An Introduction to the Application of (case 1) Best-Worst Scaling in Marketing Research

Jordan Louviere[a], Ian Lings[b], Towhidul Islam[c], Siegfried Gudergan[d], Terry Flynn[a]

[a] Centre for the Study of Choice (CenSoC), University of Technology, Sydney, PO Box 123 Broadway, NSW 2007, Australia

[b] School of Advertising, Marketing and Public Relations, Queensland University of Technology, Brisbane, GPO Box 2434, Brisbane QLD 4001, Australia

[c] Department of Marketing and Consumer Studies, University of Guelph, 50 Stone Road East, Guelph, Ontario, NTG 2W1, Canada

[d] The University of Newcastle, Social Sciences Building, University Drive, Callaghan NSW 2308, Australia

Acknowledgements: We gratefully acknowledge the assistance of Tony Marley and Andrew Kyngdon in the collection of the data on weekend getaways used in the paper.

Paper published in the *International Journal of Research in Marketing*. Please cite as:
Louviere, Jordan, Lings, Ian, Islam, Towhidul, Gudergan, Siggi, & Flynn, Terry (2013) 'An introduction to the application of (case 1) best-worst scaling in marketing research'. *International Journal of Research in Marketing*, 30(3), pp. 292-303.

An Introduction to the Application of (Case 1) Best-Worst Scaling in Marketing Research

Abstract

We review and discuss recent developments in Best-Worst Scaling (BWS) that allow researchers to measure items or objects on measurement scales with known properties. We note that BWS has some distinct advantages compared with other measurement approaches, such as category rating scales or paired comparisons. We demonstrate how to use BWS to measure subjective quantities in two different empirical examples. One of these measures preferences for weekend getaways and requires comparing relatively few objects; a second measures academics' perceptions of the quality of academic marketing journals and requires comparing a very large set of objects. We conclude by discussing some limitations and future research opportunities related to BWS.

Key words: Best-worst scaling, measurement, preference, choice

1. Introduction

Academics and practitioners in various disciplines often wish to measure an individual's strength of preference for (or level of agreement with) a number of objects (which can be statements or some other items of interest). A typical objective is to locate all the objects on a measurement scale with known mathematical properties to allow robust statistical comparisons of changes over time and/or differences between respondents. In practice this can be challenging; for example, rating scales attempt to ensure all individuals use the same numerical scale, but in practice various idiosyncrasies in response styles have been found (Auger, Devinney, & Louviere, 2007). Such idiosyncrasies can arise from individuals using rating scales in different ways, cultural differences and/or verbal ambiguities with labels (Lee, Soutar, & Louviere, 2008). Furthermore, it has been observed that individuals tend not to discriminate between response categories when they are not asked to respond in ways that elicit tradeoffs or *relative* preferences for the objects being valued, such as asking people to rate the "importance" of several factors on a rating scale. That is, respondents do not have to trade off one factor against another, with evidence that this often leads to little differences in mean ratings (e.g., Cohen & Neira 2003; Lee, Soutar, & Louviere, 2007).

An approach to dealing with such issues that has been growing in popularity in many fields is to avoid tasks that ask individuals to use numbers in favor of tasks that infer strength of preference (or other subjective, latent dimensions) from how often they choose one object over other, known objects. Such observed choice frequencies ensure that the derived numbers are on a known (choice frequency or probability) scale. However, some choice-based approaches like the method of paired comparisons require large numbers of choice questions to estimate preferences for objects. Indeed, asking individuals to choose from all possible

pairs of objects becomes infeasible in survey settings as the number of objects grow, a clear weakness of the method of paired comparisons.

The purpose of this paper is to introduce, discuss and illustrate a choice-based measurement approach that reconciles the need for question parsimony with the advantage of choice tasks that force individuals to make choices (as in real life). Prior work recognizes three choice-based measurement cases. In case 1 (the object case), individuals are asked to choose the best and worst (on some subjective scale) from a set of objects (e.g., Finn & Louviere, 1992). In case 2 (the profile case) individuals evaluate several profiles of objects described by combinations of attributes/features dictated by an underlying design; they "see" the profiles one at a time and choose the best and the worst feature/attribute levels within each presented profile (e.g., Louviere, 1994). In case 3, individuals choose the best and the worst designed profiles (choice alternatives) from various choice sets dictated by an underlying design (e.g., Marley & Pihlens, 2012).

The purpose of this paper is to introduce, discuss and illustrate case 1. We focus on case 1 because it illustrates the fundamentals of choice-based measurement in general, and what is known as "Best-Worst Scaling" (BWS) in particular. BWS was introduced by Finn & Louviere (1992), and recent advances suggest that academics and practitioners would benefit from an updated discussion of its concepts and methods. BWS is one way to avoid and overcome some of the limitations of rating-based and similar measurement methods used in marketing and in other fields. BWS Case 1 typically allows one to obtain measures for each person (respondent) on a difference scale with known properties (Marley & Louviere, 2005). Cases 2 and 3 can be viewed as extensions of case 1 in which objects or items are represented as multi-dimensional choice objects (options). However, the fundamental ideas and principles from case 1 also apply to cases 2 and 3; thus, we focus on explaining case 1 in detail because this provides a foundation for understanding cases 2 and 3.

Accordingly, the objective of this paper is to provide an introduction for academics and practitioners on how to design, implement and analyze case 1 Best-Worst Scaling studies. The case for such a paper is threefold: 1) Papers detailing the mathematical proofs of the main estimators used to implement such studies are highly technical and not easily understood by novices (Marley & Louviere, 2005; Marley, Flynn, & Louviere, 2008), hence a need for a more straightforward explanation to encourage applications. 2) Disciplines in which comprehensive 'how to do' BWS discussions have been published have seen a proliferation of empirical studies (e.g., Flynn, 2010), suggesting that a tutorial paper also should benefit marketing academics and practitioners. 3) Several methods for estimating the values of objects on underlying subjective scales have been proposed; but many of these, although easy to implement in a spreadsheet or generic statistical package, are not part of the typical 'toolbox' of methods used by academics and practitioners. Indeed, a 'user guide' paper detailing the BWS profile case (case 2) for health economists arose from requests at conferences to (among other things) 'see' what the data and regression models 'look' like (Flynn, Louviere, Peters, & Coast, 2007).

So, to provide a 'how to do' BWS tutorial, this paper is organized as follows. First we offer a conceptual framework and empirical justification for BWS. We then present two empirical studies. We emphasize how to set up, design and implement a BWS case 1 survey in practice, and how to analyze the associated results. Specifically, we present worked examples that illustrate how to use BWS for relatively small (six objects) and very large (72 objects) comparison sets. The paper ends with a discussion and conclusions section that recaps the major points of the paper, identifies some limitations and issues and suggests some potential future research directions.

2. A Conceptual Framework for BWS

BWS is underpinned by random utility theory (RUT), which also underlies discrete choice experiments used in marketing research and economics (Thurstone, 1927; McFadden, 1974). RUT assumes that an individual's relative preference for object A over object B is a function of the relative frequency with which A is chosen as better than, or preferred, to B. Thus, it requires individuals to make choices stochastically (with some error). Thurstone's (1927) paper proposed RUT, and used it to motivate and develop the method of paired comparisons, where individuals chose the 'best' object from sets of two objects. Thurstone recognized that the theory requires individuals to make errors in their choices, so that the model parameter estimates that we term 'scale values' can be derived. Scale values are measures of the locations of each object on an underlying subjective scale of interest. McFadden (1974) generalized Thurstone's RUT model to provide tractable, closed-form models that accommodate choices from sets of three or more objects. More formally, for the 'best' only case McFadden considered:

$S_A = V_A + \varepsilon_A$
$S_B = V_B + \varepsilon_B$
$S_C = V_C + \varepsilon_C$
$S_D = V_D + \varepsilon_D$

Above, the true subjective scale value (S_k) of the k-th object consists of two components, the observed value V_k, which is systematic (explainable) and the errors, ε_k which are random (unexplainable). The random component implies that one cannot predict the exact choice that a person will make, but only the probability that a person will choose each object offered (McFadden, 1974). This choice probability can be expressed as:

$P(A=best \mid A,B,C,D) = P[(V_A + \varepsilon_A) > (V_k + \varepsilon_k)]$, considering all other options are available to be chosen in the comparison set. McFadden (1974) derived what is known as the conditional logit model by assuming that the errors are distributed as independent and

identically distributed Type 1 Extreme Value. The choice probabilities for this model have the following closed form expression:

$$P(A=best \mid A,B,C,D) = \exp(V_A)/[\exp(V_A)+\exp(V_B)+\exp(V_C)+\exp(V_D)].$$

McFadden's framework relates choices from sets of multiple objects to an underlying latent scale value associated with each object, but until recently little work was available to help researchers identify and implement reasonably good ways of *collecting* choice data from individuals to implement these models. An obvious exception, of course, is the method of paired comparisons, which has been extensively studied (e.g., David, 1988). Unfortunately, the method of paired comparisons poses inherent limitations in survey applications because the number of comparisons needed increases geometrically with the number of objects to be measured. So, paired comparisons can be practical for measuring a few objects (e.g., six objects require 15 pairs), but typically are not practical for larger numbers of objects (e.g., we later study 72 objects, which would require 2556 pairs).

