The Best of Times and the Worst of Times are Interchangeable

Guy E. Hawkins[a], A.A.J. Marley[b,c], Andrew Heathcote[d], Terry N. Flynn[c], Jordan J. Louviere[c], and Scott D. Brown[d]

[a] School of Psychology, University of New South Wales, Australia

[b] Department of Psychology, University of Victoria, Canada

[c] Centre for the Study of Choice, University of Technology, Sydney, Australia

[d] School of Psychology, University of Newcastle, Australia

Abstract

We commonly determine the most preferred (best) and least preferred (worst) of a set of options, yet it is unclear whether the two choices are based on the same or different information. We examined best and worst choices using discrete choice tasks, where participants selected either the best option from a set, the worst option, or selected both the best and the worst option. One experiment used perceptual judgments of area, and another used consumer preferences for various attributes of mobile phones. In both domains, we found that the task (best, worst, or best and worst) does not alter the preferences expressed for the best (respectively, the worst) option. We also observed that the choice probabilities were consistent with a single latent dimension – options that were frequently selected as best were infrequently selected as worst, and vice versa – both within and between respondents. A quantitative model of choice and response time provided convergent evidence on those relations, with model variants that assumed an inverse relationship between the estimated parameters for best and worst choices accounting well for the data. We conclude that the diverse types of best and worst choices that we studied can be conceived as opposing ends of a single continuous dimension rather than distinct latent entities. We discuss these results in the light of rather different results for accepting (e.g., purchasing) and rejecting (e.g., not purchasing) options from a set.

Keywords: Decision-making; best choice; worst choice; best-worst scaling; state-trace analysis; mathematical model; response time.

Correspondence concerning this article may be addressed to: Guy Hawkins, School of Psychology, University of New South Wales, Sydney NSW 2052, Australia; Email: guy.e.hawkins@gmail.com.

Introduction

Suppose you seek to purchase a diamond ring for your loved one and the local jeweler has five in stock. The available options differ across multiple, independent attributes, such as size, cut, clarity, and color. When faced with this complex task, how do you make a decision? For example, you could eliminate by aspects until a single option remains (Tversky, 1972), or a subset of the five options could be generated by including the most desirable rings, and then rejecting the least favorable of the subset until left with a single option. As a result of various such decision rules, you might end up knowing which is your most preferred (best) and least preferred (worst) of the available options; for brevity and in agreement with the relevant literature, most of the time in the following we refer to these as the best and the worst option. Then you might accept (vis., purchase) the best available option, but you might not purchase (vis., might reject) it because all the rings are too expensive. This illustration shows that best and/or worst choices do not give the same information as accept and/or reject decisions. Here, we investigate whether both best and worst choices can be explained by a single latent construct or representation, and hence make use of the same underlying information. We also examine whether the selection of the worst option in a set influences which option is selected as best, and similarly whether the selection of the best option in a set influences which option is selected as worst; these are different tests of whether the two types of judgment interact.

Before proceeding to our study of best and/or worst choices, we briefly discuss the literature on accept and/or reject decisions to further support our statement that they differ; we return to this topic in our Results and Discussion of Experiment 1. Sometimes we write best choice, worst choice, or best-worst choice as generic terms.

Best-worst choice has been extensively studied in discrete choice experiments similar to our own (e.g., Finn & Louviere, 1992; Marley & Pihlens, 2012). Accepting and rejecting decisions have been studied mainly from the perspective of "framing effects" in judgment. In this paradigm, decision tasks are designed in such a way that they can lead to inconsistent preferences under different experimental conditions. For example, Shafir (1993) presented hypothetical award (accept) and deny (reject) choices regarding the two parents in an only-child-sole-custody case. The choice involved one 'impoverished' low variance option (Parent A), with mostly average attributes; and one 'enriched' high variance option (Parent B) with some excellent and some awful attributes; two sets of respondents saw the same two options, but were given different instructions. Respondents asked to which parent they would *award* sole custody tended to accept the 'enriched' option, yet respondents asked to which parent they would *deny* sole custody again tended to select the 'enriched' option. Similar patterns of preferences have been observed in investment decisions (Cheng & Chiou, 2010), choices among job candidates (Ganzach, 1995), and even in a psychophysical judgment task (Tsetsos, Chater, & Usher, 2012). Such reversals might arise because respondents focus on different attributes when accepting and rejecting (e.g., van Buiten & Keren, 2009), or perhaps the reversals are mediated by the amount of elaboration of attribute information (Ganzach & Schul, 1995; Juliano & Wilcox, 2011). The statistical principle commonly used in this research is the logic of dissociations: if one experimental manipulation (e.g., high vs. low variance choice options) has different effects under different levels of another manipulation (e.g., decisions to accept vs. reject), we might conclude that

two different cognitive processes are at work (e.g., different cognitions about accepting vs. rejecting). Although this logic is common in psychology, and can appear compelling, it can also be misleading (Loftus, 1978; Wagenmakers, Krypotos, Criss, & Iverson, 2012). Also, many of the data are from between-subject designs.

Rather than searching for differences or inconsistencies between accepted and rejected options, here we aimed to determine whether choices made under a single instruction set – namely most preferred (best) and least preferred (worst) – can be explained with recourse to only a single latent variable, such as the utility (or valence) of different attributes and options. For instance, when asked to sequentially consider a series of multi-attribute options such as automobiles or job candidates, respondents tend to arrive at the same final preference (the first of following three studies used accept/reject, the others include/exclude; Huber, Neale, & Northcraft, 1987; Levin, Jasper, & Forbes, 1998; Levin, Prosansky, Heller, & Brunick, 2001). However, explicit instruction to consider external factors, such as selection-related costs in personnel selection may result in different choice outcomes (Huber et al., 1987). We wondered whether, in the absence of manipulations designed to lead them to do otherwise, people might base best and worst choices on the same cognitive information. Specifically, we addressed two questions related to such choices. Firstly, we examined whether selecting both the best and worst option gives different data for best (respectively, worst) than when only a best (respectively, worst) choice is made. We investigated this question by testing three groups of participants with identical choice sets. One group was asked only to select their most preferred (best) option, another group was asked only to select their least preferred (worst) option, and the third group was asked to select both their most and their least preferred option. Our second question addressed whether best and worst choices are consistent with a single underlying dimension (latent variable). That is, each of the relevant parameters underlying a best choice is a monotonic (decreasing) transformation of each corresponding value in a worst choice (both relative to other available options); or are best (respectively, worst) choices based on two (or more) distinct latent variables? We examined this question by analysis of data from the group that made best and worst choices (a within-subjects comparison), and comparing the data from the group that made best choices, only, with the data for the group that made worst choices, only (a between-subjects comparison).

We performed two experiments that used *best-worst scaling*. In Experiment 1 we asked for perceptual judgments about area, and in Experiment 2 we assessed preferences for attributes of mobile phones.

Experiment 1: Perceptual Choice

In Experiment 1, participants were presented with three shaded rectangles on each trial, each with a different area. Participants were asked to select the rectangle with the largest area (as an analog to best choice) in the best-only condition, select the rectangle with the smallest area (as an analog to worst choice) in the worst-only condition, or select the rectangle with the largest and the rectangle with the smallest area in the best-worst condition. Although it may at first seem extremely unlikely that context effects, or differences between 'best' (e.g., largest) and 'worst' (e.g., smallest) choices, could arise in simple perceptual choice paradigms, there is at least one recent study that has identified just such effects (Tsetsos et al., 2012). To provide a simple example of the kind of context effect that

could lead to different 'best' and 'worst' processes, participants might have weighted the height of rectangles more heavily than the width of rectangles when making largest choices, but vice versa for smallest choices. In fact, any decision process whereby various aspects of the stimuli have different influence under best and worst choices has the potential to reject our null hypothesis at the level of options, if not at the level of aspects (attributes).

Participants

Sixty-nine first-year psychology students from the University of Newcastle participated in the experiment online in exchange for course credit, and were randomly allocated to either the best-only ($n = 23$), worst-only ($n = 24$) or best-worst ($n = 22$) conditions.

Materials and Methods

The perceptual stimuli were adapted from Trueblood, Brown, Heathcote, and Busemeyer (2013) and Hawkins et al. (in press). We factorially crossed three rectangle widths (55, 61, 67 pixels) with three heights (110, 121, 133 pixels) to create nine unique rectangular stimuli ranging in area from 6050 – 8911 pixels. The stimulus set generated a range in difficulty from simple judgments, with easily differentiable stimuli at the extremes of the set (e.g., 6050 from 8911), to difficult judgments between stimuli at one end of the range (e.g., 7370 from 7381). On each trial, three of the nine rectangles were randomly sampled without replacement. The stimuli were presented at the vertical center of the display, side-by-side in a horizontally centered row, as shown in Figure 1. All rectangles were subject to random vertical offset (± 25 pixels) to prevent the use of alignment cues in judging area. Participants in each condition completed six blocks of 100 trials.

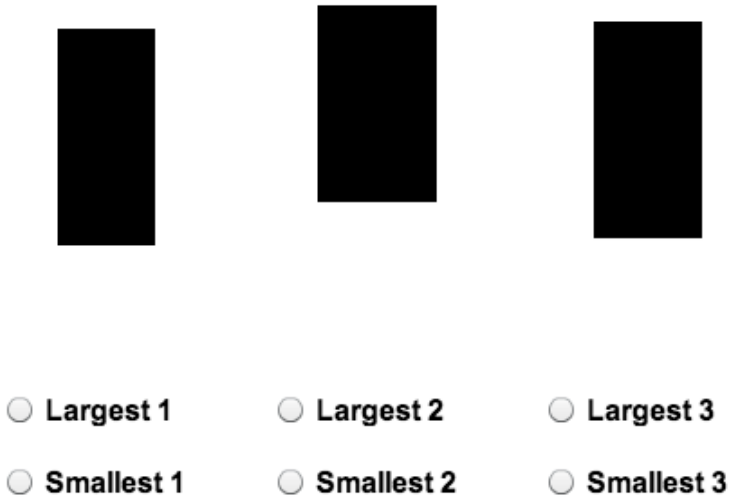

Figure 1. Illustrative example of a trial from the best-worst condition in Experiment 1.

Participants in the best-only condition were asked to identify the rectangle with the largest area. Participants in the worst-only condition were asked to identify the rectangle

with the smallest area. Participants in the best-worst condition were asked to identify the rectangle with the largest area and that with the smallest area, and were restricted from selecting the same rectangle as both the largest and smallest option. All responses were recorded with a mouse click, and in the best-worst condition the participant was free to select the ordering of the two responses; the best response could be made first and the worst response second, or vice versa. We recorded participants' choices, and the time required to make them, though we restrict our primary analyses to the choice data.

Analytic Approach

We had two primary aims: to determine whether the act of choosing a worst option alters the best choices (and vice versa), and also to determine whether best and worst choices arise from a single latent variable. These research questions do not naturally lend themselves to the traditional null hypothesis statistical testing (NHST) framework. For consistency with previous research we report NHST analyses, but our primary conclusions are drawn from Bayesian analysis and non-parametric state-trace analysis (Bamber, 1979).

