Integrating Cognitive Process and Descriptive Models of Attitudes and Preferences

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

Discrete choice experiments – selecting the best and/or worst from a set of options – are increasingly used to provide more efficient and valid measurement of attitudes or preferences than conventional methods such as Likert scales. Discrete choice data have traditionally been analyzed with random utility models that have good measurement properties, but provide limited insight into cognitive processes. We extend a well-established cognitive model, which has successfully explained both choices and response times for simple decision tasks, to complex multi-attribute discrete choice data. The fits, and parameters, of the extended model for two sets of choice data (involving patient preferences for dermatology appointments, and consumer attitudes towards mobile phones) agree with those of standard choice models. The extended model also accounts for choice and response time data in a perceptual judgment task designed in a manner analogous to best-worst discrete choice experiments. We conclude that a variety of fields can benefit from discrete choice experiments, especially when extended to include response time and combined with analyses based on evidence accumulation models.

Keywords: Preference; Decision-Making; Best-Worst Scaling; Linear Ballistic Accumulator; Evidence Accumulation; Random Utility Models.

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

Many fields rely on the elicitation of preferences. However, direct questioning methods, such as Likert scales, suffer from well-established drawbacks due to subjectivity (for a summary see Paulhus, 1991). Discrete choice – for example, choosing a single preferred product from a range of presented options – provides more reliable and valid measurement of preference in areas including health care (Ryan & Farrar, 2000; Szeinbach, Barnes, McGhan, Murawski, & Corey, 1999), personality measurement (Lee, Soutar, & Louviere, 2008), and marketing (Mueller, Lockshin, & Louviere, 2010). More efficient and richer discrete-choice elicitation is provided by best-worst scaling, where respondents select both the best option and worst option from a set of alternatives. For example, a respondent presented with six bottles of wine might be asked to report their most and least preferred bottles. Data collection using best-worst scaling has been increasingly used, particularly in studying consumer preference for goods or services (Collins & Rose, 2011; Flynn, Louviere, Peters, & Coast, 2007, 2008; Lee et al., 2008; Louviere & Flynn, 2010; Louviere & Islam, 2008; Marley & Pihlens, 2012; Szeinbach et al., 1999).

In applied fields, best-worst data are often analyzed using conditional logit (also called multinomial logit, MNL) models; the basic model is also known in cognitive science as the Luce choice model (Luce, 1959). These models assume that each option has a utility (u, also called "valence" or "preference strength") and that choice probabilities are simple (logit) functions of those utilities (Finn & Louviere, 1992; Marley & Pihlens, 2012).[1] MNL models provide compact descriptions of data and can be interpreted in terms of (random) utility maximization, but afford limited insight into the cognitive processes underpinning the choices made. Further, they do not address choice response time,[2] a measure that is increasingly easy to obtain as data collection becomes computerized.

We explore the application of evidence accumulation models – more often employed to explain simple perceptual choice tasks – to the complex decisions involved in best-worst choice between multi-attribute options. Our application bridges a divide between relatively independent advances in theoretical cognitive science (computational models of cognitions underlying simple decisions) and applied psychology (best-worst scaling to elicit maximal preference information from respondents). The result is a more detailed understanding of the cognitions underlying complex decisions, such as those involved in consumer preference, with no loss in the measurement or estimation properties relative to those of the random utility account of choices.

As summarized in the next section, previous work in this direction has been hampered by computational and statistical limitations. We show that these issues can be overcome by using the recently-developed linear ballistic accumulator (LBA: Brown & Heathcote, 2008) model. We do so by applying mathematically tractable LBA-based models to two best-worst scaling data sets: one involving patient preferences for dermatology appointments (Coast et al., 2006), and another involving preference for aspects of mobile phones (Marley

[1] The theoretical properties of MNL representations for best and/or worst choice were developed in Marley and Louviere (2005) and Marley, Flynn, and Louviere (2008); Marley and Pihlens (2012) and Flynn and Marley (submitted) summarize those results.

[2] Marley (1989) and Marley and Colonius (1992) present response time models that predict choice probabilities that satisfy the MNL (Luce) model. However, those models have limited flexibility in the form and properties of the response time distributions.

& Pihlens, 2012). In these applications, chosen to demonstrate the applicability of our methodology to diverse fields and measurement tasks, we show that previously-published MNL utility estimates are almost exactly linearly related to the logarithms of the estimated rates of evidence accumulation in the LBA model; this is the relation that might be expected from the role of the corresponding measures in the two types of models. We follow this demonstration with an application to a perceptual judgment task that uses the best-worst response procedure, with precise response time measurements, to demonstrate the benefit of response time information in understanding the decision processes involved in best-worst choice.

In the first section of this paper we describe evidence accumulation models, and the LBA model in particular. We then develop three LBA-based models for best-worst choice that are motivated by assumptions paralleling those previously used by a corresponding random utility model of choice. We show that these three LBA models provide a descriptive account of the best-worst choice probabilities equal to that of the random utility models. However, in the second section, we show that the three earlier LBA models are not supported by the response time data from the best-worst perceptual task, which provides evidence against some assumptions of MNL models. We then modify one of those LBA models to account for all features of the response time, and hence the choice, data. We conclude that response time data further our understanding of the decision processes in best-worst choice tasks, and that the LBA models we develop provide an easily applied methodology for that development that does not sacrifice the descriptive advantages of the random utility account of the choices.

Accumulator Models for Preference

Models of simple decisions as a process of accumulating evidence in favor of each response option have over half a century of success in accounting not only for the choices made but also the time taken to make them (for reviews, see: Luce, 1986; Ratcliff & Smith, 2004). When there are only two response choices, these models sometimes have only a single evidence accumulation process (e.g., Ratcliff, 1978), but if there are more options then it is usual to assume a corresponding number of evidence accumulators that race to trigger a decision (e.g., van Zandt, Colonius, & Proctor, 2000). Multiple accumulator models have provided comprehensive accounts of behavior when deciding which one of several possible sensory stimuli has been presented (e.g., Brown, Marley, Donkin, & Heathcote, 2008) and even accounted for the neurophysiology of rapid decisions (e.g., Forstmann et al., 2008; Frank, Scheres, & Sherman, 2007). Accumulator models have also been successfully applied to consumer preference, most notably decision field theory (Busemeyer & Townsend, 1992; Roe, Busemeyer, & Townsend, 2001), the leaky competing accumulator model (Usher & McClelland, 2004) and, most recently, the $2N$-ary choice tree model (Wollschläger & Diederich, 2012).

Although they provide detailed mechanistic accounts of the processes underlying decisions that is lacking in random utility models, the cognitive models suffer from practical difficulties that do not apply to classic random utility models of choice. In particular, the cognitive models do not have closed form expressions for the joint likelihood of response choices and response times, given parameter settings. When only response choices are considered, some of these models do have simple expressions for the likelihood functions.

However, when response times are included as well, the likelihood functions have to be approximated either by Monte-Carlo simulation, or by discretization of evidence and time (Diederich & Busemeyer, 2003). The Monte-Carlo methods make likelihoods very difficult to estimate accurately, and the discretization approach can prove impractical when there are many options in the choice set. More recently, Ruan, MacEachern, Otter, and Dean (2008) and Otter, Allenby, and van Zandt (2008) proposed a Poisson processes race model that made substantial progress toward the practical application of cognitive modeling to discrete (best-only) choice data. They also showed that, by using response time data, their approach provides valuable insights into consumer preferences. Although Otter et al.'s approach shares many of the advantages of the models we propose below, it is limited by the underlying Poisson accumulator model, which – in contrast to the LBA model (see Brown & Heathcote, 2008) – has been shown to provide a less-than-complete account of standard perceptual decision data (see Ratcliff & Smith, 2004).

Like other multiple accumulator models, the LBA is based on the idea that the decision maker accumulates evidence in favor of each choice, and makes a decision as soon as the evidence for any choice reaches a threshold amount. This simple "horse race" architecture makes the model simple to analyze and use, but also limits its ability to explain subtle preference reversals and context effects – a point we return to below. For the LBA model, the time to accumulate evidence to threshold is the predicted decision time, and the response time is the decision time plus a fixed offset (t_0), the latter accounting for processes such as response production. Figure 1 gives an example of an LBA decision between options A and B, represented by separate accumulators that race against each other. The vertical axes represent the amount of accumulated evidence, and the horizontal axes the passage of time. Response thresholds (b) are shown as dashed lines in each accumulator, indicating the quantity of evidence required to make a choice. The amount of evidence in each accumulator at the beginning of a decision (the "start point") varies independently between accumulators and randomly from choice to choice, sampled from a uniform distribution: $U(0, A)$, with $A \leq b$. Evidence accumulation is linear, as illustrated by the arrows in each accumulator of Figure 1. The speed of accumulation is traditionally referred to as the "drift rate", and this is assumed to vary randomly from accumulator to accumulator and decision to decision according to an independent normal distribution for each accumulator, reflecting choice-to-choice changes in factors such as attention and motivation.

Mean drift rate reflects the attractiveness of an option: a higher value means a faster rise to threshold, and therefore a more likely choice outcome. For example, in Figure 1 suppose option A has a mean drift rate of 0.6 ($d_A = 0.6$) and option B a mean drift rate of 0.4 ($d_B = 0.4$). When options A and B are presented together, a choice of response A is more likely *on average* since option A has a larger drift rate than option B, and so will usually reach threshold first. However, since there is noise in the decision process (in both start point and drift rate), the accumulator for response B will occasionally reach threshold first, leading to a choice of option B. Noise in the decision process allows the LBA to account for the observed variability in decision making, successfully predicting the joint distribution of response times and response choices across a wide range of tasks (e.g., Brown & Heathcote, 2008; Forstmann et al., 2008; Ho, Brown, & Serences, 2009). Here we employ a slightly modified LBA where drift rates are drawn from strictly positive truncated

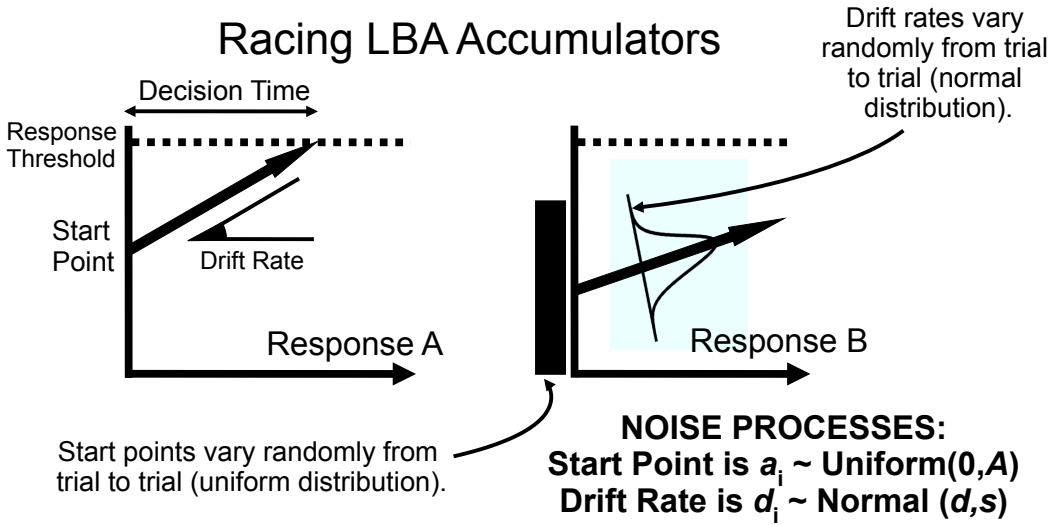

Figure 1. Illustrative example of the decision processes of the LBA. See main text for full details.

normal distributions (for details, see Heathcote & Love, 2012).[3]

Horse Race Models for Best-Worst Choice

We used the modified LBA to create three different models for best-worst scaling, derived from previously-applied random utility models of choice (see, e.g., Marley & Louviere, 2005). Each variant involves a race among accumulators representing the competing choices. In the first, which we refer to as the *ranking* model, there is one race, with the first (respectively, last) accumulator to reach threshold associated with the best (respectively, worst) choice. The second variant, the *sequential* model, has two races that occur in sequence; the winner of the first race determines the best response, and, omitting the winner of the first race, the winner of the second race determines the worst response; we constrain each drift rate for the second race to be the inverse of the corresponding rate in the first race. The third variant, the *enumerated* model, assumes a single race between accumulators that represent each possible pair of best and worst choices (e.g., 12 accumulators for 4 options). For a best-worst pair (p, q), $p \neq q$, the drift rate is the ratio $d(p)/d(q)$ of the drift rate for p versus the drift rate for q. Our later fits to data show that the estimated drift rate d for an option is, effectively, equal to the exponential of the estimated utility u for that option in the corresponding MNL model. These are not the only plausible accumulator-based best-worst models. For example, best-worst choice might be modeled by two simultaneous races driven by utilities and disutilities, respectively. However we consider these models because they correspond to the frameworks adopted by the marginal and paired-conditional ("maxdiff") random utility models most commonly used to analyze best-worst scaling data (Marley &

[3]The normal truncation requires the density and cumulative density expressions for individual accumulators in Brown and Heathcote (2008) (Equations 1 and 2) to be divided by the area of the truncated normal, $\Phi(-d/s)$.

