} are allowed to be freely determined.

5.2. Model Fit and Comparison

Tables 4 and 5 report in-sample and predictive fits for all single evaluation and sequential models with inferred and observed sequence based on the joint likelihood of the best and worst responses. The hit rates and hit probabilities are computed for fitting and predicting the best and worst responses as a paired response. We find that the presence of sequential scaling effect ψ improves the fit of the model the most, followed by the elicitation effect λ. The analysis shows that, except for the in-sample hit rate, the model SQL5 outperforms the model SQL6. The presence of the elicitation scaling factor λ that can be thought of as an overall adjustment between the two elicitation modes is a better solution for the models of best-worst decision than is the independent β_{best} and β_{worst}. Thus, the sequential logit SQL5 and probit SQP5, which incorporate both effects as well as infer the sequence of decisions, are the best fitting models in sample and out of sample, implying that the sequential evaluation in best-worst tasks is most likely and the presence of these effects is important to include in the model specification.

5.2.1. Memory Retrieval Processes. We investigated the assumptions of the error terms based on the memory retrieval processes discussed in §3. Based on the fit measures, we cannot say which process, global or episodic memory retrieval, is more representative for all respondents in the sample that we have collected. We found that both memory retrieval processes as specified by our models are equally plausible. Although it is possible to think of different reasons for the lack of the clear distinction in our samples, we would like to offer two that would allow further investigation in the future.

First, we asked respondents to self-report whether they thought about their hair care concerns in general,

Table 4 Fit of the Models—In-Sample LMD NR

	Parameters			In-sample LMD NR		In-sample hit probabilities	
Model	Preference parameter	Best-worst scaling	Sequential scaling	EV error (asymmetric)	Normal error (symmetric)	EV error (asymmetric)	Normal error (symmetric)
			Single evaluation models				
Exploded logit	$\beta_{\text{best}} = \beta_{\text{worst}}$	—	—	−13,178	—	0.3013	—
Probit with double censoring	$\beta_{\text{best}} = \beta_{\text{worst}}$	—	—	—	−13,136	—	0.2983
MaxDiff	$\beta_{\text{best}} = \beta_{\text{worst}}$	—	—	−13,179	—	0.2980	—
			Sequential evaluation models				
Inferred sequence	$\beta_{\text{best}} = \beta_{\text{worst}}$	$\lambda = 1$	$\psi = 1$	−13,201	−13,140	0.3024	0.2958
	$\beta_{\text{best}} = \lambda\beta_{\text{worst}}$	λ estimated	$\psi = 1$	−12,519	−12,388	0.3153	0.3130
	$\beta_{\text{best}} \neq \beta_{\text{worst}}$	—	$\psi = 1$	−12,349	−11,945	0.3234	0.3290
	$\beta_{\text{best}} = \beta_{\text{worst}}$	$\lambda = 1$	ψ estimated	−11,671	−11,409	0.3476	0.3593
	$\boldsymbol{\beta}_{\text{best}} = \boldsymbol{\lambda\beta}_{\text{worst}}$	**$\boldsymbol{\lambda}$ estimated**	**$\boldsymbol{\psi}$ estimated**	**−11,105**	**−11,046**	**0.3781**	**0.3726**
	$\beta_{\text{best}} \neq \beta_{\text{worst}}$	—	ψ estimated	−12,028	−11,969	0.3487	0.3446
Observed sequence	$\beta_{\text{best}} = \beta_{\text{worst}}$	$\lambda = 1$	$\psi = 1$	−13,177	−13,130	0.3039	0.2973
	$\beta_{\text{best}} = \lambda\beta_{\text{worst}}$	λ estimated	$\psi = 1$	−12,548	−12,422	0.3155	0.3120
	$\beta_{\text{best}} \neq \beta_{\text{worst}}$	—	$\psi = 1$	−12,337	−11,934	0.3237	0.3291
	$\beta_{\text{best}} = \beta_{\text{worst}}$	$\lambda = 1$	ψ estimated	−12,726	−12,610	0.3168	0.3109
	$\beta_{\text{best}} = \lambda\beta_{\text{worst}}$	λ estimated	ψ estimated	−12,489	−12,398	0.3196	0.3167
	$\beta_{\text{best}} \neq \beta_{\text{worst}}$	—	ψ estimated	−12,100	−11,769	0.3331	0.3365

Note. Values in bold are fit statistics for the best performing models.

Table 5 Fit of the Models—Out-of-Sample Hit Probabilities and Hit Rates

	Parameters			Out-of-sample hit probabilities		Out-of-sample hit rates	
Model	Preference parameter	Best-worst scaling	Sequential scaling	EV error (asymmetric)	Normal error (symmetric)	EV error (asymmetric)	Normal error (symmetric)
			Single evaluation models				
Exploded logit	$\beta_{\text{best}} = \beta_{\text{worst}}$	—	—	0.2155	—	0.3956	—
Probit with double censoring	$\beta_{\text{best}} = \beta_{\text{worst}}$	—	—	—	0.2168	—	0.3931
MaxDiff	$\beta_{\text{best}} = \beta_{\text{worst}}$	—	—	0.2165	—	0.3939	—
			Sequential evaluation models				
Inferred sequence	$\beta_{\text{best}} = \beta_{\text{worst}}$	$\lambda = 1$	$\psi = 1$	0.2184	0.2162	0.3923	0.3931
	$\beta_{\text{best}} = \lambda\beta_{\text{worst}}$	λ estimated	$\psi = 1$	0.2227	0.2228	0.3923	0.3939
	$\beta_{\text{best}} \neq \beta_{\text{worst}}$	—	$\psi = 1$	0.2219	0.2246	0.3956	0.4040
	$\beta_{\text{best}} = \beta_{\text{worst}}$	$\lambda = 1$	ψ estimated	0.2375	0.2403	0.3939	0.4015
	$\boldsymbol{\beta}_{\text{best}} = \boldsymbol{\lambda\beta}_{\text{worst}}$	**$\boldsymbol{\lambda}$ estimated**	**$\boldsymbol{\psi}$ estimated**	**0.2510**	**0.2503**	**0.4066**	**0.3998**
	$\beta_{\text{best}} \neq \beta_{\text{worst}}$	—	ψ estimated	0.2297	0.2295	0.3998	0.4032
Observed sequence	$\beta_{\text{best}} = \beta_{\text{worst}}$	$\lambda = 1$	$\psi = 1$	0.2187	0.2168	0.3973	0.3931
	$\beta_{\text{best}} = \lambda\beta_{\text{worst}}$	λ estimated	$\psi = 1$	0.2233	0.2233	0.3981	0.3931
	$\beta_{\text{best}} \neq \beta_{\text{worst}}$	—	$\psi = 1$	0.2225	0.2254	0.4032	0.3964
	$\beta_{\text{best}} = \beta_{\text{worst}}$	$\lambda = 1$	ψ estimated	0.2244	0.2229	0.4125	0.4049
	$\beta_{\text{best}} = \lambda\beta_{\text{worst}}$	λ estimated	ψ estimated	0.2242	0.2241	0.3956	0.3956
	$\beta_{\text{best}} \neq \beta_{\text{worst}}$	—	ψ estimated	0.2241	0.2267	0.3973	0.4015

Note. Values in bold are fit statistics for the best performing models.

which would correspond to global memory retrieval, or if they thought about specific episodes, which would correspond to episodic memory retrieval. The responses had a bimodal distribution, with approximately 57% of respondents indicating the global inference retrieval and 43% episodic. This almost equal division between types of perceived memory retrieval might be partially responsible for the null effect. We investigated the use of a mixture of likelihoods model where the survey responses informed the probability of using global versus episodic memory retrieval parameter (Yang and Allenby 2000), but this investigation did not result in an improvement in model fit.

Second, some psychologists (Hintzman 1986) believe that only episodic memory exists and that any abstraction or summary of episodic events is just another episode that is created at retrieval and not in the encoding of information. Also, Howard and Kahana (2002) found that semantic cues were strongest on retrieval when episodic cues were also strong. This line of thought suggests that it is likely that the two memory types work together as one system. Since the sequential logit model fits the data as well as the sequential probit, and is easier to estimate, the remainder of the paper will analyze results for the logit specification.

5.2.2. Decision Sequence (γ). The most surprising aspect of our analysis is the estimated value of the decision sequence parameter, γ. The posterior mean of γ is 2% with a posterior standard deviation of 0.9%. Thus, nearly all respondents are inferred to make the "worst" choice first. This estimate is inconsistent with the data indicating 68% of the first clicks are the "best" choices.

Table 6 Fit of Sequential Evaluation Model SQL5

	Inferred sequence	Observed sequence
LMD NR	**−11,105**	−12,489
In-sample hit probability	**0.3781**	0.3196
In-sample hit rate	**0.6204**	0.5655
Out-of-sample hit probability	**0.2510**	0.2242
Out-of-sample hit rate	**0.4066**	0.3956

Note. Values in bold are fit statistics for the best performing models.

