A Formal and Empirical Comparison of Two Score Measures for Best-Worst Scaling

A. A. J. Marley*[,1,2], T. Islam[3] and G. E. Hawkins[4]

*(corresponding author)
[1]Department of Psychology
University of Victoria
PO Box 1700 STN CSC
Victoria
BC V8W 2Y2
Canada
email: ajmarley@uvic.ca

[2]Research Professor
Institute for Choice
UniSA Business School
Level 13, 140 Arthur Street
North Sydney
NSW 2060
Australia

[3]Department of Marketing
and Consumer Studies
College of Business and Economics
University of Guelph
50 Stone Road East
Guelph
ON N1G 2W1
Canada
email: islam@uoguelph.ca

[4]Amsterdam Brain and Cognition Center
University of Amsterdam
Nieuwe Actergracht 129
1018 WS Amsterdam
The Netherlands
email: guy.e.hawkins@gmail.com

Abstract

Best-worst scaling (BWS) is a method that asks individuals to choose the most and the least preferred option from a set of available options. There has been extensive discussion and evaluation of the use of *scores* (data summaries) in the analysis of such data. Here we motivate, summarize, and compare the usefulness of two such score measures: the *analytical closed form solution* (Lipovetsky & Conklin, 2014, *Journal of Choice Modelling)* and *normalized best-worst scores* (Louviere, Flynn, & Marley, 2015, *Cambridge University Press*). We conclude that both have underlying motivations in the *maxdiff model* of best-worst choice and that the analytical closed form solution provides better fits to the aggregate choices in several best-worst choice data sets.

Keywords: best-worst scaling; best-worst choice; discrete choice experiment; scores; choice model

1 Introduction

Best-worst scaling (BWS) is a method that asks individuals to choose the most and least preferred option from a set of available options. Best-worst choice data is obtained using a design that leads to the presentation of each of several subsets of options, with each possible option being presented a number of times (within or between participants). Each participant chooses the best, and the worst, option in each presented set; thus, each option is chosen a number of times as best, as worst, or neither. Such count data have been summarized in various forms, conventionally known as *scores*. Recently, Lipovetsky and Conklin (2014; abbreviated to L&C in various places in the remainder of the paper) developed a closed-form score for best-worst choice, henceforth referred to as the *analytical best-worst (ABW) score (estimator)*. For each option x in the study (design), let N_x be the number of times x is presented, N_x^b the number of times x is selected as best, and N_x^w the number of times x is selected as worst; therefore $N_x - N_x^b - N_x^w$ is the number of times x is not selected as either best or worst. L&C develop the ABW score

$$\ln \frac{1 + \frac{N_x^b - N_x^w}{N_x}}{1 - \frac{N_x^b - N_x^w}{N_x}}, \tag{1}$$

where the quantity

$$\frac{N_x^b - N_x^w}{N_x} \tag{2}$$

is the *normalized best minus worst (NBW) score* for x. The latter closed-form measure has been used frequently in the literature on best-worst choice, where it is often linearly related to the estimated parameters of various models (Louviere, Flynn, & Marley, 2015) plus it has numerous interesting theoretical properties (summarized by Flynn & Marley, 2014). For instance, the set of (normalized) best minus worst scores is a sufficient statistic for the parameters of the *maxdiff model of best-worst choice* [below, Sect. 3, (9)].[1] Since the function in (1) is a strictly increasing function of the NBW score in (2), the set of ABW scores is also a sufficient statistic for that model.

Figure 1 shows the mathematical relation between ABW, (1), and NBW, (2), scores.[2] Our aim in this paper is to further understand the properties of, and relations between, these two measures, and to evaluate each against data. Our conclusions are that the maxdiff model of best-worst choice underlies the usefulness of each of the measures; and that the ABW measure gives better prediction of (aggregate) best (respectively, worst, best-worst) choices than does

[1] The term *maxdiff* was coined for this model as its representation can be described as a result of a decision maker choosing as the best-worst pair in a set the pair of options with the largest (random) utility difference, with the random component having an extreme value distribution.

[2] The ABW measure is clearly highly linearly related to the NBW measure for much of the latter's range. For instance, with NBW restricted to $(-.5, .5)$, the slope is 2.11 with an R^2 of .999.

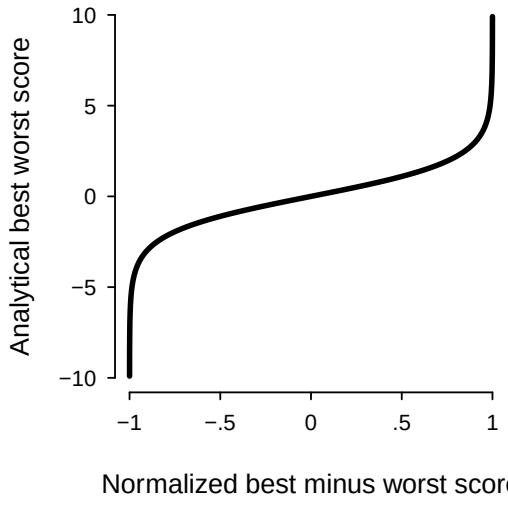

Figure 1: The relationship between the analytical best worst score (y-axis) and the normalized best minus worst score (x-axis).

the NBW score, when the (aggregate) best (resp., worst) choices are the relevant marginals of (aggregate) best-worst choices.

