Using best-worst scaling to improve psychological service delivery: An innovative tool for psychologists in organized care settings.

Leanne G. Jones, Guy E. Hawkins, & Scott D. Brown

School of Psychology

University of Newcastle, Australia

Correspondence should be addressed to:

Scott Brown

E: scott.brown@newcastle.edu.au

W: http://newcl.org/Brown

P: 0423 176 112

Abstract

With the growth of client-centred and patient-as-consumer approaches to care, understanding the preferences of psychologists' patients has never been more important. Traditional methods for measuring preference, such as Likert-type rating scales, suffer from well-known limitations including subjectivity and positive bias. "Best-worst scaling" (BWS) provides an opportunity to address some of these limitations. Despite the growing use of BWS to measure preference in other areas, BWS methods are not being used in the study of psychologists' patients. We demonstrate BWS data collection and analysis. With a sample of only 31 clients from two Australian psychology practices, we show the strength of preference for different aspects of psychologists' appointments can be measured accurately. Additionally, the inclusion of readily-available timing data from responses improved measurement sensitivity and statistical power.

Keywords: patient as consumer; measuring preference; best-worst scaling; decision-making theory.

Historically, patients as consumers of medical treatments, and psychologists' patients in particular, have had few chances to provide feedback about medical practice (Ryan & Farrar, 2000). Obtaining feedback from and consulting with patients can improve the provision of health care services, allowing better targeting of resources on a local level (e.g., individual clinics and practices) as well as a broader policy level (e.g., the types of treatments that are funded). With a growing trend to ask for feedback from patients, the use of valid and reliable tools to elicit this feedback is imperative.

Feedback has traditionally been gathered using self-report measures, such as open-ended questions, or ratings on Likert scales (e.g., rating one's agreement with a statement about health care, on a scale from "Strongly Agree" to "Strongly Disagree"). Self-report and Likert scale instruments have well-known problems with measurement accuracy and present serious difficulties in interpretation (see, e.g., Lee, Soutar & Louviere, 2008). For example, respondents often adopt a response set (over-use of a single response category), which can affect both the mean and the variance of the data (see, e.g., Baumgartner & Steenkamp, 2001). The biases often found in rating scales can be exacerbated by situational demands such as time constraints, and can be systematically influenced by irrelevant context effects such as item order or format (see Paulhus, 1991, for a review). Other well-known biases that afflict rating scales include: social desirability; a bias toward agreeing; and extreme response bias. While responses with strong positive bias can be exploited for advertising and testimonial purposes, this approach do not improve the lot of patients, or providers. For other examples of the problems arising from these biases, see: Lee et al. (2008); Schwartz and Bardi (2001); Cohen and Neira (2004, cited in Devinney, Auger & Eckhardt, 2010).

The problems associated with traditional methods for measuring preferences and opinions have led to the development of quantitative alternatives, most notably the discrete choice experiment (DCE). DCEs are questionnaires in which participants choose between

different scenarios. For example, rather than asking a patient how strongly they feel about the waiting time for an appointment with a psychologist, a DCE might ask the patient to make a hypothetical choice between two appointments, one with a short waiting time, but a high cost, and the other with a longer waiting time and lower cost. In this way, the respondent's feelings about one attribute (waiting time) can be quantitatively measured against their feelings about the other attribute (cost). At the same time, extreme positive bias is avoided by forcing respondents to choose one scenario over the other.

Traditionally, DCEs have been used to measure perceptual processes, in psychophysical studies (for dozens of examples, see Luce, 1986) and in the measurement of consumer preferences in economic decision-making. More recently, there has been an increase in the use of DCEs to explore healthcare issues, including patient and practitioner decision-making (Ryan et al., 2000). They take into account health outcomes (e.g., pain level) as well as process attributes (e.g., waiting times), whereas more traditional measures have focused on solely health outcomes (e.g., Quality Adjusted Life Years). An example of a question from a recent DCE in the health-care setting is given in Figure 1.

Which hospital would you prefer to use? (Please tick box A or B below)

	Hospital A	**Hospital B**
Travel time to hospital	1 hour	1 hour
Easy access to parking & public transport	Yes	No
Medicare Levy Extra on top of what you're currently paying (even if you're not paying anything at the moment)	$200 extra per year	$200 extra per year
Average waiting time on elective surgery	4 weeks	12 weeks
Average waiting time in casualty	2 hours	30 minutes
Complication rate from treatment	5% better than the	5% better than

	national average	the national average
	A *(Please tick one box)* B	

Figure 1. Example of a DCE question, adapted from Jan, Mooney, Ryan, Bruggemann and Alexander (2000).

A particularly useful development of DCEs is the best-worst scaling method (BWS). In a BWS procedure, respondents are presented with several hypothetical scenarios and asked to report both their most and least favoured. Although this is a seemingly small change over a traditional DCE (just asking for the *worst* in addition to the *best*), the BWS approach has several advantages (for review, see Louviere, 2008). For example, the BWS task works with, rather than against, the propensity to respond to extreme options, a factor which makes BWS tasks particularly easy for respondents. On the researcher's side, BWS tasks provide improved discrimination power and sensitivity compared to standard DCEs. A patient-friendly way of using the BWS method, reducing the cognitive load associated with comparing multiple hypothetical scenarios, is to present just a single scenario, with many different attributes, and ask the respondent for the best and worst *attribute* of that scenario.