Bayesian Analysis. We conduct all Bayesian analyses using the methods of Morey and Rouder (2013), who implemented common linear models such as ANOVA and regression within a Bayesian framework.[1] Morey and Rouder's approach produces Bayes factors for linear model effects (like those in ANOVA) indicating the weight of evidence for or against particular hypotheses. We use a targeted version of this approach to test the effects of interest to our hypotheses. Firstly, we test if there is evidence for the model with a main effect of attribute (rectangle area), using the Bayes factor BF_A. More importantly for our hypotheses, we compare this main effect model against a model with the same main effect of attribute as well as an interaction between attribute and condition (best-only vs. best in best-worst, worst-only vs. worst in best-worst, best vs. worst), and report this as a ratio of Bayes factors: $BF_{\frac{A+A\times C}{A}}$. Values of this ratio greater than one provide evidence for the inclusion of the interaction term in addition to the main effect of attribute, whereas values less than one provide evidence for exclusion of the interaction (i.e., for the main effect only model). Note that, unlike NHST, which cannot accept a null hypothesis, the Bayesian analysis is equally able to find evidence for the simpler (main effect only) or for the more complex (main effect and interaction) hypotheses.

State-Trace Analysis. State-trace analysis is designed to answer questions such as "can the observed data be explained by a single latent variable?" The analysis is non-parametric, assuming only that the latent variables (such as utility) are monotonically related to observed variables (such as choice probability). It is designed to overcome issues of scale-dependence in the analysis of range-restricted dependent variables, which can give rise to spurious dissociations. In particular, the existence of separate cognitive systems is often inferred from a dissociation. The dissociation is measured by an interaction, where the effect of one manipulation depends on the other. However, inferring the existence of such interactions is made difficult by the possibility of scale dependence. For example, if one condition leads to near-ceiling or near-floor performance, the effect of the other manipulation

[1] Morey and Rouder (2013) provide simple methods to implement the Bayesian analyses described in this manuscript in the freely available R programming environment (R Development Core Team, 2013).

will be forced to zero, leading to a spurious interaction. There are many other ways that scale dependence can lead to misleading interactions; apparent interactions that do not really indicate the presence of an underlying dissociation (see, e.g., Bamber, 1979; Dunn & Kirsner, 1988). This kind of problem can be particularly difficult in experiments on choice proportions (such as Shafir, 1993; Tsetsos et al., 2012) because of the tight bounds on the dependent variables.

State-trace analysis overcomes this shortcoming by relaxing assumptions about the measurement scales – it assumes only a monotonic relationship between the latent and observed variables. The analysis hinges on a state-trace plot, in which different dependent variables are plotted against one another. If the points of the plot can be connected with a single, monotonic (i.e., always increasing or decreasing) curve, then the data are consistent with a single latent variable (for mathematical details see Bamber, 1979). Drawing inferences about monotonicity in state-trace analysis is a difficult statistical problem. There has been recent progress using parametric bootstrapping (Dunn, Newell, & Kalish, 2012; Newell & Dunn, 2008; Newell, Dunn, & Kalish, 2010) and Bayesian approaches with order constrained hypotheses (Prince, Brown, & Heathcote, 2012). Although these approaches are promising and each have merit, we do not implement them here due to our combination of within- and between-subjects measures, and the high dimensionality of the stimuli used in Experiment 2. Rather, we use a more conventional approach based on confidence intervals (e.g., Busey, Tunnicliff, Loftus, & Loftus, 2000; Loftus, Oberg, & Dillon, 2004). In particular, we construct least–significant difference (Saville, 2003) ellipses around data points on a state-trace graph, with these ellipses based on Morey's (2008) modification of Loftus and Masson's (1994) within-subject standard errors for within-subjects comparisons. If all elliptical regions in the state-space can be connected with a single monotonic function we are unable to reject the null hypothesis that the data can be explained by a single latent dimension.

Results and Discussion

We used the response time information to screen data in a two-stage process on the basis of unusually fast or slow responses. First, we marked as extreme outliers and excluded from analysis any responses that were faster than .2 seconds or slower than 20 seconds (2.33% of total trials), and further excluded all data from one participant who had more than 20% of their responses marked as extreme outliers. Second, for the remaining 68 participants we calculated response time means and standard deviations on an individual participant basis, and excluded as outliers any trials with response times at least three standard deviations greater than the participant's mean response time. Combined across both criteria, 4.07% of trials were considered outliers and removed from analyses.

Our analyses are based on the proportion of times that each of the nine rectangles was selected as the largest option (best choice proportions) for participants in the best-only and best-worst conditions, and the corresponding proportion for smallest choices (worst choice proportions) for participants in the worst-only and best-worst conditions. For example, if a particular rectangle was included in the choice sets of 200 trials and was selected as the largest rectangle in 80 of those trials and as the smallest rectangle in 10 of those trials, then the corresponding best and worst choice proportions are $80/200 = .4$ and $10/200 =$

.05, respectively. The choice proportions for the nine rectangular stimuli were calculated separately for each participant then aggregated over participants for analysis.

Since our choice sets were randomly generated on every trial, we made a further check to examine how often there were choice sets with *dominated* or *dominating* options. A choice set has a dominating (respectively, dominated) option when one rectangle has greater (respectively, smaller) width and height than the other two rectangles. If participants attended to the manipulation of area, then a dominating option should be reliably chosen as best and rarely chosen as worst, and vice versa for dominated options. Dominating alternatives appeared in 9.6% of trials. As expected, when a dominating option was available it was selected as the largest rectangle on many trials (best-only – 90.1%, best-worst – 90.3%) and selected as the smallest rectangle on only a small proportion of trials (worst-only – 0.9%, best-worst – 4.1%). Similarly, dominated alternatives appeared in 9.5% of trials and were rarely selected as the largest option (best-only – 3.7%, best-worst – 5.1%) but frequently selected as the smallest rectangle (worst-only – 96.4%, best-worst – 86.2%). These choice proportions suggest that participants were aware of, and responsive to, our manipulation of area.

Best Choices are Unaffected by Worst Choices, and Vice Versa

We first checked whether the additional act of selecting the worst option from a set of options altered selection of the best option. We examined the best choice proportions with a two factor mixed ANOVA: 2 condition (between: best-only, best-worst) × 9 rectangle areas, using Greenhouse Geisser adjusted degrees of freedom where appropriate. There was a strong main effect of area on judgments, where larger rectangles were chosen more often as largest (best) than were smaller rectangles, $F(2.1, 89.3) = 109.7$, $p < .001$, $BF_A > 10^{146}$. In contrast, condition – best-only versus best in best-worst – had no statistically reliable effect on the proportion of best choices in the interaction effect, $F(2.1, 89.3) = .14$, $p = .88$, represented in the Bayesian analysis as evidence in favor of the main effect model over the model that includes an interaction effect, $BF_{\frac{A+A\times C}{A}} = .007$. This Bayes factor indicates that the model with no interaction is over 140 times (i.e., 1/.007) more likely than the model with an interaction.

We also conducted the corresponding analysis for worst choice proportions: whether the additional act of selecting the best option from a set of options altered selection of the worst option. The primary result for the proportion of worst choices was in the same direction as for best choices, though the evidence was not as strong. There was again a strong main effect of area, where smaller rectangles were more often chosen as smallest (worst) than were larger rectangles, $F(2.3, 97.8) = 116$, $p < .001$, $BF_A > 10^{160}$. Similar to best choice proportions, condition – worst-only versus worst in best-worst – did not have a statistically reliable interaction effect on area, $F(2.1, 89.3) = 1.9$, $p = .15$. Bayesian analysis indicated slight preference for the model with only a main effect over the interaction model, $BF_{\frac{A+A\times C}{A}} = .84$.

We used state-trace analysis to test more directly whether best choices were unaffected by the act of making worst choices. The left panel of Figure 2 shows a state-trace plot of best choice proportions for each rectangle area from the best-only condition (y-axis) and the best-worst condition (x-axis). The overlaid line in the left panel of Figure 2 demonstrates that a single, monotonically increasing curve connects all data points in the state-space,

meaning that the data are consistent with an explanation based on a single latent variable – that selection of the best option from a set is not affected when one also considers the worst option. Note that the confidence regions are quite small, indicating that a lack of power was unlikely to be the cause of this finding. For brevity we have not included in the main text the corresponding state-trace plot for worst choice proportions between the worst-only and best-worst conditions. This analysis led to the same outcome as the previous analysis, with a similar plot as the left panel of Figure 2; the relevant figure is in online supplementary material.

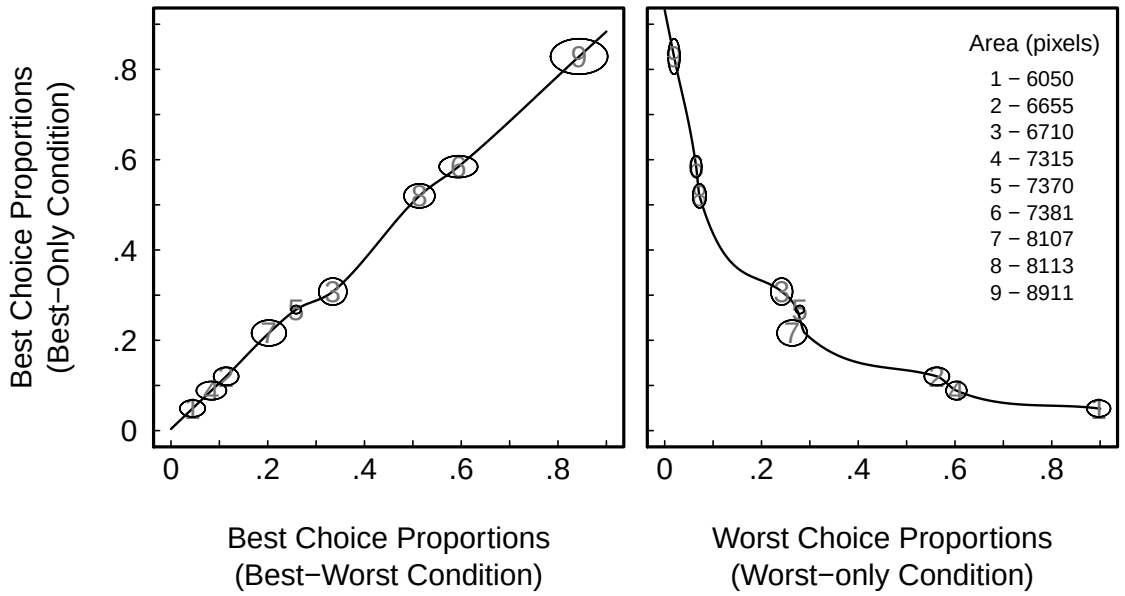

Figure 2. State-trace plots of best and worst choice proportions from selected conditions in Experiment 1. The y-axes plot best choice proportions from the best-only condition (both panels). The x-axes plot best choice proportions from the best-worst condition (left panel) and worst choice proportions from the worst-only condition (right panel). The symbols labeled $1 - 9$ represent areas of the nine unique rectangles (in pixels). Ellipses represent between-subjects least significant differences. The black curve represents a monotonic curve joining all plots in each panel – increasing in the left panel, decreasing in the right panel – consistent with the interpretation of a data generating process with a single latent variable.