One way to address the size limitation of paired comparisons is the multiple choice approach introduced by Louviere & Woodworth (1983) that relies only on 'best' choices. Although their discrete choice experiment ("DCE", also called "choice-based conjoint") approach is widely used, few researchers seem to appreciate that collecting only "first (or best) choices" provides minimal information for statistical estimation purposes. Thus, an approach that provides more statistical information than merely the first or best choice could be useful in many research applications.

BWS capitalizes on the fact that collecting 'worst' information, in a similar way to 'best' information, provides much more information. That is, BWS capitalizes on the idea that when individuals evaluate a set of three or more objects or items on a subjective scale, their choices of the top and bottom objects/items should be (all else equal) more reliable than choices of middle objects/items. Thus, BWS assumes that individuals make reliable and valid

choices of the two most extreme objects/items in a set, consistent with Adaptation Level Theory (Helson, 1964). A key advantage of BWS is that it provides information about BOTH the top ranked and the bottom ranked items in a set. Taken together, these two choices provide much more information about the ranking of the choice options in each set. Only order information matters in choices; hence, asking both top and bottom ranked choices gives much more information about the overall ranking of the objects than just the top choice.

More generally, BWS implies use of multiple comparison sets, with each set having at least three objects/items. In this respect, a BWS "experiment" is just another type of discrete choice experiment, similar to the DCEs proposed by Louviere and Woodworth (1983). To wit, they proposed constructing comparison (choice) sets from 2^J fractional factorial designs (J=the number of objects/items). However, most BWS applications design choice (comparison) sets with balanced incomplete block designs (BIBDs), such as Lee et al. (2008). A BIBD is a type of experimental design in which each choice option appears equally often, and co-appears equally often with each other choice option. Unlike 2^J designs, BIBDs ensure that choice set sizes are always equal. A type of BIBD called a "Youden" design (e.g., Raghavarao, 1988) allows one to control for order by ensuring that each object appears in every order. In our experience there is little difference in outcomes associated with order, but one can always rotate the BIBD items by superimposing a latin square design on each block to control for order.

We describe, discuss and illustrate applications of the case 1 approach below. We use two examples to illustrate relatively simple, straightforward ways to design and analyze the choice data from the BWS tasks for the following reasons: a) in many, if not most cases, one can design and analyze BWS tasks in ways that do not require complex statistical analyses; b) simple design and analysis methods make BWS accessible to many researchers with different statistical competencies; and c) simple design and analysis methods make it less likely that

researchers will make mistakes, allowing them to use BWS with more confidence. The two examples illustrate using BWS for a small and large number of objects to be measured (scaled), namely 6 and 72 objects, respectively.

3. Implementing Best–Worst Choice Tasks: Empirical example one

3.1 Empirical Issue of Interest

The first empirical example involves holiday destinations. The subjective dimension of interest is the likelihood of visiting a destination for a weekend getaway. The study population is residents of Sydney, Australia, who could choose among the following weekend getaways: 1) Central Coast (beach house), 2) Katoomba (up market hotel), 3) Barrington Tops (an isolated setting), 4) Bowral (Southern Highlands), 5) South Coast (heritage village), and 6) Sydney (up market hotel). We recruited a sample of 420 respondents from the Pureprofile online panel in Australia who satisfied the criterion that they resided in the Sydney metropolitan areas (defined by postcodes) and had taken at least one weekend getaway holiday in the previous 12 months. The Pureprofile panel has over 600,000 households recruited and maintained to match the overall Australian population on key Census demographics as closely as possible.

3.2 *Implementing BWS Tasks: Design*

The first stage in implementing a BWS survey is to choose a statistical design to construct the comparison sets. As noted earlier, researchers can choose from several statistical designs. We used a BIBD to design the comparison sets because they provide: a) constant comparison set sizes; b) the number of comparison sets increases approximately

linearly in J (number of objects/items to be measured; here J=6), such that one can often (but not always) find BIBDs for J objects/items in J or at most a few more than J sets; and c) BIBDs can be found in many sources, such as Raghavarao (1988) and Street and Street (1987). BIBDs also ensure that each of the J objects/items occur the same number of times across all sets, and co-occur the same number of times with the other (J-1) objects. These properties are important because unequal set sizes may unintentionally signal to individuals that a survey is about something unintended by the researcher and/or that they are "supposed" to choose differently in sets of different sizes, etc (i.e., if set sizes differ across a survey, it may lead to "demand artifacts"). Also, if one object appears more often than other objects it may signal that the survey is "really about" the one or more objects that appear more often.

To use a BIBD to implement a BWS survey, one numbers the objects/items of research interest from 1 to J, and replaces the same (1 to J) numbers in a BIBD table with the corresponding names, symbols or descriptions of each object to be measured. We illustrate this below with a BIBD for six objects (coded 1,2,…,6) that creates 10 comparison sets (survey questions), shown in the first four columns of Table 1.

<Table 1 here>

Next, one uses a 'find and replace' procedure to replace the object code numbers in columns 2-4 of Table 1 with the object names (here, holiday destinations, but more generally they can be items, principles, brands, etc) to make comparison sets, as shown in columns 5-7. The next step is to embed the comparison sets in a survey. One way to ask BWS questions is shown by the particular survey format shown in Table 2.

<Table 2 here>

Table 2 reveals that the BIBD used in the destination survey has the property that each destination occurs five times and co-occurs twice across the 10 sets.

3.3 BWS Response Data and Simple Analyses

A conventional BWS task requires respondents to choose the best and the worst objects in each comparison set. In this example 420 respondents were simply asked to identify the destination s/he was most likely to visit (best) and which destination s/he was least likely to visit (worst) in each of the 10 comparison sets that each contained three destinations. The analysis is based on assigning the most likely destination a '+1' and the least likely destination a '−1', and with each item appearing five times, preferences are measured on a scale bounded by −5 and +5. A simple analysis involves summary statistics like sums or means. Our survey also asked respondents to self-report the average number of trips they took to each destination during the prior 12 months (revealed preference), which allows us to compare BWS measures against these 'revealed preference' (RP) self-reports.

Columns five and six of Table 3 show one person's best and worst choices.

<Table 3 here>

Summarizing the best and worst choice data merely involves counting the occurrences of best and worst choices for each choice object, as shown in Table 4. A simple scale and semi-order (ranking) is obtained by subtracting worst counts from best counts, as shown in the last column. The advantage of collecting data about the worst choice becomes clear when considering the lower ranked objects: the scale derived from best only choice data cannot distinguish objects 5 and 6, whereas best and worst choices taken together provide rank order

information for these two objects. The reason for calculating counts is to obtain empirical estimates of the choice proportions. As noted by Louviere and Woodworth (1983), these counts contain all the statistical information in the data, and can be used to estimate the parameters of a Luce (1959) or multinomial logit model (McFadden, 1974). That is, if there are J objects, one simply estimates the intercepts or "alternative-specific constants" in the logit model:

$U_1 = \beta_1 + \varepsilon_1$,

$U_2 = \beta_2 + \varepsilon_2$,

…

$U_J = 0$

The counts corresponding to the βs above are shown in Table 4.

<Table 4 here>

Marley and Louviere (2005) show that such best count minus worst count differences (BWS 'scores') are sufficient statistics for a conditional (multinomial) logistic regression model. This implies that all the information needed for more sophisticated conditional (multinomial) logistic regression models is available from the BWS scores; so a researcher need not use more sophisticated statistical estimation methods, such as various types of regression analyses to estimate the scale values of the objects (regression model parameter estimates). Instead, one can simply use the BWS scores. In BWS Case 1, the objects being chosen are not described by attributes that vary; thus each object is represented by a single regression parameter (an "alternative-specific constant" in a conditional logistic regression model). Thus, if there are J objects to be measured (scaled), there are J-1 regression model parameters that can be identified, with the J-th parameter set to a constant (typically zero).

To summarize, there can be several dependent variables of interest in BWS:

1. For individuals.
    - One calculates the total number of times that each object of interest is chosen as best and worst across all comparison sets (choice sets).
    - The dependent variables resulting from this can be a) the difference in the best and worst counts or choice totals (i.e., best counts minus worst counts), b) a full or partial ranking obtained from the best and worst choices in each comparison set, or c) an expansion of the best and worst choices into implied choice sets as discussed by Louviere et al. (2008).
2. For aggregates of individuals.
    - One also calculates the total number of times that each object of interest is chosen as best and worst across all comparison sets (choice sets) and individuals.
    - If the data are disaggregated to represent each object in each comparison set for each person, the dependent variables resulting from this can be a) the difference in the best and worst counts or choice totals (i.e., best counts minus worst counts), b) a full or partial ranking obtained from the best and worst choices in each comparison set, or c) an expansion of the best and worst choices into implied choice sets as discussed by Louviere et al. (2008).
    - If the data are aggregated to represent each object across all comparison sets and persons, the dependent variables resulting from this can be a) the difference in the best and worst counts or choice totals (i.e., best counts minus worst counts), and b) the square root of the ratio of best counts divided by worst counts (This measure is proportional to the best counts

under the assumption that worst counts = 1/(best counts); the natural log of this quantity is the expected value of the object on the latent scale if the choice process is consistent with a conditional logit model).