To understand the discrepancy, we first analyzed the fit of models with inferred and observed sequences (Table 6). The fit of the model with inferred sequence is significantly better than the fit of the model with observed sequence in-sample and on holdout across all metrics. For example, log marginal density estimate increases from $-12,489$ with observed sequence to $-11,105$ for inferred sequence model. The in-sample hit probabilities are improved by approximately 18% (from 0.3196 to 0.3781) if we allow the latent decision sequence to be different from observed clicks, and the holdout improvement is approximately 12% (from 0.2242 to 0.2510). Thus, we speculate that the assumption that the click data represent the true mental processes in best-worst tasks might not be accurate. It is possible that people pre-make their decisions in their mind before clicking the choices on the screen.

Two additional data sets were collected that allowed us to control the order of responses to be either best-worst ($n = 290$) or worst-best ($n = 299$). The same category and items were used to collect these data sets from the same population of consumers as the original sample. We fit three models to these data: (a) a model with the sequence inferred, (b) a model that assumed the sequence according to the task instructions, and

Table 7 Model Fit for Experimental Data with Imposed Order of Selection (SQL5)

	Sample best-worst			Sample worst-best		
Measure	Inferred sequence (a)	Imposed sequence (b)	Observed sequence (c)	Inferred sequence (a)	Imposed sequence (b)	Observed sequence (c)
LMD NR	**−5,160**	−5,910	−5,918	**−5,746**	**−5,775**	−5,928
In-sample hit probability	**0.4061**	0.3171	0.3359	**0.3737**	**0.3733**	0.3549
In-sample hit rate	**0.6592**	0.6013	0.6056	**0.6277**	**0.6228**	0.6079
Out-of-sample hit probability	**0.2644**	0.2220	0.2336	**0.2492**	**0.2507**	**0.2439**
Out-of-sample hit rate	**0.3862**	0.3621	0.3690	**0.3645**	**0.3712**	**0.3746**

Note. Values in bold are fit statistics for the best performing models.

(c) a model that used the actual click data to indicate the sequence. Results are presented in Table 7. If the model with an inferred sequence is correct, then it should fit the data best and agree with the imposed sequence (column b) as indicated by the decision sequence parameter (γ). This is exactly what we found.

The right side of Table 7 displays results when respondents were asked to provide their worst responses first. The estimate of γ for this data is 6.2% with a posterior standard deviation of 2.3%, indicating the sequence worst-then-best to be most likely. For the worst-best′ data we find that the fit of the model with inferred sequence (column a) is similar to that with the imposed sequence (column b), as expected.

The left side of Table 7 displays results when respondents were asked to provide their best responses first. In contrast to the results displayed on the right side of the table, the fits of the models for this data set shows a large difference between the inferred sequence model (column a) and the other two. The estimate of γ in the best-worst data set is 5.2% with a posterior standard deviation of 2%. The results imply that respondents followed a decision sequence that was different than that imposed by the experiment or the observed click data, favoring the selection of the worst choice option first.

Our finding here is consistent with some literature in psychology. Testing the underlying processes for this result is left for future research, but we provide some speculative thoughts on that issue. Beach (1993) and Beach and Potter (1992) showed that people may engage in prescreening of alternatives before making final choices. Also, Ordóñez et al. (1999) found that items are rejected during the first phase of choices before making a "select" decision. This decision sequence was also found to be a natural tendency to make choices: results of an experimental condition in which participants were asked to "reject" were similar to the control condition where no instructions were given. Our analysis in Tables 6 and 7 provides evidence consistent with this literature and indicates that click data may represent an alternative cognitive process.

5.2.3. Sequential (ψ) and Elicitation (λ) Effects. Table 8 displays the posterior means and standard deviations of the sequential effect ψ and elicitation effects λ for the model SQL5. We included three covariates to understand the heterogeneity of parameters: involvement with, and expertise in, the category (hair care) measured on a five-point scale, as well as age. We created a dummy variable for involvement and expertise variables by separating respondents who checked the top two boxes.

Table 8 Posterior Means of Sequential and Elicitation Effects (SQL5)

Covariates	Sequential effect $\ln\psi$	Elicitation effect $\ln\lambda$
Intercept	**1.67** (0.09)	**0.27** (0.08)
Involvement	−**0.36** (0.13)	−0.17 (0.12)
Expertise	−**0.45** (0.21)	−**1.11** (0.23)
Age	−0.17 (0.06)	−**0.20** (0.06)

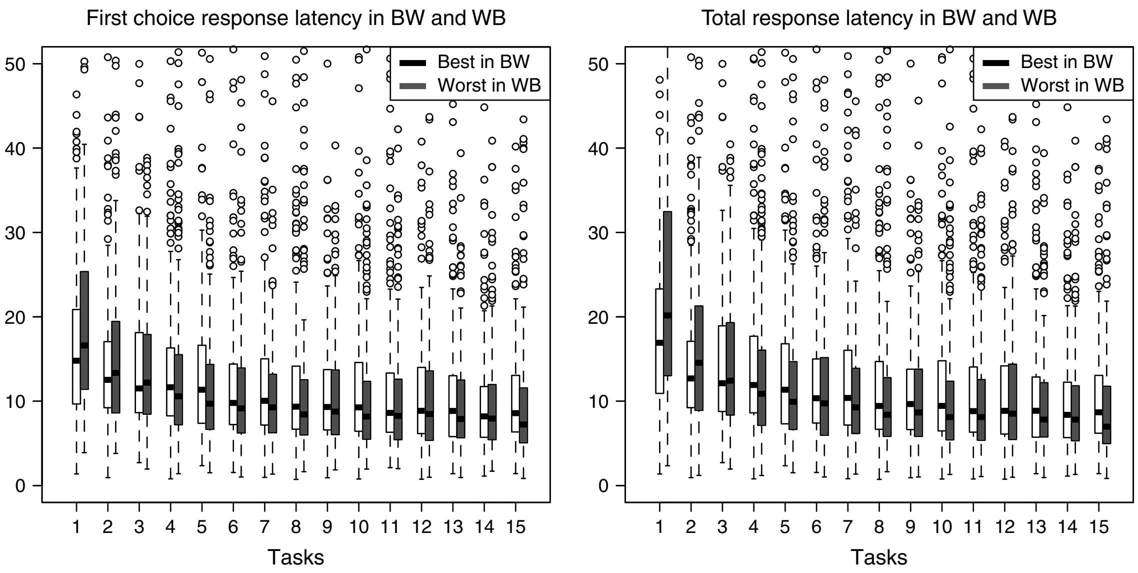
Notes. The numbers are based on reparameterization in Equations (18) and (19). Estimates in bold have more than 95% of their posterior mass away from zero. Posterior standard deviation of the mean is shown in parentheses.

The mean of sequence parameter ψ is significantly greater than 1 ($\ln\psi > 0$) with mean of $\ln\psi = 1.67$ for low involvement and low expertise respondents, indicating that the second response is more certain. The effect is reduced if a person is highly involved with the category or is an expert, but it is still positive, meaning that the second decision is less error prone for all respondents.

The mean of the elicitation scaling parameter λ is greater than 1 (mean of $\ln\lambda = 0.27 > 0$) for the low expertise group and for relatively younger respondents. This result indicates that, on average, there is more consistency in the worst choices than best choices for this group. However, expertise plays a major role in the direction of the effect. Experts in the category are more consistent with their choices in best decision than in worst.

Finally, we find that response latencies for the 15 choice tasks support the notion that worst responses are easier for respondents to provide. Figure 3 displays two measures of response latency, time until the first click and total time to provide an answer, for the data sets in which the response order was forced, or imposed. For each measure, the response times for the worst response times are shorter. This is consistent with our estimate of the elicitation effect ($\beta_{\text{worst}} = \lambda\beta_{\text{best}}$)

Figure 3 Response Latencies in the Imposed Order Best-Worst (BW) and (WB) Worst-Best Samples



reported in Table 8, where $\lambda > 1$ for younger respondents with low expertise. An estimate of $\lambda > 1$ indicates that the worst choices are surer and more consistent than the best choices, and the response latencies indicate that they are also easier for respondents to provide.

Our findings are consistent with results from the psychology literature showing that selecting the best and worst alternatives are not complementary operations (Yaniv and Schul 2000, Yaniv et al. 2002, Wilk 1997) and that worst responses are more consistent across trials (Coombs et al. 1978). Our inferences are also in line with what we saw in Figure 2. According to our best performing model of sequential evaluation, the size of the importance parameters β between the best and worst choices is modified by the two scaling factors in the most probable sequence worst-best. Although the size of β in the worst choice is increased by the elicitation scaling λ, the size of β in the best choices is increased by the sequential scaling ψ, which is much greater than the elicitation effect λ. That would make β's in best responses look larger than in the worst decision. This is exactly what we found in our initial analysis in Figure 2: estimated parameters based on the worst data are smaller than those based on the best data.