Best Choices are Monotonically Related to Worst Choices

Our second analysis had two parts. We first performed a within-subjects analysis focused solely on the best-worst condition, comparing choice proportions for the largest and smallest options, and secondly as a between-subjects analyses comparing choice proportions from the best-only and worst-only conditions. As expected, rectangles most often chosen as the largest were also least often selected as the smallest, and vice versa. Two-way repeated-measures ANOVA indicated a strong interaction between best and worst choice proportions and rectangle area in the best-worst condition, and similarly for two-way mixed

ANOVA between the best-only and worst-only conditions, $F(1.7, 36) = 114$, $p < .001$, and $F(2, 86.6) = 267$, $p < .001$, respectively. Similarly, the Bayesian analysis indicated very strong support for the interaction effect over-and-above the main effect of rectangle area in both analyses, $BF_{\frac{A+A\times C}{A}} > 2 \times 10^{134}$ and $BF_{\frac{A+A\times C}{A}} > 9 \times 10^{158}$, respectively. These results suggest that participants were acutely sensitive to rectangle area, clearly discriminating rectangles with large and small area, regardless of whether those judgments were made within or across participants.

Direct evidence that best and worst choices are based on the same latent variable comes from state-trace analysis. The state-trace plot in the right panel of Figure 2 plots best choice proportions from the best-only condition against the worst choice proportions from the worst-only condition, as a function of rectangle area. As with the previous state-trace analysis, all points in the state-space can be connected with a single monotonic curve (decreasing in this case). This suggests that selection of the largest and smallest options is consistent with an explanation based on a single latent variable, and that different processes need not be invoked for best and worst choices in this case.[2] The right panel in Figure 2 represents a more stringent test of the hypothesis that best and worst choices are based on the same information, because that test involves comparison across independent subjects (best-only and worst-only conditions). The within-subjects analysis provided even stronger support for our thesis (see supplementary online material).

Experiment 1 supported the hypothesis that best choices are not influenced by the act of making a worst choice. We also supported the hypothesis that best and worst choices are made on the same bases. It could be argued, however, that these results are unsurprising given the task at hand: if respondents paid attention solely to rectangle area, as instructed, then there is no reason to expect anything other than a single latent variable. Nevertheless, psychological science is replete with apparently simple tasks which lead to complex, sub-optimal, or irrational strategies. As a plausible hypothetical example, if judgments of rectangle area were influenced by the aspect ratio in addition to area, then the data might have differed to what we observed. For example, three rectangles with the same area but different aspect ratios might produce different choice proportions, and it is possible that aspect ratio might be differentially important for largest versus smallest judgments (height might be the most salient dimension for selecting the largest area, but not necessarily for selecting the smallest area). Nevertheless, in Experiment 2 we counter the possible criticism by replicating all key results from Experiment 1 using complex, multi-attribute stimuli in consumer judgments; a paradigm in which preference reversals and paradoxical choices are known to often occur.

Experiment 2: Consumer Choice

In Experiment 2 we aimed to replicate and extend the results of Experiment 1, moving from a perceptual choice task to a consumer judgment task. We again examined whether selecting a worst option alters best choices, and vice versa, and whether best and worst choices are consistent with a single latent dimension. In Experiment 2, a single data-

[2] A separate state-trace analysis conducted on the component attributes of the rectangles – width and height – also supported a single latent variable for best and worst choices at the level of attributes (figures not shown).

generating process might be "utility", where the strength of preference for a product might range on a single dimension from highly desirable (high utility) to highly undesirable (low utility).

In Experiment 2 we asked participants to judge mobile phones described by quantitative attributes, such as price, camera quality, and memory capacity. As in Experiment 1, participants were randomly allocated to one of three between-subjects conditions: participants either selected their best phone from each choice set (best-only), their worst phone from each choice set (worst-only), or their best and their worst phone from each choice set (best-worst). We expected that the best choice proportions from the best-only and best-worst conditions would be such that the proportion of times each level of each attribute (attribute-level, e.g., price: \$99, \$199, \$299, etc.) was chosen as best[3] would be the same in the two conditions; and that a parallel result would hold for worst choice proportions. We also expected that phones more often selected as best would be less often selected as worst, and vice versa, in a manner consistent with a single latent dimension, both within the best-worst condition, and between the best-only and worst-only conditions.

Method

Participants

Eighty-seven first-year psychology students from the University of Newcastle participated in Experiment 2 online in exchange for course credit, and were randomly allocated to either the best-only ($n = 27$), worst-only ($n = 30$), or best-worst ($n = 30$) condition.

Materials and Methods

The stimuli were adapted from Marley and Pihlens (2012) to reflect standards for current mobile phones. Participants were asked to select between mobile phones that varied on five attributes, with each attribute having three levels. Three mobile phone profiles were presented on each trial, with each phone having a single level from each of five attributes. Every phone on each trial was randomly generated from the set of possible phones. An example choice set is shown in Figure 3 and a full listing of the attributes and the levels of each attribute, which we refer to as *attribute-levels*, is shown in Table 1. The five attribute-levels that made up each phone were randomly chosen on each trial (choice set). Participants in all three conditions completed 150 choice sets, or trials, in total; six blocks each of 25 trials.

Participants in the best-only condition were asked to select their most preferred phone and participants in the worst-only condition were asked to select their least preferred phone. Participants in the best-worst condition were asked to select both their most and least preferred phone. As in Experiment 1, responses were made with a mouse click and could be provided in either order in the best-worst condition – best, then worst, or worst, then best. Participants were restricted from selecting the same phone as both the most and least preferred option. We recorded all choices, and the times taken to make them.

[3] An attribute-level is 'chosen as best' when a phone is selected as best that includes that attribute-level.

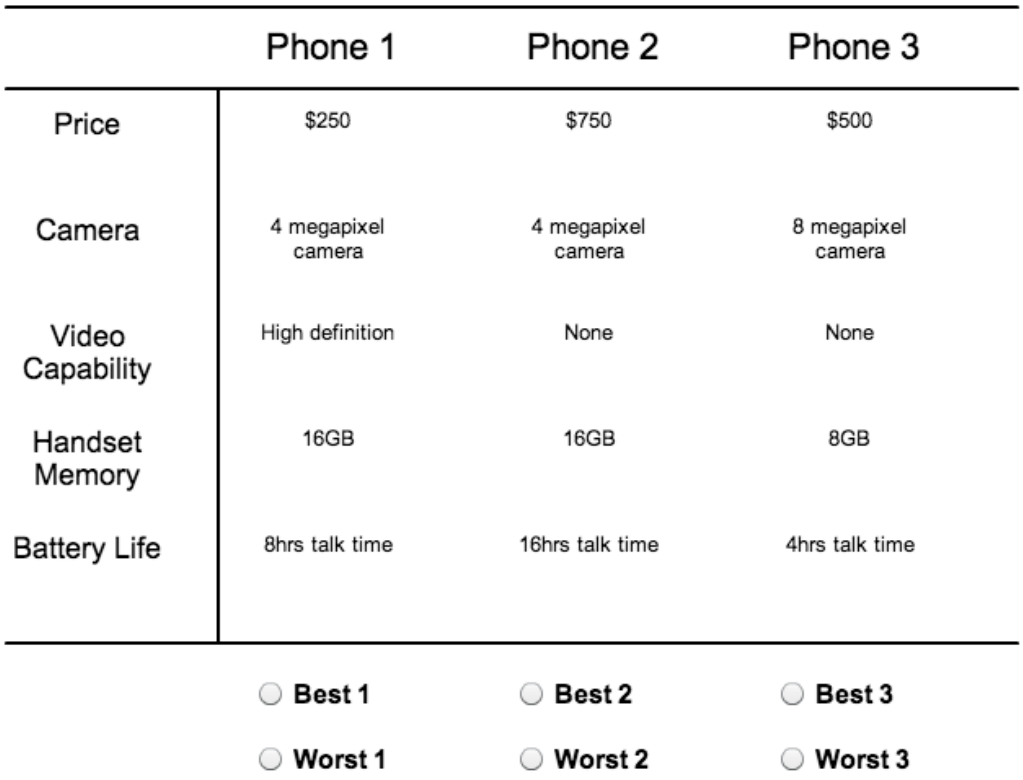

Figure 3. Illustrative example of a choice set from the best-worst condition in Experiment 2.

Results and Discussion

As in Experiment 1, we used response time information to screen data on the basis of unusually fast or slow responses. We defined as extreme outliers and excluded from analysis any responses that were faster than .35 seconds or slower than 80 seconds (2.83% of total trials), and again excluded any participant who had more than 20% of their trials marked as outliers (three participants removed). For the remaining 84 participants, we calculated means and standard deviations of response times on an individual participant basis, and excluded as outliers any trials with response times that were more than three standard deviations larger than a participant's mean response time. In total, 4.17% of trials were considered fast or slow outliers and were removed from analyses.

We calculated best choice proportions for each attribute-level as the proportion of times that the phone chosen as best contained that attribute-level, with the analogous measure for worst choice proportions. For example, if a particular attribute-level (say, \$250) appeared in 130 choice sets for a particular participant, and was in the phone selected as most preferred in 78 of those 130 choice sets, then it has the best choice proportion 78/130 = .6. As in Experiment 1, the best and worst choice proportions were calculated on an individual participant basis and then aggregated over participants for analysis.

Table 1: The five attributes and their combined 15 levels used in Experiment 2, adapted from Marley & Pihlens (2012).

Attribute	Attribute-levels
Price	\$250
	\$500
	\$750
Camera	2 megapixel camera
	4 megapixel camera
	8 megapixel camera
Video capability	None
	Standard definition
	High definition
Handset memory	8 GB
	16 GB
	32 GB
Battery life	4 hrs talk time
	8 hrs talk time
	16 hrs talk time

We calculated separate two-way ANOVAs and Bayes factors for each attribute. In the ANOVAs we compared attribute-level crossed with condition (best-only vs. best in best-worst; worst-only vs. worst in best-worst, etc.) against a main effect model including just the attribute levels. For the ANOVAs, we adopted a Bonferroni-adjusted significance level: $\alpha = .05/5 = .01$.

Best Choices are Unaffected by Worst Choices, and Vice Versa

We first confirmed that the attribute-levels had a reliable effect on performance followed by examination of whether selecting a least preferred option altered the choice of the most preferred option, and vice versa. Participants were attentive to the specific features of phones, with strong differences in preference for the different levels of each attribute (main effects). Phones were reliably more often chosen as best (i.e., larger best choice proportions; best-only and best-worst conditions) when they were cheaper, had a better camera and video capability, larger handset memory and longer battery life, all F's > 20 and all p's $< .001$ using Greenhouse-Geisser adjusted degrees of freedom, and all BF_A's $> 10^{14}$. Similarly, phones were reliably more often chosen as worst (i.e., larger worst choice proportions; worst-only and best-worst conditions) when they were more expensive, had a poorer camera, no video capability, smaller handset memory and a shorter battery life, all F's > 16 and all p's $< .001$, and all BF_A's $> 10^{18}$.

