

# Best Worst Scaling: Theory and Methods

T. N. Flynn<sup>\*,1</sup> and A. A. J. Marley<sup>1,2</sup>

*Handbook of Choice Modelling*

<sup>1</sup>T. N. Flynn  
\*(corresponding author)  
Centre for the Study of Choice  
University of Technology Sydney  
PO Box 123 Broadway  
Sydney, NSW 2007  
Australia  
email: Terry.Flynn@uts.edu.au

<sup>2</sup>A.A.J. Marley  
Department of Psychology  
University of Victoria  
Victoria  
BC V8W 3P5  
Canada  
email: ajmarley@uvic.ca

## Abstract

Best-Worst Scaling (BWS) can be a method of data collection, and/or a theory of how respondents provide top and bottom ranked items from a list. BWS is increasingly used to obtain more choice data from individuals and/or to understand choice processes. The three “cases” of BWS are described, together with the intuition behind the models that are applied in each case. A summary of the main theoretical results is provided, including an exposition of the possible theoretical relationships between estimates from the different cases, and of the theoretical properties of “best minus worst scores.” BWS data can be analysed using relatively simple extensions to maximum-likelihood based methods used in discrete choice experiments. These are summarised, before the benefits of simple functions of the best and worst counts are introduced. The chapter ends with some directions for future research.

keywords: best worst scaling; discrete choice experiments; maxdiff model; repeated best and/or worst choice: scores.

# 1 Introduction

Best-Worst Scaling (BWS) can be a method of data collection, and/or a theory of how respondents provide top and bottom ranked items from a list. We begin with a brief history, followed by motivations for the use of BWS. The three types (“cases”) of BWS will then be described in detail, before issues in the conceptualisation (modelling) and analysis of best-worst data are discussed. The chapter will end with a summary section which outlines future research issues.

## 1.1 Brief history

In 1987, whilst working at the University of Alberta, Jordan Louviere became interested in what he could do with information about the “least preferred” item from a choice set, in addition to the traditional “most preferred” item. He was primarily interested in whether a PhD student could “do” the task, and in what extra information about her utility function could be elicited. His initial focus was on “objects”, such as attitudes, general public policy goals, brands, or anything that did not require description in terms of attributes and levels. As such, his first peer-reviewed journal article examined the degree of concern the general public had for each of a set of food safety goals, including irradiation of foods and pesticide use on crops (Finn & Louviere, 1992). Figure 1 contains a BWS question similar to ones used in that study.

---

Insert Figure 1 about here

---

Finn and Louviere proposed using BWS in place of category rating scales for several reasons. First, rating scales do not force respondents to discriminate between items, allowing them to state that multiple items are of similarly high importance. Second, interpreting what the rating scale values mean is difficult. Third, the reliability and validity of rating scales are frequently unknown and unknowable. BWS addresses these issues by valuing items within a random utility framework (Thurstone, 1927; McFadden, 1974): choice frequencies provide the metric by which to compare the importance of items and the use of a model with an error theory enables predictions to be made as to how often one item might be picked over any other. Such inferences provided real life significance of the method and avoided key problems with rating scales, such as “what does 7 out of 10 mean in terms of real life choices?”

The 1992 paper modelled best and worst choices among relatively simple items, such as goals or attitudes, which Louviere generically referred to as objects. However, he had already begun applying BWS to more complex items. These were either attribute-levels describing a single alternative (profile), or were complete alternatives (profiles) of the type familiar to choice modellers. The former case, requiring respondents to identify the best attribute-level and

worst attribute-level within an alternative, was a relatively unfamiliar task to choice modellers. However, the latter case had the potential to become the most widely accepted case of BWS, by being “merely” an extension of the method to a discrete choice experiment (DCE: Louviere, Hensher & Swait, 2000; Hensher, Rose & Greene, 2005). In practice only the first case of BWS (considering objects) received any sustained interest in academia before 2005, principally in marketing among practitioners who were unhappy with rating scales. In particular, Steve Cohen won a number of best paper awards at ESOMAR conferences for his application of BWS to objects (Cohen & Neira 2003, 2004). Since the mid 2000s, there has been increasing interest in the other two cases of BWS, particularly within the fields of health and economics. This prompted Louviere, Flynn and Marley (2012) to try to standardise terminology across the fields and provide cross-disciplinary guides to BWS that included motivations for its use.