We illustrate the use of BWS methods to elicit the preferences of psychologists' patients, and to illustrate these practices directly by means of a small example study. Using patients from psychology clinics in Australia, we explored the relative importance of four different attributes of appointments with psychologists: waiting time, convenience of the appointment, the thoroughness of the appointment, and the expertise of the practitioner. We chose these particular attributes of the clinic experience following Coast et al.'s (2006) study of dermatologists' patients. Since our experiment was primarily designed to showcase best-practice with BWS data collection and analysis methods, our choice of attributes is less important than it otherwise might be.

A second aim of our study was to extend the standard analysis of data from BWS methods by the inclusion of data about the time that respondents took to make their decisions. With computerised survey instruments now the norm, collection of timing data is both easy and cost-effective, and recent research has shown that response time data can provide important insights into the decision-making process (see, e.g., Ratcliff & Smith, 2004; Trueblood, Brown & Heathcote, 2014). We recorded the time taken to make each decision, and analysed these data using a standard decision-making account from cognitive science (Brown & Heathcote, 2008). To foreshadow the result, we observed that including response time data provided greater certainty compared with the analysis of decisions without timing data.

Method

Participants

Participants were drawn from psychology practices in the Hunter Valley region of Australia. The research team never approached participants. Rather, potential participants were informed of the study by their treating psychologists, or clinic staff. Treating psychologists initially mentioned the study to some of their patients by providing them with a URL that described the study in detail. The study was not mentioned to any patients who were perceived to be vulnerable or in crisis (defined as exhibiting symptoms meeting a diagnostic criteria in the DSM IV) or where the treating psychologist believed that being approached to participate or participating in the research may have a detrimental effect on the patient, or therapeutic relationship. Potential participants could anonymously visit the URL to read more information about the study, and to choose whether to participate. A total of 54 people chose to participate further in the study, but only 31 of these gave explicit permission at the end of the questionnaire for their data to be included in our study.

Measure/materials

The BWS task was administered as an online questionnaire that presented 32 different scenarios, one at a time, and asked participants to choose the best and worst attribute from each of four attributes within each scenario. Figure 2 illustrates a typical scenario from the BWS task. The times taken by respondents to choose the best and worst attributes of each scenario were recorded, and best/worst decisions could be made in either order (i.e., best before worst or vice versa). Each scenario was defined by values on four attributes: waiting time, expertise, thorough care and convenience. These attributes were modelled after those used in a study of dermatology patients (Coast et al., 2006). The waiting time attribute took one of four levels in each scenario, while the other attributes each took one of two levels (see left two columns of Table 1), making a total of 32 (4x2x2x2) possible appointment descriptions, each of which was shown once to each respondent, with order randomised across respondents.

Best thing	**The appointment with the Psychologist**	**Worst thing**
o	You will have to wait 2 months for your appointment	o
o	The psychologist has been treating psychological complaints part-time for 1 – 2 years	o
o	Getting to your appointment will be difficult and time-consuming	o
o	The appointment will be as thorough as you would like	o

Figure 2. One of the 32 scenarios from the experiment. Respondents were instructed to indicate the best and worst aspects of the scenario, by clicking buttons on the left and right respectively.

Completing the BWS questionnaire took approximately 15 minutes on average, including a short break halfway through. After this, respondents completed the Kessler Psychological Distress Scale (K10; Kessler et al., 2003). The K10 is a short (10 item) instrument that provides a global measure of distress, with a focus on anxiety and distress experienced in the previous four weeks. Finally, participants were asked to provide some non-identifying demographic information (age, gender, income and education) along with some optional questions about their treatment such as "what stage do you think your treatment is at? (beginning, in the middle, coming to an end)", and "how severe would you rate your symptoms? (mild, moderate, severe)".

Analysis

There are many ways to analyse data collected with discrete choice and best-worst scaling questionnaires. Standard methods have been extensively studied, and have well-validated psychometric properties (for reviews, see Marley & Louviere, 2005; Regenwetter, Grofman, Marley, & Tsetlin, 2006). For expository purposes, we illustrate the use of two different methods: best-worst scores, and accumulator modelling.

Best-minus-worst scores are simple and easily-calculated descriptive statistics, without an accompanying psychological theory. The best-worst score for a particular level of a particular attribute (say, "wait 2 months") is simply the number of times this attribute was chosen as the best attribute in a scenario, minus the number of times it was chosen as worst, normalized by the number of scenarios in which the attribute appeared. In this way, the best-worst score provides a number between -1 and +1, with larger positive values for generally liked attribute levels, and larger negative values for generally disliked attribute levels.

A more modern analysis approach considers the best-worst responses as the outcomes of a simple decision-making process. This process can be modelled using a standard

theoretical account of decision-making based on "evidence accumulation". Accumulator models have an extensive history in the analysis of decision-making data, and have been applied to decisions in a wide variety of domains, from simple perceptual decision-making, to decisions about memories, and consumer choices (for a review, see Johnson & Ratcliff, 2014). More recently, accumulator models have been extended to apply to BWS tasks, using the observed choice frequencies to estimate underlying "drift rates" corresponding to each level of each attribute (see Hawkins, Marley, Heathcote, Flynn, Louviere & Brown, in press). In accumulator models, the drift rates represent how fast a respondent accumulates evidence in favour of the attribute levels presented on a particular choice occasion. A response is triggered as soon as the evidence in favour of any attribute level exceeds a criterion amount. These drift rates are analogous – and closely related – to the utility or preference strengths estimated by traditional random utility or logit analyses. While it is more complex to carry out, the accumulator model approach has two distinct advantages over other approaches: it is based on a plausible and well-studied model of the cognitive process of decision-making; and it can provide a natural account of the time taken to make each decision, which other analysis methods cannot. A new contribution of the present paper is to examine the effect that including timing data has on the results of the data analysis.