The remainder of the paper is as follows. Section 2 completely redevelops L&C's ABW score measure in a way that allows us, in Section 3, to show the measure's relationship with the likelihood of data generated by the maxdiff model of best-worst choice [below, (9)]; this link was not made in L&C. The corresponding previously known link for the NBW measure essentially puts the two measures on an equal footing. Then, in Section 4, we evaluate each measure against data. Section 5 summarizes and discusses further results and open questions.

2 Lipovetsky and Conklin's (2014) Matrix of Extended Best-Worst Data

As did L&C, we analyze best-worst choice data aggregated across participants – that is, for each choice set in the design, we calculate the proportion of participants who chose a particular pair of options as best-worst; we also calculate the appropriate marginals of those best-worst proportions to obtain the (marginal) best (resp., worst) choice proportions. Nonetheless, we describe the methods as if the data are for an individual decision maker. We present the results in terms of choice sets with four options; they are easily extended to other set sizes.

Let P denote a set of items (17 in L&C) and let $D(P)$ denote the design, that

is, the set of (sub)sets of items that occur in the study; without loss of generality, we assume each set in $D(P)$ has four items. We now present an argument for the *extended best-worst data* and *choice probabilities* of Table 1, followed by a demonstration that L&C's ABW score measures are the maximum likelihood estimates for the assumed model of the extended data. Along the way, we point out the close link between our and L&C's framework. Section 3 extends our approach to a process model of the actual best-worst data, in contrast to the modeling of the extended best-worst data considered in this section.

We assume that each person has a (common) referent item ϕ across all choice sets in the design; the referent is an inferred status quo option, not a presented option. Assume that the choice set $\{e, f, g, h\}$ is presented to the decision maker and e is chosen as best, h as worst. These observed choices are used to produce the *extended choice sets* and *implied choices For Best (resp., For Worst)* in the first column of Table 1. The underlying concept is that each option is independently compared with the referent ϕ, under a "best" condition, and, separately, under a "worst" condition. When selecting the best option, the observed choice of e best is interpreted as corresponding to e being better than ϕ, and ϕ being better than each of f, g, and h. When selecting the worst option, the observed choice of h worst is interpreted as corresponding to h being worse than ϕ, and ϕ being worse than each of e, f, and g. We then assume that, for each item x in the set $\{e, f, g, h\}$, $b(x, \phi)$ is the probability that x is better than ϕ, and $w(\phi, x)$ is the probability that ϕ is worse than x; this gives the final column of Table 1. As we discuss later, in detail, there are many possible combinations of choices that do *not* occur under the assumed expansion process; the latter combinations will be considered in Section 3, where we present a *process* interpretation of the extended best-worst data.

Table 1. Component binary choice probabilities on extended best-worst choice data on a set $\{e, f, g, h\}$, when e is chosen as best, h is chosen as worst; ϕ is a hypothetical referent item.

	Choice Set					
	e	f	g	h		
Observed Best choice	e					
Observed Worst choice				h		
Extended Choice Sets:					Implied Choices:	Choice Probability:
For Best						
e	1	0	0	0	e	$b(e, \phi)$
f	0	1	0	0	ϕ	$b(\phi, f)$
g	0	0	1	0	ϕ	$b(\phi, g)$
h	0	0	0	1	ϕ	$b(\phi, h)$
For Worst						
e	1	0	0	0	ϕ	$w(\phi, e)$
f	0	1	0	0	ϕ	$w(\phi, f)$
g	0	0	1	0	ϕ	$w(\phi, g)$
h	0	0	0	1	h	$w(h, \phi)$

Now assume that, for each item x in the design P, there is a finite real (utility) value $u(x)$, and, with no loss of generality, set $u(\phi) = 0$. We also assume that

$$b(x, \phi) = \frac{e^{u(x)}}{1 + e^{u(x)}} \text{ and } w(\phi, x) = \frac{1}{1 + e^{-u(x)}}. \tag{3}$$

These equations imply that[3]: $b(x, \phi) = w(\phi, x)$; the probabilities are in the open interval $(0, 1)$ for real values of $u(x)$; and

$$u(x) = \ln \frac{b(x, \phi)}{1 - b(x, \phi)} = \ln \frac{w(\phi, x)}{1 - w(\phi, x)}.$$

We show, below, that the assumptions in (3), with those in the *data expansion* of Table 1, lead to L&C's ABW score measure[4].

