Like It Or Not, You Are Using One Value Representation

Guy E. Hawkins[a*], Towhidul Islam[b], and A. A. J. Marley[c,d]
[a] School of Psychology, University of Newcastle, Callaghan NSW 2308, Australia
[b] Department of Marketing and Consumer Studies, College of Business and Economics, University of Guelph, Guelph ON N1G2W1, Canada
[c] Department of Psychology, University of Victoria, Victoria BC V8W 3P5, Canada
[d] Institute for Choice, University of South Australia Business School, North Sydney NSW 2060, Australia

Abstract

Do we use the same information to decide what we like and what we do not like? Best-worst scaling – where respondents select their most and their least preferred option from a set of options – is an efficient method for obtaining information of direct relevance to this question. Many best-worst scaling applications use multinomial logit (MNL) models to predict such best and worst choice data, explicitly or implicitly assuming that best and worst choices are driven by the same parameters for *utility* information. Some recent literature, however, has criticized this common practice as an overly simplistic representation of the choice process. We tested this assumption by applying three MNL-type models of increasing complexity in their parameterization to the stated best-worst choices from a total of 1200 individuals drawn from 5 data sets. Our Bayesian latent mixture modeling found clear evidence that the same utility parameters drive individuals' best and worst choices, though usually with an additional scale parameter leading to more variable worst choices. These conclusions also held for stated best-worst choices of the same individuals for the same alternatives after a 6, 12, and 18 month delay. We argue that the conclusion of several recent papers that best and worst choices are driven by different utility information or reflect different decision processes are based on inadequate data and/or data analyses.

Keywords: Best-Worst Scaling, Choice, Model, Latent Mixture, Bayesian Model Selection.

We thank Michael Lee for his assistance with the analyses reported in this manuscript. This work was supported by an Australian Research Council Discovery Early Career Researcher Award for Hawkins [Grant number DE170100177]; a Social Sciences and Humanities Research Council (SSHRC) of Canada Grant for Islam [Grant number 430199]; and a Natural Science and Engineering Research Council Discovery Grant to the University of Victoria for Marley [Grant number 8124-98]. The work was carried out, in part, whilst Marley was a Research Professor (part-time) in the Institute for Choice, University of South Australia Business School. The authors declare no competing financial interests.

*Correspondence and requests for materials should be addressed to Guy Hawkins, School of Psychology, University of Newcastle, University Drive, Callaghan NSW 2308 Australia; Email: guy.e.hawkins@gmail.com.

Introduction

Cognitive psychologists and practitioners in applied areas such as consumer choice frequently study preferences between multiattribute alternatives (e.g., computers, wines, retirement plans) using *stated preferences* – that is, the choices are hypothetical – rather than studying *revealed preferences* – that is, actual choices (such as going to a specific restaurant). There are many reasons for this practice. First, there are numerous situations where it is not possible to study revealed preferences: for instance, with an emerging technology, such as autonomous cars, there may not yet be a "market"; and in formulating policy, such as government funding of end-of-life care, the variables cannot be studied and controlled in actual choices. Second, especially for cognitive psychologists, there is the desire to control the variables and options that enter into the task – for example, in the study of *context effects* (Berkowitsch, Scheibehenne, & Rieskamp, 2014; Cohen, Kang, & Leise, 2017; Trueblood, Brown, & Heathcote, 2015, 2014; Turner, Schley, Muller, & Tsetsos, in press). There is ongoing debate of the value of stated preference data (Cherchi & Hensher, 2015). However, Louviere, Hensher, and Swait (2000) present extensive results for the external validity of stated preference methods once one allows for different variabilities in stated and revealed preference data, and Lancsar and Swait (2014) argue that external validity should be pursued from the initial conceptualization and design of any stated preference study.

Typical stated preference tasks (often called *discrete choice experiments* or *DCEs*) consider choices between products or services, for instance, preferences for mobile phones. Each mobile phone (option, profile) is defined by a number of attributes (such as price, battery life, camera quality, and so on) each of which takes one of a number of levels (e.g., the cost of the phone might be \$200, \$400, or \$600) across options. Respondents are presented with a number of choice sets with, say, four options, and for each set they are asked to make a decision, for instance, to indicate their most-preferred phone. Through careful manipulation of the attribute and level structure of such *profiles* across choice sets, the analyst can use patterns of observed choices to infer respondents' preference, or utility, parameters – that is, the subjective value respondents place on the attributes and levels that comprise the profiles; and, as required, scale (consistency, variability) parameters (Louviere et al., 2000).

Best-Worst Scaling

The efficiency of parameter estimation in DCEs depends critically on a number of experimental design factors, including the number of choice sets respondents complete, and the allocation of the attributes and levels structure to profiles and choice sets (Cherchi & Hensher, 2015; Hensher & Ho, 2016). In this light, one particularly useful extension of DCEs has been *best-worst scaling* (BWS; Louviere, Flynn, & Marley, 2015). In a typical BWS task, each respondent is presented with a number of choice sets, as in a DCE, and asked to report both their most preferred – or *best* – alternative, and their least preferred – or *worst* – alternative from those available. Although asking for the *worst* in addition to the *best* is a seemingly small change over conventional DCEs, BWS has several advantages. For example, it works with, rather than against, the propensity to respond to extreme options, which makes BWS tasks quite natural. For the analyst, BWS can provide improved discrimination power and sensitivity compared to standard DCEs (Alberni, 2012). The advantages of BWS over standard best choice is seen by the fact that Sawtooth Software, a major provider of DCE software, consulting and educational services, reported that 70% of their users used BWS methods in 2016 (Orme, 2016). For a summary of the history of BWS (sometimes referred to here as "best-worst") see Flynn and Marley (2014), and for numerous chapters on recent BWS applications see Louviere et al. (2015)[1].

A significant proportion of BWS applications to date assume that best and worst choices are based on the same utility parameters, sometimes with allowance for a different scale (variance) parameter associated with the worst choices, and that each of the best and worst choices (or, alternatively, the best-worst paired choices) satisfy forms closely related to the *multinomial logit* (MNL) model (Appendix

[1]Some material presented here is adapted from Louviere et al. (2015, Ch. 6) with permission from Cambridge University Press.

A; Louviere et al., 2000, and below); MNL models of choice are conceptually simple and commonly used in modeling stated preferences in applied areas such as consumer choice and transportation. In particular, the utility parameters that drive selection of the worst option from a set of options are assumed to be the negative of those assumed to drive selection of the best option, with possibly different scale factors (variance) for best and worst choices; this assumption implies that the process of selecting the best and worst options from a set of available options draws upon a single latent dimension representing the value of those options. Such focus on simple models of BWS has been recently criticized on empirical grounds (Dumont, Giergiczny, & Hess, 2015; Dyachenko, Walker Reczek, & Allenby, 2014; Giergiczny, Dekker, Hess, & Chintakayala, 2017; Rose, 2013). We comment on these critiques in the final section of the paper in the light of our successful application of such models to data sets with many individuals, each with a relatively small number of choices.

Here, we study three MNL-based representations that relate the utility parameters underlying best and worst choices (see Table 3, and the later model details). The names we have ascribed to each model can be interpreted in terms of the relationship that is assumed between the parameters underlying best and worst choices. The first, and simplest, representation is the *sign and scale related utility* model. It assumes that the utilities underlying selection of the worst option are exactly the 'opposite' of the utilities underlying selection of the best option; therefore, if you are highly likely to choose a particular option as the best then you are highly unlikely to choose that option as the worst. The second, slightly more complex, representation is the *sign related utility* model. It assumes that the utilities that drive selection of the worst option are generally the 'opposite' of the utilities that drive selection of the best option, but it allows for the possibility that worst choices might be more (or less) variable than best choices; the latter modeled by a nonnegative scale parameter. The third and final representation is the *independent utility* model. It assumes that the utilities that drive selection of the best and worst options are not related; in other words, there is no common information between the two types of choices.

We present data for two of three *cases* that have been studied in the literature on best-worst (Louviere et al., 2015). The *object case* is used when the question of interest is the relative value associated with each of a list of objects; these might be brands, public policy goals, or any set of objects that can be meaningfully compared. Figure 1A provides an example of the object case involving a choice between policy options for saving money in the Government budget. In contrast, the *profile case* is used when the relative value of multi-attribute options (*profiles*) is of interest; Figure 1B provides an example involving a choice between cellular phone profiles. Each profile has multiple attributes (e.g., price, memory, video, etc.), each of which has an *attribute level* (e.g., the possible prices might be \$49, \$129, \$199 or \$249). The question of interest is which profile – particular combination of attribute levels – is most preferred (best) and which (other profile) is least preferred (worst). The interested reader is referred to Louviere et al. (2015) and Flynn and Marley (2014) for details on the principles and design of the different cases.

Data sets

We applied a hierarchical Bayesian latent mixture model to each of five large-scale best-worst choice data sets; three profile cases (all drawn from Islam & Louviere, 2015) and two object cases (both drawn from Louviere et al., 2015, Ch. 2). The *profile case* data sets examined preferences for laundry detergent, pizza delivery, and toothpaste. Each of these surveys was administered to a (separate) random sample of 600 members from the Pureprofile online panel in Australia. Participants were included in the analysis if they met two screening criteria: purchase or use of the product in the previous 3-6 months, and willingness to participate in four successive survey waves of data collection over two years (described in detail below). The final sample sizes were: laundry detergent ($N = 218$), pizza delivery ($N = 186$), and toothpaste ($N = 234$). Demographic variables were collected in addition to the behavioral variables reported in this manuscript. The demographic data were relevant to the original publication (Islam & Louviere, 2015), however, they are not relevant to the current research questions so we do not report them further.

The design of the survey was identical for each of the three products, with the only difference

being the particular set of attributes that pertained to each product. Table 1 shows the attributes and attribute levels from the laundry detergent data set; the corresponding tables for the other two data sets are in the Supplementary Material. Each data set was collected in a discrete choice experiment using an experimental design suitable for testing that value (utility) functions are additive across attributes and that the underlying choice processes are well-approximated by MNL models (defined in the next section). The experimental design was created using Louviere et al.'s (2008) method (see also Louviere et al., 2015, Ch. 2, 4) and, in each domain, assigned 16 profiles to 20 choice sets, each with 4 options. For each choice set respondents first selected their most-preferred profile from the 4 available options and then selected their least-preferred profile from the remaining three options. The option selected as best remained on the screen whilst the worst option was being chosen; but, at this stage, the best option could not be changed.