The above dependent variables can be analyzed in simple ways, such as calculating them directly from the data, or one can use more sophisticated analytical methods, such as ordinal regression models and probabilistic discrete choice models. We also showed in Table 5 and Figure 1 that typically ordinary least squares applied directly to the best minus worst differences will produce reliable and valid measures of the scale positions of the objects on the latent scale. We also noted that the scale positions represent intercept terms in discrete choice models.

Researchers should be cautious about using BWS scores (best minus worst counts) to make inferences about individual respondents (Flynn, 2010); for example, in the present example, objects 3 and 4 (in the middle of the scale) are not differentiated by BWS scores. Despite this caveat, experience suggests that one does not have to aggregate choices across many individuals for average scores (for individuals in question) to perform well and correlate highly with estimates from more sophisticated models.

3.4 BWS More Sophisticated Analyses Requiring Statistical Software

We now consider two other ways to analyze the data, both of which utilize a [0,1] dependent variable (indicating whether a particular choice option was not chosen/chosen): 1) linear probability models (LPMs), which are ordinary least squares (OLS) regression models fit to the choice data; and 2) conditional logit models (McFadden, 1974). LPMs are justified by Heckman & Snyder (1997), who argue that if errors are not symmetric, LPMs are likely to better represent respondents' choices than would other models (See also Alrich & Nelson, 1984).

To estimate LPMs and conditional logit models (CLMs) we use information from the best and worst choices in each comparison set to "expand the data". By "expanding the data" we mean using the rank order information to construct several sets of implied choices for various pairs of objects associated with each comparison set. For example, if one observes best and worst choices for two objects in a set of three objects, one has the full rank ordering. One can use the observed best and worst choices to construct three pairs of implied choices (Horsky & Rao, 1984), from which one infers the implied choice that should be made for each of the three pairs of objects. Thus, for three option sets, one can "expand" best and worst choices into three implied choice pairs for each three option set (A vs B, B vs C and A vs C). For four object/option sets, best and worst choices give a semi-order; that is, one can infer the choices in most, but not all of the possible pairs. For other set sizes the idea is the same, but observing only best and worst choices gives less information about a full ranking as the number of options per set increase.

More specifically, suppose one observes best and worst choices in a set of three options - say "A", "B" and "C". Suppose further that A is chosen best and C is chosen worst. This ranking implies that A should be chosen for the pair AB and the pair AC, and B should be chosen for the third pair BC. In the case of four options, say "A", "B", "C" and "D", if a respondent chooses A best and D worst, this implies that A should be chosen in the pairs AB, AC and AD, B should be chosen in the pair BD and C should be chosen in the pair CD. Thus, the best and worst choices provide information that can be "expanded" to several pairs of choices, with the higher ranked option being the one expected to be chosen. From a choice modeling standpoint, one can "stack" the pairs, and code the option implied as the one chosen from each pair as a "1", with the other option in the pair coded as a "0". Standard conditional logit modeling estimation software can be used to estimate the model parameters. The results from our example are in Table 5.

The top part of Table 5 shows best and worst counts for each of the six destinations summed across all 420 respondents and 10 comparison sets, and the best minus worst counts (B-W) for each destination, as well as the self-reported choices (RP, or the average number of trips per person to each destination). The lower part of Table 5 contains the statistical results from estimating a conditional logit model from the Best choices and the statistical results from estimating an OLS regression model (a linear probability model) from the Best choices.

<Table 5 here>

Figure 1 plots the set of estimates from each of three regression models together with the revealed preference data against the best-minus-worst scores. The ordinary least squares regression model estimates (best-fit line), relating each of these four sets to the best-minus-worst scores are given below it. These estimates, together with associated R-squared values, suggest that all four sets are strongly linearly related to the best-minus-worst scores. In other words, all provide the same relative scale (measurement) position information about the objects. The takeaway from Figure 1 is that one typically can estimate the latent scale positions (measures) of each object with any of the methods illustrated. Thus, one may wish to use the simplest approach, by simply calculating Best counts minus Worst counts (BWS Scores). One can calculate BWS Scores for each person in a sample, and describe the resulting distribution of the scores with typical statistics, such as means, medians, standard errors, etc (as shown in Table 5 for means and associated standard errors). It is worth noting that one is unlikely to need the standard errors of the BWS scores for each person because one rarely (if ever) needs to do statistical inference for one person. Instead, one wants to summarize the statistical properties of the sample. Figure 1 also contains a graph of the relationship between the best-minus-worst (B-W) scores and the actual trips (RP). The

relationship is approximately linear, with an R-squared of 0.86, suggesting that the BWS measures match the reported choices well.

If one wishes to conduct statistical inference, one needs to be sure that the sample size is consistent with the desired test of parameter equality or differences. We discuss a general way to determine sample size requirements for best-worst studies later in the paper. The obvious caveat or limitation to the simple BWS score analysis is that averages can obscure underlying differences in measures across people. If one needs to understand individual differences, one can analyze the individual-level best-worst choices in several ways to gain such insights. That is, one can consider using various cluster analytic methods, latent class models or random parameter models. We illustrate the use of scale-adjusted latent class models (Magidson & Vermunt, 2007) later in the paper, but this is only one of several options open to analysts. The topic of capturing individual differences is vast, so in the interests of focus and space, we merely note that if one wants to explore individual differences, one will need to consider one or more methods for doing so. As a final comment, however, it is worth noting that for any research problem requiring reliable estimates of individual differences, one typically requires larger sample sizes than if one only wants to estimate sample average subjective values.

<Figure 1 here>

3.5 Examining Choice Frequencies Across the Sample.

As explained above, BWS difference scores create a rank order of preference based on sample averages. We also fit individual-level LPM models to each person's (respondent's) BWS scores. The sample statistics for these individual-level LPM estimates are in Table 6. The results suggest that the Central Coast is the most preferred destination, with Bowral in

the Southern Highlands of NSW the least preferred. Other destinations are intermediate. The average of the respondents reported most recent visits (RP) are included in the table, and the correlation between "Mean" and RP estimates is 0.91. Naturally, both the BWS scores and the RP reported visits contain error, hence the observed correlation should be taken merely as evidence that the relationship is relatively strong and in the correct direction. More generally, however, the standard errors of BWS scores can be calculated from the sample data to provide an indication of the degree of agreement in the sample on the subjective position of each object or item being measured. Lower standard errors imply more agreement in the sample.

<Table 6 here>

3.6 Sample Sizes

BWS measures are derived from multinomial frequency counts or proportions. As noted earlier, we can calculate differences in frequency counts and ratios of frequency counts. We assume that one is not interested in inferences about any one individual associated with BWS scores, as one typically is not interested in inferences about one set of rating scale values from one person. Thus, sample sizes matter only to the extent that one designs a BWS task in such a way that one can calculate the scores of interest. In most cases, and at a minimum, BWS scores give a full or partial ranking of items or objects of research interest.

Sample size considerations arise if one wants to make inferences about a population represented by a sample of people and/or if one wants to compare estimates (e.g., BWS scores) from two or more samples or different subsamples from a particular population (e.g., different "segments" in a sample). In these cases, BWS measures follow the rules for sample sizes associated with multinomial distributions. No specific methods for sample size have

been developed for B-W scaling. Sample size methods developed for multinomial proportions data can be used (e.g., Thompson 1987; Rose 2011). Several papers look at the issue of sample size requirements for multinomial proportions data (e.g., Angers, 1979, 1984 & 1989), and Thompson (1987) derived a formula to calculate sample size requirements for multinomial proportions data. Thompson showed that, like the binomial case, sample size is a function of the level of acceptable error and the degree of desired confidence required by analysts in obtaining true population proportions. He also demonstrated that the sample size requirement for multinomial proportions data is independent of the number of multinomial categories (J outcomes or choice options; i.e., items, things) but not independent of what he termed the "worst possible outcome". The worst possible outcome is defined as m of the J options having equal choice frequencies (proportions) or shares of 1/m, with the J-m remaining options having a value of zero.

Unfortunately, the value of m is not independent of the value of the confidence level, and so it must be calculated for different levels of confidence required. For example, if we use Thompson's approach and require a sample to satisfy the probability that at least 0.95 of all proportions are within 0.05 of the true population proportions and assume that m equals three, the required sample size is 510 respondents (independent of the number of options J). That is, in this example the worst possible outcome that would be observed to occur is one where three population proportions are equal to 1/3, and the rest equal zero. Unfortunately, as things stand, we do not know whether different sample size implications are associated with different analysis methods. Thus, researchers may wish to rely on sample size estimation methods for discrete choice models, such as that from Rose (2011) for the general case of best only choices. It is worth noting, however, that it is likely that sample size requirements for estimating Case 1 parameters using discrete choice models will be less than for best only choices because BWS choices provide extra choice data for any given sample size.