The probability of the pattern of extended choices in Table 1 is

$$b(e, \phi)b(\phi, f)b(\phi, g)b(\phi, h)w(\phi, e)w(\phi, f)w(\phi, g)w(h, \phi).$$

By (3), we have $w(\phi, x) = b(x, \phi)$ for each item x, so the expression becomes

$$b(e, \phi)b(\phi, f)b(\phi, g)b(\phi, h)b(e, \phi)b(f, \phi)b(g, \phi)b(\phi, h), \tag{4}$$

which equals

$$b(e, \phi)^2 b(\phi, f)b(f, \phi)b(\phi, g)b(g, \phi)b(\phi, h)^2. \tag{5}$$

Comment: The sum of (4) over all possible patterns allowed by the reasoning leading to Table 1 for the choice set $X = \{e, f, g, h\}$ can be less than 1 (see **Appendix A**). The fact that the sum can be less than one means that those expressions do not form a process model of the actual choices in a standard best-worst task, where the decision maker has to select a best-worst pair. This is not an issue at this stage as we are presenting an argument that leads to L&C's ABW score measure. Section 3 presents a process model of best-worst choice suggested by the logic of Table 1 that does not have this defect.

We now continue the reasoning based on Table 1 and show that it leads to the ABW score measure (parameter estimates) proposed by L&C.

For each (four element) set $X \in D(P)$, and $x, y \in X$, $x \neq y$, let $\widetilde{bw}_X(x, y)$ be the aggregate number of times the best-worst choice pair in X is x best and

[3]There is ongoing empirical study of when the "mirror image" property of best and worst holds, including in the binary case of (3). The currently available data are summarized and discussed in Louviere et al. (2015, Sect. 6.2) and Section 5 of the current paper considers the relevance of our work to those discussions.

[4]This is because our referent ϕ corresponds to the 0's in L&C's Table 4, and our formulae (3) correspond to L&C's pairwise logistic formulae (their (5)).

y worst[5]; clearly, $\widetilde{bw}_X(x,y) = 0$ if either x or y is not in X. Then the likelihood of the extended best-worst choice data of Table 1 is

$$\prod_{X\in D(P)} \prod_{\substack{\{e,h\}\in X \\ e\neq h}} [b(e,\phi)^2 b(\phi,f) b(f,\phi) b(\phi,g) b(g,\phi) b(\phi,h)^2]^{\widetilde{bw}_X(e,h)} \tag{6}$$

Now consider a generic option $x \in P$ and the sets $X \in D(P)$ in which x is present. In agreement with L&C's notation, let N_x be the number of times x is presented in the design $D(P)$, and let

$$N_x^b = \sum_{X\in D(P)} \sum_{r\in X-\{x\}} \widetilde{bw}_X(x,r),\ N_x^w = \sum_{X\in D(P)} \sum_{s\in X-\{x\}} \widetilde{bw}_X(s,x);$$

that is, x is best N_x^b times, worst N_x^w times, and neither best or worst $N_x - N_x^b - N_x^w$ times.

Collecting the terms for sets in which x appears in the likelihood (6), we obtain the expression

$$\begin{aligned} & b(x,\phi)^{2N_x^b} b(\phi,x)^{2N_x^w} [b(\phi,x)b(x,\phi)]^{N_x - N_x^b - N_x^w} \\ &= b(x,\phi)^{N_x + (N_x^b - N_x^w)} b(\phi,x)^{N_x - (N_x^b - N_x^w)} \\ &= b(x,\phi)^{N_x + (N_x^b - N_x^w)} [1 - b(x,\phi)]^{N_x - (N_x^b - N_x^w)} \end{aligned} \tag{7}$$

Differentiating the log of this likelihood w.r.t $b(x,\phi)$, and setting the result to zero, gives

$$\frac{N_x + [N_x^b - N_x^w]}{b(x,\phi)} = \frac{N_x - [N_x^b - N_x^w]}{1 - b(x,\phi)},$$

which, with the form of $b(x,\phi)$ from (3), gives

$$u(x) = \ln \frac{b(x,\phi)}{1-b(x,\phi)} = \ln \frac{N_x + [N_x^b - N_x^w]}{N_x - [N_x^b - N_x^w]} = \ln \frac{1 + \frac{N_x^b - N_x^w}{N_x}}{1 - \frac{N_x^b - N_x^w}{N_x}},$$

that is, the ABW estimator.

[5] With N_X the number of times the set X is presented in the design $D(P)$, we have

$$N_X = \sum_{\substack{x,y\in X \\ x\neq y}} \widetilde{bw}_X(x,y).$$

3 A Process Model Reinterpretation of Lipovetsky and Conklin's (2014) Matrix of Extended Best-Worst Data

Section 2 presented our development of L&C's ABW measure using their extended matrix form, which is based on the original best-worst data. We now discuss why their solution gives an approximation, not only to the maximum likelihood parameter estimates for a model of the best-worst data, as stated by L&C, but also to the likelihood of the original best-worst data.