Louviere, 2005).

General Framework

We now present the LBA-based models, by developing equations for the predicted choice probabilities and response times (omitting fixed offset times for processes like motor production). To begin the notation, let S with $|S| \geq 2$ denote the set of potentially available options, and let $X \subseteq S$ be a finite subset of options that are available on a single choice occasion. Assume there is a common threshold b across all options in the set of available options X. For $z \in S$ and $p, q \in S$, $p \neq q$, there are: a drift rate $d(z)$; independent random variables $\boldsymbol{U}_z$, $\boldsymbol{U}_{p,q}$ uniformly distributed on $[0, A]$ for some $0 \leq A \leq b$; and independent normal random variables $\boldsymbol{D}_z$ and $\boldsymbol{D}_{p,q}$ with mean 0 and standard deviation s. For best choices, the drift rate variable for option z is then given by $trunc(\boldsymbol{D}_z + d(z))$ where the truncation is to positive values, and for worst choices, the drift rate variable for option z is then given by $trunc(\boldsymbol{D}_z + 1/d(z))$, again with the truncation to positive values. Similarly, the drift rate variable for the best-worst pair p, q is then given by $trunc(\boldsymbol{D}_{p,q} + d(p)/d(q))$. For best choices, the probability density function (PDF) of finishing times for the accumulator for option $z \in X$ at time t, denoted $b_X(z, t)$, is given by

$$b_z(t) = \Pr\left(\frac{b - \boldsymbol{U}_z}{trunc(\boldsymbol{D}_z + d(z))} = t\right),$$

with cumulative density function (CDF)

$$B_z(t) = \Pr\left(\frac{b - \boldsymbol{U}_z}{trunc(\boldsymbol{D}_z + d(z))} \leq t\right).$$

We denote the corresponding PDF and CDF for worst choice by $w_z(t)$ and $W_z(t)$, which are given by replacing $d(z)$ with $1/d(z)$ in the above formulae. For best-worst choice they are $bw_{(p,q)}(t)$ and $BW_{(p,q)}(t)$ with $d(p)/d(q)$ replacing $d(z)$ in the above formulae. Expressions given in Brown and Heathcote (2008) can be used to derive easily computed expressions for these PDFs and CDFs under the above assumptions, and using drift rate distributions truncated to positive values as described in Heathcote and Love (2012). Then, given the assumption that the accumulators are independent – that is, each accumulator has independent samples of the start point and drift rate variability – it is simple to specify likelihoods conditional on response choices and response times. These likelihoods for each of the three best-worst LBA models are shown in the next three sub-sections.

When fitting data without response times, we made the simplifying assumptions that $s = 1$, $b = 1$ and $A = 0$, as these parameters are only constrained by latency data. Even in response time applications, fixing s in this way is common. Estimation of b can accommodate variations in response-bias and overall response speed, but without response time data, speed is irrelevant and response-bias effects can be absorbed in drift rate estimates. The relative sizes of A and b are important in accounting for differences in accuracy and decision speed that occur when response speed is emphasized at the expense of accuracy, or vice versa (Brown & Heathcote, 2008; Ratcliff & Rouder, 1998). Our setting here ($A = 0$) is consistent with extremely careful responding. We also tried a less extreme setting ($b = 2$ and $A = 1$) with essentially equivalent results.

Ranking Model

The ranking model is arguably the simplest way to model best-worst choice with a race. For a choice among n options it assumes a race between n accumulators with drift rates $d(z)$. The *best* option is associated with the first accumulator to reach threshold and the *worst* option with the last accumulator to reach threshold, as shown in the upper row of Figure 2. To link the models to data requires an expression for the probability of choosing a particular best-worst pair for each possible choice set, X, given model parameters. The PDF for a choice of option x as the best at time t, and option y as the worst at time r, where $r > t$, $bw_X(x, t; y, r)$, is given by Equation 1 shown in Figure 2, for $x, y \in X$, $x \neq y$. Equation 1 calculates the product of the probabilities of the "best" accumulator finishing at time t, the "worst" accumulator finishing at time $r > t$, and all the other accumulators finishing at times between t and r. Since the data sets do not include response times, we calculate the marginal probability $BW_X(x, y)$ of the selection of this best-worst choice pair, by integrating over the unobserved response times (t and r) – see Equation 2 in Figure 2.

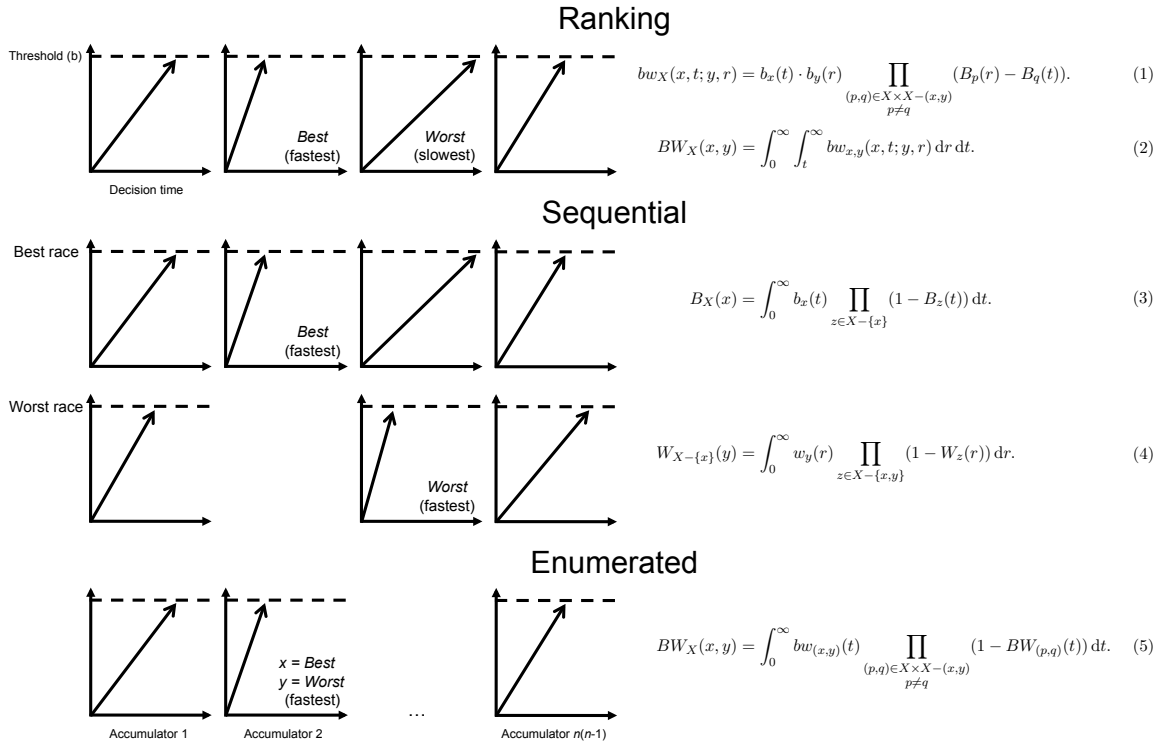

Figure 2. Illustrative example of the decision processes of the ranking, sequential and enumerated versions of the best-worst race models and their associated formulae (upper, middle and lowers rows, respectively). See main text for full details.

The ranking model predicts that the best option is chosen before the worst option, which may not be true in data. The ranking model could also be implemented in a worst-to-best form, with the first finishing accumulator associated with the worst option and the last finishing accumulator with the best option, and with each drift rate the inverse of the

corresponding drift rate in the best-to-worst version. It is difficult to separate these versions until we have response time data, which we return to later.

Sequential Model

The sequential race model assumes the best-worst decision process is broken into two separate races that occur consecutively (see middle row of Figure 2). The *best race* occurs first and selects the best option. The *worst race* then selects the worst option. The sequential model could have $2 \times n$ mean drift rate parameters – one for each option in the best race, and one for each option in the worst race. To simplify matters, we assume that, for each $z \in S$, the drift rate is $d(z)$ in the best race and $1/d(z)$ in the worst race. This ensures that desirable choices (high drift rates) are both likely to win the best race and unlikely to win the worst race. We also assume that the worst-choice race does not include the option already chosen as best so that the same option cannot be chosen as both best and worst.

The *first* accumulator to reach threshold is selected as the best option. The probability for a choice of the option x as the best, $B_X(x)$, is given in Equation 3. It is the probability density that accumulator x finishes at time t, and all the other accumulators finish at times later than t, integrated over all times $t > 0$. The worst race is run with the accumulators in the set $X - \{x\}$. The *first* accumulator to reach threshold is selected as the worst option, but each drift rate is now the inverse of the corresponding drift rates in the first race. The probability of a choice of the option y as the worst, $W_{X-\{x\}}(y)$, is given in Equation 4. The joint probability for the sequential model of choosing x as the best and y as the worst, $BW_X(x, y)$, is simply the product of Equations 3 and 4.

Clearly, a corresponding model is easily developed where the *worst race* occurs first and selects the worst option, and the *best race* occurs second and selects the best option.

Enumerated Model

The enumerated model assumes a race between each possible best-worst pair in the choice set. This is analogous to the paired conditional (also called "maxdiff") random utility model (Marley & Louviere, 2005). For a choice set with n options, the model assumes a race between $n \times (n-1)$ accumulators. This model predicts a single decision time for both responses. The probability of choosing x as best and $y \neq x$ as worst for this model is shown in Equation 5.

For a choice set with n options, the enumerated model could have $n \times (n-1)$ drift rate parameters. However, we simplified the enumerated model in a similar way to the sequential model, by again defining the desirability of options by their drift rates, and the undesirability of options by the inverse of their drift rates. In particular, we estimated a single drift rate parameter $d(z)$ for each choice option z, and set the drift rate for the accumulator corresponding to the option pair (p, q) to the ratio $d(p)/d(q)$.