The upper panels of Figure 4 plot best choice proportions for attribute-levels from the best-only condition against the equivalent proportions calculated from the best-worst condition.[4] The least significant difference ellipses for all plot points fall on the diagonal

[4]We do not show the corresponding plots for profiles – phones – as we do for the rectangles studied in Experiment 1. Since we randomized the levels of each attribute to every phone, there exist many possible

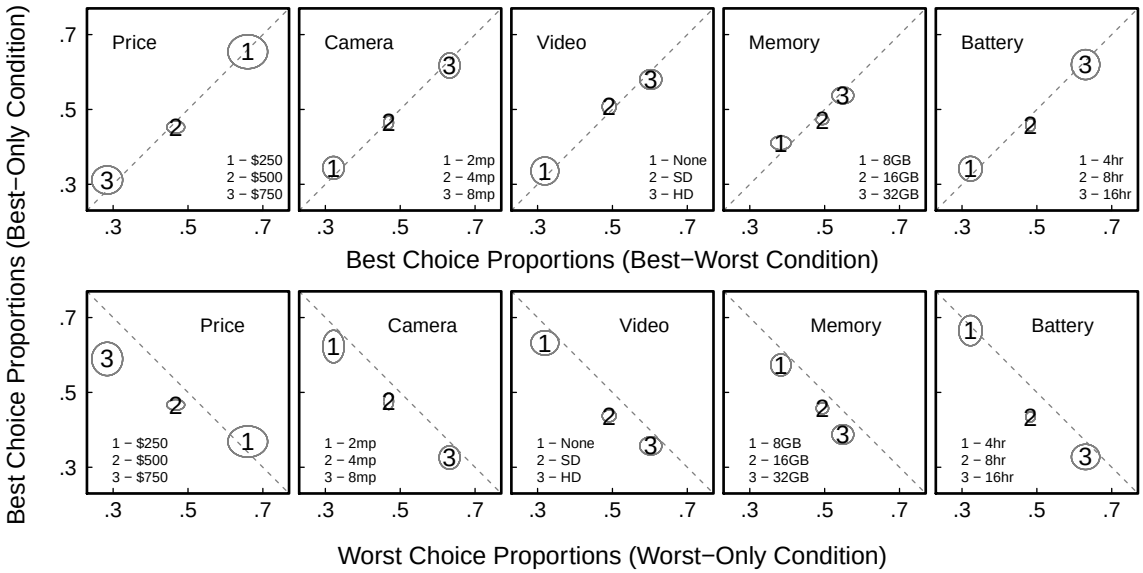

Figure 4. Attribute-level choice proportions for the mobile phone stimuli in Experiment 2. The upper panels show the attribute-level best choice proportions for the best-only condition (y-axes) against the attribute-level best choice proportions for the best-worst condition (x-axes) for each attribute, separately. The dashed gray lines show the $y = x$ line where preferences would fall if they were identically distributed across the best-only and best-worst conditions. The lower panels show the attribute-level best choice proportions for the best-only condition (y-axes) against the attribute-level worst choice proportions for the worst-only condition (x-axes). The dashed gray lines show where the attribute-level worst choice proportions would fall if they were the reverse of the attribute-level best choice proportions ($y = 1 - x$). Ellipses represent between-subjects least significant differences.

$y = x$ line, indicating that best choices were not altered by also requiring a worst choice (the corresponding figure for worst choice proportions is shown in the supplementary material). Supporting this, there were no reliable interaction effects between condition – best-only versus best in best-worst – and attribute, for any of the five attributes, all p's $> .3$. The analogous Bayes factors indicated no support for the interaction effect over and above the main effect of attribute; memory $BF_{\frac{A+A\times C}{A}} = .33$, all other $BF_{\frac{A+A\times C}{A}}$'s $< .16$. This suggests that considering the least preferred options did not alter the preference for the most preferred options. A similar result emerged for worst choice proportions, though not as strong: all p's $> .06$, and the Bayes factors for two levels provided some evidence for the alternative hypothesis (i.e., the presence of an interaction) – price $BF_{\frac{A+A\times C}{A}} = 5.19$ and camera $BF_{\frac{A+A\times C}{A}} = 1.47$ – while the remaining Bayes factors provided weak support for the null; $BF_{\frac{A+A\times C}{A}}$'s $= .13, .35$, and $.87$ for the video, memory, and battery attributes, respectively. On balance, a combined Bayes factor assuming attribute independence indicates support for the null hypothesis of no interaction, $BF_{\frac{A+A\times C}{A}} = .28$.

phone profiles such that we could not obtain reliable choice proportion data for any particular profile.

The results shown in Figure 4 strongly and unambiguously support monotonicity and hence one latent dimension within attributes. Although this is a strong conclusion, the analyses shown in Figure 4 cannot test whether all attributes map through a single ("utility") dimension. To remedy this, we also calculated across-attribute state-trace analyses, shown in Figure 5. These plots compare the attribute-level best choices from the best-only and best-worst conditions (upper triangle panels), and also compare the attribute-level worst choices from the worst-only and best-worst conditions (shown in supplementary material). Each panel makes the comparison for a pair of attributes (such as price and camera). A monotonic line in a panel indicates strong evidence that choices based on those two attributes use the same information in the different conditions (best-only and best-worst, or worst-only and best-worst). The (x, y) coordinate of each point represents the mean choice proportion for a single attribute-level for two conditions (e.g., best-only, best-worst), with ellipses representing between-subjects least significant differences (Saville, 2003).

For example, the upper row shows the price attribute, with the first panel comparing price against the camera attribute, the second panel comparing price against the video attribute, and so on. A single monotonically increasing line can connect the ellipses in all panels above the main diagonal. This means that all pairs of attributes provide evidence in favor of a single underlying decision variable. This result permits a decisive conclusion about dimensionality: the ten pairwise comparisons between attributes can be represented with a single latent variable. This conclusion is warranted given the transitive logic of state-trace analysis: if A and B are monotonically related, and B and C are monotonically related, then A and C must be monotonically related. Therefore, the joint representation of all five attributes is necessarily monotonic and arose from a single latent variable – that is, selecting the worst option from the set of options did not change the basis on which best choices were made. The corresponding analysis for worst choice proportions in the worst-only and best-worst conditions led to the same conclusion, and is presented in supplementary material.

Best Choices are Monotonically Related to Worst Choices

We next examined the relationship between attribute-level best and worst choices. Repeated-measures ANOVA between best or worst choice and the attributes indicated that the attribute-levels that appeared most often in the preferred phones also appeared least often in the phones chosen as worst, and vice versa – see lower panels of Figure 4 for best-only versus worst-only (figures plotting best vs. worst choice proportions from the best-worst condition are shown in supplementary material). This effect was reliable in the interaction between best choice against worst choice and the attribute-levels for all five attributes, across the best-only versus worst-only and best versus worst in the best-worst condition, all F's > 20 and all p's $< .001$, and in the analogous Bayes factors, all $BF_{\frac{A+A\times C}{A}}$'s $> 10^{15}$. These results suggest that participants reliably attended to the attributes that comprise mobile phone quality, clearly discriminating between phones that have positive and negative qualities (most and least preferred, respectively). They also suggest that this pattern holds regardless of whether participants made both best and worst choices, or only one of the two.

State-trace analysis provided strong evidence that a single latent dimension explains observed attribute-level best and worst preferences in Experiment 2. In Figure 5 (lower triangle panels), the best and worst choice data and their ellipses of least significant differences

can be connected in all panels with a monotonically decreasing line. This provides evidence that preferences for the best and worst attribute-levels are consistent with an explanation based on a monotonic transformation of the same underlying information – opposing ends of a single continuum where best can be considered the 'opposite' of worst, and vice versa. This relationship held whether participants selected only their most or least preferred option (Figure 5) or both (supplementary material).

Our state-trace analysis in Experiment 2 demonstrated a reciprocal relationship between best and worst preferences at the level of attributes, but we did not directly test whether the corresponding reciprocal relationship held at the level of options (i.e., mobile phones) as we did in Experiment 1. It was feasible to test the monotonicity of best and worst preferences at the level of options in Experiment 1 as there were only 9 unique rectangle areas but not in Experiment 2 due to the large number of options (243): because of this large number of options, we had insufficient data to directly estimate the best and worst drift rate for each of them. In Experiment 2 our finding of a reciprocal relationship between best and worst preferences at the level of attributes supports an explanation based on a single dimension for best and worst preferences at the level of options, if one is willing to make certain assumptions about the mapping from attribute-levels to options. These assumptions could include a representation at the option level that is a multiplicative and equally-weighted function of the representation at the attribute level. We use a model-based approach to test this assumption.

Model-Based Approaches to Testing Dimensionality

We now build on the results above to test a specific instance of a unified (one-dimensional) model for best and worst choices. If a single latent variable generated the observed data in Experiments 1 and 2, then a quantitative model that assumes a single latent mechanism should account for all aspects of the data – choices, response times, and parameter estimates consistent with the data analysis. The particular model we used to test this hypothesis was the *Linear Ballistic Accumulator* (LBA; Brown & Heathcote, 2008). We could have used any of the other accumulator-based models of consumer choice (e.g., Bhatia, in press; Busemeyer & Townsend, 1992; Otter, Allenby, & van Zandt, 2008; Roe, Busemeyer, & Townsend, 2001; Usher & McClelland, 2004), but we chose the LBA due to its computational simplicity. In previous work we demonstrated that the LBA can be fit to best-worst choices (Hawkins et al., in press). When only choices are modeled, and with certain restrictions on its parameters, the LBA is a random utility model, similar to the *multinomial logit* (Luce choice) model (Luce, 1959; Marley & Flynn, in press). Random utility models of best-worst choice typically assume that the utility, or preference strength, of an attribute-level (or, sometimes, an option) that underlies best choices also underlies worst choices, with the relevant worst value utility equal to the negative of the best utility (e.g., Marley & Pihlens, 2012).

In the following section we provide a brief overview of the LBA model in general, and the best-worst LBA in particular. We then demonstrate that the latter type of model, which assumes a single representation for best and worst utilities, provides a good account of choice and response time data. To foreshadow our results, the modeling led to three conclusions that converge with the Bayesian and state-trace analyses of Experiments 1 and 2: 1) best utilities estimated from best-only choice data are highly correlated with best

utilities estimated from best-worst choice data, and similarly for worst utilities; 2) models that allow best and worst utilities to differ in a possibly non-monotonic fashion nevertheless estimate monotonically related best and worst utilities from data; and 3) models that assume an inverse relationship between best and worst utilities provide a good account of data. In the model fits we also illustrate two methodological points: best-worst discrete choice tasks provide greater constraint on parameter estimates than traditional best-only discrete choice tasks; and, similarly, using both choices and response times places greater constraint on parameter estimates than choices, only.

The Linear Ballistic Accumulator Model

The LBA shares the evidence accumulation assumption common to most models of simple perceptual choices (e.g., Brown & Heathcote, 2005; Ratcliff, 1978; Ratcliff & Rouder, 1998; Usher & McClelland, 2001; van Zandt, Colonius, & Proctor, 2000; for review see Luce, 1986; Ratcliff & Smith, 2004). The model assumes that a decision between n options is composed of a race between n independent accumulators that collect evidence in favor of the available choice alternatives according to a drift rate, d, which represents the utility for each option. When the evidence in one accumulator reaches a pre-determined criterion – the response threshold, b – a response is triggered. The accumulator that reaches threshold determines the response, and the predicted response time is the time taken for the accumulator to reach threshold plus a fixed offset (non-decision time, t_0), which accounts for the time involved in stimulus encoding and motor processes involved in response production.