Appendix A shows that the formulae (4) do not form a complete model of the original best-worst data because the sum of those expressions over each choice set $\{e, f, g, h\}$ is not equal to one. We now extend the process that leads to (4) in the most natural way, and show that the resulting model is the *maxdiff model* of best-worst choice.[6]

The process described in Table 1 can be extended to a model of the actual best-worst data as follows: the respondent independently compares each option in $\{e, f, g, h\}$ with the referent ϕ in terms of which is better (the option or the referent); if the results of all four comparisons is exactly one option being better than ϕ, call it x, then x is noted as "best". The respondent then carries out the parallel comparison of each option in $\{e, f, g, h\}$ with the referent ϕ in terms of which is worse (the option or the referent); if the result of all four comparisons is exactly one option being worse than ϕ, call it y, then y is noted as "worst". In such a case, and $x \neq y$, these (single items) become the best, x, and worst, y, choices for that set. However, if the process does *not* result in a single best x and worst y, $x \neq y$, the respondent restarts the whole process.

Continuing with $X = \{e, f, g, h\}$, let $BW_X(e, h)$ denote the probability that e is chosen as best, and h is chosen as worst, $e \neq h$, by the above process. Then a standard argument (paralleling Marley & Louviere, 2005, Sect. 4.1.2, Case 2), with the form (5) for the (unconditional) selection of the best-worst pair (e, h), gives

$$BW_X(e,h) = \frac{b(e,\phi)^2 b(\phi,f) b(f,\phi) b(\phi,g) b(g,\phi) b(\phi,h)^2}{\sum_{\{r,s,t,u\}=X} b(r,\phi)^2 b(\phi,s) b(s,\phi) b(\phi,t) b(t,\phi) b(\phi,u)^2}. \quad (8)$$

Note that, in the denominator, every set $\{r, s, t, u\} = X = \{e, f, g, h\}$. Dividing the numerator and denominator by

$$b(\phi,e) b(e,\phi) b(\phi,f) b(f,\phi) b(\phi,g) b(g,\phi) b(\phi,h) b(h,\phi),$$

which is symmetric (i.e., independent of the positions of e, f, g, h), and nonzero by (3) with the assumption of finite valued u, gives the form

[6] The following extension of the L&C approach is not the simplest process interpretation of the maxdiff model; see Marley & Louviere (2005, Sect. 4.1.2, Case 2).

$$BW_X(e,h) = \frac{\frac{b(e,\phi)}{b(\phi,e)} \cdot \frac{b(\phi,h)}{b(h,\phi)}}{\sum_{\substack{\{r,u\}\in X \\ r\neq u}} \frac{b(r,\phi)}{b(\phi,r)} \cdot \frac{b(\phi,u)}{b(u,\phi)}}.$$

Substituting in the forms (3), this reduces to

$$BW_X(e,h) = \frac{e^{[u(e)-u(h)]}}{\sum_{\substack{\{r,u\}\in X \\ r\neq u}} e^{[u(r)-u(u)]}}, \tag{9}$$

which is the *maxdiff model of best worst choice* (Marley & Louviere, 2005).[7] The set of NBW scores for the options is a sufficient statistic for the maxdiff model (Marley & Pihlens, 2012); thus, as already noted in Section 1, the set of ABW scores is also a sufficient statistic for that model.

For each (four element) set $X \in D(P)$, and $x, y \in X$, $x \neq y$, let $\widetilde{bw}_X(x,y)$ be the aggregate number of times the best-worst choice pair in X is given by x best and y worst; also let N_X be the number of times the set X occurs in the design.[5] Then, according to the model in (8), the likelihood of the data is

$$\begin{aligned}
&\prod_{X\in D(P)} BW_X(e,h) \\
&= \prod_{X\in D(P)} \prod_{\substack{\{e,h\}\in X \\ e\neq h}} \left(\frac{b(e,\phi)^2 b(\phi,f) b(f,\phi) b(\phi,g) b(g,\phi) b(\phi,h)^2}{\sum_{\{r,s,t,u\}=X} b(r,\phi)^2 b(\phi,s) b(s,\phi) b(\phi,t) b(t,\phi) b(\phi,u)^2} \right)^{\widetilde{bw}_X(e,h)} \\
&= \prod_{X\in D(P)} \prod_{\substack{\{e,h\}\in X \\ e\neq h}} \left[b(e,\phi)^2 b(\phi,f) b(f,\phi) b(\phi,g) b(g,\phi) b(\phi,h)^2 \right]^{\widetilde{bw}_X(e,h)} \\
&\quad \times \prod_{X\in D(P)} \left(\frac{1}{\left[\sum_{\{r,s,t,u\}=X} b(r,\phi)^2 b(\phi,s) b(s,\phi) b(\phi,t) b(t,\phi) b(\phi,u)^2 \right]^{N_X}} \right)
\end{aligned}$$