4. Empirical Example 2

We now turn our attention to a second empirical example that involves very large numbers of objects or items (72). The purpose of this example is to show that relatively large set sizes can be readily incorporated into a BWS study by reducing the size of the choice set through the use of a nested BIBD design. There are two reasons why researchers may wish to consider reducing comparison set sizes: 1) If individuals are relatively consistent in their choices, large numbers of objects per choice question give insufficient data points for middle-ranked objects (i.e., zero best-minus-worst counts will be observed, giving no information on relative preferences for middle ranked objects, as was the case in Table 4). 2) Small choice sets make the task of choosing the best and worst easier for individuals who may have cognitive limitations.

There are two ways to collect enough choice data to derive an acceptable ranking when the choice set size is large: 1) After asking an individual for their best and worst objects in the set, one can ask a second round (or more) of best-worst questions to obtain additional ranking information (second best object, second worst, third best, third worst etc). Additional rounds of questions can be used to obtain a complete ranking of all objects in the choice set. This is particularly easy in web-based surveys because one can eliminate already chosen options from screens, making the task easier for respondents. If one uses this approach to collect additional best and worst choices, one must expand the choices to implied choice sets, and use more advanced analytical methods than the simple analytical methods discussed in this paper. That is, as noted by Horsky & Rao (1984), one can expand the data in each BIBD set to paired choices implied by the full or partial ranking obtained. As previously noted, the subjective values to be estimated are intercepts or alternative-specific constants in conditional

logit models or more complex choice models, such as latent class or G-MNL (Fiebig, Keane, Louviere, & Wasi, 2010).

Whether and how many additional rounds of Best and Worst questions are needed will depend on a) the size of the comparison (choice) sets, b) how many objects (total choice options) are being measured, and c) how critical it is for one to reliably and accurately measure (scale, estimate) each object of interest on the underlying latent scale for each person. For example, if one only needs to accurately discriminate the best and worst objects, and is less interested in intermediate objects (e.g., one often wants to sort potential product attributes for choice experiments into clearly important, intermediate and clearly unimportant), asking only one round of Best and Worst questions should be sufficient. More accuracy in differentiating and measuring the objects will dictate how many rounds of questions are needed, and this typically can be determined by a small pilot study in advance of the primary data collection.

A way to deal with large numbers of items to be measured is to use nested BIBD designs. That is, one first uses a suitable BIBD to generate a block of choice sets with a relatively large set size, and then uses a second suitable BIBD to reduce the number of objects in each choice set to a more manageable size. We illustrate the nested BIBD approach below and provide a worked example of how to do it.

4.1 Issue of Interest

There are several ways to 'judge' the quality of academic work. Stremersch et al (2007) examine the quality of individual articles, whilst Lehmann (2005), provides a range of criteria against which journals quality may be measured; some objective such as citations and impact factors, others more subjective. We approach the issue of measuring subjective evaluations of the quality of academic marketing journals and this forms the basis of our

second example of best worst scaling. We consulted several available journal rankings papers/reports in marketing and business to obtain a fairly comprehensive set of academic and quasi-academic marketing journals (e.g., Fry, Walters, & Scheuermann, 1985; Geary, Marriott, & Rowlinson, 2004; Sivadas & Johnson, 2005; Starbuck, 2005). We focused on marketing-specific journals, excluding non-marketing journals like *Econometrica* or *Psychometrika*, although marketing academics publish in them. This generated a list of 73 journals for BWS (see Appendix B).

4.2 Implementing BWS and Design

To generate an appropriate BIBD, we consulted Street & Street (1987) to obtain a BIBD for 73 objects that gave 72 comparison sets with 9 journals per set. We then used a second BIBD to expand each comparison set of 9 journals into 12 comparison sets with 3 journals per set. This reduced the size of comparison sets from 9 journals per set to 3 per set, giving a total of 12 x 72 = 864 comparison sets for the entire survey.

A simple example of how to pool two BIBDs in this way can be found below that involves a BIBD for 13 objects that produces 13 sets of size 9 and a BIBD for 9 objects that has 12 sets of size 3. We expand each block (row, set) in the first BIBD by using the second BIBD to make 12 new sets for each row in the first BIBD. This leads to 13 x 12 = 156 sets of size 3, as shown in Figure 2.

<Figure 2 here>

Respondents were recruited by placing notices in online newsletters via postings to the ELMAR virtual community, the European Marketing Academy, The Australia-New Zealand Marketing Academy, and The Academy of Marketing in the United Kingdom. We

also placed notices on the posting board at the Marketing Science Conference. In 2006 we published the survey online on a secure Web server, and directed respondents to a URL with the survey. Questions were added to the survey to determine respondents' academic rank, years in academia, geographic location and research specialties. To collect the BW choice data we randomly assigned each survey respondent to one block in the first BIBD, and asked the respondent to complete all 12 comparison sets from the second BIBD.

Approximately 900 respondents received an email directing them to the survey. A total of 529 responded to the recruitment invitation and provided complete data (we deleted approximately 15 people whose surveys were incomplete or unusable, including one person who did the survey 10 times).

4.3 BWS Response Data and Simple Analyses

As with example 1, the analysis of the journal ranking choices is based on assigning the best journal in the set a '+1' and the worst a '−1'. The best and worst choice data are summarized by counting the occurrence of best and worst choices for each choice object (journal). However, as respondents were randomly assigned to each of the 72 initial blocks of 12 journals before being asked to evaluate all 9 blocks of three journals per initial block of 12, there were large differences in sample sizes for various survey versions. Thus, the resulting aggregate sample probability of choosing a journal (estimated by the proportion of choices for each journal) is not independent of the probability that it is available to be chosen. To account for this, we reweighted the choice frequency counts for best and worst choices to take into account the probability of journal occurrence. The weights are calculated as follows:

$weight\ journal_i$ = average appearances for all journals / average appearances $journal_i$.

For example, suppose there are four blocks or versions of the Best-Worst task, which will occur if one randomly assigns sets of choice sets to the 4 versions in equal numbers without replacement. Each of the four versions constitutes a separate "survey". Suppose one recruits participants from an online panel and randomly assigns each person who agrees to participate to one of the versions. Let the versions be V1, V2, V3 and V4. Let the number of respondents to each version be, respectively, 35, 20, 30 and 15, for a total of 100 participants. The versions need to be weighted because in the overall sample, some of the J objects being measured occur more/less often than others. The average sample size = 25, so the weights are calculated for each version as follows:

Weight for V1 = 25/35 = 0.71
Weight for V2 = 25/20 = 1.25
Weight for V1 = 25/30 = 0.83
Weight for V1 = 25/15 = 1.67

One weights each observation in each version by the weights calculated above. We use this weighting approach in the example that follows.

The re-weighting increases the number of choices of journals appearing less often and decreases the number of choices of journals appearing more often. Using these weighted best and worst counts, we obtained a simple scale by subtracting weighted worst counts from weighted best counts. The aggregate sample results are shown in Appendix C, which also lists journals and occurrence-weighted BWS journal quality scores. The BWS scores for each journal in Appendix C gives a ratio scale of journal quality, meaning that one can conclude that a journal with a score of 1.00 has approximately twice the level of perceived quality of a journal with a score of 0.50. Because scores are ratio-scaled, they can be transformed into point systems consistent with measured levels of perceived quality.

4.4 Further Examination of the Sample.

Our sample contained enough choice data to also conducted separate analyses for each of three regions; North America (Canada and the USA = 220), Europe (94) and Australia–New Zealand (107). We tested the hypothesis that there is heterogeneity in how academics perceive (value) journal quality. To do this we estimated a scale-adjusted latent class model (SALCM) from the choice data (Magidson & Vermunt 2007). The results of the SALCM produced only one journal-quality class using the Bayesian information criterion criteria for model selection. That is, respondents tend to rank journals similarly in all three regions. Discussion and examples of the SALCM estimation technique can be found in Burke et al., (2010) and Flynn, Louviere, Peters, & Coast (2010). Despite our finding that respondent ranked journals similarly across regions, it is possible that researchers in top schools would rank journals differently to researchers in other schools. School level data were not available in this study but examining such heterogeneity across schools may provide an interesting line of enquiry for future researchers.

The top 10 journals are shown in Table 7. The sample mean quality measure for all journals is 41.5 (see Appendix C), with a standard error of 2.73; hence, differences of 5.5 scale units are significant. So, the *Journal of Marketing* and *Journal of Marketing Research* are perceived as top journals and statistically equal firsts, followed by the *Journal of Consumer Research, Journal of the Academy of Marketing Science* and *Marketing Science*, which are statistically equal seconds in perceived quality. The next four are statistically equal thirds; with the *Journal of Advertising* significantly lower than the top two in this tier.