Table 1: Attributes and attribute levels for the laundry detergent data set.

Attributes	Levels			
Brand	Radiant	Omo	Cold Power	Dynamo
Price per 100g (or 100mL)	\$0.40	\$0.60	\$0.80	\$1.00
Size (number of washes)	0.5kg/litre (10 wash)	1.0kg/litre (20 wash)	1.5kg/litre (30 wash)	2.0kg/litre (40 wash)
Form	Powder	Liquid		
Cold water wash	No	Yes		
Eco friendly	No	Yes		
Fabric softener	Not added	Added		
Colour care and remove stain	No	Yes		
Fragrance	No	Yes		

All participants across the three data sets contributed data at an additional 3 time points that were separated by 6 month intervals, where the same 20 choice sets were shown at each time point, for a total of 4 waves of data spread across 18 months. This longitudinal experimental design thus provides a unique opportunity to test the consistency of choices – and hence the predictions of the models – over time. We return to this point below.

The *object case* data sets examined preferences for attributes of government saving (resp., spending) in the budget. The object case surveys commenced with a random sample of 600 members from the Pureprofile online panel in Australia, where each participant was invited to provide responses to both surveys ($N = 561$ provided complete data sets). For these surveys, an experimental design assigned the 9 attributes to 12 choice sets each with 3 options, where an option is a particular object statement (Table 2 shows the stimuli from each survey). As above, respondents first selected their most-preferred profile from the 3 available options and then selected their least-preferred profile from the remaining 2 options. In these tasks, the option selected as best was removed from the screen before the worst option could be chosen.

Multinomial Logit Models of Choice

We model best-worst choice with a nested family based on *multinomial logit (MNL) models* (Appendix A) of best (resp., worst) choice. The models make increasingly complex assumptions about the relationship between the information used to make the best (resp., worst) choices. We begin with notation that refers to *choice options* (or *options*) without distinguishing between the object and profile cases. Appendix A provides the additional notation required for the profile case. We also present the results in terms of a numeric *utility* value associated with each choice option (and, as relevant, with each of its attribute levels).

Let S with $|S| \geq 2$ denote the finite set of potentially available choice options, and let $D(S)$ denote

Table 2: Attributes in the two object case data sets: budget saving and budget spending.

Data set	Attributes
Budget Saving	Not proceeding with the company tax cut
	Reprioritising the tax reform
	Improving fairness in the tax system
	Improving compliance measures to prevent fraud and tax evasion
	Defer spending and acquisitions in Defence
	Defer increase in foreign aid commitments
	Introducing caps and changes to eligibility for certain benefits
	Increase in the Road & User Charge for heavy vehicles
	Increase in departure tax for travellers
Budget Spending	Spreading the benefits of the boom and support for families
	Helping the most vulnerable in society
	Building an aged care system for the future
	Assisting Australian businesses while adjusting to structural changes
	Improving national infrastructure
	Building opportunities for a more productive workforce
	Continue developing a personally controlled electronic health record system
	Maintaining the current levels of bio security operations
	Continuation of Australia's presence in Afghanistan and the Middle East and supporting stability in East Timor

the *design*, that is, the set of (sub)sets of choice alternatives that occur in the study. For example, participants might be asked about their preferences for mobile phones by repeatedly asking them for choices amongst sets of four different phones: S represents the collection of mobile phones in the study, and each element of the set $D(S)$ represents the set of phones provided on one particular choice occasion. For each $X \in D(S)$, with $|X| \geq 2$, $B_X(x)$ denotes the probability that alternative x is chosen as best in X, $W_X(y)$ the probability that alternative y is chosen as worst in X, and $BW_X(x, y)$ the probability that alternative x is chosen as best in X and the alternative $y \neq x$ is chosen as worst in X.

We tested three MNL-based models of best-worst choice, which we first describe for the profile case, followed by the object case. The models are nested, with model 3 nesting model 2, which in turn nests model 1. Table 3 summarizes the forms for the scales for best (resp., worst) in the three models. For the profile case, we use boldface for the (multiattribute) options.

Table 3: Parametric forms for selecting option x as best and option y as worst under the constraints of models 1, 2, and 3.

		Parametric form	
		Best	Worst
Model 1	Sign and scale related utilities	$b(x)$	$-b(y)$
Model 2	Sign related utilities	$b(x)$	$-\alpha b(y)$
Model 3	Independent utilities	$b(x)$	$w(y)$

Profile Case

Model 3, *independent utilities,* has the form: for each $\{\boldsymbol{x}, \boldsymbol{y}\} \in X \in D(S), \boldsymbol{x} \neq \boldsymbol{y}$,

$$\begin{aligned} BW_X(\boldsymbol{x}, \boldsymbol{y}) &= B_X(\boldsymbol{x}) W_{X-\{\boldsymbol{x}\}}(\boldsymbol{y}) \\ &= \frac{e^{b(\boldsymbol{x})}}{\sum_{r \in X} e^{b(\boldsymbol{r})}} \frac{e^{w(\boldsymbol{y})}}{\sum_{\boldsymbol{s} \in X-\{\boldsymbol{x}\}} e^{w(\boldsymbol{s})}}, \end{aligned}$$

where each of b and w is an additive utility representation over m attributes; that is, there are scales b_i and w_i, $i = 1, ..., m$, such that for all $\boldsymbol{z} \in S$, $b(\boldsymbol{z}) = \sum_{i=1}^{m} b_i(z_i)$ and $w(\boldsymbol{z}) = \sum_{i=1}^{m} w_i(z_i)$.

Model 2, *sign related utilities,* is the special case of model 3 where there is a nonnegative scale factor[2] α such that for all $\boldsymbol{z} \in S$, $w(\boldsymbol{z}) = -\alpha b(\boldsymbol{z})$. The inclusion of a scale factor α at the worst (second) choice (equivalently, stage) corresponds to the assumption that the same utility information is used for best and worst choices, but worst choices may be more or less variable than best choices (smaller α indicates larger choice variability). For example, when $\alpha < 1$ worst choices are more variable[3] than best choices, and when $\alpha = 1$ model 2 becomes model 1.

Model 1, *sign and scale related utilities,* is the special case of model 3 where for all $\boldsymbol{z} \in S$, $w(\boldsymbol{z}) = -b(\boldsymbol{z})$. As noted earlier, this is one of the most frequently studied models.

Object Case

In the object case, the choice alternatives are treated as if they have no component attributes (e.g., just a brand name, or the name of a city). Therefore, we replace the boldface notation $BW_X(\boldsymbol{x}, \boldsymbol{y})$ of the profile case with $BW_X(x, y)$ and there is no decomposition of the utility functions b and w in the three models to an additive form. With these changes, models 1-3 of the object case have the same form as models 1-3 of the profile case.

Hierarchical Bayesian Latent Mixture Model

We selected between the three models with Bayesian latent mixture model analyses. Bayesian latent mixture models have many advantages over traditional model selection techniques (e.g., Akaike or Bayesian Information Criterion, as we demonstrate in Appendix C and discuss later), including the ability to simultaneously estimate participant- and population-level parameters and posterior model probabilities in a principled manner, even from relatively sparse individual participant data. Model selection via Bayesian inference also naturally accounts for model complexity; it does not rely on approximations as do the information criteria.

Conceptually, our approach involves simultaneously estimating the parameters of each of the three models separately for each participant. The descriptive adequacy of each model's account of each participant's data is balanced against the model's complexity according to Bayesian principles, which gives the marginal likelihood. The marginal likelihoods for the three models are then contrasted to assign a participant-level posterior probability to each model. This posterior probability is the model selection metric at the participant level; it indicates the probability that each model is the true account of the data, relative to the other models under evaluation and the assumption that the true model is in the set. The individual participant posterior model probabilities then inform the population-level base rate probability for each of the models, which provides the posterior probability of each model across the sample of participants.

Figure 2 shows the graphical model that implements the Bayesian latent mixture model, with standard graphical model notation (Jordan, 2004; Lee, 2008; Lee & Wagenmakers, 2013). Latent and

[2]There is an implicit scale factor with value 1 in the representation of the best choices. This constraint has to be applied for both the utility form b and the scale α to be identifiable (Swait & Louviere, 1993).

[3]The data we analyze is not suitable for deciding whether such variability is due to the type of choice (best or worst) and/or stage of choice (first or second). See Dyachenko et al. (2014) for data and theory on this point.

observed variables are represented with open and shaded nodes, respectively. Circular and square nodes represent continuous and discrete variables, respectively. Single and double-bordered nodes indicate stochastic and deterministic variables, respectively. Rectangular plates indicate independent replications over participants, choice sets, and attribute structures. The notation used in the graphical model was selected to be as similar as possible to the model specification presented in the previous sections.

The model assumes participant j's selection of the *best* option in choice set X, $B_{X,j}$, is distributed according to a categorical distribution specified by the probability vector $\widehat{B}_{X,j}$. $\widehat{B}_{X,j}$ is an i-length vector whose values are determined by the i options available in choice set X and participant j's utility for those options. In the profile case, shown in Figure 2, profile i's utility is the sum of its constituent attribute-level utilities. Specifically, when considering the *best* option, level k of attribute m has utility b_{jkm},[4] and $\boldsymbol{z}_{X,i}$ is an m-length vector of indicator variables that denotes the combination of attribute levels present in profile i of set X (and is constant across participants), which has summed utility $b_j(\boldsymbol{z}_{X,i})$. In the object case, where options do not have an attributes and levels structure, the notation simplifies to $z_{X,i}$ (i.e., bolding omitted, indicating a scalar value) and the plates representing replications over m attributes and k attribute levels are replaced with a single plate with replications over n objects.