Comparing this likelihood function with the corresponding form (6), we see that the numerator is exactly the term in (6), and the denominator does not appear in (6). Thus, L&C's derivation of their score measure does not include the terms given by the denominators of the terms in (8) in the likelihood function. Such an approach is sometimes advocated, based on the statement that the product of such denominator terms in the likelihood function can be treated as a "constant". This is incorrect, as that product depends on the values of the options in the design; and ignoring them may lead to bias in the estimation of parameters and incorrect fit measures based on log-likelihoods. Nonetheless, as we show in the next section, the ABW estimator performs extremely well

[7] The maxdiff model can also be developed in a similar manner using L&C's extended matrix method, but with best and worst extended separately, rather than together (see Marley & Louviere, 2005, Sect. 4.1.2. Case 2).

in fitting aggregate (marginal) best choice data derived from best-worst choice data.

4 Empirical Comparison of Analytical Best-Worst Scores and Normalized Best-Minus-Worst Scores

We tested the ability of the ABW and NBW scores to fit the aggregate (marginal) best choice probabilities derived from numerous sets of best-worst data obtained in discrete choice experiments; the details of the designs are provided in **Appendix B**. Four studies used Case 3 best-worst choice (choice between profiles), and two used Case 1 best-worst choice (choice between objects).[8] Table 2 shows the root mean square error (RMSE) between the observed (aggregate) best choices and the (aggregate) best choices predicted using the NBW (resp., ABW) scores as the parameters in a multinomial logit (MNL) model. In every case, the ABW scores had smaller RMSE; in almost all cases, the RMSE using the NBW scores was about twice the size of that using the ABW scores.

Table 2. Root Mean Square Error (RMSE) of in-sample prediction of aggregate (marginal) best choices using the NBW score (resp., ABW score). The RMSE of a set of L predictions (number of choice sets times alternatives in each choice set) is calculated as $RMSE = \sqrt{\frac{\sum_1^L (Best_{obs} - Best_{pred})^2}{L}}$. W_i denotes wave i of the design, where $i = 1, 2, 3, 4$.

	Product	Measure	RMSE			
			W_1	W_2	W_3	W_4
Case 3	Detergent	NBW	.095	.098	.100	.102
		ABW	.046	.039	.042	.044
	Toothpaste	NBW	.111	.111	.111	.103
		ABW	.054	.054	.054	.053
	Pizza	NBW	.101	.098	.103	.101
		ABW	.037	.036	.040	.041
	Solar	NBW	.096			
		ABW	.030			
Case 1	Budget Saving	NBW	.024			
		ABW	.014			
	Budget Spending	NBW	.021			
		ABW	.012			

Figure 2 plots a typical pair of results from Table 2. The range of predicted (marginal) best choice probabilities using NBW scores is smaller than the corresponding range using ABW scores, and of the data. This is largely due to

[8] In Case 3, a person is presented with a set of profiles (multiattribute options), such as possible vacation packages, and has to choose the best and worst of those. In Case 1, a person is presented with several objects (e.g., brand names), and has to choose the best and worst object. See Marley and Louviere (2005, Sect. 1.3) for details on the three cases of best-worst scaling.

the fact that the ABW scores are approximately two to three times larger than the corresponding NBW scores when the latter are not at the extremes of their range; that is, not near -1 or 1. (cf. Figure 1).

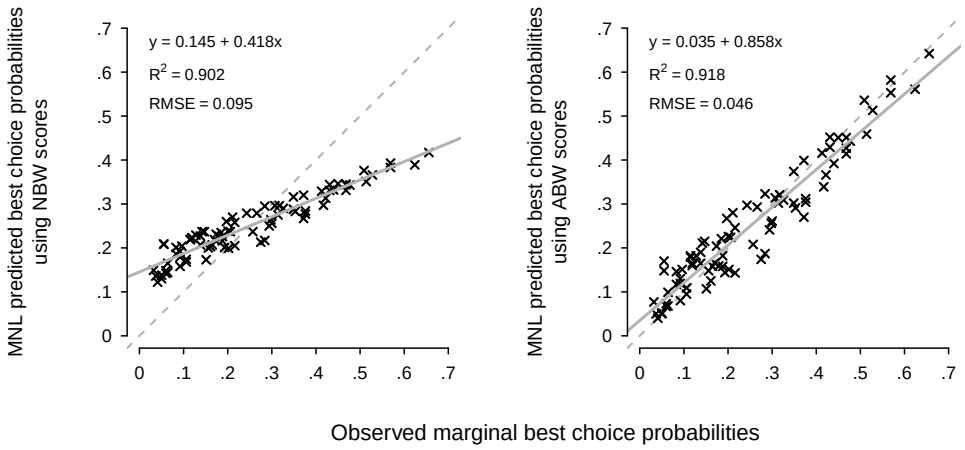

Figure 2: Observed (marginal) best choice probabilities from wave 1 of the detergent data set (x-axes) versus best choice probabilities predicted by an MNL model using the NBW scores (left panel) or ABW scores (right panel) as the (utility) parameters (y-axes). Solid lines show linear regression lines of best fit. Dashed lines show where the data would fall if the two variables were perfectly related.