<Table 7 here>

5. Discussion and Conclusions

A Google search for the terms “Best-Worst Scaling” and Maximum Difference Scaling” returned, respectively 1.96M and 3.7M “hits (April 22, 2012). Perusal of the first

several pages of hits clearly shows that the choice-based measurement approach that we call "Best-Worst Scaling" is used primarily by academics, with practitioners mainly included in the hits for "Maximum Difference Scaling". The results also suggest a growing number of academics in several different fields are adopting the BWS approach. Apart from the original Finn & Louviere (1992) paper that introduced the approach, and the theoretical treatment by Marley & Louviere (2005) that derives the formal measurement properties of various case 1 Best-Worst Choice models, there are few "how to do it" papers. Thus, the objective of this paper was to describe and discuss ways to implement, analyze and interpret case 1 (object case) Best-Worst Scaling (BWS) applications, and show how to use BWS by applying it to two empirical examples.

We showed that BWS is relatively easy to implement and analyze, even with fairly large numbers of objects (e.g., the academic journals example), making it accessible to many academic and applied researchers. We focused on simple ways to analyze BWS data (using best-minus-worst scores and simple regression models) to show that one can obtain good results with fairly simple analysis methods. A more detailed and formal treatment of the theory underlying BWS is in Marley & Louviere (2005); our aim was to give as straight forward an explanation of the theory and methods possible. Thus, we emphasized simple counts of choices (i.e., sums), expansion of best and worst choices to implied paired choices and graphical tests to allow one to assess if data are consistent with theory.

Table 8 below illustrates the steps in designing a best-worst scaling study.

<Table 8 here>

5.1 BW Scaling for Groups of Individuals

We also applied more complex regression models frequently used to analyze data from traditional discrete choice experiments. The measurement values estimated by these

conditional logit models are the natural logarithms of the classical (Luce, 1959) multiple-choice model that yields ratio scales. We caution that many of the same issues associated with those models also apply to BWS. Most notably, the assumptions underlying such models are quite strong, and have been discussed in the discrete choice modeling literature for many years (e.g., Train, 2003). Perhaps most importantly, such models theoretically apply only to single people; additional assumptions are required to extend them to aggregates of people. How well such models approximate individuals compared with aggregates of individuals remains unresolved. Recent work on choice models for single individuals by Louviere, et al. (2008) suggests that individual-level models can outperform more aggregate models. Thus, it may be that McFadden's (1974) conditional logit model or Luce's (1959) model may be a reasonable first approximation to a person's unknown choice processes for BWS choice tasks. Further work is needed to understand if and when one needs to, and how to, relax assumptions associated with these models. We set the latter issues aside here and merely note that a great deal of experience with BWS over the past five years suggests that these models seem to be reasonable first approximations to unknown individual-level choice process(es).

5.2 Unresolved Issues with BWS

There are a number of unresolved issues with BWS that can be viewed as future research opportunities. For example, because BWS relies on discrete choices, it has the limitations of random utility choice models, such as possible violations of the independence of irrelevant alternatives (IIA) property of Luce's (1959) model and McFadden's (1974) conditional logit model. Whether IIA violations exist is an empirical issue, and how serious they are in *individual-level* BWS choice data remains an open issue. Nevertheless, we note that the equal co-occurrence property inherent to BIBDs allows one to estimate violations of the IIA property of simple choice models as it ensures that the two-way interactions (cross-effects) are estimable (Lazari & Anderson 1994).

Similarly, objects or items of interest in BWS applications may exhibit various degrees of similarity and/or correlated errors, which is also an open issue. Those familiar with discrete choice models will recognize that these issues have been widely discussed in the choice-modeling literature. Hence, the issues are easy to state but complex to resolve, especially when one is modeling single people.

We also should emphasize the need to consider decision rules used by respondents. Because individuals are asked to choose the best and worst objects (i.e., largest/smallest, most/least preferred, etc), BWS is sometimes called 'maximum difference' scaling (or, 'MaxDiff' scaling). The latter nomenclature is unfortunate because the maxdiff model is only one of a number of models of the process that an individual might use to make a series of best-worst choices; in mathematical psychology the maxdiff model assumes that an individual considers all possible best-worst pairs (simultaneously) and chooses that pair that maximizes the difference between the two objects comprising the pair. Naturally, an individual might use alternative choice processes. For example, they might choose the best stimulus first, followed by choice of worst stimulus; or choose the worst stimulus first, followed by the best stimulus, and so forth. The way(s) individuals make such choices is an empirical question. However, each way implies a different possible process model of their choices and, strictly speaking, a different statistical model.

We note that BWS scales currently are relative to sets of objects studied. For example, if we offer a person the set {Hitler, Mussolini and Stalin} and ask them to choose the best and worst national leaders, they would make two choices. However, it is likely that, if asked, they would say that no one in the set was a 'good' leader. Ongoing work aims to resolve this problem by developing BWS measures that reflect absolute as well as relative positions. More generally, however, solutions to this problem require extra information external to a BWS task, and theoretically sound solutions to this problem would be welcome. Meanwhile,

one can ask people to report whether "none of the objects are good" and/or "none of the objects are bad", which can be used to anchor the scales. The latter is common practice in discrete choice experiments, dating from Louviere & Woodworth (1983). Another possibility is anchoring relative to a status quo option in each comparison set, but to our knowledge this has yet to be done.

We illustrated and discussed simple and more complex ways to analyze best-worst choice data. Examples evaluating a small and large number of objects were demonstrated, thus illustrating the generalizability of the methods. Future work should examine the extent to which BWS can out-perform traditional rating scales and investigate whether the benefits noted here are generalizable across policy and marketing areas.

Acknowledgements: We gratefully acknowledge the assistance of Tony Marley and Andrew Kyngdon in the collection of the data on weekend getaways used in the paper.

This paper was prepared with the assistance of the Services Innovation Research Program, QUT Business School, Queensland University of Technology.

References

Alrich, J. H., & Nelson, F. D. (1984). *Linear Probability, Logit, and Probit Models*. *Quantitative Applications in the Social Sciences*. California: Sage University Papers

Angers, C. (1979). Simultaneous estimation of percentile curves with application to salary data. *Journal of the American Statistical Association,* 74(367), 621-625.

Angers, C. (1984). Large sample sizes for the estimation of multinomial frequencies from simulation studies. *Simulation,* 43(4), 175-178.

Angers, C. (1989). Note on quick simultaneous confidence intervals for multinomial proportions. *The American Statistician,* 43(2), 91.

Auger, P., Devinney, T. M., & Louviere, J. J. (2007). Using best-worst scaling methodology to investigate consumer ethical beliefs across countries. *Journal of Business Ethic*, 70(3), 299-326.

Burke, P. F., Burton, C. T., Huybers, T., Islam, T., Louviere, J. J., & Wise, C. (2010). The scale-adjusted latent class model: Application to museum visitation. *Tourism Analysis*, 15(2), 147-165.

Cohen, S. H., & Neira, L. (2003). Measuring preference for product benefits across countries: Overcoming scale usage bias with maximum difference scaling. *ESOMAR 2003 Latin America Conference*, Punta del Este, Uruguay.

David, H. (1988). *The method of paired comparisons* (2nd ed.). New York: Oxford University Press.

Fiebig, D., Keane, M., Louviere, J. J., & Wasi, N. (2010). The Generalized Multinomial Logit Model: Accounting for scale and coefficient heterogeneity *Marketing Science*, 29(3), 392-421.

Finn, A., & Louviere, J. J. (1992). Determining the Appropriate Response to Evidence of Public Concern: The Case of Food Safety. *Journal of Public Policy & Marketing*, 11(1), 12-25.

Flynn, T. N. (2010). Valuing citizen and patient preferences in health: recent developments in three types of best-worst scaling. *Expert Review of Pharmacoeconomics & Outcomes Research*, 10(3), 259-267.

Flynn, T. N., Louviere J. J., Peters, T. J., & Coast, J. (2007). Best-Worst Scaling: What it can do for health care research and how to do it. *Journal of Health Economics*, 26(1), 171-189.

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

Fry, E. H., Walters, C. G., & Scheuermann, L. E. (1985). Perceived quality of fifty selected journals: Academicians and practitioners. *Journal of the Academy of Marketing Science*,13, 352-361.

Geary, J., Marriott, L., & Rowlinson, M. (2004). Journal rankings in business and management and the 2001 Research Assessment Exercise in the UK. *British Journal of Management*,15, 95-141.

Heckman, J. J., & Snyder, J. M. (1997). Linear probability models of the demand for attributes with an empirical application to estimating the preferences of legislators. *RAND Journal of Economics*, 28, S142-S189.

Helson, H. (1964). *Adaptation-Level Theory*. New York: Harper & Row.

Horsky, D., & Rao, M. R. (1984). Estimation of attribute weights from preference comparisons. *Management Science*, 30(7), 801-822.

Lazari, A. G., & Anderson, D. A. (1994). Design of discrete choice experiments for estimating both attribute and availability cross effects. *Journal of Marketing Research*, 31(3), 375-383.

Lee, J. A., Soutar, G., & Louviere, J. J. (2007). Measuring values using best-worst scaling: the LOV example. *Psychology & Marketing*, 24(12), 1043-1058.

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(4), 335-347.

Lehmann, D.R. (2005). Journal Evolution and the Development of Marketing. *Journal of Public Policy & Marketing,* 24 (1), p. 137-142

Louviere, J.J. (1994). Conjoint Analysis, In R. Bagozzi (Ed.), *Advanced Marketing Research*, Cambridge, MA: Blackwell Publishers.