5.2.4. Preference Parameters (β) and Rankings. Preference parameter estimates for single evaluation and sequential evaluation logit models are reported in Tables 9 and 10. We report the posterior means of $\bar{\beta}$ and standard deviations of the distribution of heterogeneity in (20). We find significant heterogeneity in importance weights among respondents. We also find

Table 9 Posterior Means and Standard Deviation of Distribution Heterogeneity of Importance Weights for Model SL

Item	Intercept	Involvement	Expertise	Age	SD
1. My hair is too oily.	**−2.14**	0.16	0.18	**−0.25**	1.81
2. My hair is breaking.	**−0.69**	**0.98**	**−0.61**	**−0.42**	2.03
3. My hair is stiff and resistant.	**−1.33**	**0.72**	−0.32	−0.21	1.40
4. My hair is coming out more than it used to.	0.07	**0.95**	**−1.05**	**0.03**	2.91
5. My hair is dry.	**−0.52**	**0.83**	**−1.01**	**−0.21**	1.98
6. My hair is coarse and frizzy.	**−1.35**	**1.01**	−0.36	**−0.39**	2.06
7. I have split ends.	**−1.32**	**1.03**	−0.31	**−0.37**	2.18
8. I have unruly, unmanageable hair.	**−1.03**	**0.76**	−0.52	**−0.27**	1.97
9. My hair's color is faded and dull.	**−0.44**	**0.85**	**−0.96**	**−0.35**	2.06
10. My hair seems finer.	**−0.25**	**0.42**	**−0.90**	−0.02	2.16
11. My hair is damaged from products, treatments, or sun.	**−1.01**	**1.41**	**−0.83**	**−0.47**	2.22
12. My hair has been impacted by stress and hormone fluctuations.	**−0.48**	**0.83**	**−0.62**	**−0.51**	1.78
13. My greying hair is unflattering.	**−0.41**	**1.07**	**−0.96**	**−0.54**	2.74
14. I am seeing more of my scalp or receding hairline.	**−0.50**	**0.78**	−0.67	**0.27**	3.13
15. My hair lacks shine.	**−0.37**	**0.44**	−0.43	−0.08	1.87

Notes. SD is the standard deviation for the diagonal elements of covariance matrix of the distribution of heterogeneity. Estimates in bold have more than 95% of their posterior mass away from zero.

Table 10 Posterior Means and Standard Deviation of Distribution Heterogeneity of Importance Weights for Model SQL5

Item	Intercept	Involvement	Expertise	Age	SD
1. My hair is too oily.	−**1.82**	−0.02	−0.06	−**0.31**	1.35
2. My hair is breaking.	−**0.66**	**0.83**	−**0.62**	−**0.34**	1.62
3. My hair is stiff and resistant.	−**1.10**	**0.62**	−**0.58**	−**0.25**	1.20
4. My hair is coming out more than it used to.	−**0.26**	**0.84**	−**0.65**	−0.01	2.11
5. My hair is dry.	−**0.51**	**0.73**	−**0.72**	−**0.21**	1.42
6. My hair is coarse and frizzy.	−**1.15**	**0.84**	−**0.69**	−**0.37**	1.57
7. I have split ends.	−**1.06**	**0.74**	−0.36	−**0.39**	1.67
8. I have unruly, unmanageable hair.	−**0.89**	**0.68**	−0.41	−**0.27**	1.48
9. My hair's color is faded and dull.	−**0.41**	**0.74**	−**0.59**	−**0.25**	1.38
10. My hair seems finer.	−**0.31**	**0.42**	−**0.50**	0.03	1.56
11. My hair is damaged from products, treatments, or sun.	−**0.87**	**1.16**	−**0.55**	−**0.39**	1.68
12. My hair has been impacted by stress and hormone fluctuations.	−**0.41**	**0.74**	−**0.48**	−**0.42**	1.36
13. My greying hair is unflattering.	−**0.48**	**0.85**	−0.47	−**0.35**	1.86
14. I am seeing more of my scalp or receding hairline.	−**0.63**	**0.58**	−**0.87**	**0.19**	2.33
15. My hair lacks shine.	−**0.34**	**0.42**	−0.38	−0.08	1.28

Notes. SD is the standard deviation for the diagonal elements of covariance matrix of the distribution of heterogeneity. Estimates in bold have more than 95% of their posterior mass away from zero.

that for the sequential evaluation model with sequential effect ψ and elicitation effect λ, the estimates of the items become less extreme and heterogeneity reduces. We note that the effect of expertise on preference parameters differs between the two models, which will impact what inferences managers would make with respect to what is important to experts versus nonexperts from the two models.

We now present an analysis of the aggregate top concerns of respondents as inferred from the single (SL) and sequential evaluation (SQL5) models. As an example, we examine two groups of respondents: a high involvement, high expertise, low age group (Table 11) and low involvement, low expertise, low age group (Table 12). In aggregate, the models of single and sequential evaluation agree on the set of top three or four out of five items that are of the most concern when the order is not considered. However, the sequential evaluation logit SQL5 brings up item 9 “My hair's color is faded and dull” and does not include item 4, which is “My hair is coming out more than it used to” in the first group and replaces item 10 “My hair seems finer” with item 15 “My hair lacks shine.” In addition, if the order is considered in the first group, then item 2 “My hair is breaking” would have been thought to be the most important item according to the single evaluation model, whereas it is only number five according to the best fitting model of sequential evaluation. These are different concerns and this finding would be relevant to different targeting strategies. Thus, if the company uses a single evaluation instead of a sequential evaluation model, the selection of the top five issues that drive marketing activities might be somewhat misleading.

Although aggregate results are important, the most important findings come from the analysis of the individual-level top concerns. The analysis of the individual-specific top concerns showed greater inconsistency between the single and sequential evaluation models, as shown in Figure 4. This figure shows the proportion of respondents that agree on individual ranking sets between SL and SQL5. The black dots show results that do not take into account the order of items within each set, and the square points show

Table 11 Aggregated Top 5 Concerns for High Involvement High Expertise Group

Rank	SL	SQL5
Top 1	**2 My hair is breaking.**	12 My hair has been impacted by stress and hormone fluctuations.
2	13 My greying hair is unflattering.	13 My greying hair is unflattering.
3	12 My hair has been impacted by stress and hormone fluctuations.	11 My hair is damaged from products, treatments, or sun.
4	11 My hair is damaged from products, treatments, or sun.	**9 My hair's color is faded and dull.**
Top 5	**4 My hair is coming out more than it used to.**	**2 My hair is breaking.**

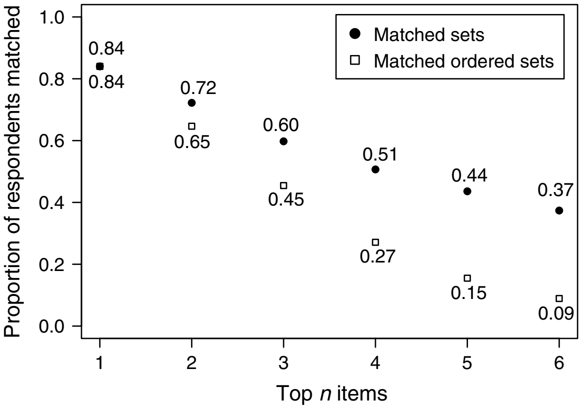
Note. Items in bold indicate differences between the model findings.

Table 12 Aggregated Top 5 Concerns for Low Involvement Low Expertise Group

Rank	SL	SQL5
Top 1	4 My hair is coming out more than it used to.	12 My hair has been impacted by stress and hormone fluctuations.
2	13 My greying hair is unflattering.	4 My hair is coming out more than it used to.
3	12 My hair has been impacted by stress and hormone fluctuations.	13 My greying hair is unflattering.
4	9 My hair's color is faded and dull.	9 My hair's color is faded and dull.
Top 5	**10 My hair seems finer.**	**15 My hair lacks shine.**

Note. Items in bold indicate differences between the model findings.

Figure 4 **Proportion of Respondents with Matching Ranking Sets Between Models of Single (SL) and Sequential Evaluation (SQL5)**



results when the order matters. We see a significant drop in the proportion of respondents with matched sets for each top *n* items. For example, for the top three items the proportion of matched individual ranking sets is only 60%. If we take into account the order of the items, then the discrepancies between the models are magnified, with the percentage reducing to 45%. These results show how the ranking of items obtained from the single evaluation models can be misleading for marketing actions, and we advise against using single evaluation models to make inferences about preference parameters from the dual-elicitation procedure best-worst.

6. Discussion

Discrete choice experiments are designed to help businesses understand consumer preferences and marketing actions. The goal of better understanding consumers and making better marketing decisions drives development of new research methods and modification of existing ones. However, it is important to investigate the decision processes in these choice experiments and what models should be used to make inferences that would allow marketers to properly address consumers' needs and wants. In this paper, we investigate the decision making process that underlies the choices in best-worst (BW) tasks and use psychological theories to drive model development. We compare the results and managerial implications of our models with that of the models that are widely used in practice.

We find evidence favoring the presence of sequential evaluation in BW tasks and believe that this decision process should be taken into account when analyzing the data from these tasks. Our proposed models for BW tasks are based on the sequential two-step evaluation of the items. They provide a better fit to the data and have better ability to predict choices out of sample than the currently used models of single evaluation such as logit, probit, and MaxDiff. In addition, we showed that the individual rankings of the top attributes needed for marketing actions would be misleading if based on the single evaluation models.