Estimating Model Parameters from Data

We fit the three race models to two best-worst choice data sets, one about patients' preferences for dermatology appointments (Coast et al., 2006), and the second about preferences for mobile phones (Marley & Pihlens, 2012). In both data sets response times were

unavailable (i.e., not recorded) and the data structure was "long but narrow" – large sample sizes with relatively few data points per participant – which is standard in discrete choice applications. Coast et al.'s data investigated preferences for different aspects of dermatological secondary care services. Four key attributes were identified as relevant to patient experiences, one of which had four levels, with the remaining three having two levels each (see Table 1). We denote each attribute/level combination (henceforth, "attribute level") with two digits, shown in parentheses in the right column of Table 1. The first digit refers to the attribute and the second digit to its level: for example, attribute level "32" refers to level 2 of attribute number 3. For all four attributes, the larger the second digit (i.e., level of the attribute) the more favorable the level of the attribute.

Table 1: The four attributes and their levels from Coast et al. (2006). The values in parentheses indicate the coding used in Figure 3 below.

Attribute	Attribute levels
Waiting time (1)	Three months (11) Two months (12) One month (13) This week (14)
Doctor expertise (2)	The specialist has been treating skin complaints part-time for 1-2 years (21). The specialist is in a team led by an expert who has been treating skin complaints full-time for at least 5 years (22).
Convenience of appointment (3)	Getting to the appointment will be difficult and time consuming (31). Getting to the appointment will be quick and easy (32).
Thoroughness of consultation (4)	The consultation will not be as thorough as you would like (41). The consultation will be as thorough as you would like (42).

Participants were given a description of a dermatological appointment that included a single level from each attribute, and asked to indicate the best and the worst attribute level. For example, on one choice occasion, a participant might be told that an upcoming appointment is: two months away (12); with a highly specialized doctor (22); at an inconvenient location (31) and not very thorough (41). They would then be asked to choose the best and worst thing about this appointment. The same 16 scenarios were presented to each participant in the study, and were chosen using design methodology that enabled all main effects to be estimated. Below we compare the parameters estimated from the race models to parameters of the MNL models for Coast et al.'s (2006) choice data reported by Flynn et al. (2008).

Marley and Pihlens (2012) examined preferences for various features of mobile phones among 465 Australian pre-paid mobile phone users in December 2007. There were nine

mobile phone attributes with a combined 38 attribute levels, described in Table 2. The values in parentheses in Table 2 code attribute levels in a manner similar to Table 1, though not all attributes have a natural preference order. Each respondent completed the same 32 choice sets with four profiles per set. Each phone profile was made up of one level from each of the nine attributes. Participants were asked to provide a full rank order of each choice set by first selecting the best profile, then the worst profile from the remaining three, and finally the best profile from the remaining two. Here, we restrict our analyses to choices of the best and the worst profile in each choice set.

Parameter Constraints

A natural scaling property of the LBA model, parallel to MNL models, is that the drift rate distributions for competing choices can be multiplied by an arbitrary scale factor without altering predicted choice probabilities – although response time predictions will be affected. We employed a modified LBA model, with truncated normal drift rate distributions. In this version, multiplying the drift rate parameters does not simply multiply all distributions. Rather, the distribution shape is altered because the amount of truncation changes depending on how far the mean of the distribution falls from zero. For this reason, the scaling property holds only approximately for the truncated-normal LBA, and the approximation depends on the size of the drift rate parameters.

For the mobile phone data, there were 38 different attribute levels, and the approximation to the regular scaling property held well enough that we were able to constrain the product of the estimated drift rates across the attribute levels to one, for each attribute. This results in 29 free drift rate parameters, and mirrors Marley and Pihlens' (2012) constraints on their MNL model parameters. The dermatology data involved more extreme choice probabilities, and so smaller drift rates for some attributes. Therefore, we were not able to exploit the usual scaling property, and we imposed no constraints: there were 10 attribute levels, and we estimated 10 free drift rates. We note that this freedom allows us, theoretically, to separately estimate mean utility parameters and the associated variance parameters, which may prove useful in future research (Flynn, Louviere, Peters, & Coast, 2010).

Model Fit

We aggregated the data across participants; thus, for each choice set in the design, we fit a single set of choice probabilities. We used two methods to evaluate the fit of the race models to data. The first compared drift rate estimates to corresponding random utility models regression coefficients reported by Flynn et al. (2008) and Marley and Pihlens (2012). In this comparison we take the logarithm of the drift rates, bringing them onto the same unbounded domain as the utility parameters. Secondly, we examined the race models' goodness of fit by comparing observed and predicted best-worst choice proportions. For each choice set, observed best-worst choice proportions were calculated by dividing the number of times a particular best-worst pair was selected across participants by how many times that particular choice set was presented across participants.

Results

Coast et al.'s (2006) Dermatology Data

Flynn et al. (2008) analyzed Coast et al.'s (2006) data using a paired model conditional logit regression, adjusted for covariates.[4] Flynn et al.'s regression coefficients were expressed as treatment-coded linear model terms (main effects for attributes, plus treatment effects for each level). For example, Flynn et al. found that the main effect for the "convenience" attribute was 0.715 with a treatment effect of 1.501 for the "very convenient" level (Table 4; Flynn et al.). For ease of comparison, we expressed the estimated drift rate parameters from the LBA models in this same coding. We referenced all parameters against the zero point defined by the *waiting time* attribute, by subtracting the mean drift rate for this attribute from all drift rates. We then calculated the main effect for each attribute as the mean drift rate for that attribute, and calculated treatment effects for each attribute level by subtracting the main effects. These calculations are independent of the parameter estimation procedure and were done solely to facilitate comparison with Flynn et al.'s results.

The upper row of Figure 3 compares the log drift rates estimated from the three race models against the corresponding parameters from Flynn et al.'s (2008) fit of the maxdiff (MNL) model; *Appendix A* gives the form of that model. The four main effect estimates are shown as bold faced single digits, and treatment effects as regular faced double digits, using the notation from Table 1. For all three model variants, there was an almost perfect linear relationship between log drift rates estimated for the race model and the parameters for the corresponding MNL model.

All of the race models provided an excellent fit to the dermatology data, as shown in the lower row of Figure 3. In those plots, a perfect fit would have all the points falling along the diagonal line. For all models there was close agreement between observed and predicted values, with all R^2's above .9. The root-mean-squared difference between observed and predicted response probabilities was 5.4%, 5.1% and 5.3% for the ranking, sequential and enumerated models, respectively. The corresponding log-likelihood values were -1379, -1406 and -1381, respectively, providing little basis to select between LBA models in this analysis. Flynn et al.'s (2008) marginal model conditional logit analysis, which has the same number of free parameters as our LBA models, produced a log-pseudolikelihood of -1944 suggesting that the LBA provides a better fit to this data.

Marley & Pihlens' (2012) Mobile Phone Data

Marley and Pihlens (2012) analyzed their full rank data using a repeated maxdiff (MNL) model;[5] *Appendix A* gives the form of that maxdiff model for the first (best) and last (worst) options in those rank orders. We compare the log drift rate parameter estimates from our race models for those first (best) and last (worst) choices against Marley and Pihlens' utility parameter estimates. The upper row of Figure 4 plots the estimated log

[4] Flynn et al. (2008) also performed a paired model conditional logit regression (without covariates) and a marginal model conditional logit analysis. The regression coefficients did not differ much across these analyses, and so we ignore those, for brevity.

[5] Marley and Pihlens (2012) also analyzed their data using other variants of the MNL framework, with little difference to the estimated coefficients. Again, for brevity, we show just the main analyses.

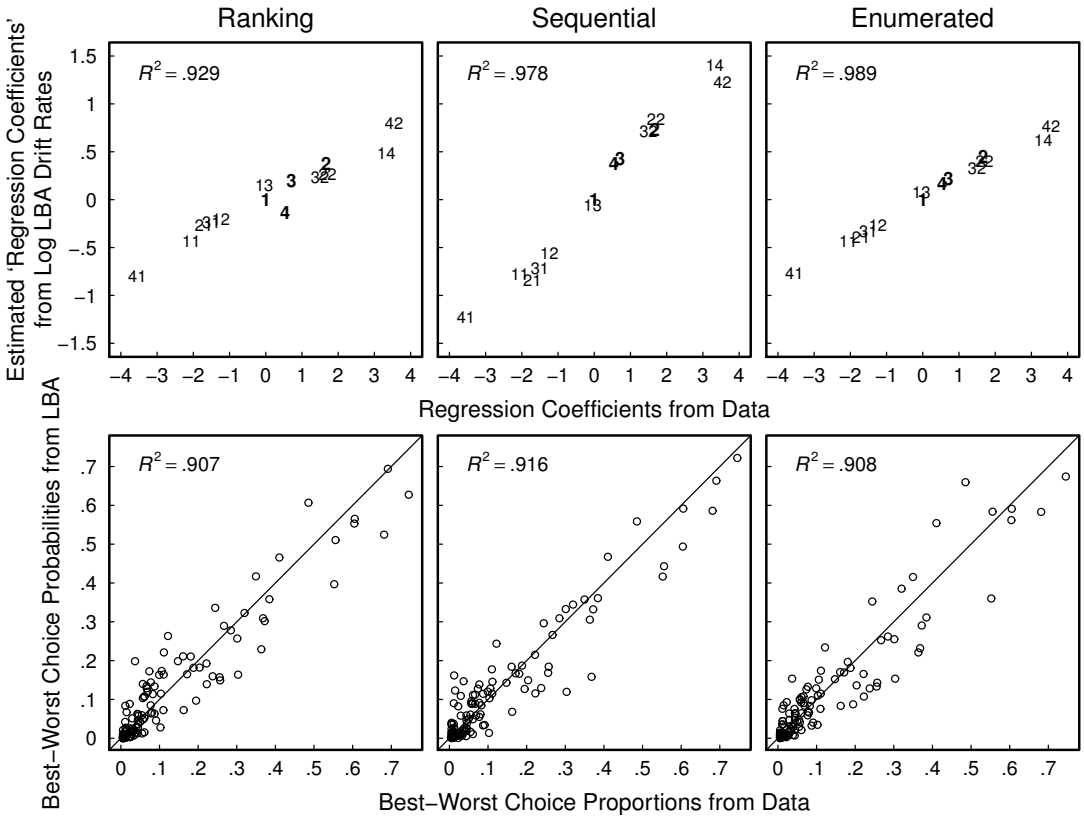

Figure 3. Log drift rate parameter estimates (upper row), plotted against Flynn et al.'s (2008) utility estimates, and goodness of fit (lower row) of the ranking, sequential and enumerated race models (columns) to Coast et al.'s (2006) data. In the upper row, bold face single digits represent main effects, double digits represent attribute levels, using the notation from Table 1. In the lower row, the x-axes display best-worst choice proportions from data, the y-axes display predicted best-worst choice probabilities from the estimated race model. The diagonal lines show a perfect fit.

drift rates from the race models against the regression coefficients from Marley and Pihlens' maxdiff model. There was again a nearly perfect linear relationship between log drift rates estimated for the race model and regression coefficients.

To further demonstrate the strength of the linear relationship between drift rates and utility estimates, we re-present the parameter values for the sequential model shown in Figure 4 separately for each attribute, in Figure 5. Figure 5 clearly illustrates the almost perfect correspondence between the rank ordering on drift rates and the rank ordering on regression coefficients. Not only do the drift rates preserve the ordering, but also differences in magnitude between levels of each attribute. For example, the *price* attribute, shown in the upper right panel of Figure 5, demonstrates that people have the strongest preference for the cheapest phones ($49) and the weakest preference for the most expensive ones ($249). However, the difference in utility (regression coefficients) is much greater between some adjacent levels than others (e.g., moving from the third to the forth level, $199 to $249).