Given the plot in Figure 1 and the results in Table 2, one would expect the following regarding the relation of each score measure to the parameters of a maximum likelihood fit of an MNL model to the best choice probabilities: For each data set, plot the ABW (resp., NBW) scores on the x-axis and the estimated parameters of the MNL model on the y-axis. Then the slope of the linear regression will be lower and closer to 1 for each ABW plot. We have carried out the relevant analyses, and this is the case for every data set in Table 2.

5 Summary, Further Results, and Open Questions

L&C focussed on predicting in-sample (marginal) best choices, as we have done thus far. However, given the close relation we have demonstrated between the ABW scores, the NBW scores, and the maxdiff model, it is worth extending the analyses to (marginal) worst choices, and best-worst choices. The results of these analyses are summarized in **Appendix B**. The in-sample fits to (marginal) worst choices are as good as those already presented for (marginal) best choices;

and, if anything, the corresponding fits to the best-worst choices are better than those to the (marginal) best or (marginal) worst choices[9]; in all cases, the ABW scores perform better than the NBW scores. Three of the best-worst data sets had four waves of data, which allowed us to test the ability of ABW and NBW scores to predict out-of-sample (marginal) best, (marginal) worst, and best-worst choices. These predictive tests supported a similar conclusion to the in-sample tests (see **Appendix B**). Thus, these results suggest that, for these aggregate data, best and worst choices are "mirror images" of each other.

Our results are very encouraging in terms of using the ABW scores as descriptive statistics for aggregate best-worst choice probabilities, and they build on earlier uses of the NBW scores (see Louviere et al., 2015). Nonetheless, we do not have a clear understanding of why these measures work so well, other than that each gives a sufficient statistic for the maxdiff model. Also, there is a growing empirical literature demonstrating that best choices do not always provide the same information (demonstrate the "same" preferences) as worst choices; examples are the model-based analyses of Giergiczny et al. (2013) and the purely data-based analyses of Greenacre et al (2015). Rose (2014) proposes that these differences are not an issue for the best-worst method if one is interested in choices processes, per se, but can be if best and worst data are naively pooled to, say, explain or predict best choices. Louviere et al. (2015, Chap. 6) summarize and discuss various of these data and the issues they raise. Also, there is an increasing tendency to model the (best-worst) choices of individual decision makers by collecting more data from each individual (e.g., Hawkins et al., 2014) and/or using Bayesian methods when there are a limited amount of data for each individual (e.g., Dumont et al., 2015). Score measures (such as ABW and NBW) play an important role in both exploratory and final analyses in these situations (see, e.g., Louviere, 2015). In particular, our results show that the ABW scores may provide a closer approximation to the relevant data than NBW scores.

As a final consideration, assume that we have a maxdiff model for choice among profiles (Case 3, choice between profiles) where the utility of an option is the sum of the (component) utilities of its attribute-levels. Count the number of times an attribute-level is "chosen"[10] as best minus the number of times it is "chosen" as worst, and normalize that number by the number of times the attribute-level appears in the design. Then Marley and Pihlens (2012) show that, given the maxdiff model, these score differences are a sufficient statistic. Future work should study how these attribute-level NBW scores (and the corresponding ABW scores) are related to maximum likelihood estimates of attribute-level utilities (results in Louviere et al., 2015, suggest each relation will be linear) and how well each score measure predicts (aggregate or individual) choice probabilities.

[9] These fits are consistent with the fact that, if the maxdiff model is the "true" model of the best-worst choices, then an MNL model is an approximation to the (marginal) best (resp., worst) choices (Marley & Louviere, 2005).

[10] Of course, in Case 3, an attribute-level is "chosen" (only) as a consequence of its being a component of a profile chosen on a given choice opportunity.

Acknowledgments

This research has been supported by Natural Science and Engineering Research Council Discovery Grant 8124-98 to the University of Victoria for Marley and by Social Sciences and Humanities Research Council (SSHRC) Grant No. 430060 for Islam. The work was carried out whilst Marley was a Research Professor (part-time) in the Institute for Choice, University of South Australia Business School.

Appendix A

To show that the sum of (4) over all possible patterns allowed by the reasoning leading to Table 1 can be less than 1, assume every binary choice probability is in the open interval $(0, 1)$ and consider a participant making the assumed *independent* choices in each of the eight (four best; four worst) extended choice sets in Table 1. Then one possible pattern of results has the probability

$$b(e,\phi)b(\phi,f)b(\phi,g)b(\phi,h)w(e,\phi)w(\phi,g)w(f,\phi)w(h,\phi).$$

However, this pattern does not arise under the logic of the extended choice sets in Table 1. Therefore, the sum of the probabilities of the terms in (4) over all the patterns allowed by the logic of Table 1 is less than 1 – as they are a subset of all the possible independent patterns. Hence those terms do not correspond to a model of the original best-worst choice data.

