WORKING PAPER

ITLS-WP-13-22

Interpreting discrete choice models based on best-worst data: A matter of framing.

By
John M. Rose

November 2013

ISSN 1832-570X

INSTITUTE of TRANSPORT and LOGISTICS STUDIES

The Australian Key Centre in
Transport and Logistics Management

The University of Sydney

Established under the Australian Research Council's Key Centre Program.

NUMBER: Working Paper ITLS-WP-13-22

TITLE: **Interpreting discrete choice models based on best-worst data: A matter of framing.**

ABSTRACT: Best worst choice response tasks have become increasingly popular as a means of increasing the amount of information captured from respondents undertaking stated preference experiments. In analysis, best worst data is often exploded to provide additional pseudo observations which may aid in model estimation. Recent studies however have questioned many of the underlying assumptions which typically accompany best worst studies, such as the symmetry of preferences across the best and worst responses as well as assumptions about equal error variances across the two response types. This paper first provides a detailed description of the various best worst tasks that have appeared within the literature before arguing that violations of preference symmetry and homogeneity of error variance should be the norm. This is because in asking respondent to choose their most and least preferred option out of a set of alternatives reflects different response frames, one positive and one negative, and behaviourally there exists no reason why one would assume that the preferences (and error variances) obtained from one type of question should precisely mirror that of the other. Using an empirical case study, the impact of the framing of these questions is examined. Finally, the argument put forward is that best-worst data should be treated in a similar to data fusion methods, where one combines two different sources of discrete choice data.

KEY WORDS: ***Best worst; question framing effects; data fusion; model interpretation.***

AUTHORS: **Rose**

CONTACT:

INSTITUTE of TRANSPORT and LOGISTICS STUDIES (C13)
The Australian Key Centre in Transport and Logistics Management

The University of Sydney NSW 2006 Australia

Telephone: +612 9114 1824
Facsimile: +612 9114 1722
E-mail: business.itlsinfo@sydney.edu.au
Internet: http://sydney.edu.au/business/itls

DATE: November 2013

1. Introduction

Stated preference methods based on discrete choice surveys have traditionally relied upon respondents being asked to select their single most preferred option out of a finite and mutually exclusive set of alternatives presented to them as part of the survey task. By varying in some way the alternatives and observing how respondents choices change (or do not change), discrete choice methods seek to establish whether there exists any link between what is being varied and the preferences respondents hold for the various alternatives under consideration. For choice tasks consisting of more than two alternatives, asking respondents to select their single most preferred alternative provides only partial ranking information as to the respondent's preferences for the full set of alternatives. By limiting the amount of information captured from each choice task, it is likely that a greater number of choice tasks will be required in order to be able to model to a desired level of precision any link between what is being varied and respondents preferences.

Previous research has sought to capture more information from each choice task by asking respondents to rank all or a subset of the alternatives presented as opposed to simply indicating their single most preferred alternative (e.g., Beggs et al. 1981, Ben-Akiva et al. 1991, Chapman and Staelin 1982, Hausman and Ruud 1987). Data from rank-order choice experiments is then typically analysed using a rank-ordered multinomial logit model or exploded logit model, where each ranking of a choice task is converted into a number of independent pseudo observations. More recently, researchers have employed what are known as best worst response tasks to obtain full or partial rankings data. In best worst tasks, respondents are asked to indicate out of the set of possible alternatives, what is their most and least preferred options.

Whilst the information captured from rankings and best worst response tasks is similar, those advocating the use of best worst responses suggest that asking respondents to indicate their most and least preferred alternatives out of a set is less difficult than asking respondents to provide rankings. This is because respondents are more likely to rank alternatives in order of their preference from most preferred to least preferred but only have to consider the most extreme alternatives in a best worst task. In considering alternatives that might be close in preference and having to rank these, so the argument goes, respondents have to expand more cognitive effort than having to deal with alternatives that are at polar opposites in terms of preferences (e.g., Auger et al. 2007, Cohen 2009, Flynn et al. 2007 or Islam 2008).

Whilst not comparing the results obtained from rankings and best worst data, past studies have examined potential issues related to both response mechanisms. Ben-Akiva et al. (1991) and Hausman and Ruud (1987) for example examined the potential for respondents to exhibit different preferences over the various pseudo ranking observations, as well as possible differences in error variances. Collins and Rose (2011) and Scarpa et al. (2011) similarly looked at potential error variance differences between the best and worst choice observations, whilst Giergiczny et al. (2013) examined both error variance and preference differences between the two choices. In all cases, differences where sought were found to exist.

Whilst it is true that best worst and rankings tasks provide similar information in terms of partial or full preference rankings of the alternatives, the implications for detecting preference and/or error variance differences between the two approaches however is not the same. Rankings tasks are designed with the sole purpose of data augmentation. Their purpose is to provide a greater number of choice observations in order to produce more robust parameter estimates for a given sample size (see e.g., Vermeulen et al. 2011). The presence of preference differences across rankings tasks therefore is particularly problematic, in that the analyst cannot simply pool the data to obtain a single vector of preference weights, and hence will likely obtain different outputs of interest including marginal willingness to pay and elasticity estimates over the alternate pseudo ranking observations. This leaves the analyst in a dilemma, having to choose which outputs to use and report post estimation. Whilst it might be tempting to simply make use of the parameters associated with the first ranked alternative, the question then arises as to why the rankings were collected in the first place. From the analysts'

perspective, the issue of error variance differences is less of a concern. Numerous econometric models are available to account for error variance differences, and provided these are accounted for correctly, the preference weights, and other associated outputs which are of interest, should be unbiased.