The model of sequential evaluation allowed us to uncover and incorporate several effects that are present in the choice decisions in BW tasks, such as the sequence and elicitation effects. These effects support the idea of preference construction, which is not consistent with the economic view that preferences preexist, the assumption that is used in simple single evaluation models of choice prevalent in practice. Additionally, collected process data supported findings on these effects from our models.

We find that respondents are more likely to select the worst item from a choice set before selecting the best item. Investigation of the discrepancy between the findings and the observed click data is left for future research, which should rely on not only quantitative but also experimental approaches. The estimated scaling effect in this sequence indicates that the item coefficients are larger in the second choice, which is consistent with the respondents' being surer of their choices. This is consistent with the work by Louviere (2001) and Fiebig et al. (2010) on scaling differences between contexts. This sequential effect contributes the most to the fit improvement as compared to the fit of single evaluation models.

We also obtain evidence of an elicitation effect. Elicitation mode matters and impacts the construction of a biased sample of evidence as respondents seek evidence congruent with the question at hand, consistent with hypothesis testing theory. We found that respondents tend to be more sure about items that describe them least than about items that describe them best. This phenomenon has been receiving attention in the consumer behavior literature (Meloy and Russo 2004, Lyubomirsky and Lee 1999), and we found evidence that the scaling factor is able to capture the mechanism of preference construction better than using unrestricted β_{best} and β_{worst}.

As an extension of the proposed sequential evaluation models, one could apply a nonlinear elicitation scaling function (i.e., $\beta_{\text{worst}} = f(\beta_{\text{best}})$). This extension would be based on the theory of differential adjustment of importance weights under different elicitation tasks, which would provide additional evidence for the confirmation bias in these tasks. We would predict, based on previous research in psychology, that positive attributes are weighted higher under the "select-the-best" condition and negative attributes are adjusted to have a stronger effect on choices under the "select-the-worst" condition. This would require data from a category of products or concerns where importance parameters are more likely to spread across positive and negative domains and, thus, provide more information on the issue of differential scaling.

Finally, we investigated the presence of episodic inference retrieval, represented by the model with extreme value error term distribution, versus global memory retrieval, represented by the normal distribution. The sample did not have enough information to reject the assumption of no distinction between the two processes. We also analyzed a likelihood mixture model to determine the probability of respondents executing one type of memory retrieval versus another but did not get improvement in model fit. However, to our knowledge, this was the first attempt to guide error term specification selection with psychological theories of decision making and not computational convenience. We hope that this effort will offer first steps in error term consideration for development of future marketing models.

Although our proposed models are better models to use to analyze BW data, this investigation of the psychological processes underlying these tasks brings us to the conclusion that this research tool, which allows for collecting more data on each respondent, does not consider the appropriateness of imposing a dual elicitation procedure on respondents. In typical market conditions consumers face choices where they try to maximize their utility by the action of selecting products. This decision is consistent with the tasks of selecting the best alternative in the choice experiments. However, rarely do people have to reject or give up an alternative explicitly, which would correspond to answering the worst choice in BW tasks. We believe that researchers should avoid collecting the worst responses unless there is a good justification that comes from the expected consumer behavior in the marketplace.

Finally, our results caution against the use of single evaluation models for the analysis of best-worst and partially ranked data. We note that partial ranks can be created by a variety of data collection methods, such as constant sum data and complete-rank data. We find that best-worst data are not well represented by a simple censoring mechanism as in models of single evaluation as discussed in §2. We speculate that other similar results may be true in data collected through other means, necessitating the use of models that better represent the true data generating mechanism.

Acknowledgments

The authors thank Bryan Orme and Sawtooth Software for helping them with the data collection for this research. They also thank Alexander Dyachenko for computing support.

Appendix A. Derivation of the Probabilities for the Least Preferred Alternative

The derivation comes from the following model:

$$z_{\text{worst}} = X_{\text{worst}}\beta + \xi_{\text{worst}}, \quad \xi_{\text{worst}} \sim \text{MinEV}(0, I).$$

The probability of selecting j's alternative during "select-the-worst" task is derived similar to the way in which the probability for traditional multinomial logit with maximum EV distribution is derived. The derivation shown here is one of the possible ways to accomplish this.

$$\begin{aligned}
\Pr(y_{\text{worst}} = j) &= \Pr(z_{\text{worst},\,j} = \min\{z_{\text{worst}}\}) \\
&= \Pr(z_{\text{worst},\,j} < z_{\text{worst},\,l} \text{ for all } l \neq j) \\
&= \Pr(x'_{\text{worst},\,j}\beta + \xi_j < x'_{\text{worst},\,l}\beta + \xi_l \text{ for all } l \neq j) \\
&= \Pr(-x'_{\text{worst},\,j}\beta - \xi_j > -x'_{\text{worst},\,l}\beta - \xi_l \text{ for all } l \neq j) \\
&= [\text{substitute } x'_{\text{worst},\,j}\beta = V_j] \\
&= \Pr(-V_j - \xi_j > -V_l - \xi_l \text{ for all } l \neq j).
\end{aligned}$$

Although there are several ways to look at this, we use a known relationship

$$-\xi = v \sim \text{MaxEV}(0, I).$$

This gives us the following:

$$\begin{aligned}
&\Pr(-V_j + v_j > -V_l + v_l \text{ for all } l) \\
&\quad = \Pr(v_l < V_l - V_j + v_j \text{ for all } l) \\
&\quad = \int_{-\infty}^{\infty} \left[\prod_{l \neq j} \int_{-\infty}^{V_l - V_j + v_j} f_v(v_l)\, dv_l \right] f_v(v_j)\, dv_j \\
&\quad = \int_{-\infty}^{\infty} f_v(v_j) \left[\prod_{l \neq j} F(v_l < V_l - V_j + v_j) \right] dv_j \\
&\quad = \int_{-\infty}^{\infty} e^{-v_j} e^{-e^{-v_j}} \left[\prod_{l \neq j} e^{-e^{-(V_l - V_j + v_j)}} \right] dv_j \\
&\quad = \int_{-\infty}^{\infty} e^{-v_j} e^{-e^{-v_j}} [e^{-e^{-v_j} \sum_{l \neq j} e^{(V_j - V_l)}}]\, dv_j \quad \left[\text{with } n = \sum_{l \neq j} e^{(V_j - V_l)}\right] \\
&\quad = \int_{-\infty}^{\infty} e^{-v_j} e^{-e^{-v_j}} [e^{-e^{-v_j} e^{\ln n}}]\, dv_j = \int_{-\infty}^{\infty} e^{-v_j} e^{-e^{-v_j} - ne^{-v_j}}\, dv_j \\
&\quad = \int_{-\infty}^{\infty} e^{-v_j} e^{-(1+n)e^{-v_j}}\, d\varepsilon_j = \frac{1}{(1+n)} e^{-(1+n)e^{-v_j}} \Big|_{-\infty}^{\infty} = \frac{1}{1+n} \\
&\quad = \frac{1}{1 + \sum_{l \neq j} e^{(V_j - V_l)}} = \frac{1}{1 + e^{V_j} \sum_{l \neq j} e^{-V_l}} \\
&\quad = \frac{e^{-V_j}}{e^{-V_j} + \sum_{l \neq j} e^{-V_l}} = \frac{e^{-V_j}}{\sum_l e^{-V_l}}.
\end{aligned}$$

This result is used in (12) for model development.

The result of this equation represents the common practice of analyzing data from best-worst experiments (Sawtooth Software 2010). The vector of x's in (11) is multiplied by -1 for the worst responses, where the $\exp(-x\beta)$ in (12) is thought of as the antilog of negative utility. Although it is mathematically equivalent, we prefer to think of (12) as a probability corresponding to the minimum of a utility measure.

Appendix B. Estimation Algorithm for Sequential Evaluation Models

1. Draw household-specific β_h using random walk *M-H*

$$\beta_h^{(n)} = \beta_h^{(o)} + \xi, \quad \xi \sim N(0, s_\beta I),$$

where s_β is selected to have an acceptance rate of approximately 35%, $\beta_h^{(n)}$ is the proposed draw of the parameter vector, and $\beta_h^{(o)}$ is the old vector. The proposed draws are accepted with probability

$$\min\left(\frac{L^{(n)} \times \exp[-(1/2)(\beta_h^{(n)} - \bar{B}'Z_h)'V_\beta^{-1}(\beta_h^{(n)} - \bar{B}'Z_h)]}{L^{(o)} \times \exp[-(1/2)(\beta_h^{(o)} - \bar{B}'Z_h)'V_\beta^{-1}(\beta_h^{(o)} - \bar{B}'Z_h)]}, 1\right),$$

where

$$L^{(n)} = \ell(y_{h,\text{best}}, y_{h,\text{worst}} \mid \beta_h^{(n)}, \psi_h, \lambda_h, \theta_h, X_h)\ell(y_{h,\text{absolute}} \mid \beta_h^{(n)}),$$
$$L^{(o)} = \ell(y_{h,\text{best}}, y_{h,\text{worst}} \mid \beta_h^{(o)}, \psi_h, \lambda_h, \theta_h, X_h)\ell(y_{h,\text{absolute}} \mid \beta_h^{(o)}).$$

Likelihood specifications $\ell(y_{h,\text{best}}, y_{h,\text{worst}} \mid \cdot)$ for sequential logit and probit models and for the absolute data $\ell(y_{h,\text{absolute}} \mid \cdot)$ are given in (15)–(17), respectively. For the sequential probit models we calculated the likelihood using GHK algorithm with $r = 500$. We modified the A_j transformation matrix compared to the traditional multinomial probit case (Rossi et al. 2005) by substituting one of 0's with -1 in the row associated with the choice index.