## 1.2 Motivation.

During the 1980s Louviere and colleagues (Louviere & Hensher, 1982; Louviere & Woodworth, 1983) pioneered discrete choice experiments. These had advantages over traditional conjoint measurement techniques in terms of sound theoretical underpinnings (random utility theory) and the need to make fewer and weaker assumptions about human decision-making (Louviere, Flynn & Carson, 2010): for example, assumptions about how people deal with numbers to answer rating scale questions were no longer required. This move away from direct quantitative valuation towards discrete choice models came at a cost: there was usually no longer enough information to estimate models for single individuals and valid inferences were typically only possible after aggregating across groups of respondents. Therefore Louviere’s initial motivation for inventing BWS was to obtain more data from a given respondent.

Louviere could have obtained more information from a respondent by simply asking her to answer more choice sets. The motivation for BWS was the following: if the respondent has already become familiar with a choice set containing 3 or more items by choosing “best”, why not simply exploit this, together with the human skill at identifying extremes (Helson, 1964), and ask her for “worst”? This would shorten the length of the DCE, certainly in terms of the number of choice sets administered, and potentially in terms of the total time taken.

Obtaining more information about a list of items was not a new idea. Ranking models were first formulated by Luce, Marley and others (Luce, 1959; Luce & Suppes, 1965; Marley, 1968). With the exception of the rank ordered logit regression model (Beggs, Cardell and Hausman, 1981; Hausman & Ruud, 1987), closed form random utility models that could be estimated using the computers available at the time were unavailable. Marley (1968) had put forward ideas on partial and complete ranking in his PhD thesis but had not taken them forward; Louviere credits Marley’s concepts of the “superior” and “inferior” items in a list as his inspiration for BWS. Indeed Louviere believed models of best-worst choices could be axiomatised, which would help establish their credentials to the wider research community, and he began collaborating with Marley to achieve

this goal (Marley & Louviere, 2005; Marley et al., 2008). In parallel to this, Louviere continued research into the use of BWS merely as a convenient way (via repeated application) to collect a complete ranking of all items in a choice set. That work was conducted primarily with respect to multi-attribute consumer goods such as mobile phones, although it could easily have been used in studies examining simpler objects of the types mentioned above.

Finally, as already mentioned briefly, Louviere was motivated to use another type of BWS (that requiring respondents to pick the best and worst attribute-levels) to address a major stumbling block encountered in discrete choice models: the confound between attribute weight and level scale (Anderson, 1970; Marley, Flynn & Louviere, 2008). Section 2.2 describes this issue in more detail but suffice to note that this confound is not addressed in traditional economic theory, and is different from the mean-variance confound that arises for all limited dependent variable (including logit and probit) models (Yatchew & Griliches, 1985).

### 1.3 BWS as a method of data collection vs BWS as a theory

BWS was initially used mainly as a method of collecting data in a cost-efficient manner, though Finn and Louviere (1992) did introduce the maximum difference (maxdiff) model for best-worst choice (see below). However, the work with Marley spurred research into BWS as a theory, explaining the processes that individuals might follow in providing best and worst data. The 2005 paper introduced sequential and maxdiff models of best-worst choices (Marley & Louviere, 2005). The former assumes the individual provides best and worst in a particular order whilst the latter is a well-established model that assumes a simultaneous choice of that pair of items that maximises the difference between them on a latent (usually utility) scale. Some applied practitioners may regard process issues as esoteric, but there is increasing evidence that these may matter when interest is in choice models for particular individuals. Moreover, what are called “maxdiff scaling” models in some estimation software are sequential, not maxdiff, ones.

Recent work by Louviere has returned to the use of BWS as a method of data collection (Louviere et al., 2008): its purpose, via repeated rounds of best-worst choices, is simply to obtain a full ranking of items in a manner that is “easy” for respondents. Analysis proceeds by expanding the data to all the implied choice sets, and then using statistical models based on first choices (either conditional or binary logistic regression). That work is not covered here as it is the subject of Chapter xx; also see the class of *weighted utility ranking models* (Marley & Islam, 2012), which are motivated, in part, by those expansion methods.

The next section describes the three types (“cases”) of BWS in detail; it is important to note that BWS can be used as a data collection method and/or a theory of process for all three cases.

## 2 BWS: The Three Cases

Louviere developed three cases of BWS, which differ in the nature and complexity of the items being chosen. Case 1 (the Object case) is the simplest, whilst Cases 2 and 3 (the Profile and Multi-profile cases) involve an attributes and levels structure that should be familiar to choice modellers. The frequent lack of clarification in many published articles as to which case is being used reflects the fact that different disciplines have tended to embrace different cases. Academic researchers from marketing, food science and personality assessment tend to be familiar with case 1 whilst those working in health are familiar with case 2 (and increasingly, case 3) and marketing industry practitioners tend to use case 3. This section concentrates on the principles and design of the three cases; Section 3 present details on the related models.