Results

Sample

Demographic details of the sample of the 31 respondents who completed the questionnaire are shown in Table 1. Eleven individuals chose not to enter demographic details, but did complete the K10 questionnaire. The remaining participants were generally well-educated, aged between 35-55, with above-average personal incomes. The K10 scores, using the scoring cut-offs suggested by Andrews and Slade (2001) indicated that our

participants were divided roughly into thirds: 13 respondents were in the "well" range (scores below 20); 8 were in the "mild" range (scores between 20 and 24); and 10 were in the "moderate" or "severe" ranges (scores above 25). Overall, the demographic and K10 data suggest that our sample was not terribly unrepresentative of psychologists' patients in general. Respondents' scores on the K10 did not correlate significantly with the best-worst scores or the drift rates estimated below for any of the levels of the different attributes.

Table 1. Age, education level, and personal income distributions for the sample of patients.

Age	**N=31**
Not reported	11 (35.5%)
18- 24	4 (12.9%)
25- 34	3 (9.7%)
35- 44	6 (19.4%)
45- 54	5 (16%)
55- 64	2 (6.5%)
Education	
Not reported	11 (35.5%)
Did not finish high school	1 (3.2%)
High school	4 (12.9%)
Other/ After year 12	3 (9.7%)
Certificate	2 (6.5%)
Bachelor degree	8 (25.8%)
Postgraduate degree	2 (6.5%)
Income	
Unknown	11 (35.5%)
No response	2 (6.5%)
Less than 25K	3 (9.7%)
25K- 50K	4 (12.9%)
50K- 80K	3 (9.7%)
80K+	8 (25.8%)

Best-Worst Scores

The average best-worst scores are presented in Table 2, with high scores corresponding to attributes that were liked and the low scores corresponding to attributes that were disliked. The scores reflect some expected patterns (e.g., progressively lower scores are associated with progressively longer waiting times), which provides some evidence that our participants appropriately read and interpreted the questionnaire.

Best-worst scores also support quantitative comparison between levels within an attribute, and even across attributes – something that is not possible with traditional Likert-scale ratings. For example, the most preferred attribute level was to have a thorough appointment (0.560) followed closely by having an expert practitioner (0.554). The difference between preference strengths for these two was not a significant difference by paired samples *t*-test ($t(30) = .06$, $p = .95$, Cohen's $d = .022$ – note that we use a Type I error rate of $p=.001$ to control family-wise error rate in this section). The least preferred attributes were those that had patients waiting more than a month for their appointments (all scores less than -0.45). The results for the attribute of expertise are interesting, for although they show that individuals like to have a professional with extensive experience, they also show no great dislike for a practitioner with much less experience (though there was still a reliable difference between levels, $t(30) = 7.59$, $p < .001$, $d = 2.77$). Similarly, respondents were not greatly influenced by the convenience, or lack of convenience, of the appointment: just 0.466 separated the best-worst scores for those two levels ($t(30) = 6.56$, $p < .001$, $d = 2.40$), compared with, say, 1.226 separating the longest and shortest levels of waiting time (which had almost double the effect size: $t(30) = 11.44$, $p < .001$, $d = 4.18$).

Table 2.

Best-worst scores for each attribute level, with standard error (across participants) in parentheses.

Attribute	**Level**	**Average BWS Score**
Time waited	You will have to wait 3 months for your appointment	-.806 (.050)
	You will have to wait 2 months for your appointment	-.698 (.059)
	You will have to wait 1 month for your appointment	-.452 (.074)
	Your appointment will be this week	.415 (.073)
Expertise	The Psychologist has been treating clients with psychological issues, part-time for 1 – 2 years	.040 (.068)
	The Psychologist is in a team led by an expert who has been treating clients with psychological issues, full-time for at least 5 years	.554 (.053)
Convenience	Getting to your appointment will be difficult and time-consuming	-.250 (.054)
	Getting to your appointment will be quick and easy	.216 (.041)
Thorough care	The consultation will not be as thorough as you would like	-.351 (.054)
	The consultation will be as thorough as you would like	.560 (.057)

Response Times

We provide a brief analysis of the timing data to demonstrate that they follow sensible trends, and are thus amenable to the accumulator modelling we describe in the following section. There were no global differences in the average time taken to provide a best or a worst response (M = 14.04 s, SE = .94, and M = 13.93 s, SE = 1.01, respectively, $t(30)$ = .22, p = .83). This suggests that, on average, participants did not adopt a deterministic response strategy but rather chose their response order on the basis of the attribute levels present in each choice set. The attribute of the level chosen as best did not influence the time taken to

make the best choice ($F(3, 110) = 1.47$, $p = .23$), but the attribute level chosen as worst had a reliable effect on the time required to make a worst choice ($F(3, 110) = 4.84$, $p = .003$). In particular, when respondents selected the waiting time attribute level as the worst option in a choice set ($M = 13.3$ s, $SE = 1.2$), they did so significantly more quickly than when they chose the experience or convenience attribute levels as the worst option ($M = 20.2$ s, $SE = 2.2$, $p = .017$ and $M = 19.6$ s, $SE = 1.7$, $p = .025$, respectively, using Bonferroni adjusted p-values). No other pairwise differences were significant. Taken together, the best-worst scores and timing data provide sensible results, but a model-based analysis is necessary to combine the two sources of data coherently.