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(1), 128-163.

Louviere, J. J. & Woodworth, G. (1983). Design and Analysis of Simulated Consumer Choice or Allocation Experiments: An Approach Based on Aggregate Data. *Journal of Marketing Research*, 20, 350-367.

Luce, R. (1959). *Individual choice behavior: A theoretical Analysis*. New York: John Wiley & Sons.

Magidson, J., & Vermunt, J. K. (2007). Removing the scale factor confound in multinomial logit choice models to obtain better estimates of preference. *Sawtooth Software Conference*.

Marley, A. A. J., Flynn, T. N., & Louviere, J. J. (2008). Probabilistic Models of Set-Dependent and Attribute-Level Best-Worst Choice. *Journal of Mathematical Psychology*, 52, 281-296.

Marley, A. A. J. & 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 multi-attribute options (profiles). *Journal of Mathematical Psychology*, 56, 24-34

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics* (105-142). New York: Academic Press.

Raghavarao, D. (1988). *Constructions and Combinatorial Problems in Design of Experiments*. New York: Dover.

Rose, J. M. (2011). Discussion of "The usefulness of Bayesian optimal designs for discrete choice experiments". *Applied Stochastic Models in Business and Industry*, 27(3), 193-196.

Stremersch, S., Verniers, I., & Verhoef, P.C. (2007). The quest for citations: Drivers of article impact. *Journal of Marketing*, 71(3), 171-193

Sivadas, E., & Johnson, M. S. (2005). Knowledge flows in marketing: An analysis of journal article references and citations. *Marketing Theory*, 5(4), 339-361.

Starbuck, W. H. (2005). How much better are the most prestigious journals? The statistics of academic publication. *Organizational Science*, 16(2), 180-200.

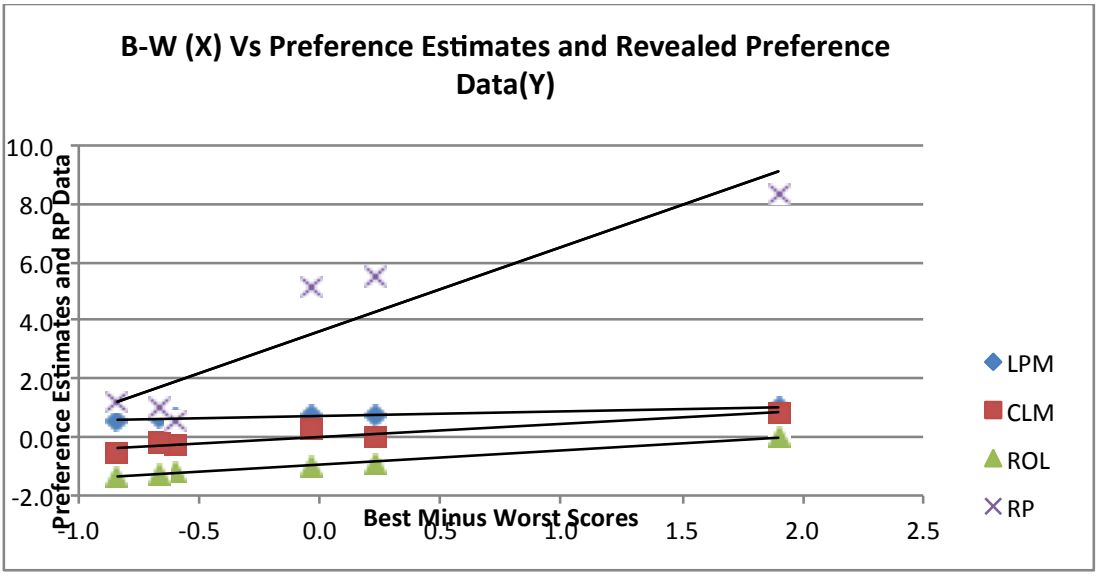
Street, D., & Street, A. P. (1987). *Combinatorics of experimental design*. Oxford: Clarendon Press.

Thompson, S. K. (1987). Sample Size for Estimating Multinomial Proportions. *The American Statistician*, 41(1), 42-46.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34, 273-286.

Train, K. (2003). *Discrete choice models with simulation*. Cambridge, UK: Cambridge University Press.

Figure 1: Comparison of BWS Measures - Getaways



This figure plots the estimates for each of three regression models and the revealed preference data against the calculated best minus worst scores (most preferred minus least preferred) for the six getaway destinations. Ordinary least squares (OLS) regressions for the linear probability model (LPM) estimates, conditional logit model (CLM) estimates, rank-ordered logit model (ROL) estimates and revealed preference (RP) data are given by:

LPM = 0.16(B-W) + 0.72; $R^2 = 0.99$
CLM = 0.45(B-W) + 0.001; $R^2 = 0.90$
ROL = 0.48(B-W) - 0.95; $R^2 = 0.99$
RP = 2.88(B-W) + 3.63; $R^2 = 0.86$

The four OLS regressions demonstrate that the easily calculated B-W scores are strongly linearly related to the estimates from three more complex regression models and the revealed preference (actual observed) data.

Figure 2: Pooling two BIBDs

BIBD 1 for 13 objects in 13 sets of size 9									
Block	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	10	6	12	3	4	13	1	9	7
2	2	10	13	9	11	3	4	8	12
3	8	13	2	6	9	7	3	1	11
4	6	2	10	4	7	11	8	12	1
5	9	7	6	12	13	5	11	10	8
6	5	11	9	1	6	8	12	3	4
7	13	5	8	11	1	4	10	7	3
8	4	3	5	7	8	2	9	6	10
9	1	8	3	13	12	10	5	2	6
10	12	9	7	8	2	1	13	4	5
11	3	12	4	2	5	6	7	11	13
12	11	4	1	5	10	9	6	13	2
13	7	1	11	10	3	12	2	5	9

BIBD 2 for 9 objects in 12 sets of size 3			
Block	**1**	**2**	**3**
1	2	4	8
2	1	4	5
3	4	7	9
4	3	4	6
5	1	2	3
6	2	5	7
7	2	6	9
8	1	8	9
9	5	6	8
10	3	7	8
11	1	6	7
12	3	5	9

Blocks: BIBD 2	**Block 1 from BIBD 1**		
1	6	3	9
2	10	3	4
3	3	1	7
4	12	3	13
5	10	6	12
6	6	4	1
7	6	13	7
8	10	9	7
9	4	13	9
10	12	1	9
11	10	13	1
12	12	4	7

…

Blocks: BIBD 2	**Block 13 from BIBD 1**		
1	1	10	5
2	7	10	3
3	10	2	9
4	11	10	12
5	7	1	11
6	1	3	2
7	1	12	9
8	7	5	9
9	3	12	5
10	11	2	5
11	7	12	2
12	11	3	9

Table 1: BIBD for 6 Objects

Set	Object codes			Object names		
1*	1	2	5	Central Coast beach house	Katoomba upmarket hotel	South Coast, heritage village
2	2	3	6	Katoomba upmarket hotel	Barrington Tops, an isolated setting	Sydney, upmarket hotel
3	3	4	2	Barrington Tops, an isolated setting	Bowral, Southern Highlands	Katoomba upmarket hotel
4	4	1	3	Bowral, Southern Highlands	Central Coast beach house	Barrington Tops, an isolated setting
5	2	5	4	Katoomba upmarket hotel	South Coast, heritage village	Bowral, Southern Highlands
6	3	5	6	Barrington Tops, an isolated setting	South Coast, heritage village	Sydney, upmarket hotel
7	4	6	5	Bowral, Southern Highlands	Sydney, upmarket hotel	South Coast, heritage village
8	1	2	6	Central Coast beach house	Katoomba upmarket hotel	Sydney, upmarket hotel
9	5	1	3	South Coast, heritage village	Central Coast beach house	Barrington Tops, an isolated setting
10	6	4	1	Sydney, upmarket hotel	Bowral, Southern Highlands	Central Coast beach house

* For example, the BIBD design states that object codes 1, 2 and 5 should appear in set 1. From the list of object names, 'Central Coast beach house' is object 1, Katoomba upmarket hotel' is object 2 and 'South Coast, heritage village' is object 5, so those are the three destinations to appear in set 1.

Table 2: Example Survey BWS Task Based on Table 1

Most likely to visit	Comparison set 1	Least likely to visit
☐	Central Coast beach house	☐
☐	Katoomba upmarket hotel	☐
☐	South Coast, heritage village	☐

Most likely to visit	Comparison set 2	Least likely to visit
☐	Katoomba upmarket hotel	☐
☐	Barrington Tops, an isolated setting	☐
☐	Sydney, upmarket hotel	☐

…

Most likely to visit	Comparison set 10	Least likely to visit
☐	Sydney, upmarket hotel	☐
☐	Bowral, Southern Highlands	☐
☐	Central Coast beach house	☐

This table merely separates the sets into a respondent-friendly format. The respondent chooses the 'most likely' and 'least likely' destination to visit in each of the ten sets.