Appendix B

B1. The Discrete Choice Experiments (DCEs)

The Laundry Detergent, Toothpaste, Delivered Pizza, and Solar Panel data sets involved DCEs (best-worst Case 3, choice between profiles) using an experimental design suitable for identifying particular forms of indirect utility function. In particular, these studies used a recently proposed approach to DCE design that involves two stages: 1) construct a set of product profiles (i.e., attribute level combinations, product descriptions) using a suitable experimental design to allow identification of particular forms of indirect utility functions of interest (e.g., additive), and 2) use a B(alanced) I(ncomplete) B(lock) D(esign) (BIBD) to assign the profiles to choice sets of fixed size (Louviere et al., 2008). For each of the four product categories, there were 9 attributes, 3 of which had 4 levels and the remaining 6 each had 2 levels. Specifically, the design used an orthogonal main effects plan to design (construct) 16 profiles (choice options); this was a purposive sample of a $4^3 \times 2^6 = 4096$ factorial that allowed estimation of attribute main effects, assuming that all interactions were non-significant. A BIBD then assigned the 16 profiles to 20 choice sets, each of which had four choice options. For each choice set, respondents first selected the best profile from the 4 available options; they then selected the worst option from the remaining 3 options. For the detergent, toothpaste, and pizza data sets there were, respectively, 218, 234, and 186 respondents who each completed four waves of

data collection across a two-year period. The solar data set consisted of 298 respondents who completed a single wave of data collection. In each of these designs, the option selected as best remained on the screen whilst the worst option was being chosen; but at this stage, the best option could not be changed. Further analyses of these data appear in Islam (2014) and Islam and Louviere (2015).

For Budget Saving and Budget Spending (best-worst Case 1, choice between objects), a BIBD was used to assign the 9 attributes of the study to 12 choice sets each having three choice options. For each choice set, respondents first selected the best from the 3 available attributes, then selected the worst from the remaining 2 attributes. The same set of 561 respondents completed the budget saving and budget spending designs. In each of these designs, the option selected as best was removed from the screen before the worst option was chosen. Further analyses of these data appear in Louviere et al (2015, Chap. 2).[11]

B2. In-Sample and Out-Of-Sample Fit to Data Using NBW (Resp., ABW) Scores

Table B1 (resp., B2) shows the RMSE for in-sample prediction of aggregate (marginal) worst (resp., best-worst) choices in a form that mirrors Table 2 of the main text. Table B1 uses *minus* the NBW (resp., *minus* the ABW) scores as the parameters in a multinomial logit (MNL) model. Table B2 estimates the utility difference in a maxdiff model for a pair of options (x, y), $x \neq y$, by the NBW score for x minus the NBW score for y (resp., by the ABW score for x minus the NBW score for y).

[11] This data set was collected at the Centre for the Study of Choice, University of Technology Sydney; we thank those involved in its collection for allowing us to analyse it.

Table B1. RMSE of in-sample prediction of aggregate (marginal) worst choices using the NBW score (resp., ABW score). The RMSE of a set of L predictions (number of choice sets times alternatives in each choice set) is calculated as $RMSE = \sqrt{\frac{\sum_1^L (Worst_{obs} - Worst_{pred})^2}{L}}$. W_i denotes wave i of the design, where $i = 1, 2, 3, 4$.

	Product	Measure	RMSE			
			W_1	W_2	W_3	W_4
Case 3	Detergent	NBW	.097	.091	.098	.092
		ABW	.043	.036	.038	.040
	Toothpaste	NBW	.126	.131	.129	.128
		ABW	.054	.053	.053	.054
	Pizza	NBW	.108	.103	.105	.102
		ABW	.032	.031	.035	.036
	Solar	NBW	.063			
		ABW	.027			
Case 1	Budget Saving	NBW	.027			
		ABW	.017			
	Budget Spending	NBW	.024			
		ABW	.013			

Table B2. RMSE of in-sample prediction of aggregate best-worst choices using the NBW score (resp., ABW score). The RMSE of a set of L predictions (number of choice sets times $m \times (m-1)$ alternatives in each choice set) is calculated as $RMSE = \sqrt{\frac{\sum_1^L (BW_{obs} - BW_{pred})^2}{L}}$. W_i denotes wave i of the design, where $i = 1, 2, 3, 4$.