Model-Based Analyses: Approach & Methods

Recent developments have integrated cognitive theories of decision-making with traditional analysis approaches for preference data (Hawkins et al., in press). These new approaches use theories of evidence accumulation – or, in our case, preference accumulation – and have several advantages over the use of best-worst scores as above. For example, the cognitive models can be estimated in hierarchical Bayesian frameworks, which provide well-known advantages in statistical inference, as well as providing a coherent approach to the difficulty of aggregating over individuals. Cognitive decision-making models also allow for the inclusion of timing data. In simple decisions (about perception or memory, for example), timing data can provide different information from choice data, providing insight into otherwise intractable problems (for an example, see Donkin, Brown, Heathcote and Marley, 2009). The easiest approach to timing data is to analyse them independently of the associated choice data, for example using ANOVAs on mean decision times. This approach can be dangerous, and can lead to incorrect conclusions when the data are influenced by un-modelled effects such as speed-accuracy tradeoffs (see Johnson & Ratcliff, 2014, for details).

Analysis based on a cognitive decision-making model addresses this problem by jointly accounting for responses and the times taken to make those responses.

We analysed timing data along with responses from the best-worst task using a decision-making theory based on "evidence accumulation". Accumulator-based decision-making models assume that the cognitive decision process is analogous to a horse race: horses race from a starting point toward a finishing line and the horse that crosses the finish line first is the winner. The different accumulators (horses) represent the different possible decision outcomes (in our case, different aspects of the scenario that could be chosen as best or worst – see Figure 3). Accumulators gather evidence (or preference strength) in favour of their associated response at a speed governed by parameters called *drift rates*, which are akin to the speed of the horse – a fast horse will reach the finish line much faster than a slow horse. Similarly, a large drift rate reflects a likely choice outcome and a small drift rate reflects an unlikely choice outcome. For more details on how accumulator models operate, and extensive mathematical treatments, see Ratcliff and Smith (2004) or Brown and Heathcote (2008).

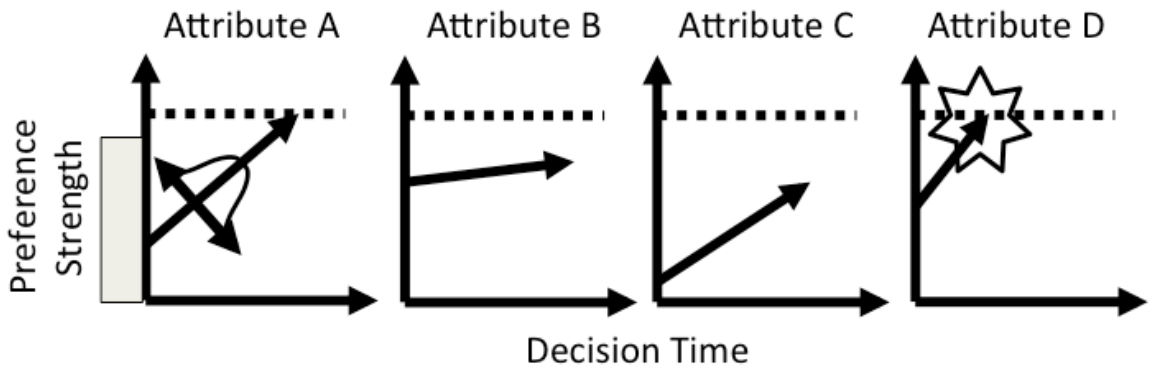

Figure 3. Illustration of an accumulator model for a choice between four attributes. Preference strength is accumulated in favour of each attribute, and the one that reaches a threshold (dashed line) first determines the choice (Attribute D, in this example, shown by

starburst). The amount of time taken to reach threshold determines the decision-making time, and the speed of accumulation is called the "drift rate". Drift rates, and the starting points of accumulation, vary between decisions with random and systematic components (illustrated by uniform and normal distribution curves in the left panel only).

We used Hawkins et al.'s (in press) methods to apply the accumulator model to data. An end product of the analysis is an estimated "drift rate" for each level of each attribute: these drift rates are akin to best-worst scores, with larger values indicating more preferred options. Since our analyses were based on a hierarchical Bayesian model, drift rates were estimated at both the population level and individually for each respondent, and all estimates come in the form of posterior distributions, from which we extracted point estimates (we used the median) and a measure of the uncertainty in that estimate (we used the widths of the 95% Bayesian credible intervals). For details of the methods used to estimate the posterior distributions, and other aspects of the hierarchical Bayesian modelling, see Turner, Sederberg, Brown and Steyvers (2013).