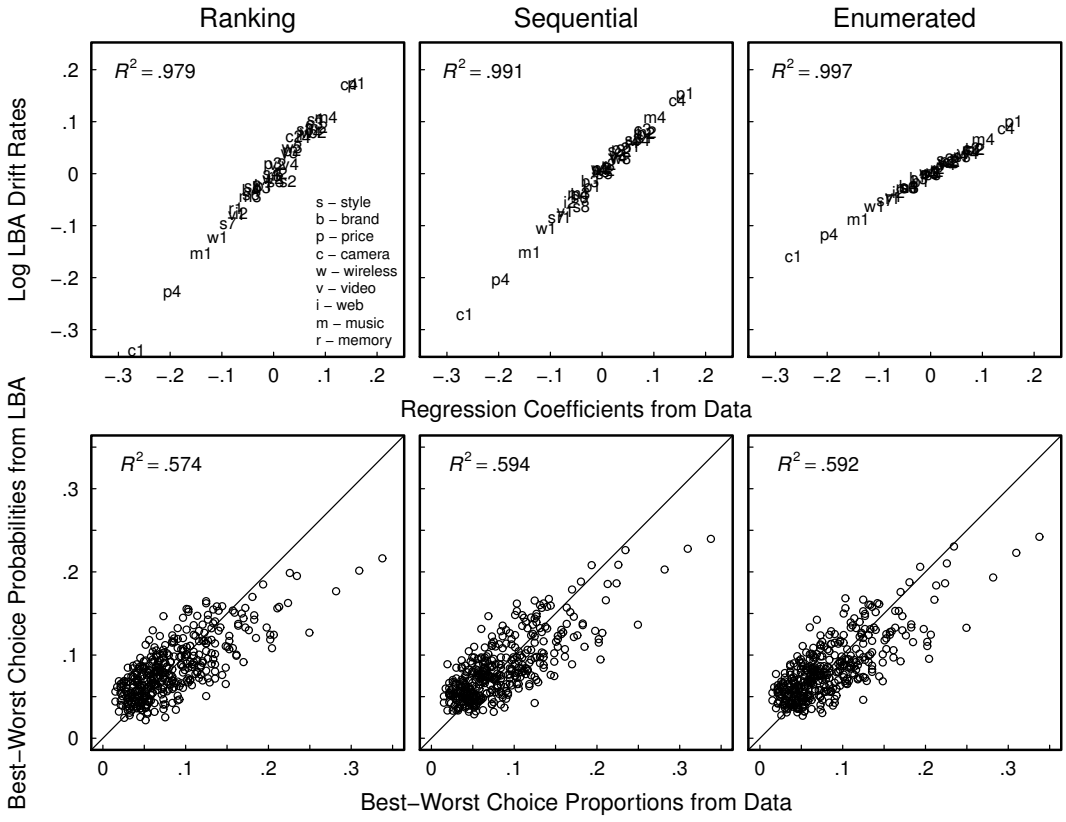

Figure 4. Log drift rate parameter estimates (upper row), plotted against Marley and Pihlens' (2012) regression coefficients, and goodness of fit (lower row) of the ranking, sequential and enumerated race models (left, middle and right columns, respectively) to Marley and Pihlens' mobile phone data. In the upper row, each point represents an attribute level, where the letter indicates the attribute and the number indicates the level of the attribute, as in Table 2. In the lower row, the x-axes display best-worst choice proportions from data, the y-axes display predicted best-worst choice probabilities from the estimated race model, and the diagonal lines represent a perfect fit.

Such a difference in magnitude also occurred in, for instance, the *camera* attribute, where a phone with no camera (level 1) was much less desirable than any phone with a camera (levels 2, 3 and 4). In all cases the estimated drift rates were sensitive to such differences in magnitude as well as the rank ordering. Sensitivity to these important outcome measures (ranking and magnitude) suggests the race models may be useful for measurement purposes.

We assessed the goodness of fit of each model by comparing observed and predicted proportions, shown in the lower half of Figure 4. As with the dermatology data, there was excellent agreement, with 3.2% root-mean-squared prediction error for all three models. Although the goodness-of-fit appears poorer in Figure 4 compared to Figure 3, this is actually due to differences in the scale of the axes across figures. The R^2's were smaller than for the dermatology data (though all are $> .57$), reflecting greater inter-individual variability in choices. To compare with the goodness of fit for the MNL models reported by

Marley and Pihlens (2012), we calculated McFadden's ρ^2 measure. McFadden's ρ^2 measures the fit of a full model with respect to the null model (all parameters equal to 1), defined as $\rho^2 = 1 - \frac{\ln \hat{L}(M_{full})}{\ln \hat{L}(M_{null})}$, where $\ln \hat{L}(M_{full})$ and $\ln \hat{L}(M_{null})$ refer to the estimated log-likelihood of the full and null models, respectively. This showed almost identical results for the race models (ranking $\rho^2 = .245$, sequential $\rho^2 = .246$, and enumerated $\rho^2 = .246$) as the MNL models (best, then worst, $\rho^2 = .244$, and maxdiff $\rho^2 = .245$).

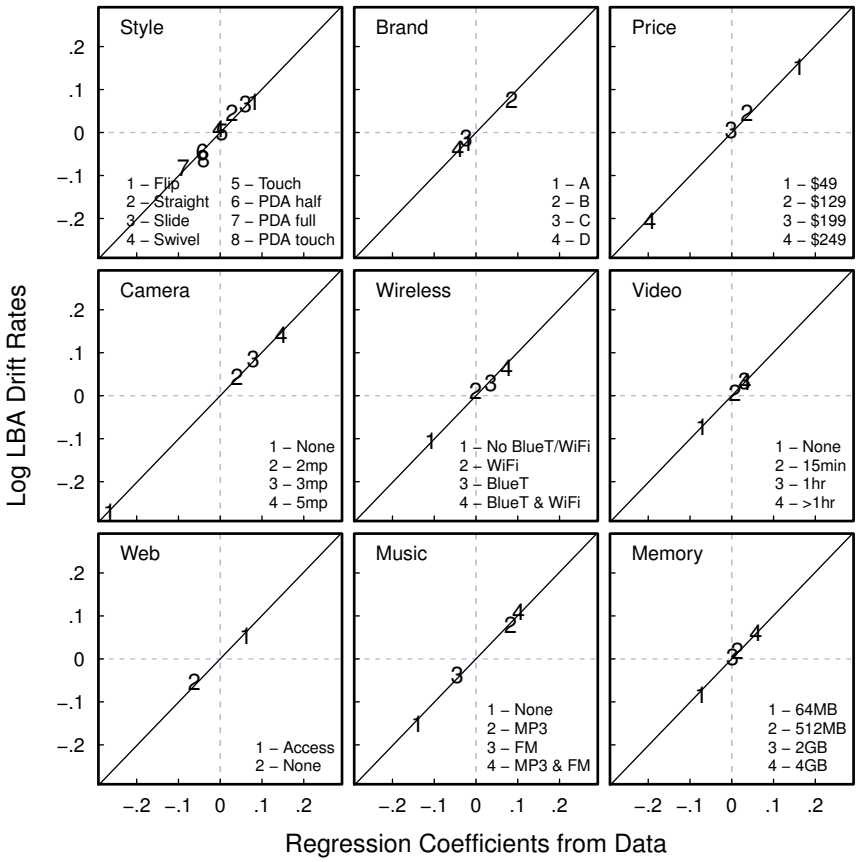

Figure 5. Fit of the sequential race model to Marley and Pihlens' (2012) mobile phone data, shown separately for each attribute. Log drift rates are plotted against Marley and Pihlens' regression coefficients. Each panel represents a different attribute. The numbers inside the nine panels represent each attribute level. The black lines in each panel represent the regression line fit to the sequential race model log drift rates and Marley and Pihlens' regression coefficients shown in the middle panel of the upper row in Figure 4. The dashed horizontal and vertical lines represent zero-reference points.

As a final comparison with existing MNL models for best-worst data, we compared the race models' drift rate parameters against "best minus worst scores" calculated from the data.[6] As with similar analyses in the literature (Finn & Louviere, 1992; Goodman, 2009;

[6]These scores are normalized differences between the number of "best" responses and "worst" responses

Mueller Loose & Lockshin, 2013), for both data sets the agreement between the best minus worst scores and the drift rates was just as strong as the relationship between drift rates and regression coefficients, reinforcing the current consensus that the best minus worst scores are a simple, but useful, way to describe data. Theoretical properties of these scores for the maxdiff model of best-worst choice are stated and proved in Flynn and Marley (submitted), Marley and Islam (2012) and Marley and Pihlens (2012).

Discussion

The MNL-inspired LBA model variants we proposed are all capable of fitting the dermatology and mobile phone data sets at least as well as the standard MNL choice models. Although the three models make different assumptions about the cognitive processes underlying best-worst choices all fit the data equally well, making them difficult to distinguish on the basis of choices alone. Response time data have the potential to tease the models apart – for example, the ranking LBA model makes the strong prediction that "best" responses will always be faster than "worst" responses. Testing such predictions against data can better inform investigations into the cognitive processes underlying preferences, paralleling similar developments in the understanding of single-attribute perceptual decisions (e.g., see Ratcliff & Smith, 2004) and best-only decisions about multi-attribute stimuli (Otter et al., 2008; Ruan et al., 2008). This illustrates the potential benefits that arise from using cognitive process-based models (such as accumulator models) for both choice and response time.

In the next section we demonstrate that a best-worst scaling task that incorporates response time measurement can aid discrimination between the LBA variants we have proposed. We show that the three LBA variants introduced above, derived from analogous MNL models, are inconsistent with the response time data from a perceptual judgment task. We propose a modification to the sequential LBA model that naturally accounts for the response latency data, demonstrating that response times provide added benefit to best-worst scaling.

Response Times in Best-Worst Scaling

The lack of latency data is commonplace in the applied discrete choice literature. Traditionally, it has not been thought necessary to record response latency as it has not been demonstrated how such measurement might usefully be included in analyses (though for recent progress see, e.g., Dellaert, Donkers, & van Soest, 2012; Haaijer, Kamakura, & Wedel, 2000; Otter et al., 2008; Ruan et al., 2008). Race models provide a natural account of decision times in best-worst scaling, so we examined the response time predictions of the previously proposed LBA models against data from a new experiment.

The ranking, sequential and enumerated race models make unique predictions about the pattern of predicted response times. For example, the ranking and sequential models predict that best responses always occur prior to worst responses. Alternatively, the two models could be instantiated in a worst-to-best fashion, in which case they predict worst responses always occur before best responses. However, compared to the ranking model, the sequential model predicts a faster distribution of worst responses – because in the sequential model, an entirely new race must be run after the "best" response is issued. If the data

elicited by each attribute level.

exhibit a mixture of response ordering – sometimes the "best" before "worst", and vice versa – then we have evidence against the assumptions of these two models, at least for their strict interpretation. We discuss below implications of response order patterns in data, and the possible inclusion of a mixture process that permits variability in the order of the races (i.e., the ranking model might sometimes occur in a worst-to-best manner and sometimes in a best-to-worst manner, or the sequential model might occur in a worst-then-best order and sometimes in a best-then-worst order). The enumerated model also makes strong predictions about response times: best and worst choices should differ only by an offset time due to motor processes, since the single enumerated race provides both the best and worst responses. Consequently, the enumerated model predicts that experimental manipulations, such as choice difficulty or choice set size, should not influence the time interval between the best and worst responses.

As a first target for investigating models of response times in best-worst scaling experiments we chose to use a simple perceptual judgment task rather than a traditional consumer choice task. Perceptual choices permit the collection of many more trials than complex consumer judgments. This enabled us to collect a large number of trials per participant, providing data that could more easily support a finer-grained analysis of response time distributions and fitting models to individual participant data. In addition, perceptual tasks permit precise stimulus control that allow testing of, for example, enumerated model predictions such as the absence of choice difficulty effects on inter-response times. We leave to future research the investigation of response time data from more typical multi-attribute discrete choice applications, such as the dermatology and mobile phone examples examined in the first section.