2. Heterogeneity in part-worth parameters (Rossi et al. 1996)

Generate $\bar{\beta} \mid B_h, V_\beta, Z \sim MVN(\tilde{\beta}, V_\beta \otimes (Z'Z + A)^{-1})$, where

$$\tilde{\beta} = \text{vec}(\tilde{B}),$$
$$\tilde{B} = (Z'Z + A)^{-1}(Z'B_h + A\bar{\bar{B}}).$$

Generate

$$V_\beta \mid B_h, Z \sim IW(\nu_\beta + H, V + S),$$
$$S = (B_h - Z\tilde{B})'(B_h - Z\tilde{B}) + (\tilde{B} - \bar{\bar{B}})'A(\tilde{B} - \bar{\bar{B}}).$$

3. Similar to the draws of β_h, generate draws of household specific ψ_h^* using random walk *M-H*

$$\psi_h^{*(n)} = \psi_h^{*(o)} + \eta, \quad \eta \sim N(0, s_\psi I),$$
$$\psi_h^{(n)} = \exp(\psi_h^{*(n)}),$$

where s_ψ is selected to have an acceptance rate of approximately 35%. The new draws are accepted with probability

$$\min\Bigg(\Bigg(\ell(y_{h,\text{best}}, y_{h,\text{worst}} \mid \beta_h, \psi_h^{(n)}, \theta_h, \lambda_h, X_h) \cdot \exp\left[-\frac{1}{2\sigma_\psi^2}(\psi_h^{(n)} - Z_h'\bar{\psi})^2\right]\Bigg) \cdot \Bigg(\ell(y_{h,\text{best}}, y_{h,\text{worst}} \mid \beta_h, \psi_h^{(o)}, \theta_h, \lambda_h, X_h) \cdot \exp\left[-\frac{1}{2\sigma_\psi^2}(\psi_h^{(o)} - Z_h'\bar{\psi})^2\right]\Bigg)^{-1}, 1\Bigg).$$

4. Heterogeneity in scaling parameter.

Generate $\bar{\psi} \mid \psi_h, Z \sim MVN(\tilde{\psi}, \sigma_\psi^2(Z'Z + T_\psi)^{-1})$, where

$$\tilde{\psi} = (Z'Z + T_\psi)^{-1}(Z'\psi_h + T_\psi\bar{\bar{\psi}}).$$

Generate

$$\sigma_\psi^2 \mid \bar{\psi}, \psi_h \sim \text{InvChiSq}\left(\nu_\psi + H, \frac{q_\psi\nu_\psi + S_\psi}{\nu_\psi + H}\right),$$
$$S_\psi = (\psi_h - Z'\tilde{\psi})'(\psi_h - Z'\tilde{\psi}) + (\tilde{\psi} - \bar{\bar{\psi}})'T_\psi(\tilde{\psi} - \bar{\bar{\psi}}).$$

Steps 3 and 4 are skipped for sequential logit models SQL1, SQL2, and SQL3 and for sequential probit models SQP1, SQP2, and SQP3.

5. Similar to the draws of β_h, generate draws of household specific λ_h^* using random walk *M-H*

$$\lambda_h^{*(n)} = \lambda_h^{*(o)} + u, \quad u \sim N(0, s_\lambda I),$$
$$\lambda_h^{(n)} = \exp(\lambda_h^{*(n)}),$$

where s_λ is selected to have an acceptance rate of approximately 35%. The new draws are accepted with probability:

$$\min\Bigg(\Bigg(\ell(y_{h,\text{best}}, y_{h,\text{worst}} \mid \beta_h, \psi_h, \theta_h, \lambda_h^{(n)}, X_h) \cdot \exp\left[-\frac{1}{2\sigma_\lambda^2}(\lambda_h^{(n)} - Z_h'\bar{\lambda})^2\right]\Bigg) \cdot \Bigg(\ell(y_{h,\text{best}}, y_{h,\text{worst}} \mid \beta_h, \psi_h, \theta_h, \lambda_h^{(o)}, X_h) \cdot \exp\left[-\frac{1}{2\sigma_\lambda^2}(\lambda_h^{(o)} - Z_h'\bar{\lambda})^2\right]\Bigg)^{-1}, 1\Bigg).$$

6. Heterogeneity in scaling parameter.

Generate $\bar{\lambda} \mid \lambda_h, Z \sim \text{MVN}(\tilde{\lambda}, \sigma_\lambda^2(Z'Z + T_\lambda)^{-1})$, where

$$\tilde{\lambda} = (Z'Z + T_\lambda)^{-1}(Z'\lambda_h + T_\lambda\bar{\bar{\lambda}}).$$

Generate

$$\sigma_\lambda^2 \mid \bar{\lambda}, \lambda_h \sim \text{InvChiSq}\left(\nu_\lambda + H, \frac{q_\lambda\nu_\lambda + S_\lambda}{\nu_\lambda + H}\right),$$
$$S_\lambda = (\lambda_h - Z'\tilde{\lambda})'(\lambda_h - Z'\tilde{\lambda}) + (\tilde{\lambda} - \bar{\bar{\lambda}})'T_\lambda(\tilde{\lambda} - \bar{\bar{\lambda}}).$$

7. Generate the vector of latent response sequence indicators θ of length H (total number of respondents), where each element is generated as follows:

$$\theta_h^{(n)} \sim \text{Bernoulli}(\gamma).$$

Use $\gamma = 0.5$ for the initial draw. The proposed draws of θ_h are accepted with probability

$$\min\left(\frac{L^{(n)}}{L^{(o)}}, 1\right),$$

where

$$L^{(n)} = \ell(y_{h,\text{best}}, y_{h,\text{worst}} \mid \theta_h^{(n)}, \cdot),$$
$$L^{(o)} = \ell(y_{h,\text{best}}, y_{h,\text{worst}} \mid \theta_h^{(o)}, \cdot).$$

8. Update the probability parameter γ as follows:

$$p(\gamma \mid \theta) \propto \left[\prod_{h=1}^{H} \gamma^{\theta_h}(1-\gamma)^{1-\theta_h}\right]\gamma^{\nu_\gamma - 1}(1-\gamma)^{q_\gamma - 1},$$
$$p(\gamma \mid \theta) \propto \gamma^{\sum_{h=1}^{H}\theta_h + \nu_\gamma - 1}(1-\gamma)^{H - \sum_{h=1}^{H}\theta_h + q_\gamma - 1},$$

meaning that

$$\gamma \sim \text{Beta}\left(\sum_{h=1}^{H}\theta_h + \nu_\gamma, H - \sum_{h=1}^{H}\theta_h + q_\gamma\right).$$

Appendix C. Calculation of Probabilities in Probit Models

For the single evaluation probit models with identity covariance matrix, the calculations of the probability of choice can be represented as follows. The probability for the best choice

$$
\begin{aligned}
&\Pr(y_{\text{best}} = i \mid \beta, X)\\
&\quad = \Pr(x_i'\beta + \varepsilon_i > x_k'\beta + \varepsilon_k \text{ for any } k \neq i,\ k = 1, \ldots, p)\\
&\quad = \Pr(\varepsilon_k < x_i'\beta - x_k'\beta + \varepsilon_i \text{ for any } k \neq i,\ k = 1, \ldots, p)\\
&\quad = \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{x_i'\beta - x_1'\beta + \varepsilon_i} \cdots \int_{-\infty}^{x_i'\beta - x_p'\beta + \varepsilon_i} \varphi(\varepsilon_p) \cdots \varphi(\varepsilon_1) \right]\\
&\qquad\qquad \cdot \varphi(\varepsilon_i \, d\varepsilon_p \cdots d\varepsilon_1 \, d\varepsilon_i\\
&\quad = \int_{-\infty}^{+\infty} \left[\prod_{k=\{1,\ldots,p,\,-i\}} \Phi(x_i'\beta - x_k'\beta + \varepsilon_i) \right] \varphi(\varepsilon_i) \, d\varepsilon_i.
\end{aligned}
$$