Participant j's selection of the *worst* option in choice set X, $W_{X,j}$, is also distributed according to a categorical distribution specified by the probability vector $\widehat{W}_{X,j}$. However, unlike best choices, for worst choices the value of the deterministic node $\widehat{W}_{X,j}$ is determined by ψ_j, a participant-level indicator variable that takes on the value of 1, 2 or 3 on the basis of a vector $\phi = (\phi_1, \phi_2, \phi_3)$ that gives the probability of a participant using model i, $i = 1, 2, 3$; thus ϕ_i is the base rate probability of model i, and $\sum_{i=1}^{3} \phi_i = 1$. There is a single ψ_j for each participant j, meaning that each participant's data is assumed to arise from the same model across all choice sets. Uncertainty around which model is correct (relative to the set of models under consideration) is represented with probabilities.

ψ_j influences the attribute-level utilities assumed for selection of the worst option, w_{jkm}. In model 1, with sign and scale related utilities ($\psi_j = 1$), they are simply the 'mirror image' of the corresponding utilities for selection of the best option; $w_{jkm} = -b_{jkm}$. In model 2, with sign related utilities ($\psi_j = 2$), they are a scaled version of the corresponding utilities for selection of the best option; $w_{jkm} = -\alpha_j b_{jkm}$, where α_j reflects participant j's sensitivity (of the utility parameters) in worst relative to best choices ($\alpha < 1$ indicates worst choices are more variable than best choices). In model 3, with independent utilities ($\psi_j = 3$), they are independent of the corresponding utilities for selection of the best option.

We performed Bayesian inference over the graphical model using Markov chain Monte Carlo (MCMC) methods in Matlab and Just Another Gibbs Sampler (JAGS; Plummer, 2003). All prior distributions on parameters were relatively vague, and are given on the right of Figure 2. We took 6000 samples from the posterior distribution of the parameters from each of 3 chains with a burn-in period of 1000 samples, for a total of 15000 samples from the posterior distribution of the parameters. Convergence to the posterior distribution was checked with the $\widehat{R}$ statistic (Brooks & Gelman, 1998).

Results

Bayesian Latent Mixture Modeling of Individual Participant Choices

We now show that our Bayesian latent mixture modeling provides consistent evidence across the five data sets: best and worst choices are most consistent with a single representation of utility (model 1 or 2); there is very little evidence to indicate that utility parameters are independent across best and worst choices (model 3). That is, the information a person uses to decide what they like and what they don't like is identical (model 1) or the same up to a scale (model 2).

Profile Case. In the three *profile case* data sets there was very strong evidence at the level of individual participants that the utility parameters that drive choices for the most-preferred option are sign related to the utilities that drive choices for the least-preferred option (i.e., model 2), with very little

[4] A separate set of b_{jkm} parameters were estimated for each model.

evidence to suggest the utility parameters are sign and scale related (model 1) or independent (model 3); see the right panel of Figure 3 which shows the posterior probability of each model (1, 2, and 3) for each participant. The left panel shows the population-level posterior probability for each model – that is, the base rate probability of membership in each of the three model classes, given the individual participant posterior probabilities. There was at least some evidence, according to the Bayes factor, that the population-level posterior probabilities were non-zero in each data set for models 2 and 3, and equal to 0 for model 1 (analysis details provided in Appendix B). Separately, Table 4 shows pairwise comparisons between the population-level posterior probabilities according to the Bayes factor (see Appendix B for details). Through visual inspection of Figure 3 and the Bayes factors shown in Table 4, we can determine there is very strong evidence for model 2 over models 1 and 3 at the population level.

Inspection of the parameter estimates for the single parameter that differentiates model 2 from model 1 – the scale factor, α – shows that choosing the least-preferred option is more variable than choosing the most-preferred option from a set of options (summary over individual-participant posterior medians for α: Laundry detergent – median = .44, IQR [.25, .95]; 76% of participants with posterior median for $\alpha < 1$; Pizza delivery – median = .43, IQR [.26, .91], 78% $\alpha < 1$; Toothpaste – median = .70, IQR [.35, 1.86], 60% $\alpha < 1$).

Table 4: Pairwise comparisons between the population-level posterior model probabilities for each data set. Evidence is shown as the log Bayes factor ($\log BF$) where a negative value provides evidence in favor of the null hypothesis (that the pair of models have equal posterior probability) and a positive value provides evidence in favor of the alternative hypothesis (that one model has a larger posterior probability than the other, though these hypothesis tests are agnostic as to which of the two models has the higher posterior probability). Model's are referred to as sign and scale related utilities (1), sign related utilities (2), and independent utilities (3).

Best-Worst Case	Data set	Model comparison	$\log BF$
Profile	Laundry Detergent	1 vs 2	29.7
		1 vs 3	-2.75
		2 vs 3	60.7
	Pizza Delivery	1 vs 2	49.5
		1 vs 3	-1.28
		2 vs 3	50.7
	Toothpaste	1 vs 2	38.0
		1 vs 3	-1.34
		2 vs 3	48.8
Object	Budget Saving	1 vs 2	-.29
		1 vs 3	.36
		2 vs 3	-.81
	Budget Spending	1 vs 2	-.23
		1 vs 3	4.1
		2 vs 3	0.52

Object Case. There is less decisive evidence in favor of one of the three models in the two *object case* data sets: at the level of individual participants there was a slight edge for model 1 over model 2, though the difference was equivocal at best, with less but not negligible evidence for model 3. The uncertainty in the individual participant posterior probabilities translated into much wider credible intervals around the population-level posterior model probabilities than was observed in the profile case data sets (left panel, Figure 3); this was confirmed via the Bayes factor whereby all three models had non-zero population-level posterior model probabilities for both data sets. The pairwise Bayes factor comparisons between models are again shown in Table 4, which suggest in the budget saving data set there was a slight edge for model

1 over model 3 but little difference between models 1 and 2, and models 2 and 3. In the budget spending data set, there was some evidence that models 1 and 2 were both superior to model 3, but evidence that models 1 and 2 performed similarly to each other.

A reasonable conclusion is that data from the object case are less informative than data from the profile case. This is because each attribute level in the profile case data sets (cf. Tables 1) appeared at least once in every choice set (i.e., once in each of 20 trials per participant). In contrast, each unique object (of 9) in the object case data sets did not appear in each trial (each of the 9 objects appeared in only 4 of 12 trials per participant). This is a large difference in terms of the influence that an object (object case) or an attribute level (profile case) can exert on the predictions from different model parameterizations (e.g., $\alpha = 1$ in model 1 vs. freely estimated α in model 2). Therefore, in these data sets, the profile case designs appear to be sufficiently informative to warrant the additional complexity of an individual-participant estimate of a scaling factor between best and worst choices (α), but this does not appear to be true for the object case. For the object case data, we conclude that each of the models that assume a relationship between best and worst utilities provide a good account, whether the model assumes that utilities driving best and worst choices are sign and scale related (model 1), or simply scale related (model 2). This conclusion is reinforced by the finding that the population-level model probabilities for the use of models 1 *or* 2 (i.e., the sum of the two) was much greater than model 3, in both the budget saving and spending data sets ($\log BF = 16.6$ and 50.4, respectively).

We confirmed the reliability of the conclusions for the profile and object cases from the latent mixture modeling analysis with an alternative approach: we estimated the parameters of model 2 from individual participant data and compared individual participant α estimates in a null model (model 1, when $\alpha = 1$) to a full model (model 2, where $\alpha \sim Gamma(2, 1)$), using the Savage-Dickey density ratio test (for details, see Appendix B). The prior distributions over the utility and α parameters in this individual participant estimation were as specified in the latent mixture model outlined in Figure 2, which did not impose a hierarchical structure over the utility and α parameters. This means that the level of updating from the prior to the posterior distribution of each participant's parameters was equivalent across the previous latent mixture modeling analysis and this individual participant analysis. This analysis produces a Bayes factor for nested model comparisons. By computing log Bayes factors for each participant and then aggregating across participants, we can report the mean log of the Bayes factor in favor of model 1 over model 2, $\log BF_{12}$, where negative values indicate evidence for model 2. In the profile case data sets we found mean evidence in favor of model 2: mean $\log BF_{12}$ across participants for laundry detergent -2.76, pizza delivery -2.36, and toothpaste -1.79. That is, model 2 was approximately $\frac{1}{\exp(-2.3)} \approx 10$ times more likely than model 1, across the three data sets. In the object case data sets there was little evidence discriminating between models 1 and 2 in the budget saving and spending data; the 95% credible interval on α included the value of 1 for many participants, and a combined log Bayes factor across participants was only marginally below 0 (mean $\log BF_{12}$ across participants: budget saving $-.20$, budget spending $-.15$; on average, model 2 was approximately $\frac{1}{\exp(-.175)} \approx 1.2$ times more likely than model 1). These results are consistent with the latent mixture modeling analysis for these data.

Prior Sensitivity of the Utility and Choice Variability Parameters

We tested the prior sensitivity of the utility (b, w) and scaling (α) parameters to ensure our conclusions were robust to prior assumptions. In addition to the relatively wide prior distribution on the utility parameters assumed in Figure 2, we also studied a prior on the utility parameters with the same mean as the wide prior, but with the variance 10 times smaller. For both the wide and narrow priors, all data sets showed strong evidence for model 1 or 2 – that is, a single utility representation for best and worst choices – and very little evidence in favor of model 3.

The α parameter of model 2 corresponds to the level of relative variability between best and worst choices: $\alpha < 1$ indicates that selecting the worst option is a more variable process than selecting the best option, and vice versa for $\alpha > 1$. To confirm that α can be reliably estimated at the individual-participant level, we performed a number of prior sensitivity and individual participant analysis checks.

First, we tested the sensitivity of α estimates under different prior assumptions, comparing the prior distribution assumed in the previous section, $\alpha \sim Gamma(2, 1)$, with two normal distributions that were both truncated to positive values and with variance $\sigma^2 = 10$, though one centered at 0 and the other at 1; $\alpha \sim N(0, 10)_{(0,)}$ and $\alpha \sim N(1, 10)_{(0,)}$. The three prior distributions on α made negligible differences to the estimated values of α or the substantive latent class assignments reported above; model 2 was preferred for almost all participants.