Experiment

We used a modified version of Trueblood, Brown, Heathcote, and Busemeyer's (in press) area judgment task. At each trial participants were presented with four rectangles of different sizes, and they were asked to select the rectangle with the largest area (i.e., an analogue to a "best" choice) and the smallest area (i.e., an analogue to a "worst" choice).

Participants

Twenty-six first year psychology students from the University of Newcastle participated in the experiment online in exchange for course credit.

Materials and Methods

The perceptual stimuli were adapted from Trueblood et al. (in press), where participants were asked to judge the area of black shaded rectangles presented on a computer display. We factorially crossed three widths with three heights to generate nine unique rectangles, with widths, heights and areas given in Table 3. The area of the rectangles at the extreme ends of the stimulus set were easily differentiable (e.g., 6050 from 8911), but those in the middle of the stimulus set were much more difficult (e.g., 8107 from 8113).

On each trial, four rectangles were randomly sampled, without replacement, from the set of nine rectangles. The stimuli were presented in a horizontal row in the center of the

screen, as shown in Figure 6. All rectangles were subject to a random vertical offset between ±25 pixels, to prevent the use of alignment cues to judge height.

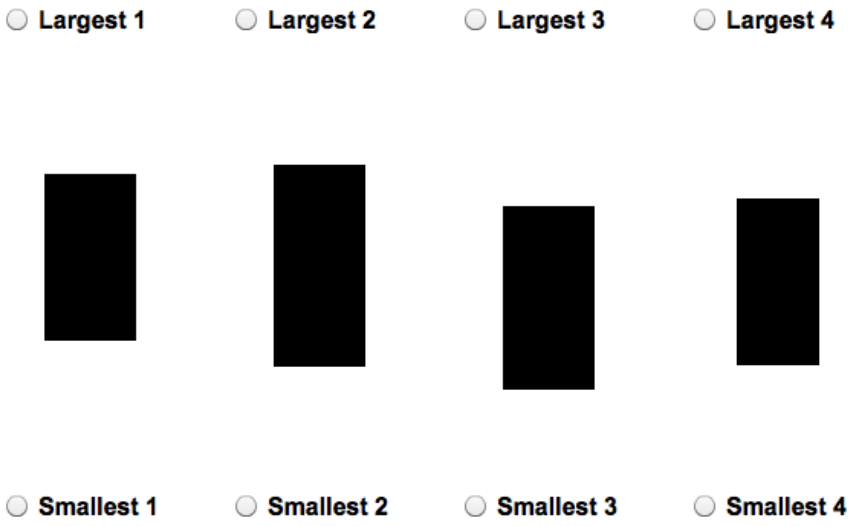

Figure 6. Illustrative example of a trial in the area judgment task. Note that if a participant selected, say, "Largest 1" as the rectangle with the largest area, then the option "Smallest 1" was made unavailable for selection.

Each participant chose the rectangle judged to have the largest area, and a different one with the smallest area. All responses were recorded with a mouse click and could be provided in either order: largest-then-smallest, or smallest-then-largest. We restricted participants from providing the same rectangle as both the largest and smallest option, by removing the option selected first as a possibility for the second response. On each trial we recorded the rectangle chosen as largest and the latency to make the choice, and the rectangle chosen as smallest and the time to make that choice. Participants completed 600 trials each, across three blocks.

Results

We excluded trials with outlying responses that were unusually fast or slow, defined as faster than .5 seconds or slower than 25 seconds. We also excluded two participants who each had more than 10% of their trials marked as outliers. Of the remaining participants' data, outliers represented only 0.9% of total trials.

We first report the proportion of correct classifications – correct selection of the largest (resp., smallest) rectangle in the stimulus display. We follow this analysis by considering the effect of response order – whether participants responded in a largest-then-smallest, or smallest-then-largest, manner – on both choice proportion and response latency data.

Correct Classifications

Our first step in analysis was to determine whether the area judgment manipulation had a reliable effect on performance. The left and middle panels of Figure 7 display the

proportion of correct responses for largest (resp., smallest) judgments as a function of choice difficulty. We operationalized difficulty as the difference in area between the largest (resp., smallest) and second largest (resp., second smallest) rectangle presented at each trial, which we refer to as the max-vs-next (resp., min-vs-next) difference. A small max-vs-next (resp., min-vs-next) difference in area makes it difficult to resolve which is the largest (resp., smallest) rectangle.

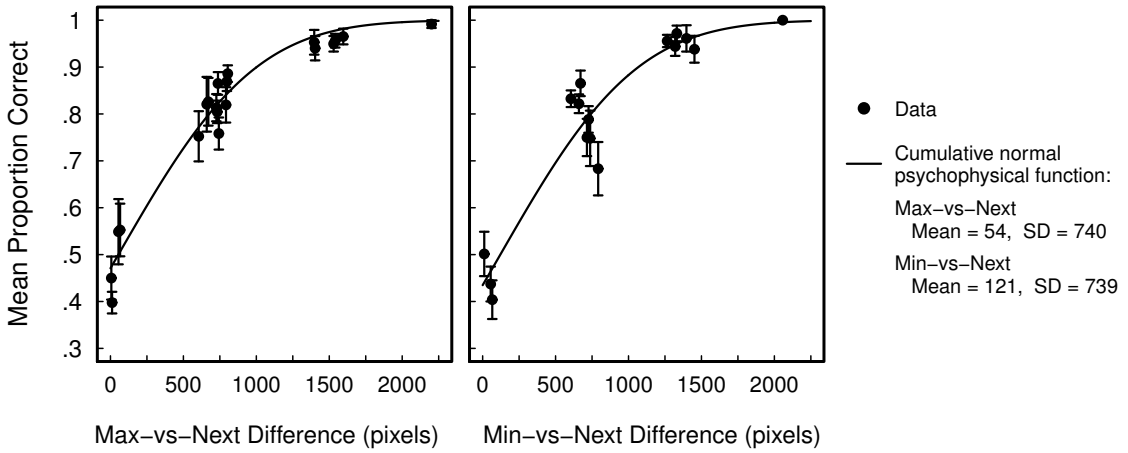

Figure 7. The left panel shows the proportion of times that the rectangle chosen as best was the largest rectangle (i.e., correct choice) as a function of the difference in area between the largest and second largest rectangles in the display (in pixels). The middle panel shows the proportion of times that the rectangle chosen as worst was the smallest rectangle as a function of the difference in area between the smallest and second smallest rectangles in the display. The overlaid lines represents the best fitting cumulative normal psychophysical functions, fit separately to both panels, with legend shown in the right panel.

Performance was well above chance even for the smallest max-vs-next and min-vs-next differences (6 and 11 pixels, respectively), yielding 45% and 50% correct selections of the largest and smallest rectangles in the display, respectively (chance performance is 25%). As expected, when the max-vs-next and min-vs-next difference increased, so to did the proportion of correct responses. We have separately overlaid on the max-vs-next and min-vs-next difference scores the best-fitting cumulative normal psychophysical functions. The good fit is consistent with the notion that participants' decisions were sensitive to a noisy internal representation of area.

Response Order Effects

To connect the following model and data with earlier material, we refer to *best* (resp., *worst*) rather than largest (resp., smallest). There was considerable variability across participants in the proportion of best-before-worst versus worst-before-best responses (Figure 8), ranging from almost completely worst-first to almost completely best-first. Over participants, the majority of first responses were for the best option ($M = .73$), which was significantly different from chance according to one sample t-test (i.e., test against $\mu = .5$),

$t(23) = 3.18$, $p = .004$. These patterns suggest that models which strictly impose a single response ordering, such as the ranking and sequential models, require modification.

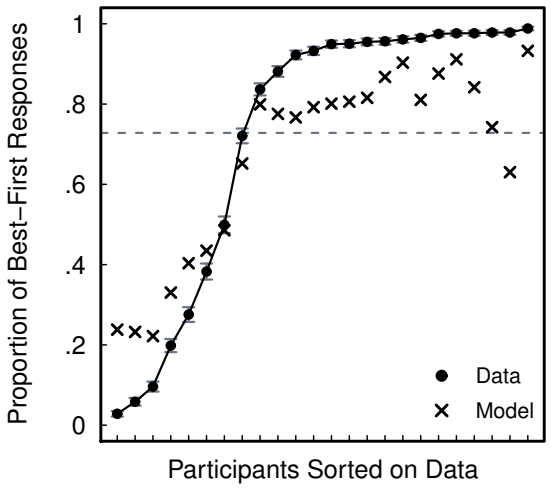

Figure 8. Proportion of trials on which the best response was made before the worst response, shown separately for each participant. Circular symbols and crosses represent data and predictions of the simultaneous race model, respectively. Error bars represent the standard error of a binomial proportion, according to the formula: $se(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n}$, where $\hat{p}$ is the proportion of best-first choices in data and n is the number of trials. The dashed horizontal line represents the mean proportion of best-first responses across participants.

We next examined whether choice difficulty influenced response order. We defined the difficulty of best and worst choices respectively using the max-vs-next and min-vs-next criteria described above, but we collapsed these into three exhaustive difficulty categories: hard (less than 250 pixels), medium (500–1000 pixels), and easy (greater than 1250 pixels; no difference scores fell between 250–500 or 1000-1250 pixels). The mean proportion of best-first responses reliably decreased as the discrimination of the largest rectangle became more difficult, $F(1.4, 31.9) = 4.07$, $p = .04$, using Greenhouse-Geisser adjusted degrees of freedom (all subsequent ANOVAs also report Greenhouse-Geisser adjusted degrees of freedom). This effect suggests that when there is an easy-to-see largest rectangle, that response was likely to be made first: easy $M = .739$ (within-subjects standard error = .006), medium $M = .726$ (.003) and hard $M = .719$ (.006). The analogous result was observed for worst-first choices, but larger in effect size – an easy-to-see smallest rectangle made a worst-first response more likely, $F(1.6, 35.7) = 10.03$, $p < .001$; easy $M = .289$ (.006), medium $M = .275$ (.004) and hard $M = .256$ (.006).

We also examined the effect of choice difficulty on the time interval between the first and second responses – the inter-response time. For each participant we calculated the average inter-response time for best-first trials as a function of the difficulty of the worst (second) response (easy, medium, difficult), and for worst-first trials as a function of the difficulty of the best (second) response. As the second judgment became more difficult, the latency between the first and second responses increased, approximately half a second

across the three difficulty levels, $F(1.2, 24.8) = 10.81$, $p = .002$;[7] easy $M = 1.51s$ (.12), medium $M = 1.75s$ (.08) and hard $M = 1.95s$ (.109).

Discussion

Data from the best-worst perceptual choice experiment exhibited three effects relevant to testing the models: large differences between participants in the preference for best-then-worst versus worst-then-best responding; changes in the proportion of best-first and worst-first responding as a function of choice difficulty; and changes in inter-response times due to choice difficulty. These effects are inconsistent with all three MNL-derived race models in their original forms. Firstly, the three models cannot accommodate within-participant, across-trial variability in best-first or worst-first responding. They might be able to account for the response order effects through the addition of a mixture process. On a certain proportion of trials, defined by a new parameter, the ranking race or the sequential races could be run in reverse order, or the enumerated model could execute its responses in opposite orders (Marley & Louviere, 2005, considered such mixture models for best-then-worst and worst-then-best choice).

The mixture approach adds a layer of complexity to the model – an extra component outside the choice process itself – which is unsatisfying. Putting that objection aside, these changes would not be able to account for the effect of choice difficulty on the proportion of best- and worst-first responses, or on the inter-response times, because the mixture process is independent of the choice process. For these reasons we do not explore the fit to response time data of the ranking, sequential or enumerated models when augmented with a mixture process. Instead we propose a modification to the sequential model, preserving the idea that there are separate best-choice and worst-choice races, but assuming that they occur simultaneously rather than sequentially.