Whilst in the past, best worst data have been primarily used for the same purpose, we argue that best worst data should be consider much more than simply a data augmentation tool. Indeed, we argue that the best and worst choices reflect different response frames, one positive and one negative, and as such, there exists no reason why one would assume that the preferences (and error variances) obtained from one type of question should precisely mirror that of the other. This is similar to the framing effects described by Prospect Theory (Kahneman and Tversky 1979) which states in part that individuals will have asymmetrical preferences around positive and negatively framed questions. As such, any differences in the outputs of best and worst choice questions should be interpreted as differences arising from framing effects, and not as a problem with the technique itself. Given the above interpretation, we argue that the best worst method should be used as a data augmentation technique if and only if the preferences are found to be symmetrical for the best and worst responses, after accounting for any possible error variance differences.

The remainder of this paper is organised as follows. In the next section, we describe three cases of best worst response mechanisms that have been used within the literature. As well as discussing each case, we provide a brief treatise on the specifics of the data formats required for analysis. Next, specific assumptions made about the analysis of best worst data are discussed before an empirical example is used to examine these assumptions and their impact upon how to properly interpret models obtained from best worst data. Finally, concluding comments are provided.

2. Best worst examples

Prior work has recognised three unique approaches to best worst survey response mechanisms. In case 1, respondents are asked to choose the most and least preferred object from a set of objects (e.g., Louviere et al. 2013). In case 2, the task consists of respondents viewing a set of attributes, each described by a series of attribute levels, and being asked to select the most and least preferred attribute or level out of the set shown (e.g., Beck et al. 2013). Case 3 involves respondents viewing a set of alternatives, each described by a number of attributes and levels, and being asked to select the best and worst alternative from those shown (e.g., Rose and Hensher 2013). In this section, we discuss with examples each case in more detail, including how the data format requirements.

2.1 Best worst case 1

Best worst case 1 involves respondents being presented with a number of alternatives and being asked to select their most and least preferred alternative out of the set shown. In case 1, only the names of the alternatives (mode, brand, etc.) are shown with no other information about possible attributes and associated attribute levels presented (see Figure 1 for an example). Respondents complete multiple tasks in which different subsets of alternatives are presented, where typically the number of alternatives presented in each task is fixed across the entire experiment. As such, best worst case 1 is analogous to the original availability type experiments used in the 1980s and 90s, differing only in assuming a fixed choice set size and that information as to the least preferred alternative is also captured (e.g., Lazari and Anderson 1994, see Rose et al. 2013 for a review of such experiments).

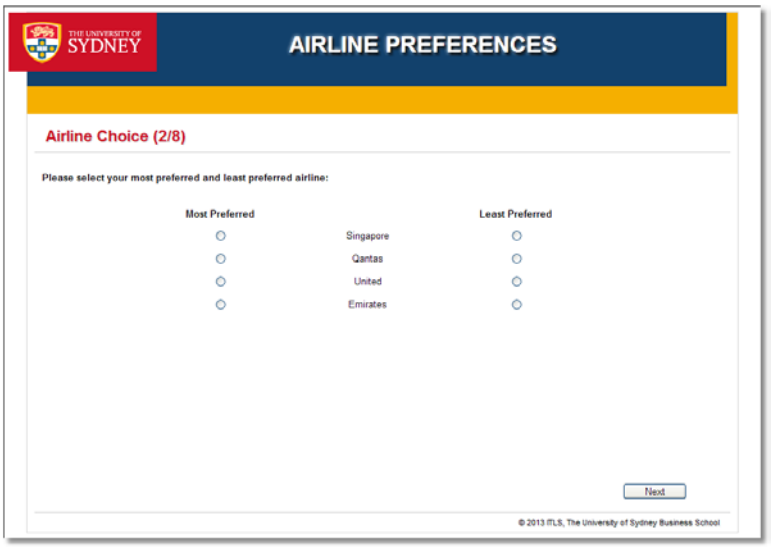

Figure 1: Example case 1 best worst task

For the example shown in Figure 1, each respondent was asked to complete eight tasks, providing 16 choices observations in total (eight best and eight worst). In each task, respondents saw sets of four airlines out of a total of 16, with the four airlines presented in each task determined by an underlying experimental design. Within each task, an airline may appear only once (e.g., Singapore Airlines will never appear against itself in the same task) however across tasks, the positioning of an alternative may differ (e.g., Singapore Airlines appears first in Figure 1, however it may appear in any position the next time it is present within a task).

The data format for case 1 is similar to that of an unlabeled choice experiment. The alternatives in effect reflect a specific position within the task (e.g., top, second from the top, second from the bottom, bottom), whilst the objects, in this case the various airlines, are represented as the attributes. For each task, two observations are constructed, one reflecting the best choice, and one the worst. For the best choice observations, the objects presented are dummy coded 1 if the alternative is present or 0 otherwise. For the worst choice task, the variables are simply the negative of the best values (i.e., -1, if the alternative is present or 0 otherwise). It is suggested that in setting up the data for the second answered question in a task, most commonly assumed to be the worst choice, that the alternative chosen first (say the best) be removed from for the second pseudo observation. This is because the probability of choosing the same alternative simultaneously as being both the best and worst present should be zero, however if all alternatives are present in both observations, then it is likely that the model will naively assign a non zero probability to this outcome.

2.2 Best worst case 2

In case 2, respondents are asked to evaluate combinations of attributes, the levels of which are determined by some underlying experimental design. As such, unlike case 1 where the objects being evaluated are labelled alternatives, in case 2, the objects being evaluated are attributes, the combinations of which form a single generic ‘alternative’. Rather than choose from a set of labelled alternatives, respondents are then asked to select for the generic ‘alternative’, which attribute is the most preferred and least preferred. One way in which case 2 has been used in the past is to gather attitudinal type questions, where the generic ‘alternative’ represents the respondent’s attitude towards that particular object (the generic ‘alternative’). Figure 2 presents an example of a case 2 choice task in which respondents were asked their perceptions as to how ready Australian households are to make more sustainable travel behaviour choices, and what government incentives might induce such a change in the future.