Second, when model 2 was applied to each data set in isolation (i.e., not as a latent mixture over the three models), the three prior distributions produced very similar posterior α estimates. For example, in the pizza delivery data set, the Gamma, and Normal with $\mu = 0$, resp., $\mu = 1$ prior distributions led to median α estimates across participants of .47, .41, and .41, respectively.

Bayesian Analyses of Choices Pooled Across Participants

To confirm that our per-participant model selection inferences were not unduly biased by the small number of observations per participant, we repeated the comparison across the three models by more conventional methods in the discrete choice literature: analyzing data pooled across participants, which is equivalent to assuming that there was zero across-participant heterogeneity in preferences, or that all data were produced by a single participant who made very many choices.

We first attempted to analyze the pooled data using the same latent mixture model approach as above. However, there was evidence of poor mixing of the MCMC chains; every iteration of every chain for each data set favored model 2. Even when we pooled across a smaller subset of participants and implemented model priors heavily biased against model 2, the resultant posteriors still indicated very strong evidence for model 2.

To confirm that the results of this pooled analysis were valuable – that is, not due to potential sampling problems in the latent mixture modeling analysis – we also performed model selection based on independent applications of each model to the pooled data, again using the Savage-Dickey density ratio test (Appendix B). We performed two key comparisons: model 1 vs model 2, and model 1 vs model 3; by transitivity we can compute the Bayes factor for model 2 vs model 3.

i. For the profile case, the log of the Bayes factors in favor of model 1 over model 2, $\log BF_{12}$, where negative values indicate evidence for model 2, were: laundry detergent $\log BF_{12} = -1084$, pizza delivery $\log BF_{12} = -992$ and toothpaste $\log BF_{12} = -1113$, each indicating very strong evidence for model 2 over model 1. Similar results were obtained for the object case data sets: budget saving $\log BF_{12} = -496$ and budget spending $\log BF_{12} = -606$.

ii. For the profile case data, the log of the Bayes factors in favor of model 1 over model 3, where negative values indicate evidence for model 3, were: laundry detergent $\log BF_{13} = 2$, pizza delivery $\log BF_{13} = 13$ and toothpaste $\log BF_{13} = 19$. Similar results were again obtained for the object case data sets: budget saving $\log BF_{13} = 11$ and budget spending $\log BF_{13} = 11$.

Taking i. and ii. together, since model 2 was favored over model 1, and model 1 was favored over model 3, the set of Bayes factors provides strong evidence in favor of model 2 over each of models 1 and 3 when choices are pooled across participants.

Out-Of-Sample Prediction of Future Choices

All participants across the 3 profile case data sets contributed data at an additional 3 time points that were separated by 6 month intervals, where the same 20 choice sets were shown at each time point, for a total of 4 waves of data spread across 18 months. The previous analyses modeled only the first wave of data collection, so this longitudinal experimental design provides a unique opportunity to test the consistency of choices – and hence the predictions of the models – over time.

We used the longitudinal design to take a different perspective on model selection by using out-of-sample prediction of future behavior. Independently for each participant and model, we used Bayesian parameter estimation to obtain the posterior distribution of the parameters for the wave 1 data (there was *not* a latent mixture component in this analysis). We then used the posterior distribution of the

parameters estimated at wave 1 as the prior predictive distribution for best-worst choices at waves 2, 3 and 4, separately for each participant and model; that is, the prior predictive probability for each best-worst choice, given the wave 1 parameters estimated from each model. This procedure is analogous to calculating the marginal likelihood of the wave 2, 3 and 4 data given the posterior distribution of the parameters estimated from wave 1. Where the prior predictive probability was estimated as 0 (i.e., no posterior samples from wave 1), it was assumed a single sample from a single MCMC chain was associated with that choice to give non-zero probability. We then took the logarithm of the prior predictive probability for each observed best-worst choice and summed across choices within participants, to obtain the log-likelihood of each participant's best-worst choices separately for each model and wave. By comparing these summed log-likelihoods between models, this analysis allowed us to test which model most accurately predicted future choices given the previous choices that were observed.

Figure 4 shows the distribution across participants of differences in log-likelihood for the prediction of out-of-sample best-worst choices at wave 2, separately for the set of 3 pairwise model comparisons (rows) and 3 data sets (columns). The corresponding predictions for waves 3 and 4 are shown in Supplementary Material (Figures D1 and D2). The arrows in each row indicate how to interpret the difference distribution; in the top row, for example, positive values indicate that model 2 (sign related utilities) outperformed model 1 (sign and scale related utilities), and vice versa for negative values. The value above each panel shows the sum of the distribution of log-likelihood differences across participants. This provides a summary of the out-of-sample predictive performance for each pairwise model comparison at the population level.

All three data sets showed the same pattern: model 2 more accurately predicted future best-worst choices than model 1 (top row), and models 1 and 2 both outperformed model 3 (middle and lower rows, respectively). This pattern was observed across the three waves of data. The simplest conclusion to draw is that the simpler models that were considered – those that assume a systematic relationship between best and worst utility parameters – provide a better prediction of individual level future choices, with an edge for sign related utilities (model 2) over sign and scale related utilities (model 1). This finding is most likely due to the most complex model (model 3, with independent utilities) over-fitting the wave 1 choice data, thereby reducing its capacity to predict future choice data. The result is consistent with the latent mixture modeling of the same three models for the observed wave 1 data for the profile case. Taken together, model 2 – with sign related utilities – provides the best in-sample and out-of-sample account of best-worst choice data.

Discussion and Conclusions

We investigated whether selecting the most- and least-preferred options from a set of options – known as best-worst scaling – relies on a single latent preference scale; that is, whether people draw on the same information to decide what they like and what they don't like. Using principled model selection techniques, we found convincing evidence that choosing the best and the worst options from a set is driven by a single latent preference scale. This finding held across complex, multi-attribute options, or profiles – like preferences for pizza or laundry detergent – and objects that do not have a specified attribute structure – like preferences for spending in government budgets – across 5 data sets each with a large sample size (N range from 186 to 561).

When deciding between multi-attribute options, the evidence overwhelmingly supported a model that assumes the preference-related information driving selection of the worst option is a scaled form of the preference-related information driving selection of the best option; the scaling was such that observed worst choices were more variable than best choices.[5] This result was supported when the models were

[5]Model 2 has two interpretations that are formally equivalent and equally valid given the architecture of the model. The first interpretation is that the *attribute-level utilities* for 'worst' choices are *scaled* versions of the attribute-level utilities for 'best' choices; the psychological interpretation is that people are less sensitive to attribute-level information in 'worst' relative to 'best' choices (if $\alpha < 1$). The second interpretation is that the *attribute-level utilities* have the *same* values across 'best' and 'worst' choices, except for sign, and the overall (summed) utility when selecting the worst option from a choice set is

evaluated using in-sample techniques – individual-participant posterior model probabilities obtained via hierarchical Bayesian latent mixture modeling – and predictive out-of-sample techniques – parameter estimates from one time point used to predict best-worst choices up to 18 months into the future.

When deciding between objects, the evidence was more equivocal: our model selection techniques could not convincingly discriminate between two of the three models: one that assumes preference-related information driving worst choices is a 'mirror' image (model 1) and a 'noisier' version (model 2) of the preference-related information driving best choices (cf. Figure 3). There was little to no evidence in favor of the model that assumes independent preference-related information across best and worst choices in the object case (i.e., model 3).

The equivocal model selection results regarding the two unidimensional preference scale models in the object case reflect a sparsity of information in these designs: for each participant, each object was only available to be selected (as best or worst) in 4 trials, from a total of 12 trials. In contrast, in the profile case data sets each attribute level appeared at least once in every one of the 20 trials each participant completed. The net result of the contrasting object and profile case designs is that the utility parameter estimates in the profile case are more constrained by data than in the object case, and it follows that there is little information available to discriminate models 1 and 2 ($\alpha = 1$ vs. freely estimated α) in the object case. The safest conclusions to draw from this result is that there is insufficient information in typical object case designs to support a model that assumes independent preference-related information in best and worst choices (i.e., model 3); models that assume a relationship between the utilities driving best and worst choices (models 1 and 2) provide a good account of the data.

Our methods and results differ from other recent work that has concluded that model 2 (sign related utilities) is unable to account for best-worst choices in various data sets (Dumont et al., 2015; Dyachenko et al., 2014; Giergiczny et al., 2017; Rose, 2013). It is possible that the discrepancy between our conclusions and previous studies is due to the analysis of different data sets. Further to differences in data sets, we propose that there are three primary reasons for the discrepancy.

First, other recent work has focused on sample-level analyses based on deterministic or stochastic scale heterogeneneity (e.g., in heteroskedastic MNL or mixed MNL models), whereas we have focused on individual-level analyses. An exception is Dumont et al. (2015) who estimated individual-level models in a Bayesian fashion; however, they only modeled best choice, even though their data sets included best-worst choice, and hence they cannot validly conclude that best and worst choices are inconsistent with a single latent preference scale. In contrast, our Bayesian latent mixture modeling of best-worst choice provides a direct estimate of the most parsimonious account of each participant's data.

Second, the different studies used different model selection techniques. Our use of hierarchical Bayesian latent mixture modeling is the state-of-the-art method for performing individual-participant model selection. It provides a direct metric for interpretation: the posterior probability that a model is true given the observed (individual participant) data, relative to the set of models under comparison and the assumption that the true model is in the set. The Bayesian latent mixture modeling approach naturally incorporates model complexity in terms of both the model's number of parameters and its functional form, via the prior distributions placed over the parameters of each model. This is superior to conventional techniques such as the Akaike or Bayesian Information Criteria (AIC, BIC) which incorporate model complexity only in terms of the number of model parameters. In Appendix C we show that these conventional techniques do not appropriately balance model complexity; briefly, when the three models are estimated on group-aggregated data and compared with AIC or BIC – a method that has been used in previous studies (e.g., Giergiczny et al., 2017) – both metrics indicate strong evidence in favor of the most complex model across all five data sets. This result not only conflicts with our primary model comparison, but also our analysis of each model's ability to predict out-of-sample future choices, where the most complex model performed the most poorly. We believe that a guiding principle in model selection

scaled by parameter α; the psychological interpretation is that people are equally sensitive to the attribute-level information across 'best' and 'worst' choices (i.e., the same latent preference scale) but the choice process is more variable when selecting the worst option from a set (if $\alpha < 1$).

ought to be out-of-sample prediction, so based on this criterion we believe our results indicate that AIC and BIC did not select the most appropriate model for our data. This perspective is consistent with contemporary literature on model comparison in cognitive science (e.g., Annis & Palmeri, 2018; Shiffrin, Lee, Kim, & Wagenmakers, 2008).