The upper row of Figure 6 gives an example of an LBA decision between options A, B, and C, represented as independent accumulators that race against each other. The vertical axes represent the quantity of evidence, horizontal axes the passage of time. At the beginning of a decision, the amount of evidence in each accumulator (the 'start point') varies independently across accumulators and randomly from choice-to-choice, sampled from a uniform distribution between zero and A (a parameter of the model). Evidence accumulates linearly, represented as the arrows within each accumulator. The rate of evidence accumulation is defined by the 'drift rate', which is assumed to vary randomly across accumulators and from choice-to-choice according to independent samples from a Gaussian distribution: $N(d, s)$, truncated to nonnegative values in our implementation. Mean drift rate (d) reflects the attractiveness (or utility) of an option: a larger value gives a faster rise to threshold and a more likely choice outcome, on average. Variability in drift rates is assumed to reflect decision-to-decision changes in extraneous factors, such as motivation or fluctuations in attention. The general LBA framework has proven successful in accounting for the observed variability in the joint distribution of choices and response times across a range of perceptual choice paradigms (e.g., Brown & Heathcote, 2008; Forstmann et al., 2008; Ho, Brown, & Serences, 2009; Ludwig, Farrell, Ellis, & Gilchrist, 2009).

We previously explored a number of ways to extend the LBA model to account for best-worst choice tasks (Hawkins et al., in press). In the preferred extension, which we called the *parallel best-worst LBA*, we assumed that best and worst choices were the result of two parallel races: a best race and a worst race, shown in the upper and lower rows of Figure 6. This model accounted for the choices and response times from a perceptual choice task, and the choice data (when response times were unavailable) from two applied best-worst scaling data sets. The parallel best-worst LBA made a similar assumption about

drift rates as the random utility representation of the *maxdiff model* for best-worst choices (see Marley & Louviere, 2005; Marley & Pihlens, 2012): if $d(x)$ is the drift rate for option x in the race to decide the best option, then $1/d(x)$ is the drift rate for the same option in the race to decide the worst option. This inverse assumption on drift rates implies that those options frequently chosen as best are also infrequently chosen as worst, and vice versa. Our state-trace analyses of Experiments 1 and 2 support this assertion, but explicitly fitting the model to data provides a more precise test of the same hypothesis. If there is a non-monotonic relationship between best and worst choices – if selecting an option as best is qualitatively different to selecting an option as worst – then a model which enforces this inverse relationship will qualitatively misfit the data. Furthermore, we can fit models that assume no relationship between best and worst drift rates (i.e., where all drift rates are free parameters) and use model selection techniques to determine whether the inverse relationship between best and worst drift rates provides the most parsimonious (and well-fitting) account of the data.

We fit the standard LBA to the best-only and worst-only conditions, and the best-worst LBA to the best-worst conditions, of Experiments 1 and 2. We first fit the models to choice proportions only, by integrating over the distribution of predicted response times to obtain marginal choice probabilities (for detail see Hawkins et al., in press). We then fit the models to the joint distribution of choices and response times to the data from the best-only, worst-only, and best-worst conditions of Experiments 1 and 2.

Estimating Model Parameters from Data

We fit the models to individual participant data from Experiments 1 and 2 via maximum-likelihood estimation using differential evolution optimization algorithms. In these model fits we were primarily interested in estimating the drift rates; those rates measure the attractiveness of different attribute-levels and, consequently, of options.[5] In Experiment 1 we estimated nine drift rate parameters, one for each rectangle area.[6] In Experiment 2 there were too many possible phone profiles (243) to estimate a drift rate for each, so we assumed that the drift rate for a profile was defined by the product of drift rate contributions from its attributes, consistent with Marley and Pihlens' (2012) constraints on their multinomial logit model parameters. The reciprocal relationship assumed between best and worst drift rates at the level of attributes, and the multiplicative and equal-weighted representation for the drift rate for an option in terms of the attribute-level drift rates, leads to a reciprocal relationship between best and worst drift rates at the option level. The validity of this assumption is therefore evaluated by how well the model fits the data. Hence, in Experiment 2 we estimated 10 drift rates, corresponding to the 15 different attribute-levels, with one level in each attribute arbitrarily constrained such that the product of the drift rates for the levels within each attribute was equal to one (cf. Marley & Pihlens, 2012).

When jointly fitting to choices and response times we set $s = 1$ to fix a scale for

[5]For choice models that can be written in the standard (additive) random utility form – for instance, the maxdiff model for best-worst choice – results to date suggest that the utility estimates for such forms are the log of the drift rate estimates in the analogous LBA model (Hawkins et al., in press).

[6]We did not estimate attribute-level drift rates for rectangle width and height in Experiment 1 as we did for the attributes studied in Experiment 2.

the evidence accumulation process, and estimated from data the range in the start-point (A), response threshold (b), and non-decision time (t_0). In the best-worst condition we estimated a separate response threshold for the best and worst races (b_{best}, b_{worst}). We also assumed a fixed probability (2%) of a uniform contaminant mixture process (Ratcliff & Tuerlinckx, 2002), interpreted psychologically as a small probability of a failure of attention to the required judgment across trials, which also serves to stabilize maximum likelihood parameter estimates. When fitting to choice proportions only, the time-related parameters of the model are not identifiable, and were arbitrarily fixed at $s = 1$, $b = 1$, $A = 0$ and $t_0 = 0$ (for details see Hawkins et al., in press). Thus, we estimated 9 (Experiment 1) or 10 (Experiment 2) drift rates when fitting to choices-only, and an additional three (best-only, worst-only) or four (best-worst) parameters when fitting the joint distribution of choices and response times.

We evaluated goodness of fit using two methods. We first compared observed and predicted attribute-level choice proportions, separately for best and worst choices. That is, for each participant we calculated the proportion of times that each attribute-level was chosen (that is, was in the phone that was chosen) relative to the number of times each attribute-level was available as a to-be-selected option, both in data and model predictions. In Experiment 1 each participant therefore provided 9 best or worst choice proportions in the best-only and worst-only conditions, one for each rectangle area, and those in the best-worst condition provided 9 best and 9 worst choice proportions. Similarly, each participant in Experiment 2 provided 15 best (and/or worst) attribute-level choice proportions. We calculated these values from individual participant data and fits to individual participant data, but present both types of results averaged over participants.

When fitting to choices and response times, we examined the correspondence between observed and predicted distributions of response times and choice proportions. As with the choice proportions, even though we fit the models to individual participant data, to simplify exposition we display the fit of the model predictions aggregated over participants. After the models were fit to individual participants, we aggregated observed and predicted response time distributions by calculating separately for each condition and participant the $1^{st}, 5^{th}, 10^{th}, 15^{th}, \ldots, 90^{th}, 95^{th}, 99^{th}$ percentiles of the response time distribution, and then averaging these individual participant quantiles to form an aggregate distribution of response times. Quantile averaging of this sort conserves distribution shape, under the assumption that individual participant distributions differ only in scale and location (Gilchrist, 2000). We then converted the averaged quantiles to display distributions in a histogram-like form.

Model Fit to Data

The models provided a good fit to response time and response choice data from the best-only, worst-only, and best-worst conditions of Experiments 1 and 2 – see Figure 7. The model captured all qualitative and most quantitative trends in the aggregate response time distributions: a sudden onset with a sharp peak and a long positively skewed tail. Some aspects of the model fit are not quantitatively precise, such as smaller variance in data than model predictions for the best-only condition of Experiment 1, and the best-worst

condition of Experiment 2, but all general trends were captured.[7] The fits to response time distributions illustrate two important points. Firstly, best-worst choice tasks do not result in unusual response times. Secondly, this is one of the first demonstrations that a single quantitative model can account for markedly different response latency data, ranging from approximately one second up to 30 seconds, from substantially different tasks – perceptual judgments of area and consumer-like preferences for mobile phones. We consider this generality a strength of the LBA models presented here.

When fit to choices and response times, the models provided a good account of choice proportions from both Experiments, albeit with some underestimation of large choice proportions – see inset panels of Figure 7. In Experiment 1 there was strong agreement between model and data for the best-only condition ($R^2 = .95$, root-mean-squared prediction error 7.97%), worst-only condition ($R^2 = .98$, 7.64%), and best-worst condition (best – $R^2 = .95$, 9.93%, worst – $R^2 = .96$, 8.47%). The models also provided a good account of choice data in Experiment 2 (best-only: $R^2 = .88$, 5.02%; worst-only: $R^2 = .88$, 5.00%; best-worst: best – $R^2 = .87$, 5.49%, worst – $R^2 = .90$, 4.81%). When the models were fit to choice proportions only, they also provided an excellent account of the choice data.[8] The fit to choice data in Experiment 2 was slightly worse than in Experiment 1 due to the complex, multi-attribute mobile phone stimuli and the more subjective nature of the decisions, which produce inherently noisier responses than the simple rectangles used in Experiment 1. Despite this, the models provided a good account of all response data.

Parameter Estimates

Figure 8 displays parameter estimates for the model fits to choices and response times in the best-worst conditions. The upper panels of Figure 8 display the nine estimated drift rates for the rectangular stimuli separated into the height and width dimensions. As expected, as each dimension increases in size so does the distribution of drift rate estimates. This results in a more likely best response for rectangular stimuli with larger area, and conversely a more likely worst response for rectangles with smaller area. Similarly, in Experiment 2 the estimated drift rates followed the trends expected from the choice data. For example, as the price of a phone increased, the drift rate decreased.

All of the other model fits resulted in similar parameter estimates. Regardless of whether one considered data from the best-only, worst-only, or best-worst conditions, the 'utility' conclusions drawn from the attribute-level drift rate estimates led to similar conclusions; there were strong correlations between the median parameter estimates from each combination of conditions across both experiments – see Table 2.

That the best-worst models provide a good fit to data suggests that the 'inverse' assumption on drift rates – the drift rate associated with selecting the worst option is the reciprocal of the drift rate associated with selecting the best option – is reasonable. It also suggests that the multiplicative and equal-weighted representation for option-level

[7] A more detailed fit could be made by setting the drift rate variability to one for best choices, and estimating it for worst choices.

[8] Goodness of fit to choice data when the models were fit to choice proportions, only: Experiment 1 best-only, $R^2 \approx 1$ (RMSE: .7%); worst-only, $R^2 \approx 1$ (.48%); best-worst, best – $R^2 = .98$ (3.70%); worst – $R^2 = .95$ (6.07%); Experiment 2 best-only, $R^2 = .98$ (1.71%); worst-only, $R^2 = .97$ (2.17%); best-worst, best – $R^2 = .95$ (2.65%), worst – $R^2 = .96$ (2.28%).