Following Hawkins et al. (in press), we model the joint decisions about best and worst attributes using two separate races: in the *best race* the accumulators gather evidence until one of the options reaches the threshold to trigger a *best* response, and in the *worst race* the accumulators gather evidence until one of the options reaches the threshold to trigger a *worst* response. The drift rate for each accumulator in the worst race is set to be the reciprocal of the drift rate for the corresponding accumulator in the best race. In this way, strongly preferred responses in the best race (which have large drift rates) will be unlikely to win the worst race (where their drift rates will be small). The drift rate therefore reflects the preference strength, of an attribute-level: attractive attribute-levels (e.g., a thorough

appointment) will quickly trigger a *best* response and unattractive attribute-levels (e.g., long waiting time) will quickly trigger a *worst* response. This reciprocal relationship is a simplifying assumption, and other approaches are possible (although, at least in some consumer and perceptual choices, there is empirical evidence in favour of the reciprocal assumption: see Hawkins, Marley, Heathcote, Flynn, Louviere & Brown, in press; Islam, Hawkins, & Marley, in preparation).

Model-Based Analyses: Results

We estimated the parameters of the decision-making model twice: once using only the choices people made (equivalent to observing only which horse won the race), and a second time using the choices people made and the time taken to make them (equivalent to finding out which horse won the race and the duration of the race). Our analysis focuses on the drift rate parameters, as those measure the quantity of key interest, preference strength. We assumed the other parameters of the model, such as the height of the decision threshold and the amount of time taken for non-decision processing, were fixed across attributes and choice sets. Hence, those parameters cannot convey information about the differences between attributes and levels of attributes so we do not consider them further.

Population-level estimates of the drift rates from the choice-only analysis agreed closely with the average best-worst scores in Table 1 ($r = .95$, $t(8) = 8.6$, $p < .001$). When timing data were also included in the analysis, there was a similarly strong relationship between the best-worst scores and drift rates ($r = .96$, $t(8) = 9.4$, $p < .001$). These analyses confirm that the model-based analyses provide convergent conclusions with the best-worst scores above, whether or not timing data are included.

We investigated whether including timing data in the analyses was beneficial, by investigation of the precision of the drift rate estimates. The precision of each drift rate

estimate can be inferred from the width of the posterior distribution for that estimate (the marginal posterior distributions for all drift rate estimates are in the endnote). We found that the inclusion of timing data led to more precise drift rate estimates, corresponding to greater sensitivity in the measurement of respondents' preference strengths. For example, Figure 4 plots the width of the credible interval for each population-level drift rate parameter when estimated with timing data against the corresponding width when estimated without timing data. The diagonal ($y = x$) line indicates where the points would fall if the analyses with and without timing data yielded equally large credible intervals. About half of the drift rate parameters were estimated with very high precision (very small 95% credible intervals, the cluster of points in the lower-left corner of Figure 4) and for those parameters, the inclusion of timing data made little difference. However, for all other drift rate parameters, the incorporation of timing data in the analyses resulted in smaller 95% credible intervals. Smaller credible intervals imply improved measurement sensitivity and statistical power, obtained without any extra cost from participants' point of view; just by including otherwise-discarded timing data in the analyses.

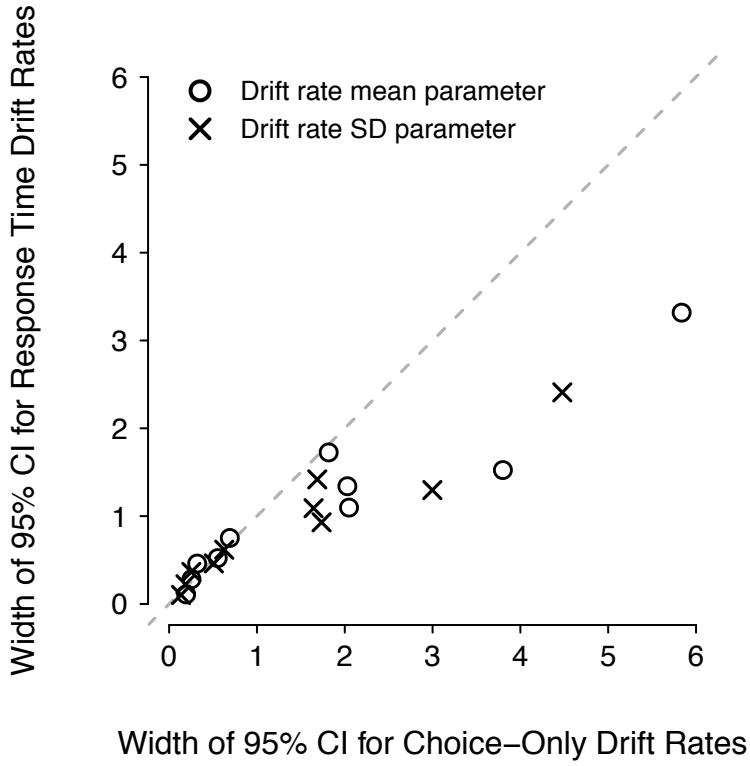

Figure 4. Width of the 95% credible intervals for drift rates in the accumulator model. The *x*-axis shows the credible interval when the model took account solely of the choices people made, and the *y*-axis shows the credible interval when the model simultaneously considered the choices people made and the time taken to make them. The circles and crosses represent the credible intervals calculated on the population mean and standard deviation of the drift rates. The diagonal ($y = x$) line indicates where the credible intervals from the two analyses would lie if they were equivalent according to the two methods.