Table 3: One respondent's answers to the BIBD for 6 Objects

Set	Object number codes			Best	Worst
1*	1	2	5	1	5
2	2	3	6	2	6
3	3	4	2	2	3
4	4	1	3	1	3
5	2	5	4	2	5
6	3	5	6	3	5
7	4	6	5	4	5
8	1	2	6	1	6
9	5	1	3	1	5
10	6	4	1	1	6

* For example, in set 1 the hypothetical respondent picked 'Central Coast beach house' (object 1) as best (most likely destination to visit) and 'South Coast, heritage village' (object 5) as the worst (least likely destination to visit); in set 2 the person chose the 'Katoomba upmarket hotel' (object 2) as the best (most likely destination to visit) and the 'Sydney, upmarket hotel' (object 6) as the worst (least likely destination to visit).

Table 4: Total Frequency Counts, Best and Worst Choices

Object	Best Count	Worst Count	B-W Difference
1*	5	0	5
2	3	0	3
3	1	1	0
4	1	1	0
5	0	3	-3
6	0	5	-5

* For example, 'Central Coast beach house' (object 1) was chosen as best (most likely to visit) by the hypothetical respondent five times but was never chosen as worst (least likely to visit).

Table 5: Results for Getaways Study

Destinations	N	Calculated from Best and Worst Choice Totals										RP (Ave Reported Trips)
		Best Choices			Worst Choices			B-W Statistics				
		Sum	Mean	SE	Sum	Mean	SE	B-W	Mean	SE	Stdev	
Sydney	490	939	1.916	.090	951	1.941	.095	-12	-.025	.175	3.864	5.138
South Coast	490	768	1.567	.068	656	1.339	.064	112	.228	.120	2.658	5.534
Bowral	490	488	.996	.057	899	1.835	.067	-411	-.839	.111	2.465	1.186
Barrington	490	643	1.312	.072	970	1.980	.083	-327	-.668	.143	3.163	0.988
Katoomba	490	576	1.176	.063	868	1.771	.070	-292	-.595	.121	2.688	0.593
Central Coast	490	1347	2.749	.084	417	.851	.063	930	1.898	.136	3.001	8.300

Destinations	N	Estimated from The Best Choices							
		Cond Logit Model Results				Linear Prob Model Results			
		Est	SE	Wald	Sig	Est	SE	t	Sig
Sydney	490	-.053	.025	4.329	.037	.490	.018	27.156	.000
South Coast	490	.064	.024	7.187	.007	.536	.018	29.625	.000
Bowral	490	-.169	.025	45.098	.000	.447	.018	25.103	.000
Barrington	490	-.171	.026	44.434	.000	.447	.018	24.862	.000
Katoomba	490	-.077	.025	9.744	.002	.480	.018	27.009	.000
Central Coast	490	.407	.025	16.28	.000	.702	.018	38.091	.000

Table 6: Individual-Level LPM Estimates

Destination	N	Mean	StdErr	StdDev	RP
Sydney	420	0.734	0.030	0.624	5.138
South Coast	420	0.728	0.024	0.501	5.534
Bowral	420	0.575	0.022	0.453	1.186
Barrington	420	0.613	0.028	0.580	0.988
Katoomba	420	0.644	0.022	0.450	0.593
Central Coast	420	1.019	0.028	0.574	8.300

These results expand on the LPM ones of Table 5

Table 7: BWS Scores for Top 10 Journals

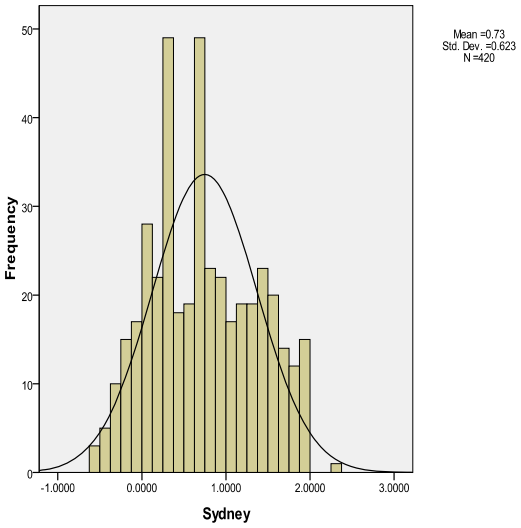
Rank	Journal	BWS Score
1	Journal of Marketing	100.00
	Journal of Marketing Research	98.01
2	Journal of Consumer Research	91.19
	Journal of the Academy of Marketing Science	90.06
	Marketing Science	89.47
3	Journal of Retailing	80.30
	Journal of Business Research	78.69
	Journal of Consumer Psychology	75.27
	International Journal of Research in Marketing	74.72
4	Journal of Advertising	71.74

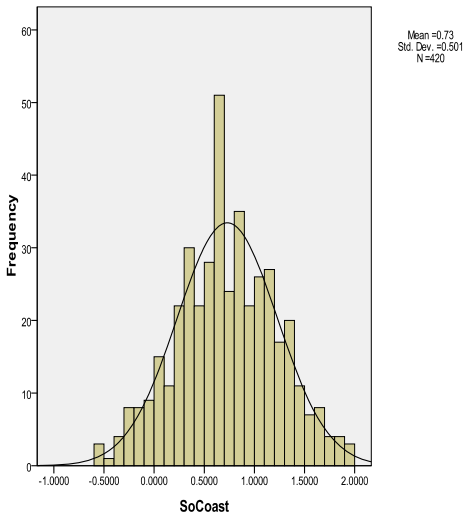
These scores represent ratio-scaled percentage quality scores for the top ranked journals.

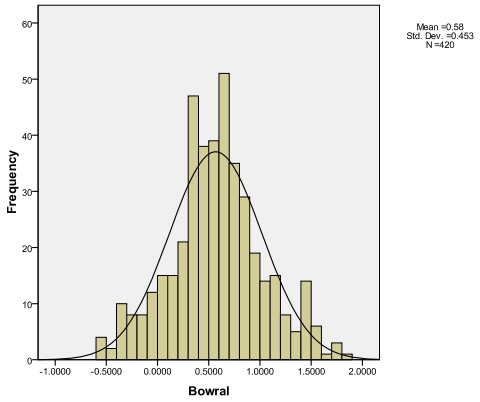
Objects to be evaluated	Theoretically any number of items can be compared with each other. It is important to note that this approach produces relative importance scores for the objects compared, indicating which are better, or worse, than others. BWS does not provide a subjective evaluation of how good or bad these objects are in absolute terms.
How to compare the objects	Small numbers of objects can be evaluated using paired comparisons, but this quickly becomes impractical as the number of comparison objects increases Moderate numbers of objects are better compared in sets created by BIBDs that are readily available, produce comparison sets of constant size and usually produce a number of comparison sets that are roughly equal to the total number objects to be compared. For large numbers of objects a single BIBD design typically produces insufficient data to rank middle-valued items. In this case, one can consider either collecting additional best and worst information or using a nested BIBD.
Designing the survey	Number the objects to be compared sequentially, and simply replace the numbers from the BIBD with the objects to be compared. Ask respondents to choose the best and worst from the list.
Choosing the sample size	Refer to Thompson (1987) or Rose (2011)
Pre-Examine the data	Check that respondents are consistent in their choices. For example, a histogram of the individual choices of each object provides insights into engagement with the survey. At an aggregate (sample) level the consistency of choices across the sample gives insights into consistency of views about the relative rank of objects.
Analyzing the data	Simple counts of best (+1) and worst (-1) provide insights into the preference of individuals. Best – worst (B-W) scores correlate well with revealed preferences and predict real behavior comparably with more sophisticated regression models.
Reporting the results	Results can be used to create a hierarchy of preferences – as in the examples in the paper, and/or to predict behavior, as shown in the regression examples.