	Product	Measure	RMSE			
			W_1	W_2	W_3	W_4
Case 3	Detergent	NBW	.051	.048	.053	.053
		ABW	.037	.031	.034	.036
	Toothpaste	NBW	.064	.066	.066	.063
		ABW	.043	.045	.046	.045
	Pizza	NBW	.053	.053	.054	.053
		ABW	.031	.031	.034	.034
	Solar	NBW	.040			
		ABW	.028			
Case 1	Budget Saving	NBW	.031			
		ABW	.022			
	Budget Spending	NBW	.025			
		ABW	.017			

We also examined out-of-sample prediction error according to RMSE. We used multiple time origins (i.e., waves) to produce one, two, and three step ahead predictions. The first time origin is wave 1 and from this origin one step ahead (six months) to three steps ahead (one and half years) predictions are made. The time origin is advanced by one wave and a set of one and two step ahead predictions are generated. Finally, the origin is advanced to wave 3 and a final one step ahead prediction is generated. Therefore, from the three data sets that each have four waves we can generate three one-step, two two-step, and one three-step ahead predictions. Tables B3, B4 and B5 show the mean RMSE averaged over the three one-step (resp., two two-step, and one three-step) ahead predictions according to NBW and ABW scores for each data set, separately for (marginal) best, (marginal) worst, and best-worst choices. The NBW and ABW scores are used in Table B3 (resp., B4, B5) in exactly the way they are used in Table 2 (resp., Table B1, B2).

Table B3. RMSE for out-of-sample prediction of aggregate (marginal) best choices using the NBW score (resp., ABW score).

Product	Measure	RMSE		
		One step	Two step	Three step
Detergent	NBW	.098	.099	.096
	ABW	.049	.055	.060
Toothpaste	NBW	.113	.115	.111
	ABW	.061	.065	.067
Pizza	NBW	.103	.102	.107
	ABW	.048	.050	.053

Table B4. RMSE for out-of-sample prediction of aggregate (marginal) worst choices using the NBW score (resp., ABW score).

Product	Measure	RMSE		
		One step	Two step	Three step
Detergent	NBW	.098	.097	.103
	ABW	.049	.053	.058
Toothpaste	NBW	.131	.134	.131
	ABW	.060	.065	.065
Pizza	NBW	.110	.110	.116
	ABW	.049	.047	.049

Table B5. RMSE for out-of-sample prediction of aggregate best-worst choices using the NBW score (resp., ABW score).

Product	Measure	RMSE		
		One step	Two step	Three step
Detergent	NBW	.052	.050	.052
	ABW	.036	.037	.040
Toothpaste	NBW	.067	.069	.072
	ABW	.045	.044	.043
Pizza	NBW	.054	.053	.053
	ABW	.035	.033	.034

References

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

Flynn, T. N. & Marley, A. A. J. Best-worst scaling: practice and theory. Invited chapter in S. Hess & A. Daly (Eds.) *Handbook of Choice Modelling*. Edward Elgar Publishing, 2014, pp. 178-201.

Giergiczny, M., Chintakayala, P., Dekker, T., & Hess, S. (2013). Testing the consistency (or lack thereof) between choices in best-worst surveys. Paper presented at the *Third International Choice Modelling Conference*, Sydney, July 3-5.

Greenacre, L., Dunn, S., & Mocanu, A. (2015). Heterogeneity in the consistency of best-worst scale responses. *Australian Marketing Journal*. 23, 227-234.

Hawkins, G. E., Marley, A. A. J., Heathcote, A., Flynn, T. N., Louviere, J. J., & Brown, S. D. (2014). The best of times and the worst of times are interchangeable. *Decision*, 1, 192-214.

Islam, T. (2014). Household level innovation diffusion model of photo-voltaic (PV) solar cells from stated preference data. *Energy Policy*, 65, 340-350.

Islam, T., & Louviere, J. J. (2015). The stability of aggregate-level preferences in longitudinal discrete choice experiments. In J. J. Louviere, T. N. Flynn & A. A. J. Marley (2015). *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge University Press. Cambridge: UK. Chap. 13, pp. 265-277.

Lipovetsky, S., & Conklin, M. W. (2014). Best-worst scaling in analytical closed-form solution . *The Journal of Choice Modelling*, 10, 60-68.

Louviere, J. J. (2015). Using alternative-specific DCE designs and best and worst choices to model choices. In J. J. Louviere, T. N. Flynn & A. A. J. Marley (2015). *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge University Press. Cambridge: UK. Chap. 15, pp. 297-315.

Louviere, J. J., Flynn, T. N. & Marley, A. A. J. (2015). *Best-Worst Scaling: Theory, Methods and Applications. Cambridge University Press. Cambridge: UK.*

Louviere, J. J., Street, D. J., Burgess, L., Wasi, N., Islam, T., & Marley, A. A. J. (2008). Modelling the choices of single individuals by combining efficient choice experiment designs with extra preference information. *Journal of Choice Modelling*, 1, 128-163.

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.

Rose, J. M. (2014). Interpreting discrete choice models based on best-worst data: A matter of framing. *Transportation Research Board 93rd Annual Meeting. Washington, DC. Jan. 12-16.*