Furthermore, model selection metrics that do not rely on parameter counts, such as the Deviance Information Criterion (DIC) or marginal likelihood, which are computed for a model's account of a multi-participant data set, also have shortcomings – most importantly, while these methods allow participant parameter values to vary within the constraints of specified population-level distributions, their model structure is constrained to a single form (i.e., in our analyses, this would force every participant to satisfy the same model – say, model 1, 2, or 3). A consequence of this approach is that it is not possible to accommodate a set of participants where a few show a complex pattern of behavior – say, model 3 is their true model – whereas the majority of participants show a simpler pattern of behavior – model 1 or 2 might be their true model. However, all participants can be accommodated within the more complex model, with some parameter redundancies (since model 3 nests models 1 and 2). This means the model selection technique might determine the more complex model to be the most appropriate model of the data, since it can account for the participants with a complex pattern of behavior and the (more numerous) participants with a simpler pattern of behavior. The crux of the issue is that the complexity penalty of the model selection technique can sway the above example to prefer the simple model for the sample (i.e., when there is a relatively strict complexity penalty) or the more complex model (i.e., when there is a relatively lenient complexity penalty). These issues do not arise when modeling individual participant data, and the latent mixture modeling approach we adopted here provides a principled way to perform model selection simultaneously at the participant and group levels.

Third, previous studies have had a limited number of data sets. In contrast, we tested multiple data sets across multiple waves for two best-worst scaling cases (profile and object), which permits stronger claims about generalizability. Furthermore, re-analysis of summary data from previous work suggests that if those authors had explored the utility-scaled model 2 that we tested, they might have come to a different conclusion about the relationship between best and worst utilities. For example, Dyachenko et al. (2014) collected best and worst choices in an object case study reflecting concerns associated with hair care. They examined numerous models including a 'baseline' that was related (though not equivalent) to our model 3. Specifically, Dyachenko et al.'s baseline model involved fitting separate MNL models to the best and to the worst choices – that is, they estimated utility values using the best choices and (independently) estimated utility values using the worst choices, each on the full choice set; this is not a model of the best-worst choices, nor is it intended to be. However, this method produced best (resp., worst) utility estimates for studied objects that were clearly positively related (cf. Figure 2 of Dyachenko et al.). From these values we can estimate a scale value[6] $\alpha = 0.38$ for a population level version of our model 2. This α estimate is remarkably similar to the α estimates we observed in some of our profile case data sets. Interestingly, in their discussion section, Dyachenko et al. say "... we found that the scaling factor[7] is able to capture the mechanism of preference construction better than using unrestricted[8] β_{best} and β_{worst}", which is consistent with the results we report here.

Turning to limitations of our current work, it is important to note we are *not* claiming that our model 2 is uniformly the 'best' model for best-worst choice data, only that, for the data we considered, it provided the most parsimonious account between models 1, 2, and 3. For example, Dyachenko et al. (2014) develop, and provide evidence for, models that assume a person uses a cognitive strategy of mixing best-then-worst choices with worst-then-best choices, even when responses are required in a best-then-worst order; our model 3 is easily extended in a similar manner. Another limitation of model 3 is that the component for best choices cannot handle context effects. This is because every MNL model can be

[6] Calculated as the range of the utilities for best divided by the range of the utilites for worst.

[7] We think that *scaling factor* here refers to Dyachenko et al.'s (2014) measure of variability due to type of choice (best or worst) and/or stage of choice (first or second).

[8] The βs denote weights in the standard linear-in-parameters framework.

interpreted as a random utility model and therefore predicts, in particular, that *regularity* holds, which is not always the case (Rieskamp, Busemeyer, & Mellers, 2006). However, motivated by the fact that MNL models of choice can be derived as special cases of specific linear ballistic accumulator models of choice and response time (Hawkins et al., 2014; Marley & Regenwetter, 2017), we can replace the utility parameters $u(z)$ for an option z in a choice set X by a context dependent form $u_X(z)$ such as that in the *multiattribute linear ballistic accumulator* (Trueblood et al., 2014). Hancock, Hess, and Choudhury (submitted) successfully apply such a model to best choice; future work could apply it, with our Bayesian methods, to best-worst choice.

References

Alberni, A. (2012). Repeated questioning in choice experiments: Are we improving statistical efficiency or getting repondents confused? *Journal of Environmental Economics and Policy*, *1*, 216–233.

Annis, J., & Palmeri, T. J. (2018). Bayesian statistical approaches to evaluating cognitive models. *Wiley Interdisciplinary Reviews: Cognitive Science*, *9*, e1458.

Berkowitsch, N. A., Scheibehenne, B., & Rieskamp, J. (2014). Rigorously testing multialternative decision field theory against random utility models. *Journal of Experimental Psychology: General*, *143*, 1331–1348.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*, 434–455.

Cherchi, E., & Hensher, D. A. (2015). Workshop synthesis: Stated preference surveys and experimental design, an audit of the journey so far and future research perspectives. *Transportation Research Procedia*, *11*, 154–164.

Cohen, A. L., Kang, N., & Leise, T. L. (2017). Multi-attribute, multi-alternative models of choice: Choice, reaction time, and process tracing. *Cognitive Psychology*, *98*, 45–72.

Dumont, J., Giergiczny, M., & Hess, S. (2015). Individual-level models vs. sample-level models: Contrasts and mutual benefits. *Transportmetrica A: Transport Science*, *11*, 465–483.

Dyachenko, T., Walker Reczek, R., & Allenby, G. M. (2014). Models of sequential evaluation in best–worst choice tasks. *Marketing Science*, *33*, 828–848.

Flynn, T. N., & Marley, A. A. J. (2014). Best-worst scaling: Theory and methods. In S. Hess & A. Daly (Eds.), *Handbook of Choice Modelling* (pp. 178–201). Massachusetts, USA: Edward Elgar Publishing.

Giergiczny, M., Dekker, T., Hess, S., & Chintakayala, P. K. (2017). Testing the stability of utility parameters in repeated best, repeated best–worst and one–off best–worst studies. *European Journal of Transport and Infrastructure Research*, *17*, 457–476.

Hancock, T., Hess, S., & Choudhury, C. (submitted). A comparison of the suitability of accumulation models in travel behaviour modelling.

Hawkins, G. E., Marley, A. A. J., Heathcote, A., Flynn, T. N., Louviere, J. J., & Brown, S. D. (2014). Integrating cognitive process and descriptive models of attitudes and preferences. *Cognitive Science*, *38*, 701–735.

Hensher, D. A., & Ho, C. (2016). Identifying a behaviourally relevant choice set from stated choice data. *Transportation*, *43*, 197–217.

Islam, T., & Louviere, J. J. (2015). The stability of aggregate–level preferences in longitudinal discrete choice experiments. In J. J. Louviere, T. N. Flynn, & A. A. J. Marley (Eds.), *Best worst scaling: Theory, methods and applications* (pp. 265–277). Cambridge, UK: Cambridge University Press.

Jordan, M. I. (2004). Graphical models. *Statistical Science*, *19*, 140–155.

Lancsar, E., & Swait, J. (2014). Reconceptualising the external validity of discrete choice experiments. *Pharmacoeconomics*, *32*, 951–965.

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, *15*, 1–15.

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. New York: Cambridge University Press.

Louviere, J. J., Flynn, T. N., & Marley, A. A. J. (2015). *Best–worst scaling: Theory, methods and applications*. Cambridge, UK: Cambridge University Press.

Louviere, J. J., Hensher, D. A., & Swait, J. D. (2000). *Stated choice methods: Analysis and applications*. Cambridge, UK: Cambridge University Press.

Louviere, J. J., Street, D., Burgess, L., Wasi, N., Islam, T., & Marley, A. A. J. (2008). Modelling the choices of individual decision-makers by combining efficient choice experiment designs with extra preference information. *Journal of Choice Modelling*, *1*, 128–164.

Marley, A. A. J., Flynn, T. N., & Louviere, J. J. (2008). Probabilistic models of set-dependent and attribute-level best-worst choice. *Journal of Mathematical Psychology*, *52*, 281–296.

Marley, A. A. J., & Louviere, J. J. (2005). Some probabilistic models of best, worst, and best–worst choices. *Journal of Mathematical Psychology*, *49*, 464–480.

Marley, A. A. J., & Pihlens, D. (2012). Models of best–worst choice and ranking among multiattribute options (profiles). *Journal of Mathematical Psychology*, *56*, 24–34.

Marley, A. A. J., & Regenwetter, M. (2017). Choice, preference, and utility: Probabilistic and deterministic representations. In W. Batchelder, H. Colonius, E. Dzhafarov, & J. Myung (Eds.), *New handbook of mathematical psychology. volume 1: Measurement and methodology* (pp. 374–453). Cambridge, UK: Cambridge University Press.

Orme, B. (2016). *Results of the 2016 sawtooth software users survey.* Retrieved 25 October 2017, from `https://www.sawtoothsoftware.com/about-us/news-and-events/news/1693-results-of-2016-sawtooth-software-user-survey`

Plummer, M. (2003). JAGS: A program for the analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing.*

Rieskamp, J., Busemeyer, J. R., & Mellers, B. A. (2006). Extending the bounds of rationality: Evidence and theories of preferential choice. *Journal of Economic Literature*, *44*, 631–661.

Rose, J. M. (2013). *Interpreting discrete choice models based on best–worst data: A matter of framing* (Working Paper No. ITLS-WP-13-22). University of Sydney.

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, *32*, 1248–1284.

Swait, J. D., & Louviere, J. J. (1993). The role of the scale parameter in the estimation and comparison of multinomial logit models. *Journal of Marketing Research*, *30*, 305–314.