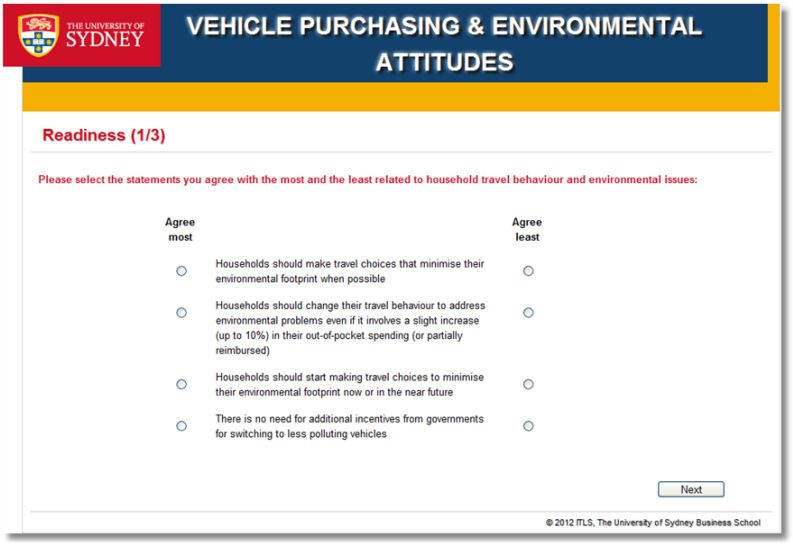

Figure 2: Example case 2 best worst task

In the example experiment shown, respondents saw three tasks consisting of four attributes, each described by three levels (the attributes and levels are presented in Table 1). In the survey, the location of each attribute was fixed across the three tasks with only the levels varying, although it should be noted that best practice would be to randomize the location of each attribute across tasks. An experimental design was then used to select the attribute levels that where shown in any given task.

Unlike case 1, the data for case 2 surveys should be treated similar to data collected from a labelled choice experiment. Each 'attribute' from a case 2 task maps to a specific alternative in the data (e.g., in the above example, the first alternative represents the attribute, *change in behaviour*, the second *change in spending*, etc.), whilst the variables within the data relate to the levels of the attributes presented (e.g., for the first alternative, *change in behaviour*, the three levels shown in Table 1 will represent the variables to be modelled). As with case 1, the variables (each attribute level) are dummy coded 1 if the level is present in a given task or 0 otherwise. For the worst choice task, the variables are simply the negative of the best values (i.e., -1, if the level is present or 0 otherwise). Further, similar to case 1, the alternative observed to have been chosen first for a given task (often assumed to be the best), should be removed when setting up the observation for the counter choice (the worst if the respondent first provided information as to their most preferred alternative).

Table 1: Attributes and attribute levels of example best worst task 2 exercise

Level	Change in behaviour	Change in spending	Delayed change	Government incentives
1	Households cannot be expected to further change their travel habits	Households should change their travel behaviour to help address environmental issues when change does not cost them anything (or is fully reimbursed)	There is no need for households to adjust their travel behaviour to minimise their environmental footprint within the foreseeable future	There is no need for additional incentives from governments for switching to less polluting vehicles
2	Households should make travel choices that minimise their environmental footprint when possible	Households should change their travel behaviour to address environmental problems even if it involves a slight increase (up to 10%) in their out-of-pocket spending (or partially reimbursed)	Households should be ready to adjust their travel behaviour to minimise their environmental footprint within 5-10 years	Governments should provide some non-financial incentives (parking priority, etc.) to households switching to less polluting vehicles
3	Households should make major changes to their day-to-day travel behaviour	Households should change their travel behaviour to address environmental problems even if it involves moderate increases (10-20%) in out-of-pocket spending	Households should start making travel choices to minimise their environmental footprint now or in the near future	Governments should compensate households for excess expenditure associated with responsible travel choices

2.3 Best worst case 3

Best worst case 3 differs substantially from cases 1 and 2. Case 3 corresponds closely with traditional discrete choice experimental (DCE) tasks in that respondents observe a series of alternatives, each described by a set of attributes which are further described by attribute levels. Unlike traditional DCEs however, respondents are not asked to select only their most preferred alternative, but rather their most preferred and least preferred out of the set shown. For experiments involving three alternatives, capturing information as to the most and least preferred alternatives will provide full preference rankings. For experiments with four or more alternatives, only partial preference rankings will be captured. As such, follow up questions asking for second best, second worst, etc., are often asked to provide full preference ranking information. An example of a best worst case 3 experiment involving five capturing full preference rankings for five alternatives is presented in Figure 3.

Several potential data formats are possible for case 3 experiments, each modelling different possible decision processes. Early experiments using case 3 tended to treat the data as if the experiment were binary, exploding the data to create pseudo observations for all possible pairwise combinations of alternatives (see e.g., Marley and Islam 2012 for a discussion of this approach as applied to best worst data, or Beggs et al. 1981 for traditional rankings tasks). For example, consider an experiment involving five alternatives, where the respondent selected alternative 2 as being the most preferred, alternative 4 as the second most preferred, alternative 1 as the third most preferred and alternative 4 as the fourth most preferred (hence alternative 3 is the least preferred). Panel (a) of Figure 4 demonstrates this data structure. More recent studies tend to explode the data by retaining all alternatives minus the previous selected best (or worst) alternative in the subsequently constructed pseudo choice observation. Even within such an approach, differences exist, reflecting the pattern of how the choices were made within the task. For example, if a respondent is thought to choose the best alternative, then the second best alternative, followed by the next best alternative, etc. then the data might be set up in a more traditional rank explosion process, such that the variables for the observation reflecting the next best choice are coded exactly the same as those for the previous choice observation, with the only exception being the deletion of the previously selected best alternative (panel (b) of Figure 4). If on the other hand, respondents are believed to have selected the best alternative, followed by the worst, followed by the second best, etc. then the rank explosion process differs yet again. In this case, the pseudo observations for the worst cases remove the previously

chosen best alternative, whilst the values for the variables are the negatives of the best values. The next pseudo observation reflecting the second best choice then removes the previously chosen best and worst alternatives, whilst restoring the signs of the variables. This process is then repeated until only two alternatives remain (see panel (c) of Figure 3). It is possible to construct a hybrid rank explosion scheme in cases where respondents select the best and worst alternatives in a less systematic way. In any case, it is advisable that analysts record the precise pattern in which the choices were made; otherwise an assumption will need to be made in order to set-up the data.