The Simultaneous Model

The simultaneous model makes predictions consistent with the choice and response time patterns observed in data. Where the sequential model assumes *consecutive* best, then worst, races, the simultaneous model assumes *concurrent* best and worst races. The best option is associated with the first accumulator to reach threshold in the best race, and the worst option is associated with the first accumulator to reach threshold in the worst race. We present, and test, the simplest version of this model which allows the same option to be selected as both best and worst, which was not allowed in our experiment. The model also allows for vanishingly small inter-response times, which are not physically possible. These predictions affect a sufficiently small proportion of decisions for our rectangle data that we neglect them here, for mathematical convenience.

The probability of a choice of the option x as best at time t and option y as worst at time r, where no constraint exists between t and r, is given as the product of the individual likelihoods of the best and worst races,

$$bw_X(x, t; y, r) = b_x(t) \prod_{z \in X - \{x\}} (1 - B_z(t)) \cdot w_y(r) \prod_{z \in X - \{y\}} (1 - W_z(r)).$$

[7] Data from three participants were removed from this analysis due to incomplete data in all cells.

To calculate marginal probability $BW_X(x, y)$ from the simultaneous race model in the absence of response time data, the individual likelihoods of the best and worst races are integrated over all times $t > 0$ and $r > 0$, respectively. When fit in this manner to choices, only, the simultaneous model provides an account of the dermatology and mobile phone data sets equal in quality to the three LBA models described in the first section (dermatology – correspondence between MNL model regression coefficients and log estimated LBA drift rates, $R^2 = .97$, close agreement between observed and predicted choice proportions, $R^2 = .92$, and 5% and root-mean-squared prediction error; mobile phones – $R^2 = .99$, $R^2 = .60$, and 3.7%, respectively).

The simultaneous model overcomes the drawbacks of the three previous models by accounting for all general choice and response time trends observed in data. For instance, the model is able to capture inter- and intra-individual differences in response style – those participants that prefer to respond first with the best option, or first with the worst option – by allowing separate threshold parameters for the best and worst races. For example, a participant who primarily responds first with the best option will be described by a lower response threshold in the best race than in the worst race. This means that, on average, an accumulator in the best race reaches threshold prior to an accumulator in the worst race.

The simultaneous race model also accounts for the effect of choice difficulty on best- and worst-first responses, via differences in drift rates across rectangles. Very easy discrimination of the largest rectangle tends to occur when there is a large area for one rectangle, with a correspondingly large drift rate and so a fast response and a largest-before-smallest response order. Similarly, the simultaneous model predicts that difficult judgments rise to threshold more slowly than easy judgments. Therefore, irrespective of whether the best or worst race finishes first, the slower of the two races will still exhibit an effect of choice difficulty on latency. Since the best and worst races are (formally) independent, by extension the difficulty of the slower (second) judgment will also affect the inter-response time.

Estimating Simultaneous Model Parameters from Perceptual Data

We fit the simultaneous model to individual participant data from the best-worst area judgment task. Our methods were similar to those used previously with the multi-attribute data. We estimated nine drift rate parameters, one for each rectangle stimulus. We fit the model twice, once where we ignored response time data (as before) and once where we used those data. When ignoring response times, we arbitrarily fixed $A = 0$, $b = 1$, $s = 1$ and $t_0 = 0$ for the best and worst races. When fitting the model to response times, we estimated a single value of the start-point range, A, and non-decision time, t_0, parameters, with separate response thresholds for the best and worst races, b_{best} and b_{worst} (we again fixed $s = 1$, which serves the purpose of fixing a scale for the evidence accumulation processes). Therefore, when the simultaneous model was fit to response time data it required four additional free parameters compared to fits to choice-only data. Regardless of the data type – choice-only, or choices and response times – the approach to parameter optimization was the same: for each participant and trial we calculated the log-likelihood given model parameters, summed across trials, and maximized. We provide code to fit the simultaneous model to a single participant's data to choices and response times in the freely available R language (R Development Core Team, 2012) in the "publications" section of the authors' website at `http://www.newcl.org/`.

Model Fits

Although we fit the model to data from individual participants, for ease of exposition we primarily report the fit of the models at the aggregate level (i.e., results summed over participants). Unlike the previous fits, where we compared drift rate estimates to the corresponding random utility model regression coefficients, here we compare drift rate estimates to the area of the rectangles to demonstrate that the model recovers sensible parameter values. As above, we first assess goodness of fit by comparing observed and predicted choice proportions. For each participant we calculated the number of times that each rectangle area was chosen as best (resp., worst), and then normalized by the number of trials on which each rectangle area was presented. For the fits to response time data, we also examine goodness of fit by assessing the observed and predicted distribution of best and worst response times at the aggregated level. To demonstrate that the model captures individual differences, we also present examples of model fits to individual participant response time distributions. Finally, we compared predictions of the model to the response order data (choice proportions and response times).

Choice Proportions

The upper panels of Figure 9 plot mean estimated drift rate from the simultaneous model against rectangle area. Whether based on fits to choices-only or to choices and response times, the estimates followed a plausible pattern – mean drift rate increased as a sigmoidal function of rectangle area, which is the standard pattern in psychophysical judgments (e.g., Ratcliff & Rouder, 1998). There was a very strong effect of rectangle area on mean estimated log drift rate, for the fits to choice-only, $F(2.1, 49.3) = 161$, $p < .001$, and response time, $F(2.1, 48.8) = 143$, $p < .001$, fits.

The middle and lower panels of Figure 9 show the goodness of fit of the simultaneous model to choice data, separately for both methods of fitting the model. Choice-only fits provided an excellent account of the best and worst choice proportions – both R^2's $> .98$ and root-mean-squared difference between observed and prediction choice proportions of 1.7% and 4%, respectively. When the model was forced to accommodate response times as well as response choices, it still provided a good fit to choice proportion data: R^2's $> .95$ and root-mean-squared prediction error of 7.1% and 7.5% for best and worst proportions, respectively. The slight reduction in goodness of fit is expected, since the latter model fits are required to account for an additional aspect in the data (response times) – predictions of response times and response choices are not independent, and the data contain measurement noise.

Response Times

When fit to response times, the simultaneous model provides a good account of best and worst response time distributions. We first consider group level data where we used a quantile averaging approach to analyse aggregate response time distributions. Quantile averaging conserves distribution shape, under the assumption that individual participant distributions differ only by a linear transformation (Gilchrist, 2000; Figure 11 suggests this was generally the case in our data). For each participant and separately for best and

worst responses we calculated the $1^{st}, 5^{th}, 10^{th}, 15^{th}, \ldots, 90^{th}, 95^{th}, 99^{th}$ percentiles of response time distributions, and then averaged the individual participant percentiles to form an aggregate distribution of response times. Finally, we converted the averaged percentiles to histogram-like distributions, shown in the upper panel of Figure 10. The averaged data demonstrate the stereotypical properties of latency distributions: sharp onset of the leading edge closely followed by a single peak and a slow decline to form a long, positively skewed tail. In the aggregate distributions, best responses were faster than worst responses, as expected from the proportion of best-first responses across participants (Figure 8). Importantly, the simultaneous race model provides a good account of the best and worst response time distributions, capturing each of the characteristic distribution trends just described.

We next consider the fit of the simultaneous model to individual participant latency data. The lower half of Figure 10 shows data from nine individual participants, and demonstrates that the simultaneous race model provides a very good fit to individuals, whether they prefer best-first responding, worst-first responding, or a mixture (see Figure 11 in *Appendix B* for model fits to the response time distributions of all 24 participants).

Response Order Effects

Figure 8 shows the proportion of best-first responses for each participant. Overlaid on those data are model predictions of the expected proportion of best-first responses for each participant. The model captures the qualitative trends in the pattern of best-first preference data across participants, with a smooth shift from predominantly worst-first participants through to predominantly best-first participants. However, the model does not capture the strength with which some participants prefer a best- or worst-first pattern of responding.

The simultaneous model can also capture the effect of the difficulty of the area judgment on the proportion of best- and worst-first responses and the mean latency between the first and second responses. We calculated separately for each participant the proportion of best- and worst-first responses as a function of choice difficulty in both data and model predictions, and examined the relationship by aggregating across these proportions. Again, the model provided a good account of the response proportion data across difficulty levels with R^2's of .87 and .9 for the best- and worst-first responses, respectively. Similarly, the latency between the first and second responses was also influenced by the difficulty of the area judgment of the *second* choice – as the difficulty of the second judgment increased, so too did the time inter-response latency. The simultaneous model predicts this qualitative trend in data: increased inter-response time with increased difficulty, with a predicted increase of approximately .4 seconds for each increase in difficulty level.

General Discussion

Accurately eliciting preferences is important in a wide variety of applied fields from public policy to marketing, as exemplified by the dermatology and mobile phone data that we examined. Discrete choice, and in particular best-worst scaling, provide more robust and efficient measures of preference than alternative approaches, particularly when combined with random utility analyses. One drawback of random utility models of choices, only, is that they have limited interpretability as mechanistic accounts of the cognitions underlying decision making. On the other hand, accumulator (or race) models, which involve both the choices made and the time to make them, have proven useful in illuminating

the cognitive and neurophysiological processes underpinning simple single-attribute decisions. Over the last several decades there have been increasingly successful and practical applications of accumulator models to multi-attribute preference data, including decision field theory (Busemeyer & Townsend, 1992, 1993; Roe et al., 2001), the leaky competing accumulator model (Usher & McClelland, 2004), the Poisson race model (Otter et al., 2008; Ruan et al., 2008), and most recently the $2N$-ary choice model (Wollschläger & Diederich, 2012). Our proposal builds on these developments and applies them to a more complex choice task, best-worst scaling, with the aim of producing a model that has both tractable statistical properties and a plausible cognitive interpretation. This is an example of "cognitive psychometrics", which aims to combine psychological insights from process-based modeling with the statistical advantages of measurement approaches (see, e.g., Batchelder, 2009; van der Maas, Molenaar, Maris, Kievit, & Borsboom, 2011).

In the first section of this paper we demonstrated how a simplified accumulator model (LBA: Brown & Heathcote, 2008) can make race models practical for the analysis of the complex, multi-attribute best-worst decisions typically required in many applications. The three MNL-inspired LBA variants we examined were all capable of fitting choice data at least as well as the best random utility models of choice. The parameter estimates from the race models were closely related to the parameter estimates from the random utility models, providing further confidence in the use of the race models to describe data.

In the second section of this paper we demonstrated the benefit that response times add to understanding the cognitive processes underpinning best-worst choice. In the context of a best-worst response procedure implemented in a simple perceptual experiment, we provided evidence against the assumptions of the three LBA models proposed in the first section, and by extension provided evidence against some assumptions of the MNL models from which they were derived. We then modified one of the LBA models to develop a new simultaneous best race and worst race model. The simultaneous model provided a good account of data at the individual participant and aggregate levels, for both choices and response times, including a range of newly identified phenomena relating to the effect of decision difficulty on the order and speed of best and worst responses.

Further work remains to determine whether the simultaneous model can account for the longer scale response times produced by complex multi-alternative choices, and whether a single model can accommodate best-only and best-worst choices with these types of stimuli. For both simple and complex choices there are challenges remaining related to the fine-grained measurement and modeling of the time to make both best and worst choices. Methodologically it would be desirable to use a faster response method than moving a pointer with a mouse and clicking a target to minimize the motor component of inter-response time. Recent approaches using eye movements appear promising in this regard (Franco-Watkins & Johnson, 2011a, 2011b). A complimentary approach would be to elaborate the simultaneous model to accommodate the time course of motor processes and to address paradigms like ours that by design do not allow the same best and worst response.