Table 2: Correlation coefficients (r) between median drift rate estimates from the LBA model fits to the best-only (B), worst-only (W), and best-worst (BW) conditions of Experiments 1 and 2 (all p's $< .001$; RT = response time).

	Model fit	B vs. BW	W vs. BW	B vs. W
Experiment 1	Choices-only	.96	−.94	−.97
	Choices & RT	.94	−.89	−.95
Experiment 2	Choices-only	.97	−.91	−.94
	Choices & RT	.94	−.92	−.91

drift rates in terms of attribute-level drift rates used in Experiment 2 is acceptable, which supports an inverse relationship between best and worst preferences for options. However, we can more directly test the inverse assumption on drift rates by allowing the model to estimate different drift rate (utility) parameters for the best and worst races. This allows, for example, the possibility that some options might be frequently chosen as best *and* worst (as in Shafir, 1993). This assumption is not easily checked without response time data, because the choice-only models come close to saturation (i.e., nearly as many free parameters as data points). Therefore, for both Experiments 1 and 2, we compared the fit of the best-worst LBA model to both choices and response times when implementing the inverse assumption (as reported above) to the fit of a model that had a separate set of drift rates for the best and worst races, and hence 9 (Experiment 1) or 10 (Experiment 2) additional free parameters in fitting data.

The more complex model that allows separate utilities for best and worst choices must necessarily fit the data better than the simple inverse model which it nests. To balance goodness of fit against model complexity, we used the Bayesian Information Criterion (BIC; Schwarz, 1978) and the Akaike Information Criterion (AIC; Akaike, 1974). AIC is much more lenient than BIC in its treatment of complexity, so will tend to prefer more complex models. For the perceptual stimuli used in Experiment 1, the simpler inverse model was preferred for 14 of 20 participants according to BIC, but only 2 of 20 participants according to AIC. The mean difference in selection criteria between the inverse and free models was equivocal for BIC (Δ BIC $M = -4.4$, $SD = 38.7$, where negative mean difference indicates support for the inverse model) but not for AIC (Δ AIC $M = 34.7$, $SD = 38.7$). For the phone stimuli used in Experiment 2, the inverse model was strongly preferred for all 30 participants according to BIC (Δ BIC $M = -48.1$, $SD = 16.1$) but the evidence was more equivocal for AIC, with the inverse model preferred for 17 of 30 participants (Δ AIC $M = -3.7$, $SD = 16.1$).

The AIC and BIC comparisons above are biased in favor of the complex model in the sense that they test just one particular version of the constrained model (a version where the drift rate for the worst choice is exactly the reciprocal of the drift rate for the best choice). In a more general sense, the drift rate estimates from the free model were highly consistent with a constrained explanation that assumes a single underlying source for best and worst choices. The median best and worst drift rates in the 'free' model were strongly negatively correlated for both Experiments 1 and 2, $r = -.99$, $t(7) = 18.9$, $p < .001$ and

$r = -.91$, $t(13) = 7.8$, $p < .001$, respectively, suggesting that best and worst drift rates provide similar information. These results provide yet more support that best and worst choices can be explained with a single latent dimension.

Constraining Parameter Estimates with Best-Worst Scaling and Response Times

It is plausible that, relative to traditional (best-only) discrete choice tasks, the two responses elicited in best-worst choice tasks provide greater information (Vermeulen, Goos, & Vandebroek, 2010). One reason the two responses may provide greater constraint is because previously empty cells in the design matrix are filled with useful information – options that are undesirable – translating into more precise parameter estimates in the best-worst compared to best-only and worst-only conditions. We directly tested this hypothesis. For each fitting mode (choices-only, choices and response times), condition (best-only, worst-only, best-worst) and experiment we conducted separate repeated measures ANOVAs with the estimated drift rates for each of the attribute-levels as the repeated measures factor. We used these ANOVAs to define the precision of parameter estimates as the amount of variance explained by the attribute-levels: $R^2 = 1 - \frac{SS_{error}}{SS_{total}}$. This approach was based on the assumption that additional data (a second response, and/or the response times) will reduce the proportion of unexplained variance in the parameter estimates. When fit to choice proportions only, drift rate estimates for the rectangular stimuli in Experiment 1 were more strongly related to the attribute-levels in the best-worst condition ($R^2 = .72$) compared to the best-only condition ($R^2 = .54$), but not the worst-only condition ($R^2 = .78$). A similar pattern was observed for the mobile phone data from Experiment 2: the best-worst condition ($R^2 = .46$) provided an improvement over the worst-only, but not best-only, condition ($R^2 = .37$ and $R^2 = .46$, respectively).

Similarly to the addition of a second choice in best-worst scaling, we expected that fitting models to choices and response times would provide further constraint on parameter estimates. This is exactly what happened for all three conditions in Experiment 1 – adding response times reduced error in the parameter estimates compared to fitting choice proportions, only (best-only – $R^2 = .73$ vs. .54; worst-only – $R^2 = .83$ vs. .78; best-worst – $R^2 = .81$ vs. .72). For Experiment 2 data, fitting the model to choices and response times increased error in the parameter estimates compared to choices, only, for the best-only condition ($R^2 = .40$ vs. $R^2 = .46$), led to a mild improvement for the worst-only condition ($R^2 = .41$ vs. $R^2 = .37$), but provided no change relative to the best-worst condition ($R^2 = .46$).

Our analysis suggests that obtaining a second different type of response (worst or best), and/or collecting response time information, generally provides greater constraint on model parameter estimates. It appears that there might be an approximate upper limit on the proportion of variance that relatively simple models such as the LBAs fitted here can explain in parameter estimates (e.g., approximately $R^2 \approx .80$ and $R^2 \approx .45$ in Experiments 1 and 2, respectively). Our results suggest that adopting either approach (collecting latency data or a second response) can reduce the proportion of unexplained variance and lead to more stable parameter estimates, which permits more reliable conclusions to be drawn.

General Discussion

We examined selection of the best and the worst option through the lens of discrete choice tasks, and best-worst scaling in particular, with the aim of investigating whether the two modes of judgment can arise from a single cognitive construct. Evidence from two experiments – one perceptual and one consumer – analyzed with three statistical approaches – Bayesian analysis, state-trace analysis, and cognitive modeling – converged on a single thesis. In particular, the deliberation involved in selecting the worst option from a set does not influence preferences for the best option, and, equally, selecting which option is best does not influence preferences for the worst option. Furthermore, best and worst choices appear to reflect judgments that arise from a single latent variable.

Our finding that patterns of best and worst choices can be explained by a single latent dimension may appear to be at odds with previous demonstrations that accepting and rejecting can be inconsistent (e.g., Shafir, 1993; Tsetsos et al., 2012); for example, these authors report that the same option can be both accepted and rejected under different task framing manipulations. However, in the Introduction, we have already discussed the fact that best/worst and accept/reject do not give the same information.

A further key difference between these previous studies and the research reported here is that we did not manipulate the variability in the attributes comprising the choice options. For example, in Tsetsos et al.'s (2012) task, streams of numbers were generated from two Gaussian distributions, one with a higher variance than the other, and respondents were asked to evaluate the mean value of each stream. When decisions were separated over time, respondents tended to both reject *and* accept the high variance option when given different task goals. This procedure has two important differences from our tasks, and from many standard choice experiments. Firstly, in our tasks decisions, when best and worst choices were made, they were made to each choice set, and therefore not as widely separated over time as in Shafir (1993) and Tsetsos et al. (2012). Secondly, our stimulus sets were not manipulated to contain high and low variance options (equivalently, 'enriched' or 'impoverished' options, Shafir, 1993). Furthermore, these studies used dissociation logic to support the thesis that accepting is inconsistent with rejecting. Future research is required to determine whether manipulations such as Tsetsos et al.'s produce inconsistency between best and worst choices when analyzed with a methodology that addresses the issue of confounding by scale-dependent interactions. Here, we directly addressed this issue with state-trace analysis, which assesses the ordinal relationships between experimental factors to overcome the effects of range-restricted response variables. With this analysis we demonstrated consistency between best and worst choices.

Another possible explanation for the discrepancy between Shafir's (1993) and Tsetsos et al.'s (2012) and our results is the divergent goals of researchers using process-based versus measurement-based models. For example, Busemeyer and Rieskamp (2013) describe how researchers using measurement models aim to apply statistical models to large samples of (discrete) choice data aggregated over people, to efficiently estimate parameters – which can be important to economists – while process-focused researchers aim to understand the cognitions underlying choice by individuals – which can be important to psychologists. This focus has led psychologists to construct particular experimental situations that yield interesting context effects, which in turn reveal subtle cognitive biases. In contrast, economists

tend to focus on measurement using designs that avoid such effects.

In this paper we aimed to bring together the occasionally opposing goals of choice modelers and psychologists, by applying a cognitive model typical of the psychological literature to an experimental task typical of the discrete choice literature. The LBA model applied in this paper does not naturally account for context effects, though it can be extended to do so (Trueblood, Brown, & Heathcote, in press). It is likely that the best-worst LBA implemented here can similarly be extended to account for known context effects in best-worst choice. For example, Shafir's (1993) finding that the sum of the proportion of times that an option is accepted under one instruction and rejected under a second instruction can be greater than 1 suggests accepting and rejecting are not consistent; we should also remember that Shafir (1993) manipulated the two instruction sets between subjects. However, framing (context) effects such as Shafir's (1993) might be fit by a weighted multi-attribute model where the weights change across tasks, but the underlying utilities do not. Thus, if respondents are guided to use different decision rules for accepting and rejecting, we may find different results for the two tasks. However we have found that in unguided tasks, with standard best, respectively worst, instructions, such effects tend not to arise.

In our results, the proportion of times that each stimulus option was selected as best did not differ across the best-only and best-worst choice conditions, for either perceptual or consumer stimuli (Experiments 1 and 2). This result is important since the best response is usually of greater interest to experimenters, particularly in applied domains. That is, for a marketer it is likely more profitable to know which product consumers might purchase, rather than the (potentially many) products they likely will not purchase. Our results suggest that data on best, or worst, choices can be used interchangeably to yield similar conclusions about preference strengths, though the best-worst procedure likely produces more reliable parameter estimates for model-based inference.

Acknowledgments

This research has been supported by Natural Science and Engineering Research Council Discovery Grant 8124-98 to the University of Victoria for Marley, and by Australian Research Council grants FT120100244 and DP12102907 to the University of Newcastle for Brown and Heathcote. The work was carried out, in part, whilst Marley was a Distinguished Professor (part-time) in the Centre for the Study of Choice, University of Technology, Sydney.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Bamber, D. (1979). State-trace analysis: A method of testing simple theories of causation. *Journal of Mathematical Psychology*, *19*, 137–181.

Bhatia, S. (in press). Associations and the accumulation of preference. *Psychological Review*.

Brown, S., & Heathcote, A. (2005). A ballistic model of choice response time. *Psychological Review*, *112*, 117–128.

Brown, S., & Heathcote, A. J. (2008). The simplest complete model of choice reaction time: Linear ballistic accumulation. *Cognitive Psychology*, *57*, 153–178.

Busemeyer, J. R., & Rieskamp, J. (2013). Psychological research and theories on preferential choice. In S. Hess & A. Daly (Eds.), *Handbook of Choice Modelling.* Edward Elgar Publishing.