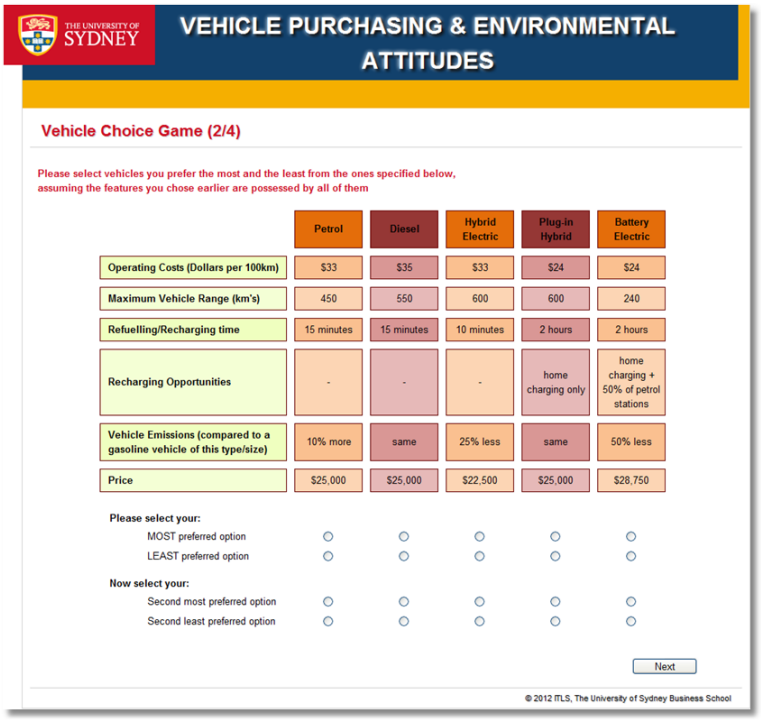

Figure 3: Example case 3 best worst task

3. Modelling assumptions

As indicated in the previous section, when respondents are asked to choose their best and worst alternatives from a set of objects, the typical data format adopted assumes that the variables for the worst alternative be coded as the negative of the best, such that $x_{nsjk|W} = -x_{nsjk|B}$. Two exceptions to this data format exist for the best worst case 3 format, involving the pairwise explosion (panel (a) of Figure 4) and the approach approximating the traditional rank explosion data construction method where respondents are assumed to answer the questions asked systematically in the order of their most preferred to least preferred alternative (panel (b) of Figure 4). Both of these data structures assume that respondents treat the task as a traditional rankings exercise, in contrast to the third data structure which assumes that respondents provide their answers in a sequence of best then worst choices (panel (c) of Figure 4). It is this last data structure, and its corresponding assumed underlying decision process, that this current paper is most concerned with.

When respondents are assumed to make their choices in a best then worst sequences of choices, then the coding of the variables for the worst alternative as the negative of the best implies a sign reversal for the parameters across the best and worst responses as $\beta_{k|W} x_{nsjk|W} = \beta_{k|W} \left(-x_{nsjk|B}\right) = -\beta_{k|W} x_{nsjk|B}$.

Behaviourally, this implies that an attribute that makes an alternative more attractive relative to other alternatives present in a choice situation (and hence increase the probability that that alternative will be selected as best), is less likely to make it unattractive if a respondent were asked to select their least

preferred alternative from the same set of alternatives. Such an assumption holds for all best worst cases.

Although not strictly necessary, most studies involving best worst data make further assumptions about the preference parameters as well as about the error variances exhibited across the best worst choices. An often made assumption within the literature is that the magnitude of the parameters for the best and worst choices are equal such that $\left|\beta_{k|W}\right| = \left|\beta_{k|B}\right|$. Such a strong assumption is necessary if and only if the best worst task is being used for the purposes of data augmentation. That is, the analyst may collect best worst data to obtain more information per individual respondent without having to increase the overall number of choice tasks shown to any given respondent. This might be motivated by a desire to either obtain enough observations to estimate individual specific models (e.g., Louviere et al. 2013) or simply as a data enrichment process to add observations in order to aid in the estimation of a traditional choice type model estimated on pooled data (e.g., Rose and Hensher 2013). In the former case, there likely won't exist enough observations to obtain meaningful estimates at an individual level that differ for the best and worst choices. In the later case, the additional pseudo observations are used purely to enrich the existing best choice, with the aim of increasing the precision of the parameter estimates. Although testable, with the exception of a single study (Giergiczny et al. 2013), relaxation of this assumption has been tended to be overlooked within the literature, most likely as a result of how models of best worst have been interpreted in the past.

	A	B	C	D	E
1	Pseudo observation	Alternative	choice	X1	X2
2	1	1	0	1	2
3	1	2	1	8	9
4	2	1	1	1	2
5	2	3	0	4	5
6	3	1	0	1	2
7	3	4	1	2	0
8	4	1	1	1	2
9	4	5	0	1	5
10	5	2	1	8	9
11	5	3	0	4	5
12	6	2	1	8	9
13	6	4	0	2	0
14	7	2	1	8	9
15	7	5	0	1	5
16	8	3	0	4	5
17	8	4	1	2	0
18	9	3	0	4	5
19	9	5	1	1	5
20	10	4	1	2	0
21	10	5	0	1	5

(a)

	A	B	C	D	E	F	G
1	Set	Explosion	Altij	# alternatives present	Choice	X1	X2
2	1	1	1	5	0	1	2
3	1	1	2	5	1	8	9
4	1	1	3	5	0	4	5
5	1	1	4	5	0	2	0
6	1	1	5	5	0	1	5
7	1	2	1	4	0	1	2
8	1	2	3	4	0	4	5
9	1	2	4	4	1	2	0
10	1	2	5	4	0	1	5
11	1	3	1	3	1	1	2
12	1	3	3	3	0	4	5
13	1	3	5	3	0	1	5
14	1	2	3	2	0	4	5
15	1	2	5	2	1	1	5