The main advantage of the LBA approach over earlier process models is its mathematical tractability. However, a disadvantage when compared with more complete models, such as decision field theory and the leaky competing accumulator model, is that the LBA models we have developed do not describe various context effects that occur in preferential choice. This happens because the LBA models belong to the class of "horse race" ran-

dom utility models (Marley & Colonius, 1992), which are known to fail at explaining many context effects (see Rieskamp, Busemeyer, & Mellers, 2006). However, Trueblood, Brown, Heathcote, and Busemeyer (in preparation) have recently extended the LBA approach to include contexts effects in a way that retains its computational advantages.

We conclude by proposing that – even though further development is desirable – the simultaneous race model as it stands is a viable candidate to replace traditional random utility analysis of data obtained by best-worst scaling. The simultaneous model provides an account of choice data equivalent to the MNL choice models, but in addition it accounts for various patterns in response time data and provides a plausible explanation of the latent decision processes involved in best-worst choice.

Acknowledgments

This research has been supported by Natural Science and Engineering Research Council Discovery Grant 8124-98 to the University of Victoria for Marley. The funding source had no role in study design, or in the collection, analysis and interpretation of data. The work was carried out, in part, whilst Marley was a Distinguished Professor in the Centre for the Study of Choice, University of Technology, Sydney.

Appendix A: The Maxdiff Model for Best-Worst Choice

Using the generic notation of Figure 2, $BW_X(x, y)$ is the probability of choosing x as best and $y \neq x$ as worst when the available set of options is X. The *maxdiff model for best-worst choice* assumes that the utility of a choice option in the selection of a best option (u) is the negative of the utility of that option in the selection of a worst option ($-u$) and that

$$BW_X(x, y) = \frac{e^{[u(x)-u(y)]}}{\sum_{\substack{\{p,q\}\in X \\ p\neq q}} e^{[u(p)-u(q)]}} \qquad (x \neq y). \tag{1}$$

For the Flynn et al. (2008) data and analyses, X is a set of attribute levels, and so x and y are the attribute levels selected as best and worst, respectively. Marley et al. (2008) present mathematical conditions under which all the attribute levels are measured on a common difference scale; in this case, the utility of one attribute level can be set to zero.

For the Marley and Pihlens (2012) best-worst data and analyses, X is a set of multi-attribute profiles, and so x and y are the profiles selected as best and worst, respectively. Marley and Pihlens also assume that each profile has an additive representation over the attribute levels; that is, assuming that each profile has m attributes, then there are utility scales u_i, $i = 1, ...m$, such that if $z = (z_1, ...z_m)$, with z_i the attribute level for z on attribute i, then

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

Marley and Pihlens (2012) present a set of mathematical conditions under which the utility of a profile is such a sum of the utilities of its attribute levels, and each attribute is measured on a separate difference scale; in this case, one level on each attribute can have its utility set to zero.

Appendix B: Simultaneous LBA Model Fits to Individual Participant Response Time Distributions

In this Appendix we show the fit of the simultaneous race model to response time distributions at the level of individual participant data, for all 24 participants. Each panel of Figure 11 shows the fit of the race model to a separate participant. As described in the main text, there are clear individual differences in the pattern of responding – best-first, worst-first, or no preference for either best- or worst-first – as well as the time course of decisions – some participants made much faster responses than others. For both patterns of individual differences the simultaneous model provides a good account of the latency distributions from the majority of participants.

References

Batchelder, W. H. (2009). Cognitive psychometrics: Using multinomial processing tree models as measurement tools. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model based measurement.* American Psychological Association Books.

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

Brown, S., Marley, A. A. J., Donkin, C., & Heathcote, A. (2008). An integrated model of choices and response times in absolute identification. *Psychological Review*, *115*, 396–425.

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

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic–cognitive approach to decision making. *Psychological Review*, *100*, 432–459.

Coast, J., Salisbury, C., de Berker, D., Noble, A., Horrocks, S., Peters, T. J., & Flynn, T. N. (2006). Preferences for aspects of a dermatology consultation. *British Journal of Dermatology*, *155*, 387–392.

Collins, A. T., & Rose, J. M. (2011). Estimation of stochastic scale with best–worst data. *Manuscript, University of Sydney.*

Dellaert, B. G. C., Donkers, B., & van Soest, A. (2012). Complexity effects in choice experiment-based models. *Journal of Marketing Research*, *49*, 424–434.

Diederich, A., & Busemeyer, J. R. (2003). Simple matrix methods for analyzing diffusion models of choice probability, choice response time, and simple response time. *Journal of Mathematical Psychology*, *47*, 304–322.

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.

Flynn, T. N., Louviere, J. J., Peters, T. J., & Coast, J. (2007). Best–worst scaling: What it can do for health care research and how to do it. *Journal of Health Economics*, *26*, 171–189.

Flynn, T. N., Louviere, J. J., Peters, T. J., & Coast, J. (2008). Estimating preferences for a dermatology consultation using best–worst scaling: Comparison of various methods of analysis. *BMC Medical Research Methodology*, *8*(76).

Flynn, T. N., Louviere, J. J., Peters, T. J., & Coast, J. (2010). Using discrete choice experiments to understand preferences for quality of life. *Social Science & Medicine*, *70*, 1957–1965.

Flynn, T. N., & Marley, A. A. J. (submitted). Best-worst scaling: Theory and methods. In S. Hess & A. Daly (Eds.), *Handbook of Choice Modelling.*

Forstmann, B. U., Dutilh, G., Brown, S., Neumann, J., von Cramon, D. Y., Ridderinkhof, K. R., & Wagenmakers, E.-J. (2008). The striatum facilitates decision–making under time pressure. *Proceedings of the National Academy of Science*, *105*, 17538–17542.

Franco-Watkins, A. M., & Johnson, J. G. (2011a). Applying the decision moving window to risky choice: Comparison of eye–tracking and mouse–tracing methods. *Judgment and Decision Making*, *6*, 740–749.

Franco-Watkins, A. M., & Johnson, J. G. (2011b). Decision moving window: Using interactive eye tracking to examine decision processes. *Behavior Research Methods*, *43*, 853–863.

Frank, M. J., Scheres, A., & Sherman, S. J. (2007). Understanding decision making deficits in neurological conditions: Insights from models of natural action selection. *Philosophical Transactions of the Royal Society, Series B*, *362*, 1641–1654.

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

Goodman, S. N. (2009). An international comparison of retail wine consumer choice. *International Journal of Wine Business Research*, *21*, 41–49.

Haaijer, R., Kamakura, W., & Wedel, M. (2000). Response latencies in the analysis of conjoint choice experiments. *Journal of Marketing Research*, *37*, 376–382.

Heathcote, A., & Love, J. (2012). Linear deterministic accumulator models of simple choice. *Frontiers in Psychology*, *3*, n/a. doi: 10.3389/fpsyg.2012.00292.

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

Lee, J. A., Soutar, G. N., & Louviere, J. J. (2008). The best–worst scaling approach: An alternative to Schwartz's values survey. *Journal of Personality Assessment*, *90*, 335–347.

Louviere, J. J., & Flynn, T. N. (2010). Using best–worst scaling choice experiments to measure public perceptions and preferences for healthcare reform in Australia. *The Patient: Patient–Centered Outcomes Research*, *3*, 275–283.

Louviere, J. J., & Islam, T. (2008). A comparison of importance weights and willingness-to-pay measures derived from choice–based conjoint, constant sum scales and best–worst scaling. *Journal of Business Research*, *61*, 903–911.

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

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

Marley, A. A. J. (1989). A random utility family that includes many of the "classical" models and has closed form choice probabilities and choice reaction times. *British Journal of Mathematical and Statistical Psychology*, *42*, 13–36.

Marley, A. A. J., & Colonius, H. (1992). The "horse race" random utility model for choice probabilities and reaction times, and its competing risks interpretation. *Journal of Mathematical Psychology*, *36*, 1–20.

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., & Islam, T. (2012). Conceptual relations between expanded rank data and models of the unexpanded rank data. *Journal of Choice Modelling*, *5*, 38–80.

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.

Mueller, S., Lockshin, L., & Louviere, J. J. (2010). What you see may not be what you get: Asking consumers what matters may not reflect what they choose. *Marketing Letters*, *21*, 335–350.

Mueller Loose, S., & Lockshin, L. (2013). Testing the robustness of best worst scaling for cross-national segmentation with different numbers of choice sets. *Food Quality and Preference*, *27*, 230–242.

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.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). San Diego, CA: Academic Press.

R Development Core Team. (2012). *R: A language and environment for statistical computing.* Vienna, Austria. (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.

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.

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.

Ruan, S., MacEachern, S. N., Otter, T., & Dean, A. M. (2008). The dependent Poisson race model and modeling dependence in conjoint choice experiments. *Psychometrika*, *73*, 261–288.

Ryan, M., & Farrar, S. (2000). Using conjoint analysis to elicit preferences for health care. *British Medical Journal*, *320*, 1530–1533.

Szeinbach, S. L., Barnes, J. H., McGhan, W. F., Murawski, M. M., & Corey, R. (1999). Using conjoint analysis to evaluate health state preferences. *Drug Information Journal*, *33*, 849–858.

Trueblood, J. S., Brown, S. D., Heathcote, A., & Busemeyer, J. R. (in preparation). The multi-attribute linear ballistic accumulator model of context effects in multi-alternative choice.

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

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

van der Maas, H. L. J., Molenaar, D., Maris, G., Kievit, R. A., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. *Psychological Review*, *118*, 339–356.

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.

Wollschläger, L. M., & Diederich, A. (2012). The 2N-ary choice tree model for N-alternative preferential choice. *Frontiers in Psychology*, *3*, n/a. doi: 10.3389/fpsyg.2012.00189.

Table 2: The nine attributes and their combined 38 levels from Marley and Pihlens (2012). The values in parentheses indicate the coding used in Figures 4 and 5 below.

Attribute	Attribute levels
Phone style (s)	Clam or flip phone (1)
	Candy bar or straight phone (2)
	Slider phone (3)
	Swivel phone (4)
	Touch screen phone (5)
	PDA phone with a HALF QWERTY keyboard (6)
	PDA phone with a FULL QWERTY keyboard (7)
	PDA phone with touch screen input (8)
Brand (b)	A (1)
	B (2)
	C (3)
	D (4)
Price (p)	\$49.00 (1)
	\$129.00 (2)
	\$199.00 (3)
	\$249.00 (4)
Camera (c)	No camera (1)
	2 megapixel camera (2)
	3 megapixel camera (3)
	5 megapixel camera (4)
Wireless connectivity (w)	No bluetooth or WiFi connectivity (1)
	WiFi connectivity (2)
	Bluetooth connectivity (3)
	Bluetooth and WiFi connectivity (4)
Video capability (v)	No video recording (1)
	Video recording (up to 15 min; 2)
	Video recording (up to 1 h; 3)
	Video recording (more than 1 h; 4)
Internet capability (i)	Internet access (1)
	No internet access (2)
Music capability (m)	No music capability (1)
	MP3 music player only (2)
	FM radio only (3)
	MP3 music player and FM radio (4)
Handset memory (r)	64 MB built-in memory (1)
	512 MB built-in memory (2)
	2 GB built-in memory (3)
	4 GB built-in memory (4)

Table 3: The nine rectangular stimuli generated by factorially crossing three rectangle widths with three rectangle heights. All measurements are in pixels.