Trueblood, J. S., Brown, S. D., & Heathcote, A. (2014). The multi-attribute linear ballistic accumulator model of context effects in multi–alternative choice. *Psychological Review*, *121*, 179–205.

Trueblood, J. S., Brown, S. D., & Heathcote, A. (2015). The fragile nature of contextual preference reversals: Reply to Tsetsos, Chater, and Usher (2015). *Psychological Review*, *122*, 848–853.

Turner, B. M., Schley, D. R., Muller, C., & Tsetsos, K. (in press). Comparing theories of multi-alternative, multi-attribute preferential choice. *Psychological Review*.

Wagenmakers, E.-J., Lodewyckx, T., Kiruyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*, 158–189.

Appendix A
Multinomial Logit Models and Representation of Profiles

Let X be the set of currently available options and z an option in X. The most general form of *multinomial logit (MNL) model* that we consider has the form

$$P_X(z) = \frac{e^{\alpha u(z)}}{\sum_{r \in X} e^{\alpha u(r)}},$$

where $P_X(z)$ is the probability of z being "selected" from X. P_X is replaced by B_X (resp., W_X) when we are dealing with best (resp., worst) choice; the *scale factor* α can depend on the *choice stage*; and the *utility* u can depend on whether the choice is best or worst.

This notation is all we need to state theoretical results for the *object case*. However, for the *profile case* we need the following additional notation. There are m attributes, usually with $m \geq 2$, and we let $M = \{1, ..., m\}$. Attribute i, $i = 1, ..., m$, has $q(i)$ levels; we call these *attribute levels* and sometimes let p, q denote typical attribute levels, with the context making clear which attribute is involved. A *profile* (traditionally called a *multiattribute option)* is an m-component vector with each component i taking on one of the $q(i)$ levels for that component. Given a set P of such profiles, let $D(P)$ denote the *design*, i.e., the set of (sub)sets of profiles that occur in the study. We denote a typical profile by

$$\boldsymbol{z} = (z_1, ..., z_m), \tag{1}$$

where z_i, $i = 1, ..., m$, denotes the level of attribute i in profile $\boldsymbol{z}$. For the object case, we assume that each object x has a scale value $u(x)$; it follows from the results in Marley and Louviere (2005) that u is a difference scale, i.e., unique up to an origin[9]. For the profile case, we assume that each profile $\mathbf{z}$ has a scale value $u^{(m)}(\mathbf{z})$ with $u^{(m)}$ a difference scale. We also assume the additive representation

$$u^{(m)}(\boldsymbol{z}) = \sum_{i=1}^{m} u_i(z_i),$$

where each u_i is a separate (different) difference scale.

[9]That paper uses a representation in terms of $b = e^u$, and b is shown to be a ratio scale, i.e., unique up to a muliplicative scale factor. This implies that u is a difference scale, i.e., unique up to an additive constant (or origin). This relation holds for all the results stated in this paper as having been demonstrated in Marley and Louviere (2005), Marley, Flynn, and Louviere (2008) or Marley and Pihlens (2012).

Appendix B
Savage-Dickey Density Ratio Test

This appendix provides a brief explanation of the Savage-Dickey density ratio test used in the main text to compare population-level posterior model probabilities in our latent mixture model (cf. Table 4) and nested models; a complete explanation of the test can be found in Wagenmakers, Lodewyckx, Kiruyal, and Grasman (2010) and Lee and Wagenmakers (2013).

Comparing population-level posterior model probabilities

These hypothesis tests examined whether there was evidence at the population level that the posterior model probabilities differed between the three models. They focused on the ϕ parameter of the latent mixture model (see Figure 2 of the main text), which gives the base rate probability of a respondent making decisions consistent with models 1, 2 or 3.

The hypothesis tests were based on pairwise comparisons between elements of the posterior distribution of ϕ, similar in spirit to the paired samples $t-$test. Namely, we took the difference between two elements of the posterior distribution of ϕ and examined whether the difference was equal to 0 (null hypothesis) or different to 0 (alternative). Specifically, the *Savage-Dickey density ratio* – a method for computing a Bayes factor – is the ratio of the density of the posterior difference distribution at 0 (i.e., the point value of relevance to the null hypothesis, with the latter meaning that the posterior probability is equal for the two models) relative to the density at 0 of the difference distribution under the prior. As in our estimation of the latent mixture model, we used an uninformative Dirichlet distribution as the prior over the elements of ϕ; we obtained the prior distribution for the difference between two elements of ϕ via simulation.

The Savage-Dickey density ratio is computed as the ratio of the prior to the posterior at 0, so, for example, the Bayes factor for the comparison of models 1 and 3 in the laundry data set is approximately $1/15.57 \approx .064$ (cf. Table 4 of the main text). This ratio is a Bayes factor that gives the relative odds that the data were generated by the alternative hypothesis relative to the null hypothesis, where values greater than 1 indicate support for the alternative and values less than 1 indicate support for the null. In the main text we report log Bayes factors, which are more straightforward to interpret: positive values support the alternative, negative values support the null.

Nested model comparison

These hypothesis tests examined whether the evidence favored model 1 over model 2, and/or model 1 over model 3. Results of these comparisons are reported in the main text.

Model 1 vs Model 2. The ratio of the density of the prior distribution to the posterior distribution for α at the critical point $\alpha = 1$, where model 2 reduces to the nested model 1, is the Bayes factor between the models. We approximated this ratio by computing the density of best-fitting Gaussian distributions to the posterior samples of α.

Model 1 vs Model 3. We placed a prior on a parameter δ_{ij} to code for the "effect size" difference between the utility for best choices and worst choices for the i^{th} attribute at its j^{th} level. Both the prior distribution, which is the same for all δ_{ij}, and the posterior distribution, which varies for each δ_{ij} according to the model and data, were monitored. This allowed us to compute an "average data" Bayes factor between models 1 and 3, using the Savage-Dickey method: The ratio of the prior to posterior for the set of δ's at the critical point where $\delta_{ij} = 0$ for all i and j (i.e., where model 3 reduces to the nested model 1) is the Bayes factor between the models. We approximated this via the density of best-fitting Gaussian distributions to the posterior samples, and by assuming that the density of the high-dimensional δ posterior is well approximated by the product of the marginals.

To ensure that the assumption of independence across dimensions did not unduly influence the outcome of this analysis, we conducted a second analysis that estimated δ_{ij} as before except with the

additional assumption that the set of δ's were drawn from a population-level Gaussian distribution with mean (μ) and variance (σ^2) estimated from data. We calculated the Savage-Dickey density ratio on the population-level mean parameter at the critical point $\delta^{\mu} = 0$. This analysis led to identical conclusions as the first analysis, reported in the main text, for the profile and object case data sets.

Appendix C
Model Comparison via Akaike and Bayesian Information Criteria

This appendix contrasts the model comparison results reported in the main text with a different method for quantitatively comparing models. Here, we compared the three models for each of the five data sets reported in the main text with data aggregated over participants, model parameters estimated via maximum likelihood, and model comparison conducted with the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Similar methods have been used in previous comparisons of the stability of utility parameters across choice formats or stages (such as best and worst choice; e.g., Giergiczny et al., 2017).

Table C1 shows that AIC and BIC both indicate very strong evidence in favor of model 3 in 4 of the data sets, and equivocal evidence in the final data set. We interpret this result to mean that AIC and BIC do not appropriately account for the complexity of the model, given that model 3 demonstrated much poorer out-of-sample prediction than models 1 and 2 (cf. Figure 4 of the main text, Figures D1 and D2 of the supplementary material) and as such it is unlikely to be the best model for the data.

Table C1: AIC- and BIC-based comparison between the three models discussed in the main text: sign and scale related utilities (1), sign related utilities (2), and independent utilities (3). Bold face indicates the AIC- and BIC-preferred model for each data set.

Data set (N participants, N choice sets)	Model	N parameters	Log-likelihood	AIC	BIC
Laundry Detergent (218, 16)	1	15	-9611.6	19253.2	19355.9
	2	16	-9579.7	19191.4	19301.0
	3	30	-9470.6	**19001.2**	**19206.7**
Pizza Delivery (186, 16)	1	15	-7910.8	15851.6	15951.9
	2	16	-7878.8	15789.6	15896.7
	3	30	-7794.4	**15648.8**	**15849.5**
Toothpaste (234, 16)	1	15	-9587.0	19204.0	19307.8
	2	16	-9554.9	19141.8	19252.5
	3	30	-9275.0	**18610.0**	**18817.6**
Budget Saving (561, 12)	1	8	-10067.2	20150.4	20210.5
	2	9	-10023.5	20065.0	20132.6
	3	16	-9989.9	**20011.8**	**20131.9**
Budget Spending (561, 12)	1	8	-10334.5	20685.0	20745.1
	2	9	-10286.0	20590.0	20657.6
	3	16	-10233.5	**20499.0**	**20619.1**

Appendix D
Supplementary Material

Table D1: Attributes and attribute levels for the pizza delivery data set.

Attributes	Levels			
Brand	Pizza Hut	Domino's	Eagle Boys	Pizza Haven
Price	$12	$14	$16	$18
Delivery time (minutes)	10	20	30	40
Number of toppings	1	3		
Free delivery	No	Yes		
Type of crust	Regular	Thin		
Free dessert	No	Yes		
Free drink	No	Yes		
Free salad	No	Yes		

Table D2: Attributes and attribute levels for the toothpaste data set.