(b)

	A	B	C	D	E	F	G	H
1	Set	Explosion	Best worst	Altij	# alternatives present	Choice	X1	X2
2	1	1	1	1	5	0	1	2
3	1	1	1	2	5	1	8	9
4	1	1	1	3	5	0	4	5
5	1	1	1	4	5	0	2	0
6	1	1	1	5	5	0	1	5
7	1	2	-1	1	4	0	-1	-2
8	1	2	-1	3	4	1	-4	-5
9	1	2	-1	4	4	0	-2	0
10	1	2	-1	5	4	0	-1	-5
11	1	3	1	1	3	0	4	2
12	1	3	1	4	3	1	2	0
13	1	3	1	5	3	0	1	5
14	1	2	-1	1	2	0	-1	-2
15	1	2	-1	5	2	1	-1	-5

(c)

Figure 4: Alternate data formats for case 3 best worst tasks

The second assumption, which has received much wider attention within the literature, is the assumption that the error variances will be constant between the best and worst choice exercises. Despite this assumption, a number of studies have found differences between the error variances between the various choices made (e.g., Collins and Rose 2011, Giergiczny et al. 2013, Scarpa et al. 2011). Given the current state of econometric modelling, any empirical differences between the variances of the best and worst choices is less of an issue, as these can (although often are not) be taken into account during the modelling process.

4. Empirical case study

To demonstrate the correct interpretation of models estimated on best worst data, we make use of data collected using the best worst case 2 survey task described previously. As part of a wider survey dealing with alternative fuelled vehicles, 204 respondents completed three best worst case 2 tasks related to their beliefs about how ready Australian households are to adopt more environmentally friendly travel behaviour and what government incentives might be required to induce such changes. A screen capture of an example survey task is provided in Figure 2, whilst the attributes and attribute levels are provided in Table 1. A Bayesian D-efficient design was generated under the assumption of uniform distributed prior parameters which was used to allocate the attribute levels to each of the choice tasks and the final design consisted of 12 choice tasks, blocked into four blocks of size three. The data were collected from households living in Sydney Australia drawn from the Pure Profile panel (www.pureprofile.com) in October 2011. An overview of the socio-demographics of the final sample is provided in Table 2.

Table 2: Descriptive statistics of the socio-demographic characteristics of the sample

	Household size	**Income**	**Gender**	**Age**
Average	3.57	52.33	0.52	46.03
Median	3.00	40.00	1.00	45.00
Std dev.	1.25	35.51	-	14.88

Table 3 reports four models estimated on the data using both Nlogit and Python BIOGEME (Bierlaire, 2003). The use of two software packages was purely done as a cross validation exercise. The first model reported, M1, is an MNL model assuming symmetrical preferences and constant error variances across the best and worst choice questions. Models 2 (M2) and 3 (M3) are heteroskedastic MNL (HMNL) models assuming symmetrical preferences, but allowing for scale differences (where scale is inversely related to the error variance of the model) between the best and worst choice observations (e.g. Dellart et al. 1999; Hensher et al. 1999; Louviere et al. 2000; Swait and Louviere 1993). The utility specification of the HMNL is given as

$$U_{nsj} = \left(1 + \sum_{q}^{Q} \delta_q w_q\right) \sum_{k=1}^{K} \beta_k x_{nsjk} + \varepsilon_{nsj} \qquad (1)$$

and where δ_q is a parameter associated with covariate w_q. The 1 in Equation (1) is required if $\sum_{q}^{Q} \delta_q w_q$ enters the equation linearly as shown above (see e.g., Dellart et al. 1999).

In Model M2, w_q is assumed to be a constant whereas in Model M3, w_q represents the household size and age of the respondent. The final model, Model 4 is also a HMNL model, however in this model, asymmetry in preferences across the best and worst choices was allowed for. In Models 2 to 4, the scale was kept constant for the worst choices, hence the resulting estimates reflect the scale of the best choice observations relative to the worst.

For Model M2, the scale parameter was found to be statistically insignificant which would suggest that there exists no differences in the error variances between the best and worst choice tasks under

the assumption of symmetrical preferences, however when scale is made a function of age and household size, as in Model M3, the model suggests that older respondents have a lower scale (increased error variances) for the best choices relative to their worst choices, whilst respondents who belong to larger households have a higher scale (lower error variances) for the best choices relative to the worst choices when compared to respondents from households with smaller numbers of residents.

Similar to M3, model M4 allows scale to be a function of covariates. In this model, however, only age was found to be statistically significant. Unlike models M1 to M3, the preferences for the best and worst choices were allowed to be asymmetrical in magnitude. The approach adopted here treats the best and worst choices as if they are two separate data sets with common sets of attributes and applies similar methods as those used to combine different discrete choice data. By forcing at least one parameter to be generic (symmetrical in this case) across the two choices, scale differences alongside preference asymmetry can be explored for the remaining attributes. T-tests, Wald tests, or log-likelihood ratio tests are then applied to the best and worst parameters to test for differences. Where no differences are found, the model is re-estimated assuming that preferences are symmetrical across the best and worst choices for that attribute. After extensive testing, only one attribute level, '*Households should make travel choices that minimise their environmental footprint when possible*' associated with the *Change in behavior* attribute was found to be statistically different in magnitude across the best and worst choice observations. In this case, the attribute level was found to be statistically significant for the best choices but insignificant for the worst choices (see the bolded values in Table 2).

For all four models, alternative specific constants (ASC) were estimated for the first three alternatives representing the attributes, *Change in behavior*, *Change in spending*, and *Government incentives*. These ASCs therefore reflect the average of the unobserved effects for the attributes, relative to the last attribute *Government incentives*, as well as any top to bottom bias that might exist given the non-rotation of the attributes across the choice tasks. The negative and statistically significant ASC for the *Change in spending* attribute for example suggests that after accounting for any level effect, the model predicts respondents to be less likely to select this attribute as the best alternative, and more likely as the worst, all else being equal.