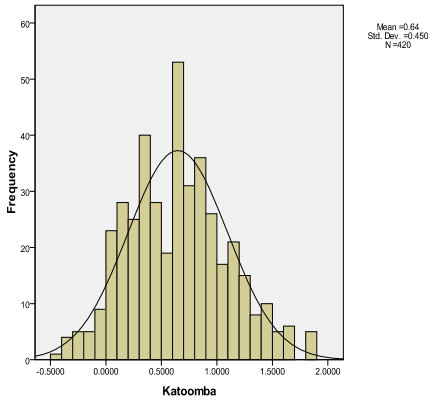
Table 8: steps in the design of a best-worst study

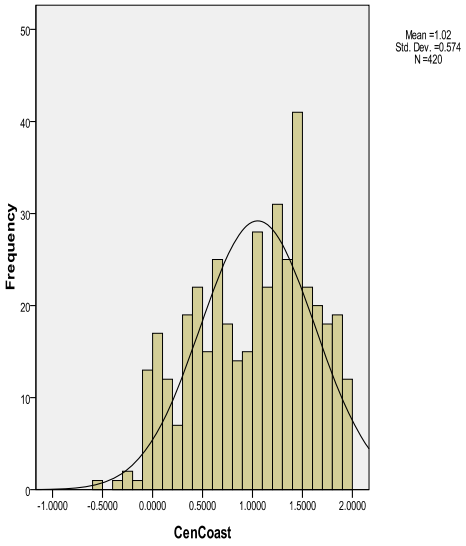
Appendix A: Histograms for individual-level LPM estimates

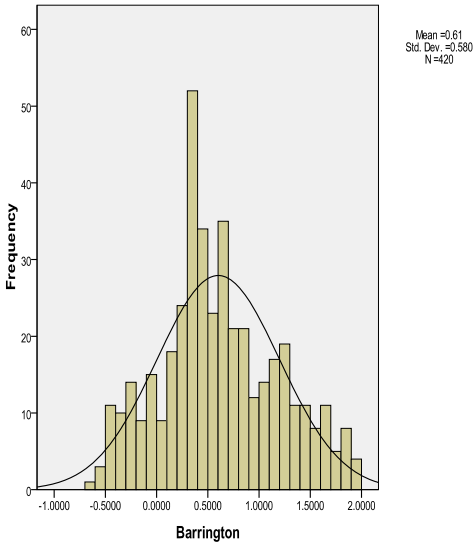












Appendix B: Journals – in alphabetical order

Academy of Mkting Science Review	**J. of MacroMkting**
Academy of Mkting Studies J.	**J. of Mkt-Focused Mgmt**
Advances in Consumer Research	**J. of Mkting**
Advances in Intnl Mkting	**J. of Mkting Channels**
Asia-Pacific J. of Mkting and Logistics	**J. of Mkting Communications**
Australasian Mkting J.	**J. of Mkting Education**
Australian J. of Mkt Research	**J. of Mkting for Higher Education**
European J. of Mkting	**J. of Mkting Mgmt**
Industrial Mkting Mgmt	**J. of Mkting Research**
Intnl J. of Advertising	**J. of Mkting Theory and Practice**
Intnl J. of Bank Mkting	**J. of Non Profit and Public Sector Mkting**
Intnl J. of Mkt Research	**J. of Personal Selling and Sales Mgmt**
Intnl J. of Nonprofit and Voluntary Sector Mkting	**J. of Product and Brand Mgmt**
Intnl J. of Research in Mkting	**J. of Public Policy and Mkting**
Intnl J. of Retail and Distribution Mgmt	**J. of Retailing**
Intnl J. of Service Industry Mgmt	**J. of Retailing and Consumer Services**
Intnl Mkting Review	**J. of Service Research**
J. of Advertising	**J. of Services Mkting (The)**
J. of Advertising Research	**J. of Strategic Mkting**
J. of Brand Mgmt	**J. of Targeting, Measurement and Analysis for Mkting**
J. of Business and Industrial Mkting	**J. of the Academy of Mkting Science**
J. of Business Research	**J. of Vacation Mkting**
J. of Business-to-Business Mkting	**Mkting Bulletin**
J. of Consumer Behaviour	**Mkting Education Review**
J. of Consumer Mkting	**Mkting Health Services**
J. of Consumer Psychology	**Mkting Intelligence and Planning**
J. of Consumer Research	**Mkting Letters**
J. of Current Issues and Research in Advertising	**Mkting Mgmt**
J. of Customer Behaviour	**Mkting Research**
J. of Database Mkting	**Mkting Science**
J. of EuroMkting	**Mkting Theory**
J. of Fashion Mkting and Mgmt	**Mkting Week**
J. of Global Mkting	**Psychology and Mkting**
J. of Interactive Mkting	**Qualitative Mkt Research: An Intnl J.**
J. of Intnl Consumer Mkting	**Services Mkting Quarterly (formerly J. of Professional Services Mkting)**
J. of Intnl Mkting	**Sport Mkting Quarterly**
J. of Intnl Mkting and Mkting Research	

Appendix C: Academic Marketing Journals-Overall Scores and Regional Scores

Overall Sample	North America	Aust / NZ	Europe	Journal
100.00	100.00	98.89	93.89	Journal of Marketing
98.01	91.83	100.00	100.02	Journal of Marketing Research
91.19	93.20	86.89	81.80	Journal of Consumer Research
90.06	89.87	97.21	73.89	Journal of the Academy of Marketing Science
89.47	88.25	97.96	77.94	Marketing Science
80.30	83.31	67.95	85.72	Journal of Retailing
78.69	86.90	69.09	70.00	Journal of Business Research
75.27	77.73	73.28	73.97	Journal of Consumer Psychology
74.72	61.98	81.77	78.11	International Journal of Research in Marketing
71.74	69.98	65.25	77.83	Journal of Advertising
68.47	63.97	73.10	66.72	Journal of Advertising Research
67.74	59.07	83.75	61.62	European Journal of Marketing
66.08	67.57	66.74	61.66	Journal of Service Research
65.67	69.11	73.04	47.21	Psychology and Marketing
63.65	65.19	87.83	53.84	Marketing Letters
59.30	60.08	52.73	61.14	Advances in Consumer Research
57.17	68.80	49.58	27.66	Journal of Public Policy and Marketing
53.96	41.26	60.26	61.51	International Journal of Market Research
53.43	49.15	51.04	63.88	Journal of International Marketing
52.19	49.99	41.44	58.68	Industrial Marketing Mgmt
51.14	40.11	67.00	45.48	Academy of Marketing Science Review
50.59	59.20	30.53	22.64	Journal of Personal Selling and Sales Mgmt
50.21	54.48	52.77	25.44	Journal of Services Marketing
48.42	47.14	57.93	45.71	Journal of Product and Brand Mgmt
48.00	41.63	65.88	42.41	Journal of Consumer Behavior
46.71	53.23	24.98	39.92	Journal of Macromarketing
46.15	40.52	41.70	56.25	Marketing Theory
44.35	37.36	50.05	46.83	International Marketing Review
43.73	46.39	38.74	35.05	Journal of Marketing Theory and Practice
43.21	36.73	44.71	46.48	Journal of Marketing Mgmt
42.14	35.63	44.88	29.76	Journal of Strategic Marketing
39.38	40.65	41.29	29.62	Journal of Business-to-Business Marketing
38.52	33.14	41.90	42.76	Journal of Retailing and Consumer Services
37.95	44.35	29.87	23.74	Journal of Consumer Marketing
37.12	28.56	35.78	49.82	Journal of Business and Industrial Marketing
36.86	28.52	44.89	54.99	International Jl Retail & Distribution Mgmt
36.34	22.97	38.58	41.09	Jl of International Marketing and Marketing Research
35.80	44.30	27.06	19.30	Journal of Marketing Education
35.16	30.38	34.30	35.61	International Journal of Advertising
34.95	27.24	49.86	26.65	Marketing Research

34.19	42.76	30.85	16.39	Journal of Brand Mgmt
33.17	28.22	33.01	29.82	International Journal of Service Industry Mgmt
32.36	28.89	22.44	29.51	Journal of Interactive Marketing
32.36	32.81	34.35	24.43	Journal of Global Marketing
31.98	20.80	37.74	35.92	Marketing Intelligence and Planning
29.13	19.91	26.46	18.85	Journal of Customer Behavior
29.11	31.86	28.44	9.60	Services Marketing Qtrly (aka Jl Prof'l Services Mktg)
28.07	33.53	16.24	12.35	Journal of Current Issues and Research in Advertising
28.03	28.14	23.35	18.79	Jl of Targeting, Measurement & Analysis for Mktg
28.02	27.92	27.48	19.87	Marketing Mgmt
27.25	22.52	38.16	13.79	Qualitative Market Research: An International Journal
26.96	26.08	21.69	20.03	Advances in International Marketing
26.67	27.13	23.66	19.92	Journal of Marketing Channels
26.50	20.57	28.28	17.26	Journal of International Consumer Marketing
26.48	22.28	34.45	13.89	Journal of Marketing Communications
24.46	13.74	40.89	16.57	Australasian Marketing Journal
23.95	17.75	26.86	12.38	Journal of Non Profit and Public Sector Marketing
23.21	21.56	22.56	20.65	Academy of Marketing Studies Journal
22.07	19.61	35.72	10.09	Australian Journal of Market Research
21.55	24.97	15.06	10.43	Marketing Education Review
20.77	20.43	14.14	18.88	Journal of Database Marketing
18.89	9.61	34.11	10.74	Asia-Pacific Journal of Marketing and Logistics
17.81	20.71	13.94	16.77	International Jl of Nonprofit & Voluntary Sector Mktg
17.41	19.25	9.86	14.62	Journal of Market-Focused Mgmt
16.91	16.60	16.27	10.23	Marketing Health Services (aka Jl Health Care Mktg)
16.31	14.52	11.68	14.95	Journal of Marketing for Higher Education
16.01	19.73	10.53	0.65	Journal of Euromarketing
12.35	6.95	15.77	4.75	Marketing Bulletin
10.10	8.26	5.48	17.03	International Journal of Bank Marketing
9.59	11.83	3.87	-3.12	Sport Marketing Quarterly
7.73	1.60	6.37	3.11	Journal of Fashion Marketing and Mgmt
7.32	4.90	2.22	0.00	Journal of Vacation Marketing
0.00	0.00	0.00	-8.11	Marketing Week