The probability for the worst choice

$$
\begin{aligned}
&\Pr(y_{\text{worst}} = i \mid \beta, X)\\
&\quad = \Pr(x_i'\beta + \varepsilon_i < x_k'\beta + \varepsilon_k \text{ for any } k \neq i,\ k = 1, \ldots, p)\\
&\quad = \Pr(\varepsilon_k > x_i'\beta - x_k'\beta + \varepsilon_i \text{ for any } k \neq i,\ k = 1, \ldots, p)\\
&\quad = \int_{-\infty}^{+\infty} \left[\int_{x_i'\beta - x_1'\beta + \varepsilon_i}^{+\infty} \cdots \int_{x_i'\beta - x_p'\beta + \varepsilon_i}^{+\infty} \varphi(\varepsilon_p) \cdots \varphi(\varepsilon_1) \right]\\
&\qquad\qquad \cdot \varphi(\varepsilon_i) \, d\varepsilon_p \cdots d\varepsilon_1 \, d\varepsilon_i\\
&\quad = \int_{-\infty}^{+\infty} \left[\prod_{k=\{1,\ldots,p,\,-i\}} (1 - \Phi(x_i'\beta - x_k'\beta + \varepsilon_i)) \right] \varphi(\varepsilon_i) \, d\varepsilon_i.
\end{aligned}
$$

The probability for the best-worst pair of choices

$$
\begin{aligned}
&\Pr(y_{\text{best}} = i, y_{\text{worst}} = j \mid \beta, X)\\
&\quad = \Pr(x_j'\beta + \varepsilon_j < x_k'\beta + \varepsilon_k < x_i'\beta + \varepsilon_i\\
&\qquad\qquad\qquad \text{for any } k \neq i, j,\ k = 1, \ldots, p)\\
&\quad = \Pr(x_j'\beta - x_k'\beta + \varepsilon_j < \varepsilon_k < x_i'\beta - x_k'\beta + \varepsilon_i\\
&\qquad\qquad\qquad \text{for any } k \neq i, j,\ k = 1, \ldots, p)\\
&\quad = \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{x_i'\beta - x_j'\beta + \varepsilon_i} \left[\int_{x_j'\beta - x_1'\beta + \varepsilon_j}^{x_i'\beta - x_1'\beta + \varepsilon_i} \cdots \int_{x_j'\beta - x_k'\beta + \varepsilon_j}^{x_i'\beta - x_k'\beta + \varepsilon_i} \varphi(\varepsilon_k) \cdots \right.\right.\\
&\qquad\qquad \left.\left. \cdot \varphi(\varepsilon_1) \, d\varepsilon_1 \cdots d\varepsilon_k \right] \varphi(\varepsilon_j) \, d\varepsilon_j \right] \varphi(\varepsilon_i) \, d\varepsilon_i\\
&\quad = \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{x_i'\beta - x_j'\beta + \varepsilon_i} \left\{ \prod_{k=\{1,\ldots,p,\,-i-j\}} [\Phi(x_i'\beta - x_k\beta + \varepsilon_i) \right.\right.\\
&\qquad\qquad \left.\left. - \Phi(x_j'\beta - x_k'\beta + \varepsilon_j)] \right\} \varphi(\varepsilon_j) \, d\varepsilon_j \right] \varphi(\varepsilon_i) \, d\varepsilon_i.
\end{aligned}
$$

We used a trapezoid rule to evaluate these integrals for single evaluation probit models. We ran sensitivity analysis to set the cutoff and step size values for the integration. We set it so that the marginal improvement from increased range and step size would be less than 10^{-4} for the average correct hit probability measure.

For sequential evaluation models, we used GHK algorithm (Keane 1994, Hajivassiliou et al. 1996) with 300 draws to evaluate the multivariate normal probabilities. Matrix A_j was used to transform the region of integration into a rectangular one as described in Rossi et al. (2005). The same algorithm was used for the probabilities associated with the best and worst likelihood components, although we multiply the X matrix by -1 for the "worst" likelihood part, which is a result of the symmetry property of normal distribution.

Appendix D. Calculation of Fit Measures

D.1. Log-Marginal Density (LMD)

LMD for in-sample data is calculated as a harmonic mean of the likelihood values at posterior draws of the parameters (Newton and Raftery 1994).

$$
\log p(y_{\text{best}}, y_{\text{worst}} \mid M) = \left(\frac{1}{R - \text{burnin}} \sum_{r=1}^{R-\text{burnin}} \frac{1}{\ell(B_r \mid M)} \right)^{-1},
$$

where $\ell(B_r \mid M)$ is the log-likelihood vector of values calculated with posterior draws r of parameters, B_r is a set of parameters in the model, and burn-in is the number of draws not used to allow for convergence. In single evaluation models, B includes only preference parameters β, and sequential evaluation models include β, θ, λ, and ψ.

D.2. Correct Hit Probabilities

Correct hit probabilities are the average predicted choice probabilities of making the observed choice in the sample.

$$
\text{HP}(M) = \frac{1}{n_{\text{obs}}} \sum_{m=1}^{n_{\text{obs}}} \bar{\Pr}(y_{m,\,hh,\,\text{best}} = i, y_{m,\,hh,\,\text{worst}} = j \mid M),
$$

where n_{obs} is the number of BW choice tasks in the sample for one respondent, p is the number of alternatives per task, and $\bar{\Pr}(y_{m,\,hh,\,\text{best}} = i, y_{m,\,hh,\,\text{worst}} = j \mid M)$ is the posterior means of probabilities of choosing alternative i as best and j as worst. These predictive probabilities are computed for the holdout sample over the MCMC draws R of parameters excluding burn-in draws. Hit probabilities (HPs) for all respondents are calculated by averaging HP over all respondents.

$$
\begin{aligned}
&\overline{\Pr}(y_{m,\,hh,\,\text{best}} = i, y_{m,\,hh,\,\text{worst}} = j \mid M)\\
&\quad = \frac{1}{R - \text{burnin}} \sum_{r=1}^{R-\text{burnin}} P(y_{m,\,hh,\,\text{best}} = i, y_{m,\,hh,\,\text{worst}} = j \mid B_r, X).
\end{aligned}
$$

D.3. Hit Rates

Hit rates (HRs) are obtained by finding the percent of times the model predicted the observed choice correctly.

$$
\text{HR}(M) = \frac{1}{n_{\text{obs}}} \sum_{m=1}^{n_{\text{obs}}} I(y_m = \hat{y}_m \mid M),
$$

where

$$
I(y_m = \hat{y}_m \mid M) = \begin{cases} 1, & \text{if } y_m = \hat{y}_m, \\ 0, & \text{if } y_m \neq \hat{y}_m, \end{cases}
$$

and $\hat{y}_m = (i, j)$ if

$$
\begin{aligned}
&\overline{\Pr}(y_{m,\,\text{best}} = i, y_{m,\,\text{worst}} = j \mid M)\\
&\quad = \max\{\bar{\Pr}(y_{m,\,\text{best}}, y_{m,\,\text{worst}} \mid M)\colon m = 1, \ldots, p\}.
\end{aligned}
$$

For the heterogeneous case the hit rate is calculated as follows:

$$
\text{HR}(M) = \frac{1}{N_{hh} n_{\text{obs}}} \sum_{hh=1}^{N_{hh}} \sum_{m=1}^{n^{\text{obs}}} I(y_{hh,\,m} = \hat{y}_{hh,\,m} \mid M).
$$

Figure E.1 MCMC Traceplots for SQL5 (Sequential Logit with Inferred Sequence) Estimation—Simulation Study

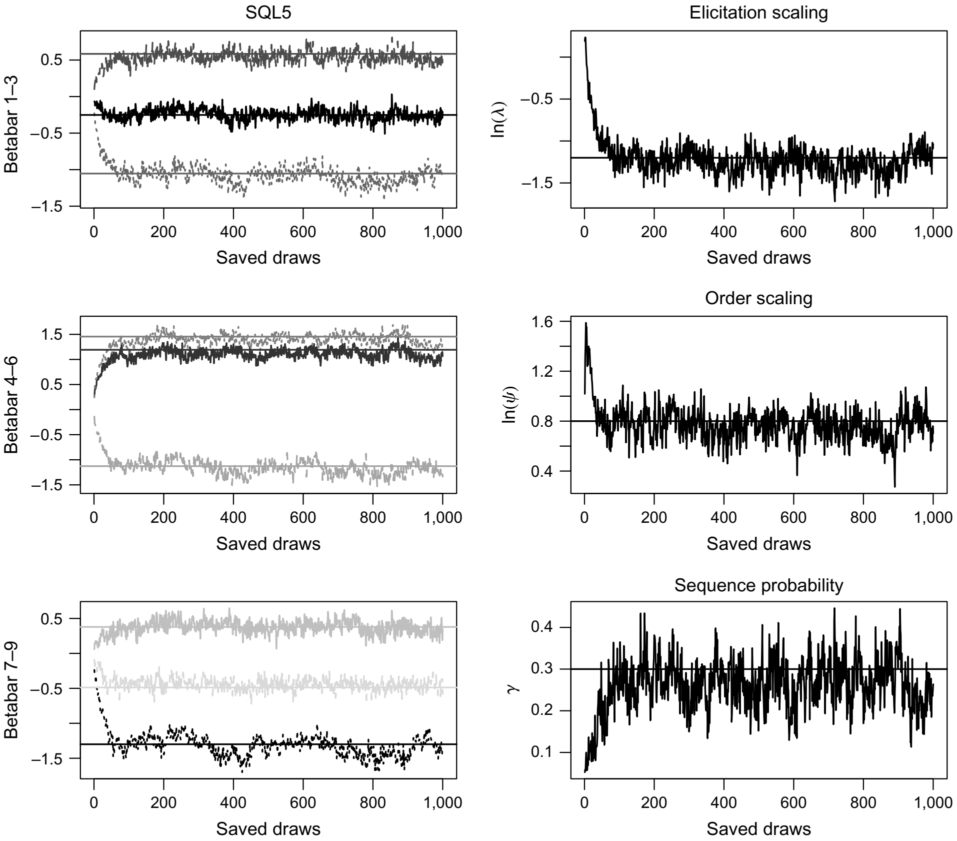


Figure E.2 MCMC Traceplots for SQL5s (Sequential Logit with Observed Sequence) Estimation—Simulation Study