Table 3: Model results

	M1: MNL		**M2: HMNL 1**		**M3: H MNL 2**		**M4: HMNL 3**	
	Par.	(***rob. t*-rat.**)	**Par.**	(***rob. t*-rat.**)	**Par.**	(***rob. t*-rat.**)	**Par.**	(***rob. t*-rat.**)
Scale	-	-	-0.013	(-0.06)	-	-	-	-
Age	-	-	-	-	-0.013	(-2.26)	-0.007	(-2.32)
Household size	-	-	-	-	0.167	(1.91)	-	-
Attitude attributes								
Change in behaviour	-0.106	(-1.28)	-0.106	(-1.27)	-0.084	(-1.02)	-0.253	(-2.09)
Households cannot be expected to further change their travel habits	-0.383	(-2.29)	-0.386	(-2.23)	-0.347	(-2.02)	-0.493	(-2.44)
Households should make travel choices that minimise their environmental footprint when possible	0.702	(4.27)	0.707	(3.48)	0.743	(4.02)	0.817	(4.03)
Households should make major changes to their day-to-day travel behaviour	-0.015	(-0.09)	-0.014	(-0.08)	0.058	(0.33)	-0.016	(-0.08)
Change in spending	-0.282	(-3.27)	-0.282	(-3.24)	-0.275	(-3.24)	-0.43	(-3.43)
Households should change their travel behaviour to help address environmental issues when change does not cost them anything (or is fully reimbursed)	0.698	(4.21)	0.702	(3.57)	0.698	(3.85)	0.770	(3.79)
Households should change their travel behaviour to address environmental problems even if it involves a slight increase (up to 10%) in their out-of-pocket spending (or partially reimbursed)	0.079	(0.48)	0.079	(0.48)	0.119	(0.74)	0.048	(0.24)
Households should change their travel behaviour to address environmental problems even if it involves moderate increases (10-20%) in out-of-pocket spending	0.101	(0.56)	0.101	(0.56)	0.160	(0.88)	0.055	(0.26)
Delayed change	-0.438	(-4.88)	-0.44	(-4.60)	-0.419	(-4.53)	-0.65	(-4.56)
There is no need for households to adjust their travel behaviour to minimise their environmental footprint within the foreseeable future	-0.482	(-2.88)	-0.487	(-2.61)	-0.456	(-2.50)	-0.668	(-3.13)
Households should be ready to adjust their travel behaviour to minimise their environmental footprint within 5-10 years	0.316	(1.90)	0.318	(1.82)	0.385	(2.18)	0.319	(1.60)
Households should start making travel choices to minimise their environmental footprint now or in the near future	0.506	(2.74)	0.508	(2.63)	0.559	(3.01)	0.484	(2.15)
Government incentives	0	-	0	-	0	-	0	-
There is no need for additional incentives from governments for switching to less polluting vehicles	-0.608	(-3.52)	-0.610	(-3.59)	-0.574	(-3.34)	**-1.490**	**(-2.98)**
Governments should provide some non-financial incentives (parking priority, etc.) to households switching to less polluting vehicles	0.273	(1.58)	0.276	(1.46)	0.322	(1.80)	0.344	(1.67)
Governments should compensate households for excess expenditure associated with responsible travel choices	0	-	0	-	0	-	0	-
Attitude attributes (specific to worst choices only)								
There is no need for additional incentives from governments for switching to less polluting vehicles	-	-	-	-	-	-	**-0.346**	**(-1.54)**
Model fit								
LL (β)	-1441.083		-1441.081		-1438.015		-1437.060	
LL(0)	-1507.430		-1507.430		-1507.430		-1507.430	
ρ^2	0.044		0.044		0.046		0.047	
Adj. ρ^2	0.033		0.032		0.033		0.034	
Normalised AIC	2.378		2.379		2.376		2.374	
Normalised BIC	2.436		2.442		2.443		2.441	
Sample								
Number of respondents	204		204		204		204	
Number of observations	1224		1224		1224		1224	

The remaining estimates reflect the marginal utilities associated with each of the statements. Given that all variables within the data are coded as 0 and 1 for the best choices, and 0 and -1 for the worst, the parameters are directly comparable. For attributes associated with a negatively signed coefficient, *holding all else equal*, the probability that that attribute will be selected as being the best decreases as the magnitude of the coefficient increases, whilst the probability that the same attribute will be chosen as the worst attribute out of the set increases. It is worth noting that given the non-linear nature of the logit model, the choice probabilities for an attribute for both best and worst outcomes will not be symmetrical.

Perhaps the easiest way to understand and interpret the estimates derived from best worst type data is to examine the choice probabilities derived under the two different data assumptions associated with the best and worst choice frames (i.e., $x_{nsjk|B}$ and $x_{nsjk|W} = -x_{nsjk|B}$). Table 4 presents the choice probabilities for the best and worst choice outcomes computed under two different sets of four statements for all four models. In the first scenario, we compute the choice probabilities assuming the first statement for all four attributes are present, whilst in the second scenario, keeping the first statement for the first attribute, we recompute choice probabilities assuming the second statement for the attributes two and three and the last statement for the fourth attribute (see Table 1 for the statements). In scenario one, the probability that the first statement will be selected as best ranges from 0.2 to 0.225 depending on the model used, whilst the same statement has a probability of being selected as worst ranging from 0.242 to 065. As can be seen in scenario 2, in the presence of an alternate set of statements, the probability that the same statement will be chosen as the best statement drops to 0.179 to 0.187, whilst the probability that it will be selected as the worst increases to between 0.325 and 0353. Despite providing a range of values, the analyst in reality would select one model and hence one set of outcomes. In the current case, based on a log-likelihood ratio test, model M4 is statistically a better model fit for the data relative to models M1 and M2, and is a slight improvement on model M3 in terms of the AIC and BIC statistics (it has the same number of parameters as M3 and hence a formal statistical test cannot be completed to compare these last two models).