Busemeyer, J. R., & Townsend, J. T. (1992). Fundamental derivations from decision field theory. *Mathematical Social Sciences*, *23*, 255–282.

Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence–accuracy relation in recognition memory. *Psychonomic Bulletin & Review*, *7*, 26–48.

Cheng, P. Y., & Chiou, W. B. (2010). Rejection or selection: Influence of framing in investment decisions. *Psychological Reports*, *106*, 247–254.

Dunn, J. C., & Kirsner, K. (1988). Discovering functionally independent mental processes: The principle of reversed association. *Psychological Review*, *95*, 91–101.

Dunn, J. C., Newell, B. R., & Kalish, M. L. (2012). The effect of feedback delay and feedback type on perceptual category learning: The limits of multiple systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 840–859.

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

Forstmann, B. U., Dutilh, G., Brown, S., Neumann, J., von Cramon, D. Y., Ridderinkhof, K. R., et al. (2008). Striatum and pre–SMA facilitate decision–making under time pressure. *Proceedings of the National Academy of Science*, *105*, 17538–17542.

Ganzach, Y. (1995). Attribute scatter and decision outcome: Judgment versus choice. *Organizational Behavior and Human Decision Processes*, *62*, 113-122.

Ganzach, Y., & Schul, Y. (1995). The influence of quantity of information and goal framing on decision. *Acta Psychologica*, *89*, 23–36.

Gilchrist, W. (2000). *Statistical modelling with quantile functions.* London: Chapman & Hall/CRC.

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

Ho, T., Brown, S., & Serences, J. (2009). Domain general mechanisms of perceptual decision making in human cortex. *Journal of Neuroscience*, *29*, 8675–8687.

Huber, V. L., Neale, M. A., & Northcraft, G. B. (1987). Decision bias and personnel selection strategies. *Organizational Behavior and Human Decision Processes*, *40*, 136–147.

Juliano, L., & Wilcox, K. (2011). Choice, rejection, and elaboration on preference-inconsistent alternatives. *Journal of Consumer Research*, *38*, 229–241.

Levin, I. P., Jasper, J. D., & Forbes, W. S. (1998). Choosing versus rejecting options at different stages of decision making. *Journal of Behavioral Decision Making*, *11*, 193–210.

Levin, I. P., Prosansky, C. M., Heller, D., & Brunick, B. M. (2001). Prescreening of choice options in 'positive' and 'negative' decision–making tasks. *Journal of Behavioral Decision Making*, *14*, 279–293.

Loftus, G. R. (1978). On interpretation of interactions. *Memory & Cognition*, *6*, 312–319.

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within–subject designs. *Psychonomic Bulletin & Review*, *1*, 476–490.

Loftus, G. R., Oberg, M. A., & Dillon, A. M. (2004). Linear theory, dimensional theory, and the face–inversion effect. *Psychological Review*, *111*, 835–863.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York: Wiley.

Luce, R. D. (1986). *Response times*. New York: Oxford University Press.

Ludwig, C. J. H., Farrell, S., Ellis, L. A., & Gilchrist, I. D. (2009). The mechanism underlying inhibition of saccadic return. *Cognitive Psychology*, *59*, 180–202.

Marley, A. A. J., & Flynn, T. N. (in press). Best and worst scaling: Theory and application. In J. D. Wright (Ed.), *International encyclopedia of the social and behavioral sciences* (2nd ed.). Elsevier.

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.

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology*, *4*, 61–64.

Morey, R. D., & Rouder, J. N. (2013). BayesFactor: Computation of Bayes factors for simple designs [Computer software manual]. Available from `http://CRAN.R-project.org/package=BayesFactor` (R package version 0.9.4)

Newell, B. R., & Dunn, J. C. (2008). Dimensions in data: Testing psychological models using state-trace analysis. *Trends in Cognitive Sciences*, *12*, 285–290.

Newell, B. R., Dunn, J. C., & Kalish, M. (2010). The dimensionality of perceptual category learning: A state-trace analysis. *Memory & Cognition*, *38*, 563–581.

Otter, T., Allenby, G. M., & van Zandt, T. (2008). An integrated model of discrete choice and response time. *Journal of Marketing Research*, *45*, 593–607.

Prince, M., Brown, S., & Heathcote, A. (2012). The design and analysis of state-trace experiments. *Psychological Methods*, *17*, 78–99.

R Development Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from `http://www.R-project.org` (ISBN 3–900051–07–0)

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two–choice decisions. *Psychological Science*, *9*, 347–356.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two–choice reaction time. *Psychological Review*, *111*, 333–367.

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, *9*, 438–481.

Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multi–alternative decision field theory: A dynamic artificial neural network model of decision–making. *Psychological Review*, *108*, 370–392.

Saville, D. J. (2003). Basic statistics and the inconsistency of multiple comparison procedures. *Canadian Journal of Experimental Psychology*, *57*, 167–175.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

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

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

Trueblood, J. S., Brown, S. D., Heathcote, A., & Busemeyer, J. R. (2013). Not just for consumers: Context effects are fundamental to decision–making. *Psychological Science*, *24*, 901–908.

Tsetsos, K., Chater, N., & Usher, M. (2012). Salience driven value integration explains decision

biases and preference reversal. *Proceedings of the National Academy of Science*, *109*, 9659–9664.

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review*, *79*, 281–299.

Usher, M., & McClelland, J. L. (2001). On the time course of perceptual choice: The leaky competing accumulator model. *Psychological Review*, *108*, 550–592.

Usher, M., & McClelland, J. L. (2004). Loss aversion and inhibition in dynamical models of multialternative choice. *Psychological Review*, *111*, 757–769.

van Buiten, M., & Keren, G. (2009). Speakers' choice of frame in binary choice: Effects of recommendation mode and option attractiveness. *Judgment and Decision Making*, *4*, 51–63.

van Zandt, T., Colonius, H., & Proctor, R. W. (2000). A comparison of two response time models applied to perceptual matching. *Psychonomic Bulletin & Review*, *7*, 208–256.

Vermeulen, B., Goos, P., & Vandebroek, M. (2010). Obtaining more information from conjoint experiments by best–worst choices. *Computational Statistics and Data Analysis*, *54*, 1426–1433.

Wagenmakers, E.-J., Krypotos, A.-M., Criss, A. H., & Iverson, G. (2012). On the interpretation of removable interactions: A survey of the field 33 years after Loftus. *Memory & Cognition*, *40*, 145–160.

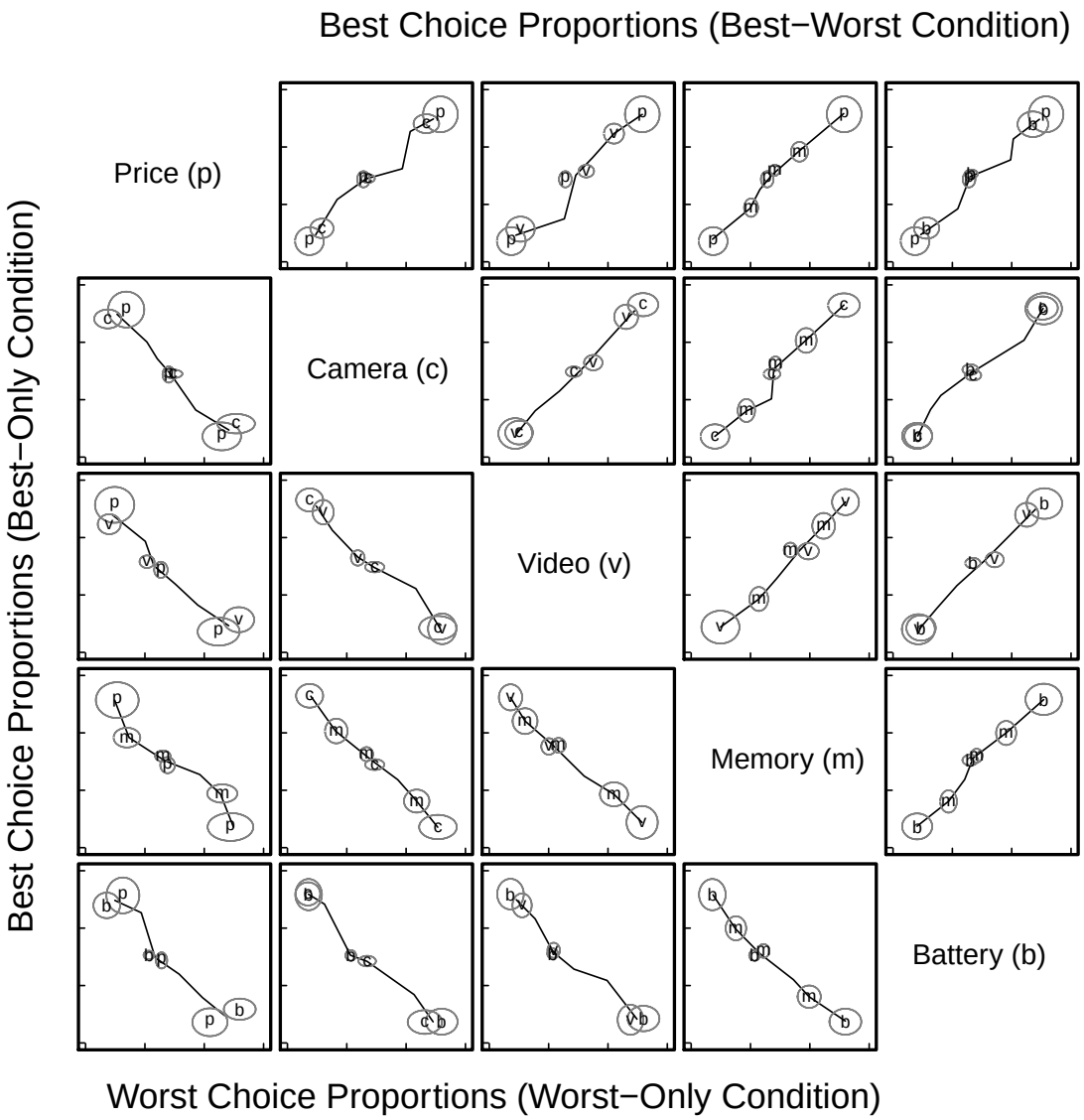

Figure 5. State-trace plots for the mobile phone judgments in Experiment 2. The upper right panels plot attribute-level best choice proportions between the best-only (y-axes) and best-worst (x-axes) conditions. The lower left panels plot attribute-level best choice proportions for the best-only condition (y-axes) against attribute-level worst choice proportions for the worst-only condition (x-axes). Each panel displays the co-variation of attribute-level mean choice proportions between the levels of two attributes, with the figure depicting all pairwise comparisons of attributes. Ellipses represent between-subjects least significant differences. Monotonic curves added to aid visualization only and do not represent 'best-fitting' monotonic functions to data. Axis scaling omitted for clarity. Note that axes are not equal in scale (see Figure 4).