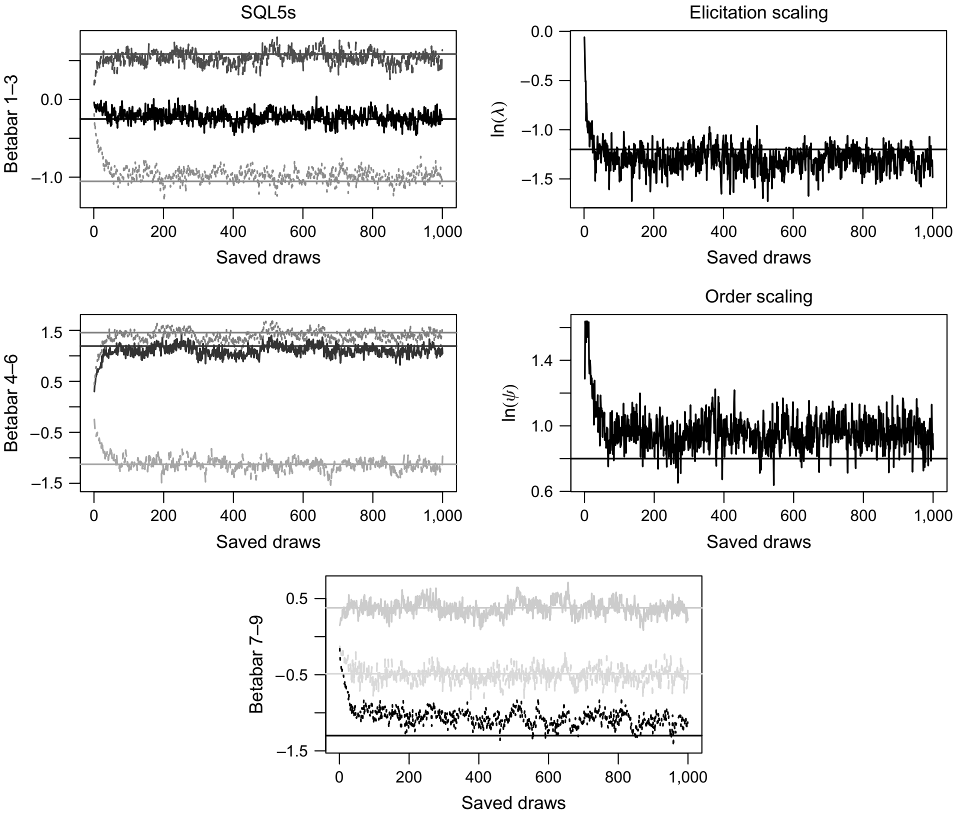


Figure E.3 MCMC Traceplots for SQP5 (Sequential Probit with Inferred Sequence) Estimation—Simulation Study

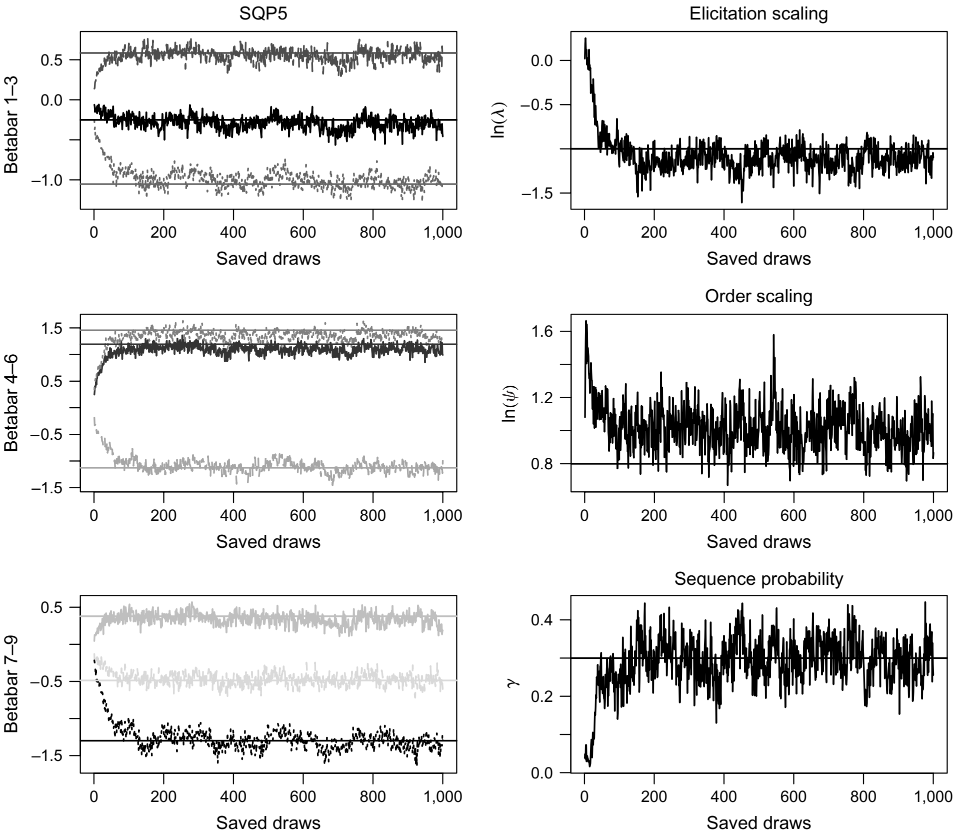


Figure E.4 MCMC Traceplots for SQP5s (Sequential Probit with Observed Sequence) Estimation—Simulation Study

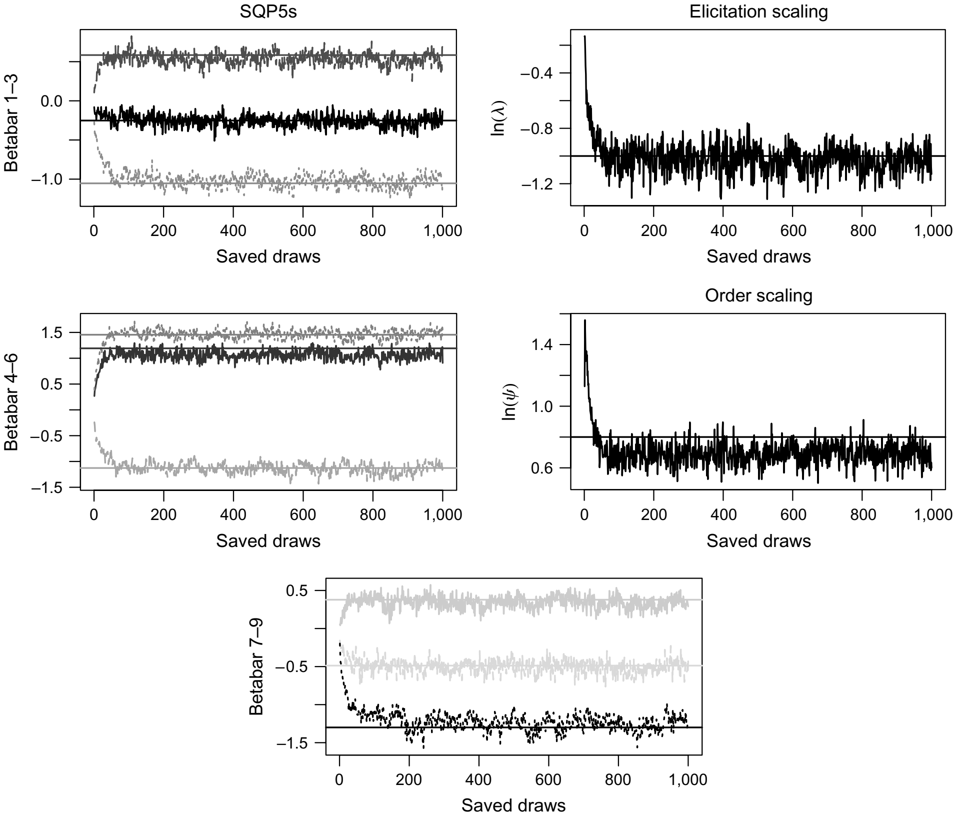


Table E.1 Recovery of Parameters in Simulation Studies

Model	Preference parameters β number of recovered	Sequence parameter ψ true (HPD region)	Elicitation parameter λ true (HPD region)	Sequence probability γ true (HPD region)
Estimated 95% HPD region of the posterior distribution				
Inferred sequence				
Sequential logit (SQL5)	15/15	0.8 (0.531, 0.964)	−1.2 (−1.382, −1.003)	0.3 (0.157, 0.381)
Sequential probit (SQP5)	15/15	0.8 (0.754, 1.253)	−1.0 (−1.559, −0.870)	0.3 (0.201, 0.404)
Observed sequence				
Sequential logit (SQL5s)	14/15	0.8 (0.782, 1.131)	−1.2 (−1.533, −1.113)	NA
Sequential probit (SQP5s)	15/15	0.8 (0.555, 0.812)	−1.0 (−1.215, −0.860)	NA
Estimated 99% HPD region of the posterior distribution				
Inferred sequence				
Sequential logit (SQL5)	15/15	0.8 (0.459, 1.028)	−1.2 (−1.653, −0.912)	0.3 (0.130, 0.415)
Sequential probit (SQP5)	15/15	0.8 (0.700, 1.333)	−1.0 (−1.449, −0.831)	0.3 (0.176, 0.440)
Observed sequence				
Sequential logit (SQL5s)	15/15	0.8 (0.715, 1.169)	−1.2 (−1.656, −1.060)	NA
Sequential probit (SQP5s)	15/15	0.8 (0.516, 0.869)	−1.0 (1.259, −0.814)	NA