Discussion

Our primary aim was to demonstrate the effectiveness of BWS in the elicitation and analysis of patients' preferences. A secondary research aim was to evaluate the use of response timing data in this process. We found that respondents were readily able to complete the BWS task, and that – even with data from only 31 respondents – the method provided accurate estimates of patients' preferences. These estimates were made even more accurate

by the inclusion of timing data in analyses based on a cognitive model of decision-making. The improved measurement accuracy gained from the inclusion of otherwise-discarded timing data was a reduction in credible interval width of about 40% for the most uncertain estimates. Assuming an approximately square-root relationship between sample size and the standard deviation of the posterior distributions over parameters, the inclusion of timing data provided the equivalent of more than doubling the sample size.

The need for further investigation of the preferences of psychologists' patients is highlighted by differences between our results and those reported by Coast et al. (2006) using similar methodology with dermatologists' patients: for example, the psychologists patients' placed greater emphasis on thoroughness of the appointment and less emphasis on its convenience than the dermatologists' patients. Our results also highlight the quantitative nature of the BWS results, which can be very useful for decision-making at a policy level (e.g., by addressing trade-off questions such as "how many weeks of waiting time are equivalent to an appointment with a more expert practitioner?").

An advantage of using BWS methods to investigate preference is the ease with which BWS questionnaires can be delivered online. Online delivery can sometimes reduce participation rates, but has other important advantages, including low cost and improved anonymity for respondents. Online delivery also allows respondents to participate in the study at a location of their choosing, which can be important for anxious or otherwise vulnerable populations. Finally, depending on the platform chosen for the software, online delivery often allows for the easy and accurate collection of timing data.

While our example study investigates patients from private psychology practices, the method could prove useful in a much wider range of situations. BWS can be applied to any research questions that hinge on the measurement of preference or attitude strength, and can be particularly useful in organised care settings. For example, BWS methods could be used to

explore the attitudes of convicted offenders, or new military recruits' values toward combat, or returned veterans towards their reintegration into society. In each case, the methodological and analysis approaches we illustrate can be applied, but the particular attributes and levels will need to be tailored to the research question.

Limitations and Future Directions

Our respondents represent a very small sample, just 31 people, drawn from only two psychology practices in New South Wales, Australia. Even though the BWS methods allowed for accurate estimation of preference strength in this small sample, it would be unwise to draw population-wide conclusions about the preferences of psychologists' patients from these data. Of course, the primary goal of our work – to illustrate the use of BWS methods and analyses in psychological service delivery – is not limited by this small sample.

A difficult question in the measurement of preference strengths regards absolute versus relative preference. Indeed, this problem applies to even such fundamental cognitive processes as the psychological estimation of physical magnitudes (see Brown, Marley, Dodds & Heathcote, 2009). The BWS approach measures preferences for each attribute-level relative to all the other attribute-levels used in the survey. So, for example, if our survey was adjusted such that the 'three months' attribute-level changed status from the longest wait time to the shortest (with new, even longer, wait times introduced), the best-worst scores and drift rates estimated for the three-month-wait attribute-level would reflect this change. Given the ubiquity of context effects in psychology, it is difficult to imagine a method that could measure the strength of preference for some attribute-level in a truly absolute sense. For example, consumers' purchasing preferances are reliably altered by irrelevant contextual changes, such as the inclusion of unwanted alternative options (Trueblood, Brown, Heathcote & Busemeyer, 2013; Trueblood et al., 2014). Context effects in BWS will be reduced by

using larger sets of choice attributes and wider ranges for the levels of those attributes, but it will remain important to keep in mind the relativity of the results.

Conclusions

We second the assertion that "the use of discrete choice experiments is feasible to ascertain patient's preference of aspects of primary care consultation" (Longo et al., 2006, p.35; see also Jan et al., 2000, p.64, for similar statements). Discrete choice experiments have high levels of external validity and are easy for respondents to navigate. Comparisons of measures from DCE experiments with market choices have shown strong relationships between the experimental measures and real choices (Louviere et al., 2008). Coupled with mathematical modelling based on decision-making theories from cognitive science, DCEs are an efficient way to understand patient preferences while avoiding many of the problems that are associated with other methods.

References

Andrews, G., & Slade, T. (2001). Interpreting scores on the Kessler Psychological Distress Scale (k10). *Australian and New Zealand Journal of Public Health, 25,* 494-497. http://dx.doi.org/10.1111/j.1467-842X.2001.tb00310.x

Baumgartner, H., & Steenkamp, J.B. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38*, 143-156. http://dx.doi.org/10.1509/jmkr.38.2.143.18840

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology, 57*, 153-178. http://dx.doi.org/10.1016/j.cogpsych.2007.12.002

Brown, S.D., Marley, A.A.J., Dodds, P., & Heathcote, A.J. (2009) Purely relative models cannot provide a general account of absolute identification. *Psychonomic Bulletin & Review, 16*, 583-593. http://dx.doi.org/10.3758/PBR.16.3.583

Coast, J., Salisbury, C., De Berker, D., Noble. A., Horrocks, S. A., Peters, T. J., & Flynn, T. N. (2006). Preferences for aspects of a dermatology consultation. *British Journal of Dermatology, 155*, 387-392. http://dx.doi.org/10.1111/j.1365-2133.2006.07328.x

Devinney, T. M., Auger, P., & Eckhardt, G. M. (2010). *The myth of the ethical consumer.* Cambridge University Press: United Kingdom.