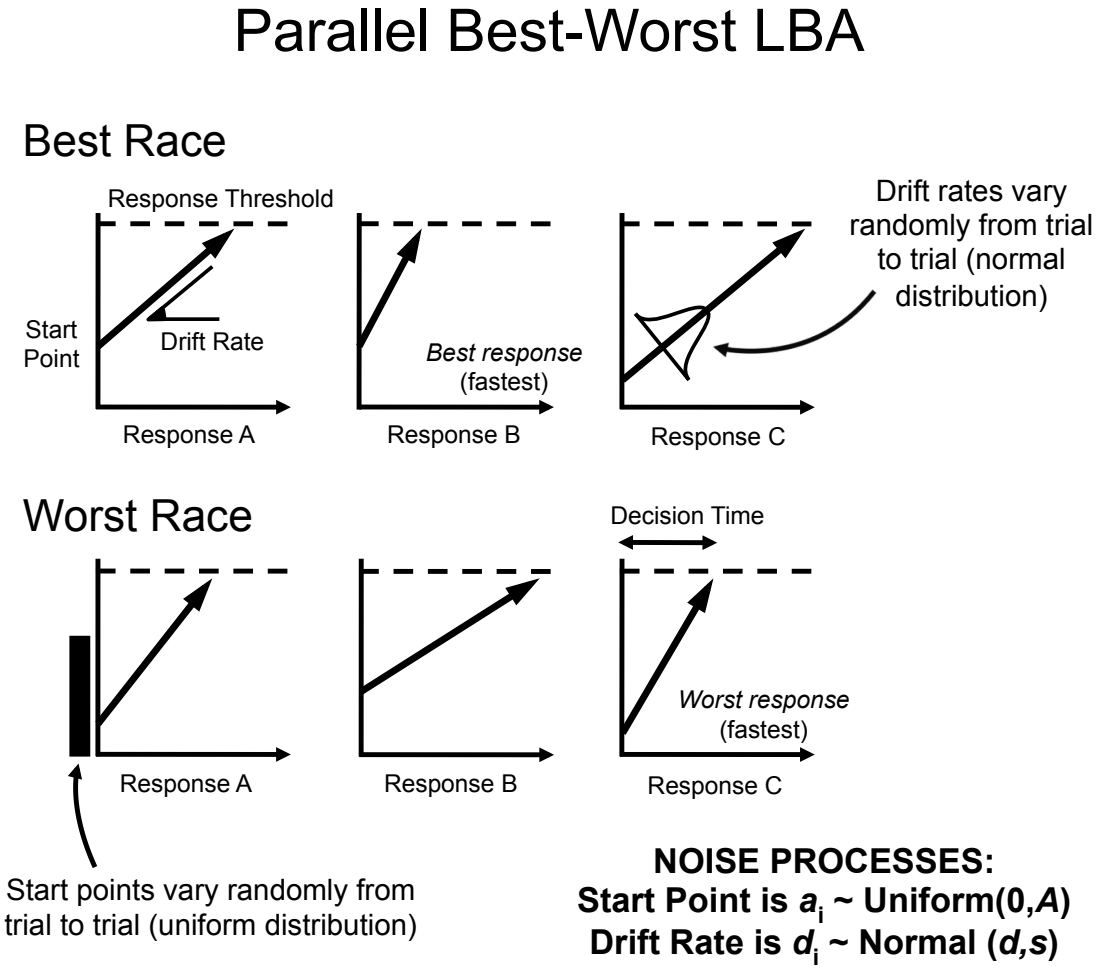

Figure 6. Illustrative example of the decision processes of the parallel best-worst linear ballistic accumulator model. See main text for full details.

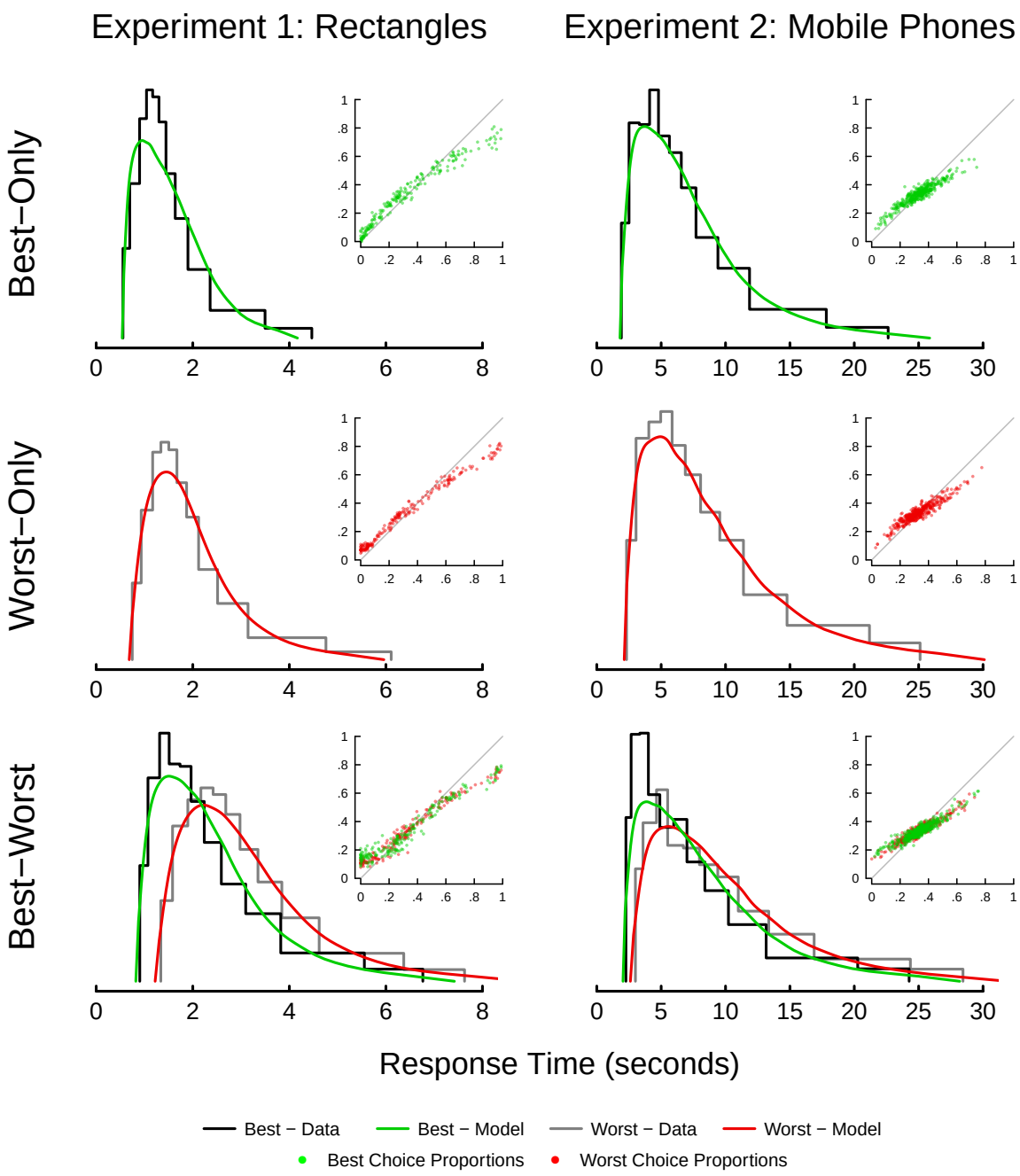

Figure 7. Aggregate response time distributions and attribute-level choice proportions from the best-only, worst-only, and best-worst conditions (upper, middle and lower rows, respectively) from Experiments 1 and 2 (left and right columns, respectively). Black and gray stepped histograms represent best and worst response times in data, respectively. Green and red smoothed density curves represent model predictions for best and worst response times, respectively. The inset panels represent attribute-level choice proportions observed in data (x-axes) and predicted by the models (y-axes), and the $y = x$ lines show a perfect fit.

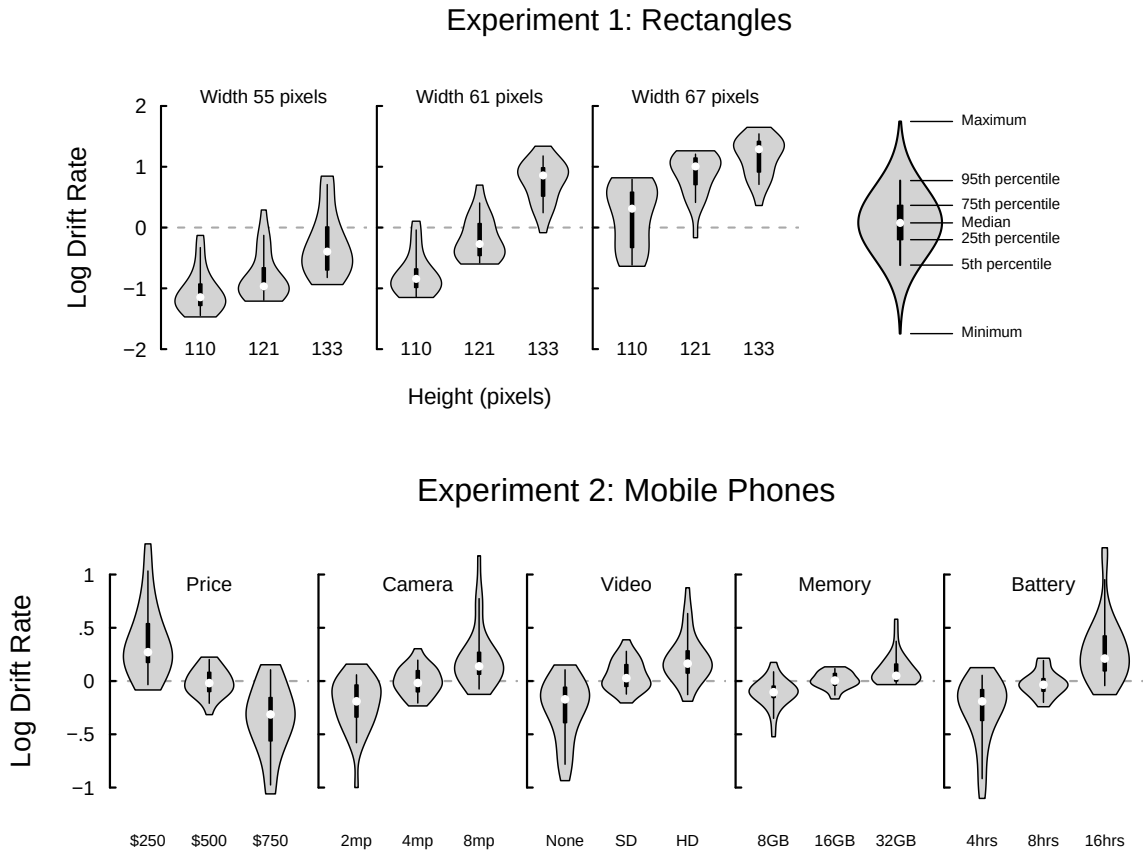

Figure 8. Violin plots demonstrating variability across participants in log LBA drift rate estimates in the best-worst condition of Experiments 1 and 2 (upper and lower panels, respectively). Each panel shows the distribution of drift rate estimates for the levels comprising each attribute. The legend in the upper right panel describes the marked regions that comprise each violin plot.