Attributes	Levels			
Brand	Colgate	Macleans	Oral-B	Home Brand
Price per 100g	$1.90	$2.60	$3.30	$4.00
Size	110g	130g	150g	170g
Container	Tube	Pump		
Freshens breath	No	Yes		
Strengthen teeth and enamel	No	Yes		
Removes stains and whitens	No	Yes		
Flavour	Mild	Mint		
Fluoride protection	No	Yes		

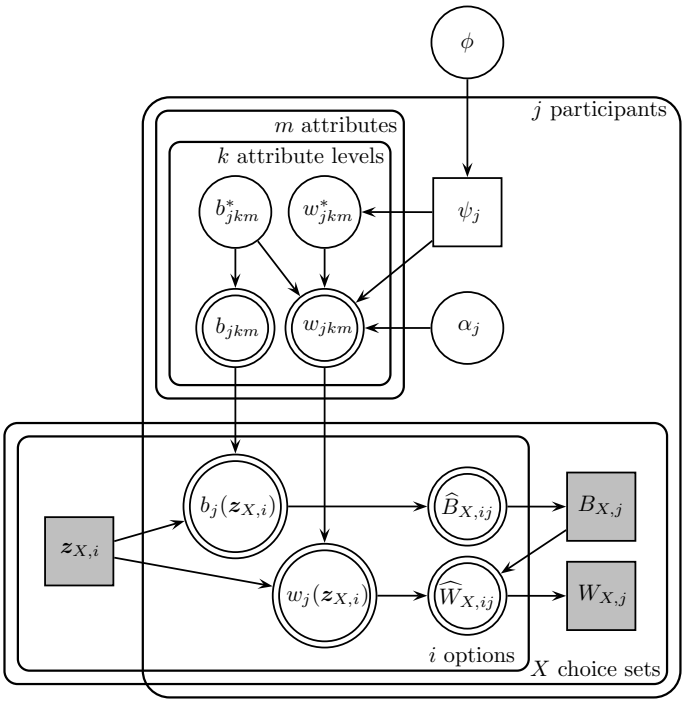
A

Best	Budget saving measure	Worst
✓	Improving fairness in the Tax System	
	Defer spending and acquisitions in Defence	
	Improving compliance measures to prevent fraud and tax evasion	✓

B

	Phone 1	Phone 2	Phone 3	Phone 4
Phone Style				
	Clam or flip phone	Candy Bar or straight phone	Swivel flip	PDA phone with touch screen input
Handset Brand	A	B	C	D
Price	$49.00	$199.00	$249.00	$129.00
Built-in Camera	No camera	5 megapixel camera	2 megapixel camera	3 megapixel camera
Wireless Connectivity	No Bluetooth or WiFi connectivity	Bluetooth and WiFi connectivity	WiFi connectivity	Bluetooth connectivity
Video Capability	No video recording	Video recording (up to 1 hour)	Video recording (more than 1 hour)	Video recording (up to 15 minutes)
Internet Capability	Internet Access	Internet Access	No Internet access	No Internet access
Music Capability	No music capability	MP3 Music Player only	FM Radio only	MP3 Music Player and FM Radio
Handset Memory	64 MB built-in memory	2 GB built-in memory	512 MB built-in memory	4 GB built-in memory

Figure 1. Example screenshots of the two best-worst choice cases. (A) demonstrates the object case in a study about budget saving measures. The respondent is asked to imagine the described budget saving measures and tick which measure would be best and which would be worst. (B) demonstrates the profile case in a study of cell phone preferences (reproduced with permission from Marley & Pihlens, 2012). The respondent is asked to imagine the described phone profiles and select the phones they like most and least.



$$\phi \sim \text{Dirichlet}(1,1,1)$$
$$\psi_j \sim \text{Categorical}(\phi)$$
$$\alpha_j \sim \text{Gamma}(2,1)$$

$$b^*_{jkm} \sim N(0,10) \quad \text{independently}$$
$$b_{jkm} = b^*_{jkm} - \overline{b^*_{jm}} \quad \text{for } \psi_j = 1, 2, \text{ or } 3$$

$$w_{jkm} = \begin{cases} -b_{jkm} & \text{if } \psi_j = 1 \\ -\alpha_j b_{jkm} & \text{if } \psi_j = 2 \\ w^*_{jkm} \sim N(0,10) & \\ w^*_{jkm} - \overline{w^*_{jm}} & \text{if } \psi_j = 3 \end{cases}$$

$$b_j(\boldsymbol{z}_{X,i}) = \textstyle\sum_{l=1}^{m} b_{jm}(\boldsymbol{z}_{X,il}),$$
$$w_j(\boldsymbol{z}_{X,i}) = \textstyle\sum_{l=1}^{m} w_{jm}(\boldsymbol{z}_{X,il}),$$
where $\boldsymbol{z}_{X,i} = (z_{X,i1}, ..., z_{X,im})$.

$$\widehat{B}_{X,ij} = \frac{e^{b_j(\boldsymbol{z}_{X,i})}}{\sum_{r \in X} e^{b_j(\boldsymbol{z}_{X,r})}}$$

$$B_{X,j} \sim \text{Categorical}(\widehat{B}_{X,j})$$

$$\widehat{W}_{X,ij} = \begin{cases} 0 & \text{if } B_{X,j} = i \\ \frac{e^{w_j(\boldsymbol{z}_{X,i})}}{\sum_{r \in X - \{B_{X,j}\}} e^{w_j(\boldsymbol{z}_{X,r})}} & \text{if } B_{X,j} \neq i \end{cases}$$

$$W_{X,j} \sim \text{Categorical}(\widehat{W}_{X,j})$$

Figure 2. The graphical model used to select between the three MNL models in the profile case data sets. See main text for details. Note: Different node sizes are purely for clarity of the node labels.

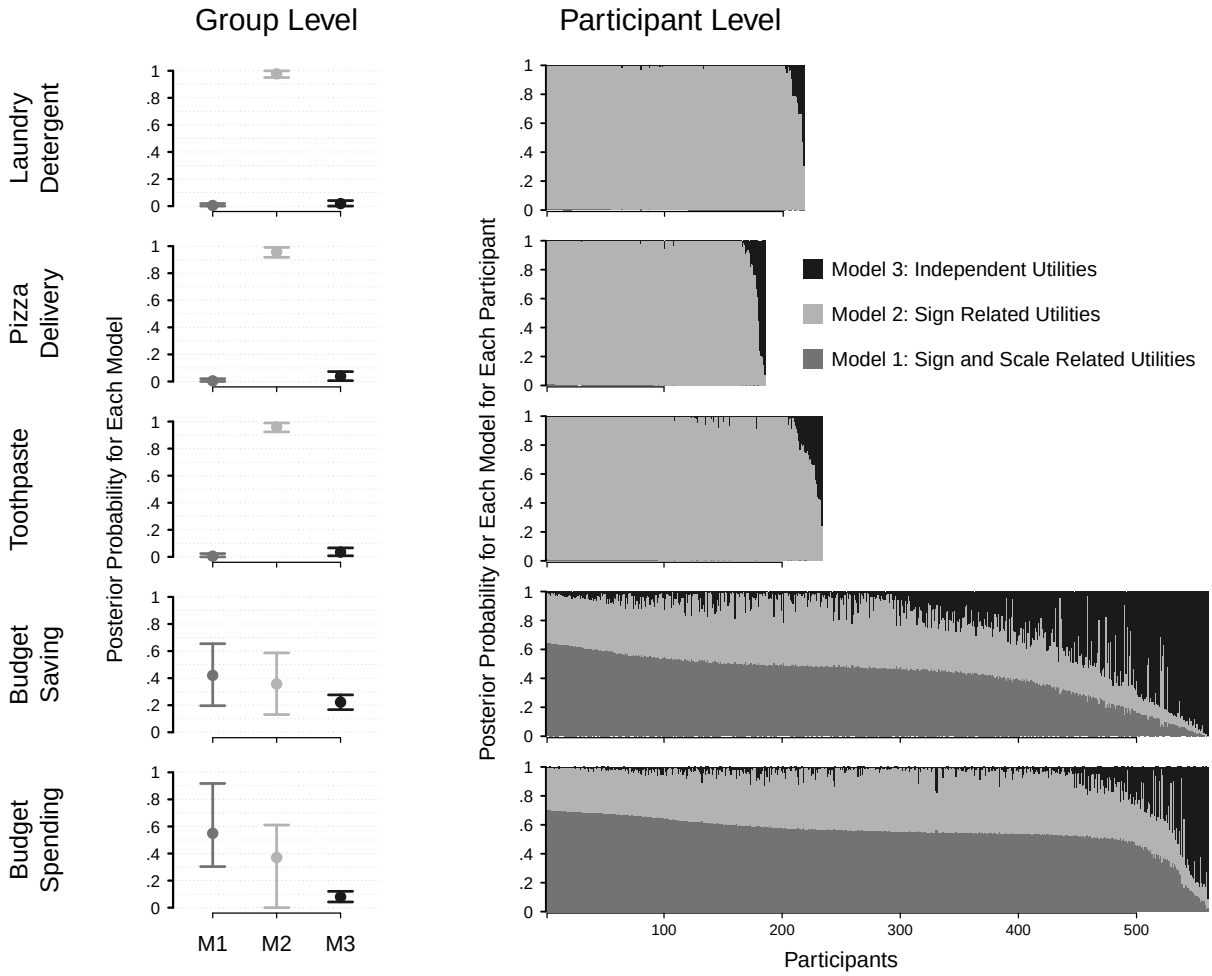


Figure 3. Model selection results from the Bayesian latent mixture model analysis. Rows represent the five data sets; the top three come from the profile case and the lower two from the object case. The left column shows the population-level posterior distribution for the base rate probability of model membership, where the dot and error bars represent the mean and 95% highest density interval of the population-level posterior distribution, respectively. The right column shows participant-level posterior probabilities of model membership. Each vertical bar of each panel represents model selection probabilities for a single participant, where the proportion of a given shade in each bar represents the posterior probability for that model. The intermediate shade represents model 1 (sign and scale related utilities), the lightest shade represents model 2 (sign related utilities), and the darkest shade represents model 3 (independent utilities).