Donkin, C., Brown, S. D., Heathcote, A., & Marley, A. A. J. (2009) Dissociating speed and accuracy in absolute identification: The effect of unequal stimulus spacing. *Psychological Research, 73,* 308-316. http://dx.doi.org/10.1007/s00426-008-0158-2

Farrar, S., Ryan, M., Ross, D., & Ludbrook, A. (2000). Using discrete choice modelling in priority setting: an application to clinical service developments. *Social Science and Medicine, 50,* 63-75. http://dx.doi.org/10.1016/S0277-9536(99)00268-3

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. http://dx.doi.org/10.1186/1471-2288-8-76

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*, 457–472. http://dx.doi.org/10.1214/ss/117701113

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*. http://dx.doi.org/10.1111/cogs.12094

Hawkins, G. E., Marley, A. A. J., Heathcote, A., Flynn, T. N., Louviere, J. J., & Brown, S. D. (in press). The best of times and the worst of times are interchangeable. *Decision.*

Jan, S., Mooney, G., Ryan, M., Bruggemann, K., & Alexander, K. (2000). The use of conjoint analysis to elicit community preferences in public health research: A case study of hospital services in South Australia. *Australia and New Zealand Journal of Public Health, 24,* 64-70. http://dx.doi.org/10.1111/j.1467-842X.2000.tb00725.x

Johnson, J., & Ratcliff, R. (2014). Computational and Process Models of Decision Making in Psychology and Behavioral Economics. In P. W. Glimcher & E. Fehr (Eds). *Neuroeconomics (2nd ed).* New York: New York University Press. http://dx.doi.org/10.1016/B978-0-12-416008-8.00003-6

Kessler, R. C., Barker, P. R., Colpe, L. J., Epstein, J. F., Gfroerer, J. C., Hiripi, E., et al. (2003). Screening for serious mental illness in the general population. *Archives of General Psychiatry, 60*, 184-189. http://dx.doi.org/10.1001/archpsyc.60.2.184

Lee, J. A., Soutar, G. & Louviere, J. J. (2008). The best–worst scaling approach: An alternative to Schwartz's values survey. *Journal of Personality Assessment, 90*, 335-347. http://dx.doi.org/10.1080/00223890802107925

Longo, M. F., Cohen, D. R., Hood, K., Edwards, A., Robling, M., Elwyn, G., & Russell, I. T. (2006). Involving patients in primary care consultations: Assessing preferences using discrete choice experiments. *British Journal of General Practice, 56,* 35–42.

Louviere, J. J., Street, D., Burgess, L., Wasi, N., Islam. T., & Marley, A. A. J. (2008). Modeling the choices of individual decision-makers by combining efficient choice experiment designs with extra preference information. *Journal of choice modelling, 1*, 128-164. http://dx.doi.org/10.1016/S1755-5345(13)70025-3

Luce, R. D. (1986). *Response Times.* NY: Oxford University Press.

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. http://dx.doi.org/10.1016/j.jmp.2005.05.003

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*. NY: Academic. http://dx.doi.org/10.1016/B978-0-12-590241-0.50006-X

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review, 111,* 333-367. http://dx.doi.org/10.1037/0033-295X.111.2.333

Regenwetter, M., Grofman, B., Marley, A. A. J., & Tsetlin, I. (2006). *Behavioral social choice: probabilistic models, statistical inference, and applications.* Cambridge: Cambridge University Press.

Ryan, M., & Farrar, S. (2000). Using conjoint analysis to elicit preferences for health care. *British Medical Journal, 320,* 1530-1533. http://dx.doi.org/10.1136/bmj.320.7248.1530

Schwartz, S. H., & Bardi, A. (2001). Value hierarchies across cultures: Taking a similarities perspective. *Journal of Cross-Cultural Psychology, 32,* 268- 290. http://dx.doi.org/10.1177/0022022101032003002

Trueblood, J., Brown, S. D., & Heathcote, A. (2014). The multi-attribute linear ballistic accumulator model of context effects in multi-alternative choice. *Psychological Review 112*(2), 179-205. http://dx.doi.org/10.1037/a0036137

Trueblood, J., Brown, S.D., Heathcote, A. & Busemeyer, J. (2013) Not just for consumers: Context effects are fundamental to decision-making. *Psychological Science, 24*(6), 901-908. http://dx.doi.org/10.1177/0956797612464241

Turner, B., Sederberg, P., Brown, S. D., & Steyvers, M. (2013). A note on efficiently sampling from distributions with correlated dimensions. *Psychological Methods, 18*, 368-384. http://dx.doi.org/10.1037/a0032222

Endnote: MCMC Methods and Results.

We estimated the posterior distributions of the parameters of the Best-Worst LBA model using differential evolution Markov chain Monte Carlo methods (DE-MCMC: Turner et al., 2013). The subject-level parameters of the model were drawn from hyper-distributions, corresponding to the population of respondents, which we assumed were normal distributions truncated to positive values. We placed relatively uninformative prior distributions over population mean (μ) and standard deviation (σ) parameters of the truncated normal distributions. When we estimated the model with and without response times, we assumed the prior distributions of the $i = 10$ drift rate (d) parameters of the model were distributed as

$$\mu_i^d \sim Normal_{(0,Inf)}(3,3)$$

$$\sigma_i^d \sim Gamma(1,.1),$$

and the standard deviation of the distribution of drift rates was fixed at $s = 1$. When estimating the model with response times we assumed relatively uninformative prior distributions for the start-point (A), response thresholds (b, estimated separately for the best and worst decisions), and non-decision time (t_0) parameters of the LBA:

$$\mu^A, \mu^b \sim Normal_{(0,Inf)}(20,50)$$

$$\mu^{t_0} \sim Normal_{(0,Inf)}(3,3)$$

$$\sigma^A, \sigma^b, \sigma^{t_0} \sim Gamma(1,.1).$$

We allowed for a large range across the parameter space of the time-related parameters since we have little previous research to govern the expected range in response times in complex, multi-attribute decisions in best-worst scaling tasks.