Rather than compute the choice probabilities under various scenarios, an alternative is to compute the marginal effects (or elasticities for continuous variables assuming case 3) and work directly with these. The marginal effects for each statement are reported in Table 5 for the best and worst choices for all four models. The marginal effects reflect how much the choice probabilities can be expected to change over the sample as a percentage given a unit change in the variable of interest, holding everything else constant. Once more, the analyst would need to select a specific model and rely on only one set of outputs, however what is clear from this exercise is that their exists asymmetry between the best and worst choices in terms of the impacts of an attribute level being chosen as best or worst.

It is worthwhile noting that for model M4, the fact that the preference parameters are allowed to be asymmetrical for some of the attributes across the best and worst choices does not change the interpretation of the marginal effect results. Best worst data should be interpreted in terms of both questions frames, and not just from the perspective of the most preferred question. Given the non-linear nature of the logit model, it matters little whether the parameters are symmetrical or not across the best and worst choices in terms of the impact upon the modelled choice probabilities. Where concerns are warranted with the finding of asymmetrical preferences however is only when the technique is used purely as a data augmentation technique, else such a finding reflects a question framing effect.

5. Concluding remarks

This paper explores issues surrounding the use of best worst choice data. As well as detailing the various best worst response formats, we have argued that best worst choice data may potentially have been misused in the past. The emphasis upon using best worst data appears to have been on treating the data as augmentation tool, that is, to provide additional data for modelling purposes. As argued herein, best worst data should only be used for this purposes when preferences for the best worst

responses can be shown to be symmetrical, and then only after accounting for any potential scale differences. In this light, best worst data should be treated as if they are two separate data sets, and methods commonly used to combine data discrete choice data sets should be applied to such data. Unfortunately, most studies do not undertake such tests, or at least they are not reported. We therefore make a similar call to that made by Giergiczny et al. (2013) who in finding preference asymmetries as well as scale differences when analysing four different best worst case 3 data sets concluded that their paper acts as "a clear warning to the continued reliance on BW approaches without questioning the consistency...". Where this paper differs to the conclusions of Giergiczny et al. (2013 however is in concluding that the presence of any differences is not an issue with the method itself if the model results are used to explore framing effects, only if one uses the technique to naively pool the data and explore simply the best choice outcomes.

Table 4: Choice probabilities for different combinations of statements based on Model M1

(a) Scenario 1

	M1			**M2**			**M3**			**M4**		
Attribute and statement	Par.	Prob(best)	Prob(worst)	Par.	Prob(best)	Prob(worst)	Par.	Prob(best)	Prob(worst)	Par.	Prob(best)	Prob(worst)
Change in behaviour	-0.106	0.200	0.246	-0.106	0.200	0.246	-0.084	0.195	0.242	-0.253	0.225	0.265
Statement 1	-0.383			-0.386			-0.347			-0.493		
Change in spending	-0.282	0.493	0.099	-0.282	0.492	0.099	-0.275	0.522	0.103	-0.430	0.493	0.089
Statement 1	0.698			0.702			0.698			0.770		
Delayed change	-0.438	0.130	0.379	-0.440	0.130	0.379	-0.419	0.117	0.377	-0.650	0.149	0.469
Statement 1	-0.482			-0.487			-0.456			-0.668		
Government incentives	0.000	0.177	0.276	0.000	0.178	0.276	0.000	0.166	0.279	0.000	0.132	0.177
Statement 1	-0.608			-0.610			-0.574			-1.490		

(b) Scenario 2

	M1			**M2**			**M3**			**M4**		
Attribute and statement	Par.	Prob(best)	Prob(worst)	Par.	Prob(best)	Prob(worst)	Par.	Prob(best)	Prob(worst)	Par.	Prob(best)	Prob(worst)
Change in behaviour	-0.106	0.185	0.327	-0.106	0.185	0.328	-0.084	0.179	0.325	-0.253	0.187	0.353
Statement 1	-0.383			-0.386			-0.347			-0.493		
Change in spending	-0.282	0.246	0.246	-0.282	0.246	0.246	-0.275	0.245	0.246	-0.430	0.242	0.246
Statement 2	0.079			0.079			0.119			0.048		
Delayed change	-0.438	0.267	0.226	-0.440	0.267	0.226	-0.419	0.282	0.218	-0.650	0.252	0.233
Statement 2	0.316			0.318			0.385			0.319		
Government incentives	0.000	0.302	0.200	0.000	0.301	0.200	0.000	0.294	0.211	0.000	0.319	0.168
Statement 3	0.000			0.000			0.000			0.000		

Table 5: Marginal effects for model M1

	M1		M2		M3		M4	
	Best	**Worst**	**Best**	**Worst**	**Best**	**Worst**	**Best**	**Worst**
Change in behaviour								
Statement 1	-0.368	0.315	-0.371	0.317	-0.333	0.284	-0.474	0.418
Statement 2	0.600	-0.658	0.604	-0.664	0.632	-0.697	0.680	-0.778
Statement 3	-0.014	0.012	-0.013	0.011	0.054	-0.049	-0.015	0.014
Change in spending								
Statement 1	0.603	-0.670	0.606	-0.674	0.608	-0.669	0.644	-0.745
Statement 2	0.075	-0.071	0.075	-0.070	0.113	-0.107	0.045	-0.043
Statement 3	0.095	-0.094	0.094	-0.094	0.150	-0.150	0.051	-0.052
Delayed change								
Statement 1	-0.475	0.414	-0.480	0.417	-0.449	0.390	-0.662	0.573
Statement 2	0.297	-0.300	0.299	-0.302	0.360	-0.366	0.300	-0.306
Statement 3	0.473	-0.486	0.475	-0.488	0.522	-0.537	0.455	-0.469
Government incentives								
Statement 1	-0.591	0.420	-0.593	0.421	-0.559	0.397	-1.478	0.616
Statement 2	0.244	-0.235	0.247	-0.238	0.289	-0.278	0.294	-0.292