Appendix E. Simulation Results of the Sequential Models

We simulated several data sets of 300 respondents with 15 attributes and 15 choices, which corresponded to the size of the smallest sample we collected, according to the sequential evaluation models with inferred and observed sequence of decisions—SQL5, SQP5, as well SQL5s and SQP5s, respectively. Figures E.1–E.4 show traceplots of the MCMC draws for these models. Estimations are based on 50,000 draws that saved every 50th draw. Recovery of parameters for each model is shown in Table E.1 with burn-in 400.

Both proposed sequential logit (SQL5) and probit (SLP5) models with inferred sequence recover all contextual effect parameters: sequential effect ψ, elicitation effect λ, and sequence probability γ, as well as preference parameters.

References

Bacon L, Lenk P (2012) Augmenting discrete-choice data to identify common preference scales for inter-subject analyses. *Quant. Marketing Econom.* 10(4):453–474.

Beach LR (1993) Broadening the definition of decision making: The role of prechoice screening of options. *Psych. Sci.* 4(4):215–220.

Beach LR, Potter RE (1992) The pre-choice screening of options. *Acta Psychologica* 81(2):115–126.

Ben-Akiva M, Morikawa T, Shiroishi F (1992) Analysis of the reliability of reference ranking data. *J. Bus. Res.* 24(2):149–164.

Bettman JR, Luce MF, Payne JW (1998) Constructive consumer choice processes. *J. Consumer Res.* 25(3):187–217.

Böckenholt U (2004) Comparative judgments as an alternative to ratings: Identifying the scale origin. *Psych. Methods* 9(4):453–465.

Carlston DE (1980) The recall and use of traits and events in social inference processes. *J. Experiment. Soc. Psych.* 16(4):303–328.

Chapman RG, Staelin R (1982) Exploiting rank ordered choice set data within the stochastic utility model. *J. Marketing Res.* 19(3):288–301.

Coombs CH, Donnel ML, Kirk DB (1978) An experimental study of risk preference in lotteries. *J. Experiment. Psych.: Human Perception Perform.* 4(3):497–512.

Fiebig DG, Keane MP, Louviere J, Wasi N (2010) The generalized multinomial logit model: Accounting for scale and coefficient heterogeneity. *Marketing Sci.* 29(3):393–421.

Finn A, Louviere JJ (1992) Determining the appropriate response to evidence of public concern: The case of food safety. *J. Public Policy Marketing* 11(2):12–25.

Fitts PM, Deininger RL (1954) *S-R* compatibility: Correspondence among paired elements within stimulus and response codes. *J. Experiment. Psych.* 48(6):483–492.

Fitts PM, Seeger CM (1953) *S-R* compatibility: Spatial characteristics of stimulus and response codes. *J. Experiment. Psych.* 46(3):199–210.

Gumbel EJ (2004) *Statistics of Extremes* (Dover Publications, Inc., Mineola, NY).

Hajivassiliou V, McFadden D, Ruud P (1996) Simulation of multivariate normal rectangle probabilities and their derivatives theoretical and computational results. *J. Econometrics* 72(1–2): 85–134.

Hintzman DL (1986) "Schema abstraction" in a multiple-trace memory model. *Psych. Rev.* 93(4):411–428.

Hoch SJ, Ha Y-W (1986) Consumer learning: Advertising and the ambiguity of product experience. *J. Consumer Res.* 13(2):221–233.

Howard MW, Kahana MJ (2002) When does semantic similarity help episodic retrieval? *J. Memory Language* 46(1):85–98.

Johnson NL, Kotz S (1995) *Continuous Univariate Distributions*, Vol. 2, Wiley Series in Probability and Statistics (Wiley, Chichester, UK).

Keane MP (1994) A computationally practical simulation estimator for panel data. *Econometrica* 62(1):95–116.

Louviere JJ (1991) Best-worst scaling: A model for the largest difference judgments. Working paper, University of Alberta, Edmonton, Alberta.

Louviere JJ (2001) What if consumer experiments impact variances as well as means? *J. Consumer Res.* 28(3):506–511.

Lyubomirsky S, Lee R (1999) Changes in attractiveness of elected, rejected, and precluded alternatives: A comparison of happy and unhappy individuals. *J. Personality Soc. Psych.* 76(6):998–1007.

Machin J (2006) Choosing by selecting or rejecting: The implications of decision strategy for consumer satisfaction. Dissertation, Wharton School of Business, University of Pennsylvania, Philadelphia.

Manski CF (1977) The structure of random utility models. *Theory Decision* 8(3):229–254.

Marley AAJ, Louviere J (2005) Some probabilistic models of best, worst, and best-worst choices. *J. Mathematical Psych.* 49(6):464–480.

Marley AAJ, Pihlens D (2012) Models of best-worst choice and ranking among multiattribute options (profiles). *J. Math. Psych.* 56(1):24–34.

Meloy MG, Russo JE (2004) Binary choice under instructions to select versus reject. *Organ. Behav. Human Decision Processes* 93:114–128.

Newton MA, Raftery AE (1994) Approximate Bayesian inference with the weighted likelihood bootstrap. *J. Royal Statist. Soc. Ser. B (Methodological)* 56(1):3–48.

Ordóñez LD, Benson L III, Beach LR (1999) Testing the compatibility test: How instructions, accountability, and anticipated regret affect prechoice screening of options. *Organ. Behav. Human Decision Processes* 78(1):63–80.

Rossi PE, Allenby GM, McCulloch RE (2005) *Bayesian Statistics and Marketing*, Wiley Series in Probability and Statistics (Wiley, Chichester, UK).

Rossi PE, McCulloch RE, Allenby GM (1996) The value of purchase history data in target marketing. *Marketing Sci.* 15(4):321–340.

Sawtooth Software (2010) The MaxDiff/web system technical paper. Technical report, Sawtooth Software, Salt Lake City.

Sawtooth Software (2012) Report on conjoint analysis usage among Sawtooth Software customers, Sawtooth Software, Salt Lake City.

Shafir E (1993) Choosing versus rejecting: Why some options are both better and worse than others. *Memory Cognition* 21(4): 546–556.

Shafir E, Simonson I, Tvesky A (1993) Reason-based choice. *Cognition* 49(1–2):11–36.

Slovic P (1995) The construction of preference. *Amer. Psychologist* 50(5):364–371.

Snyder M (1981) Seek and ye shall find: Testing hypotheses about other people. Heiman CP, Higgins ET, Zanna MP, eds. *Social Cognition: The Ontario Symposium on Personality and Social Psychology* (Lawrence Erlbaum Associates, Hillside, NJ), 277–303.

Snyder M, Swann WB (1978) Hypothesis-testing processes in social interaction. *J. Personality Soc. Psych.* 36(11):1202–1212.

Tanner MA, Wong WH (1987) The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* 82(398): 528–540.

Tulving E (1972) Episodic and semantic memory. Tulving E, Donaldson W, eds. *Organization of Memory* (Academic Press, New York), 381–402.

Tversky A, Simonson I (1993) Context-dependent preferences. *Management Sci.* 39(10):1179–1189.

Wilk R (1997) A critique of desire: Distaste and dislike in consumer behavior. *Consumption Markets Culture* 1(2):175–196.

Yang S, Allenby G (2000) A model for observation, structural, and household heterogeneity in panel data. *Marketing Lett.* 11(2):137–149.

Yaniv I, Schul Y (2000) Acceptance and elimination procedures in choice: Noncomplementarity and the role of implied status quo. *Organ. Behav. Human Decision Processes* 82(2):293–313.

Yaniv I, Schul Y, Raphaelli-Hirsch R, Maoz I (2002) Inclusive and exclusive modes of thinking: Studies of prediction, preference, and social perception during parliamentary elections. *J. Experiment. Soc. Psych.* 38(4):352–367.