We drew 4,000 samples from the posterior distributions of each of 40 chains and discarded the first 3,000 samples as burn in, providing a total of 40,000 samples from the posterior distributions of the parameters. We examined convergence of the MCMC chains

using the Gelman-Rubin convergence statistic, $\hat{R}$, which compares sample variances between chains to the sample variances within chains (Gelman & Rubin, 1992). When chains have converged the between- and within-sample variances are equal and $\hat{R} = 1$, whereas values of $\hat{R} > 1.1$ indicate inadequate conference. All MCMC chains showed good convergence: all $\hat{R}$'s < 1.01 for the model fitted to choices-only and choices and response times. Figure A1 shows the marginal posterior distributions over the drift rate parameters for each attribute-level.

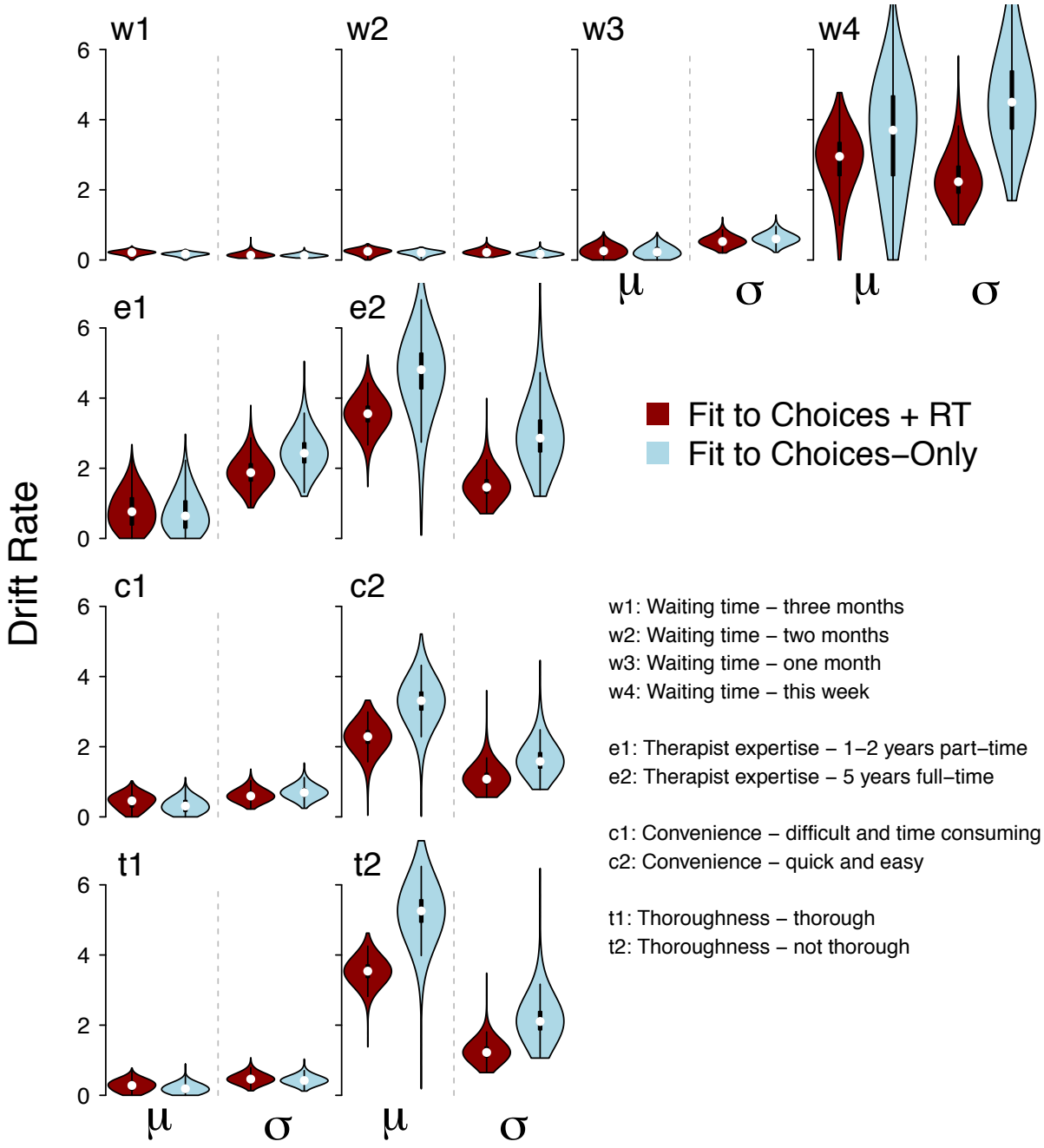

Figure A1. Posterior distributions over drift rate parameters for each attribute-level. Red and blue violin plots were estimated with and without response time data, respectively. Each plot panel corresponds to a single attribute level (e.g., top left panel corresponds to the longest level of the waiting time attribute). Within each panel, the left and right halves show posterior distributions over the location (μ) and standard deviation (σ) parameters for the group-level distribution of drift rates.