In the current paper we admit to having relied on relatively simple models, however we believe our arguments extend to more advanced models, though we note that performing tests on preference symmetry and scale differences may increase the time required to properly analyse best worst data. Nevertheless, failure to do so may result in biased estimates which will impact on the model outputs in indeterminate ways. The use of fairly simple econometric models in our analysis however was a deliberate choice. The main objective of the current paper is not about the modelling of best worst data, but rather about appropriate interpretations of models estimated on best worst data. Whilst it is possible to estimate more advanced models that allow for heterogeneity of preferences and or scale, the overall message would remain the same, though the empirical findings might change somewhat. That is, best worst choice data should be examined from the perspective of framing effects, and be used as a data enrichment technique only under strict conditions. This message should be independent of the econometric model estimated.

Further, although we have used as an empirical example, an experiment utilising a best worst case 2 response format, we note that the above message applies equally to all three cases of best worst data. As noted above Giergiczny et al. (2013) found preference asymmetries as well as scale differences across four different best worst case 3 data sets. Nevertheless, we would argue that the mere presence of preference asymmetries and scale differences is not a problem for the approach, provided the resulting models are able to account for such affects and the analyst is prepared to interpret the results appropriately; that is in terms of the positive and negative framing effects of how the questions are asked.

Finally, it is possible that in treating the best and worst outcomes as two different discrete choice data sets and testing for preference asymmetries and difference in error variances, that not only will the preference asymmetry be found, but that other outputs of interest including welfare measures such as marginal rates of substitution, will also differ across the two responses. From a purely policy perspective, such an outcome may pose difficulties, as decision makers will most likely have to select outputs associated with either the best or worst responses in making their final decision. Nevertheless, such an dilemma is not a reflection of the method, but rather a reflection on the behaviour of people, in that how we frame our questions does matter.

References

Auger, P., Devinney, T.M. and Louviere, J.J. (2007) Best-Worst Scaling Methodology to Investigate Consumer Ethical Beliefs Across Countries, *Journal of Business Ethics*, 70, 299–326.

Beggs, S., Cardell, S. and Hausman, J. (1981) Assessing the potential demand for electric cars, *Journal of Econometrics*, 16, 1-19.

Bierlaire, M. (2003) BIOGEME: A free package for the estimation of discrete choice models, *Proceedings of the 3rd Swiss Transportation Research Conference*, Ascona, Switzerland.

Ben-Akiva, M., Morikawa, T. and Shiroishi, F. (1991) Analysis of the reliability of preference ranking data, *Journal of Business Research*, 23(3), 253-268.

Chapman, R.G. and Staelin, R. (1982) Exploiting rank ordered choice set data within the stochastic utility model, *Journal of Marketing Research*, 19, 288-301.

Cohen, E. (2009) Applying Best-Worst Scaling to Wine Marketing, *International Journal of Wine Business Research*, 21(1), 8–23.

Collins, A.C. and Rose, J.M. (2011) Estimation of stochastic scale with best-worst data, *International Choice Modelling Conference*, July 4-6, Leeds, United Kingdom.

Dellart, B.G.C., Brazell, J.D. and Louviere, J.J. (1999) The Effect of Attribute Variation on Consumer Choice Consistency, *Marketing Letters*, 10(2), 139–147.

Flynn, T.N., Louviere, J.J., Peters, T.J. and 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.

Hausman, J.A. and Ruud, P.A. (1987) Specifying and testing econometric models for rank-ordered data, *Journal of Econometrics*, 34, 83-104.

Hensher, D.A., Louviere, J.J. and Swait, J. (1999) Combining sources of preference data, *Journal of Econometrics*, 89(1–2), 197–221.

Kahnemann, D. and Tversky, A. (1979) Prospect theory: an analysis of decisions under risk, *Econometrica*, 47(2), 263-91.

Lazari, A.G. and Anderson, D.A. (1994) Designs of Discrete Choice Experiments for Estimating Both Attribute and Availability Cross Effects, *Journal of Marketing Research*, 31, 375-383.

Louviere, J.J., Hensher, D.A. and. Swait, J. (2000) Stated Choice Methods: Analysis and Applications in Marketing, Transportation and Environmental Valuation, Cambridge University Press, Cambridge, UK.

Louviere, J.J. and 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.

Louviere, J.J., Lings, I, Islam, T., Gudergan, S. and Flynn, T. (2013) An introduction to the application of (case 1) best–worst scaling in marketing research, *International Journal of Marketing Research*, 30, 292-303.

Louviere, J.J., Street, D., Burgess, L., Wasi, N., Islam, T. and 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(1), 128-164.

Marley, A.A.J. and Islam, T. (2012) Conceptual relations between rank data and models of the unexpanded rank data, *Journal of Choice Modelling*, 5(1), 38-80.

Meyer, R.K., and Nachtsheim, C.J. (1995) The coordinate-exchange algorithm for constructing exact optimal experimental designs, *Technometrics*, 37(1), 60-69.

Rose, J.M. and Hensher, D.A. (in press) Tollroads are only part of the overall trip: the error of our ways in past willingness to pay studies, *Transportation*.

Rose, J.M., Louviere, J.J. and Bliemer, M.C.J. (2013) Efficient stated choice designs allowing for variable choice set sizes, *International Choice Modelling Conference*, Sydney, Australia, 3rd-5th July 2013.

Scarpa, R., Notaro, S., Louviere, J.J. and Raffelli, R. (2011) Exploring Scale Effects of Best/Worst Rank Ordered Choice Data to Estimate Benefits of Tourism in Alpine Grazing Commons, *American Journal of Agricultural Economics*, 93(3), 813–828.

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

Vermeulen, Goos, P. and Vandebroek, M. (2011) Rank-order conjoint experiments: efficiency and design, *Journal of Statistical Planning and Inference*, 141(8), 2519–2531.