Width	Height	Area
55	110	6050
55	121	6655
55	133	7315
61	110	6710
61	121	7381
61	133	8113
67	110	7370
67	121	8107
67	133	8911

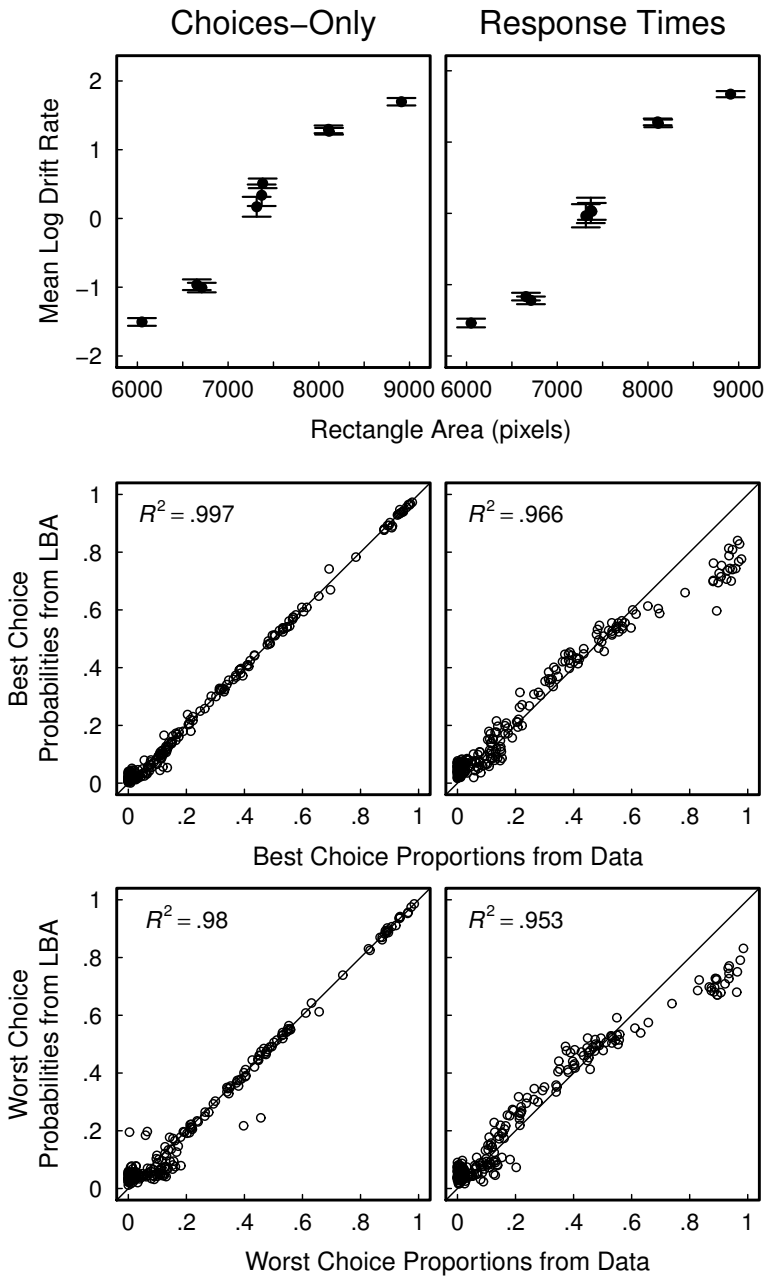

Figure 9. Estimated drift rates and goodness of fit to data for the simultaneous race model when fit to choices, only, and choices and response times (left and right columns, respectively). The upper panels show mean estimated log drift rates as a function of rectangle area. Error bars indicate within-subjects standard errors of the mean. The middle and lower panels show the goodness of fit to the experimental data for best and worst responses, respectively. The x-axes display choice proportions from data, the y-axes display predicted choice probabilities from the simultaneous race model, and the diagonal lines show a perfect fit. In the lower panels, each participant contributed 9 data points to each panel – one for each rectangle area.

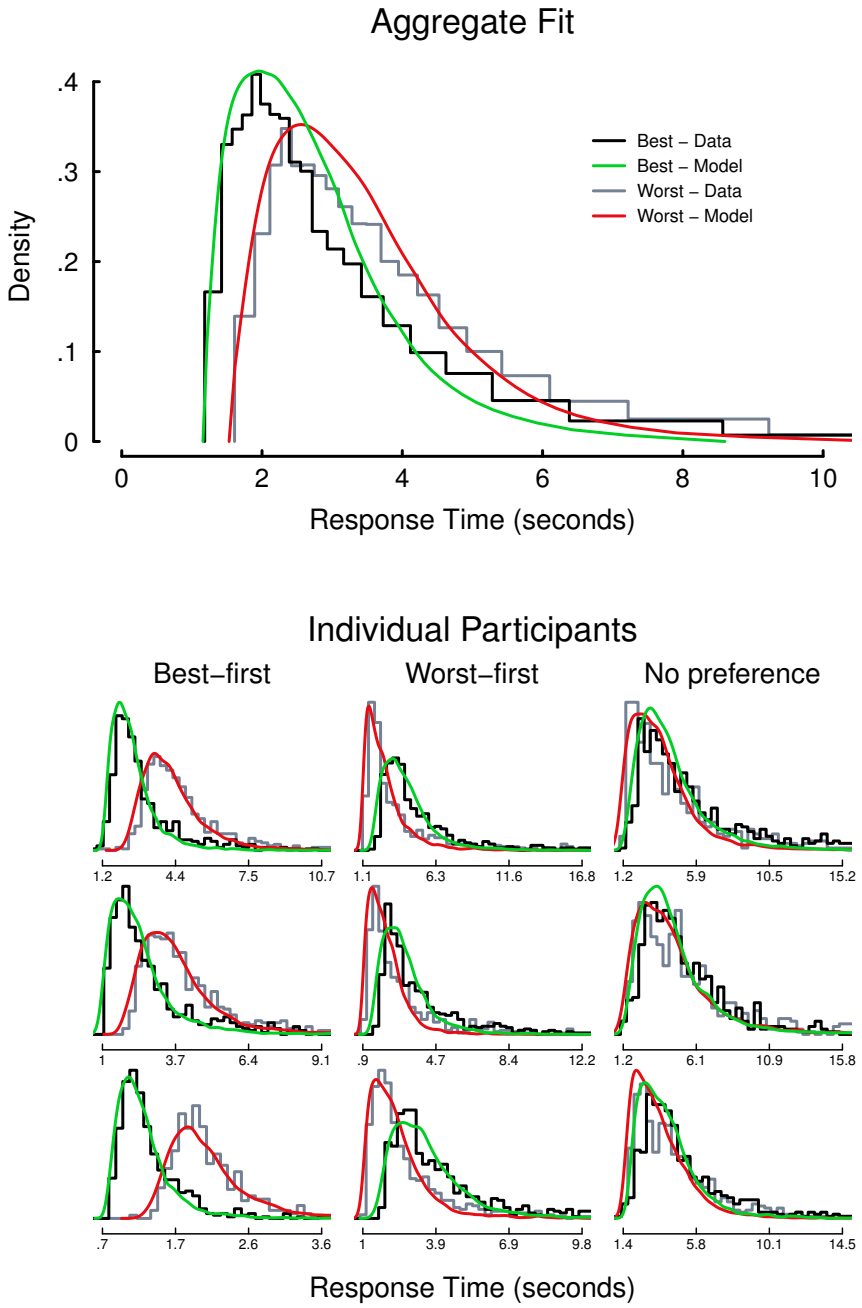

Figure 10. Response time distributions for experimental data and predictions from the simultaneous race model. The upper panel shows quantile averaged data as stepped histograms with best and worst responses shown in black and gray, respectively. The smooth density curves show model predictions with best and worst predictions shown in green and red, respectively, averaged in the same way as the data. The lower panels display model fits to a selection of individual participant data. We show participants broadly classified into three categories of responders: those who tended to respond with the best option first, the reverse who tended to respond with the worst option first, and those participants that demonstrated no strong preference for either best or worst first responding. For model fits to all 24 participants see Figure 11 in *Appendix B*.

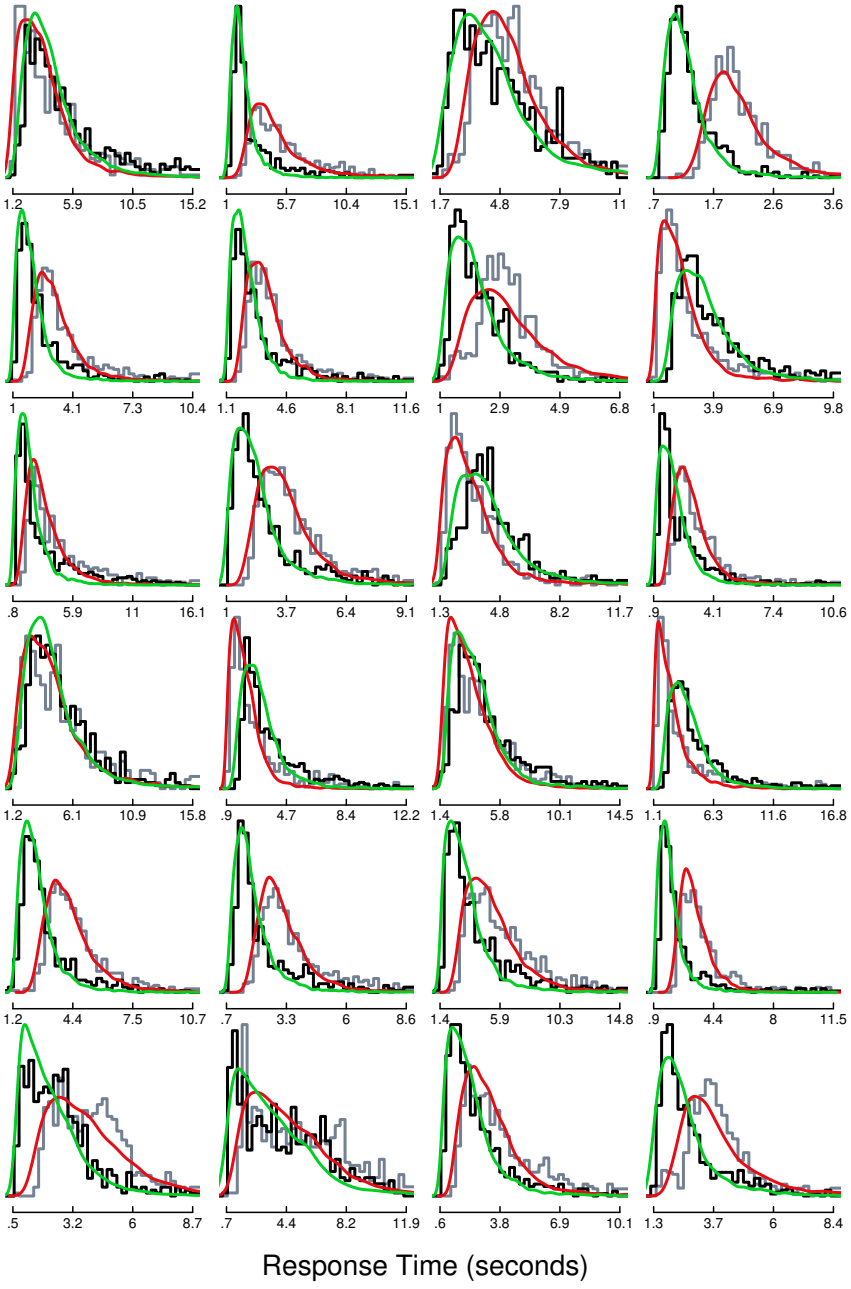

Figure 11. Response time distributions for experimental data and predictions from the simultaneous race model. Each panel shows a separate participant. Data are shown as stepped histograms with best responses in black and worst responses in gray. Model predictions are shown as smooth density curves with best predictions in green and worst in red.