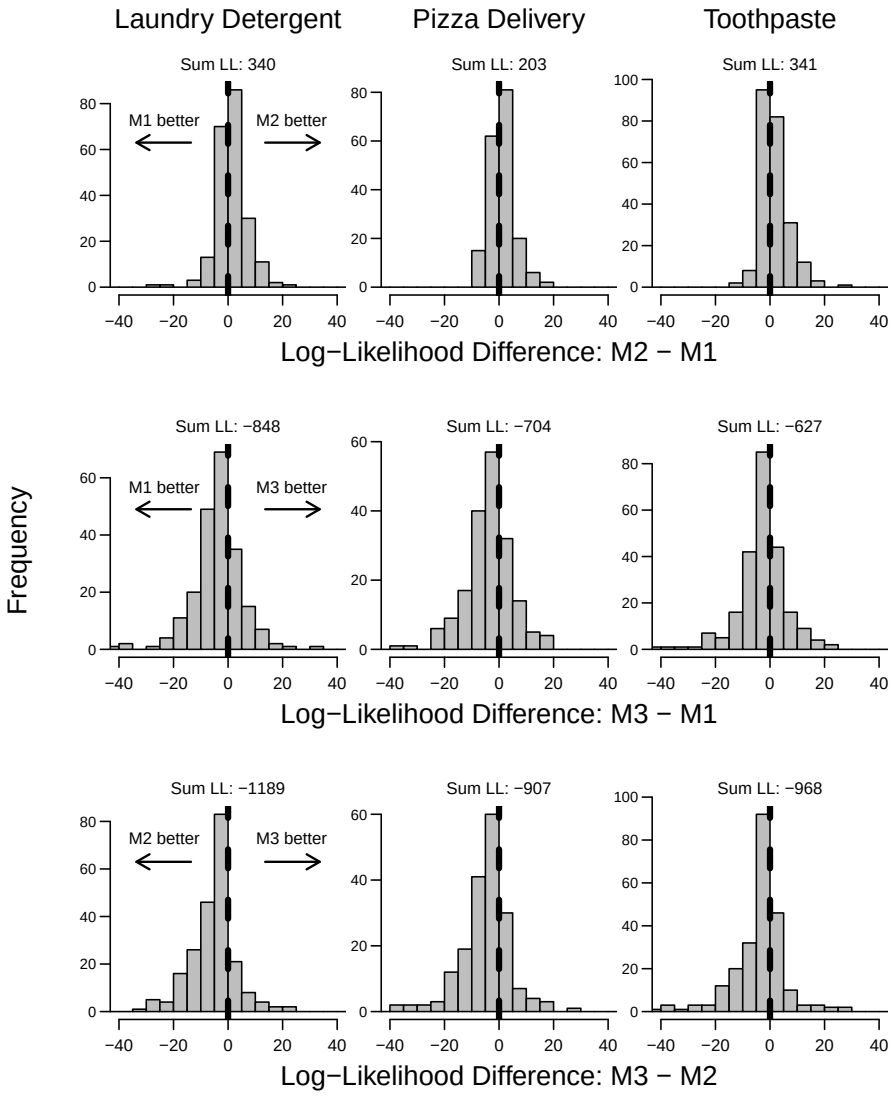


Figure 4. Prediction of best-worst choices at wave 2 in the profile case data sets given parameters estimated from wave 1. Columns represent the three profile case data sets. Rows show the three pairwise comparisons between the models with sign and scale related utilities (M1), sign related utilities (M2) and independent utilities (M3). Histograms show the across-participant distribution of the difference in log-likelihood predicted probability of best-worst choice data at wave 2, given the posterior predicted probability of their choices at wave 1 (i.e., observed data). Distributions that fall below zero indicate that the left-pointing model (e.g., M1 in the top row) provides a better account (larger log-likelihood) of participant choices at wave 2, given parameters parameters estimated from choices at wave 1, and similarly for distributions that fall above zero for the right-pointing model (e.g., M2 in the top row). The value above each panel shows the summed log-likelihood difference corresponding to each distribution of log-likelihood differences.

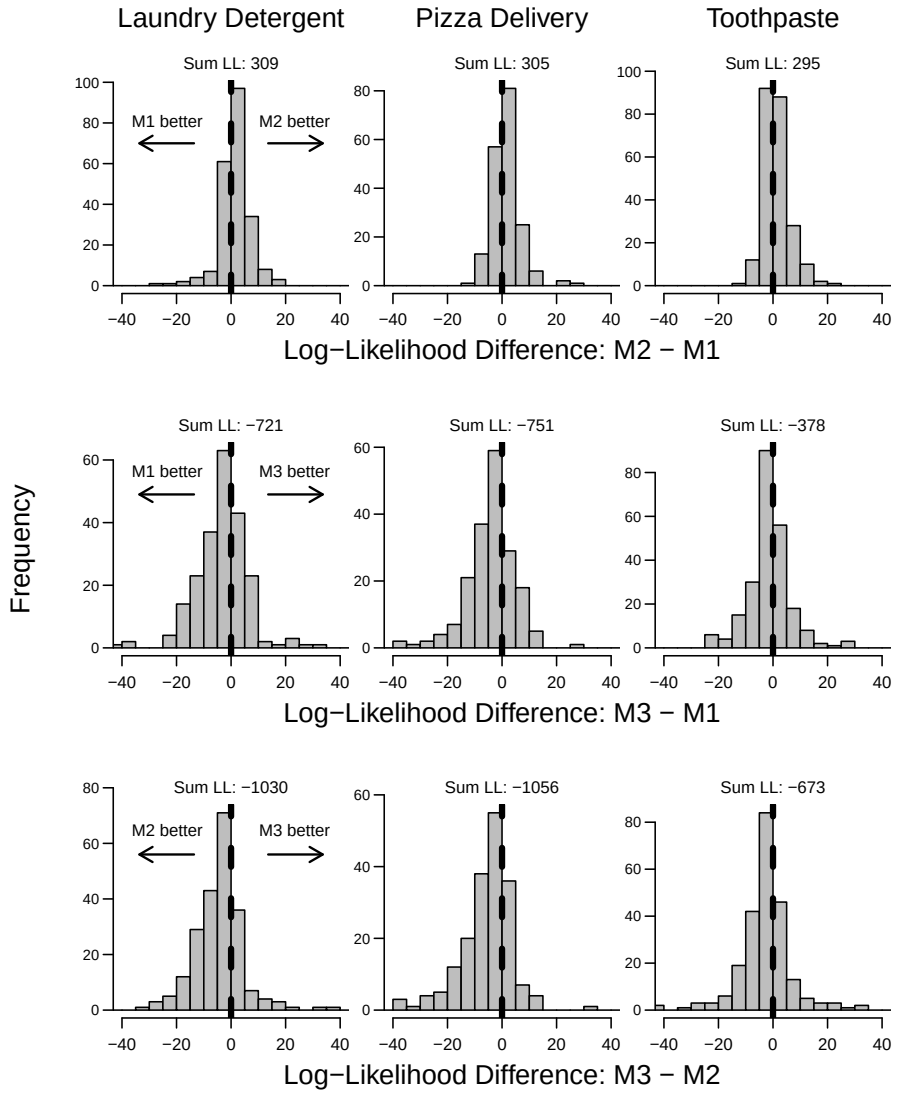


Figure D1. Prediction of best-worst choices at wave 3 in the profile case data sets given parameters estimated from wave 1. All other details are as described in Figure 4 of the main text.

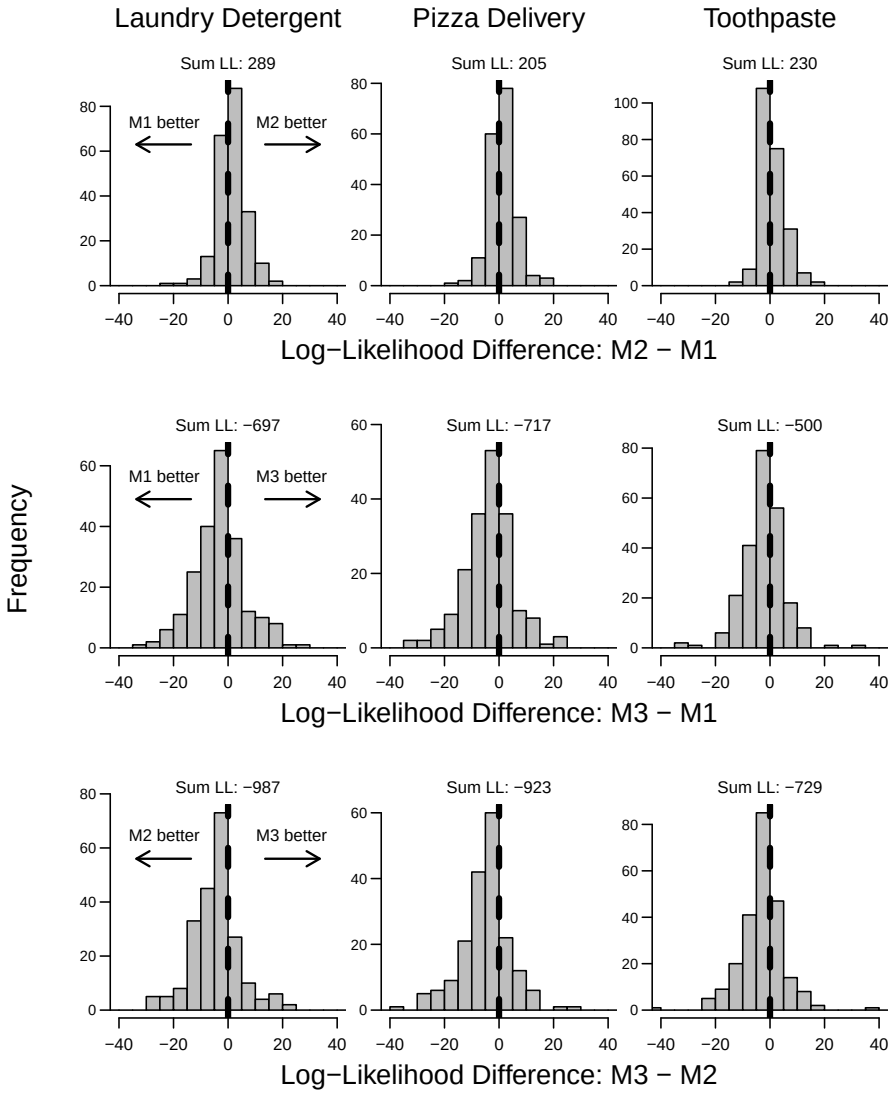


Figure D2. Prediction of best-worst choices at wave 4 in the profile case data sets given parameters estimated from wave 1. All other details are as described in Figure 4 of the main text.