### 2.1 Case 1 (Object case)

Case 1 BWS is appropriate when the researcher is interested in the relative values associated with each of a list of objects. These might be brands, public policy goals, or any set of objects that can be meaningfully compared. Generally, these will not be described in terms of an attribute and level structure. However, if the researcher is interested in valuing items such as brands, (s)he must recognise that respondents might infer particular levels of key attributes when considering these: instructions to respondents must be carefully worded to standardise any such inferences if estimates of (for example) airline carrier are not to be confounded with estimates of assumed levels of service. As mentioned above, the first peer-reviewed case 1 study investigated food safety. It made it clear that the latent scale of interest does not have to be “utility”; degree of concern was key and other metrics may be of relevance, depending on the application. Indeed many applications of category rating scales are amenable to replacement by best-worst questions.

Once the researcher has chosen the list of objects, (s)he must present choice sets of these to respondents to obtain best and worst data. Choice sets here serve a similar purpose to those in traditional DCEs: statistical designs are implemented that include (some or all) subsets of all possible items which, with suitable assumptions, facilitate inferences about the value associated with the wider list of objects. More specifically, choice frequencies across all sets are used to estimate the relative values associated with objects. Since there is no attribute and level structure to consider, Case 1 designs are typically less complex (and less problematic) than those for DCEs. Early work by Louviere utilised  $2^J$  designs, extending his seminal work on DCEs (Louviere & Hensher, 1982; Louviere & Woodworth, 1983). Such designs are so-called because for  $J$  objects, there are  $2^J$  distinct choice sets possible. Table 1 gives all 16 choice sets for 4 objects - there is one full set of four, four triples, six pairs, four singletons and the null (empty) set.

---

Insert Table 1 about here

---

Fractions of a  $2^J$  design can be used to keep the number of choice sets small, using similar principles to those used in DCEs (for example main effects designs). The potential problems with these designs are psychological rather than statistical in origin. The size of the choice set is not constant and respondents may infer things that have no relevance for the outcome of interest: they might decide that objects in small choice sets “must be somehow important to the researcher so I’ll pay more attention to those”.

Today, Balanced Incomplete Block Designs (BIBDs) are more common. A BIBD ensures that occurrence and co-occurrences of objects is constant, helping minimise the chance that respondents can make unintended assumptions about the objects based on aspects of the design. For example, a study of importance of nine short-term investment priorities in Sydney’s public transport systems used a BIBD which presented 12 sets of size three. Each object appeared four times across the design and each pair of objects appeared once. BIBDs are available from design catalogues, such as that in Street & Street (1987). Unfortunately there are some numbers for which there are no BIBDs, whilst other numbers have two or more possible BIBDs (varying in the number and size of choice sets). In the former case, it is best to include some “realistic but irrelevant” objects to make the number up to one for which there is a BIBD; an alternative strategy of using a statistical algorithm to produce a “nearly” balanced design risks problems similar to those above in terms of what the respondents are assuming. Furthermore, some of the attractive analysis methods to be discussed become problematic.

An attractive feature of BIBDs is that the number of choice sets is often not markedly different from the number of objects being valued. This, together with the fact they force respondents to discriminate between objects, makes case 1 BWS attractive in comparison with category rating scale surveys. Those BIBDs that ensure that every object appears in every possible position the same number of times are called Youden designs and represent the most robust defence against any propensity of the respondent to “read too much into” the size or composition of the choice sets on offer.

## 2.2 Case 2 (Profile case)

Case 2 BWS is largely unknown outside of the field of health economics, to which it was introduced by McIntosh and Louviere in a conference paper (2002). It is easiest to describe using an example (Figure 2) based on a dermatology study (Coast et al., 2006; Flynn et al., 2008).

---

Insert Figure 2 about here

---

The set looks like a single profile (alternative) from a DCE or conjoint study. However, the choices the respondent is asked to make do not require him/her to consider the value of the profile as a whole. Instead, (s)he must consider the attribute levels that describe it, choosing the one that is best (most attractive) and the one that is worst (least attractive). Case 2 BWS is most popular in health because (1) the systems of many industrialised countries do not typically give opportunities for patients to become “experienced consumers” and (2) health care goods/services can be complicated and even pairs of specifications (in a simple DCE) may lead to unacceptable cognitive burden, particularly among vulnerable patient groups.

In some respects, Case 2 is merely Case 1 with the objects grouped into an attribute and level structure. However, what makes Case 2 unique is that attribute-levels are only meaningful when forming a profile. Thus, if meaningful profiles are to be presented, two levels of the same attribute cannot compete with one another; each level competes against a level from every other attribute. This means that designs for Case 1 are generally inappropriate for Case 2. However, Case 2 design is relatively easy for those researchers who would generate a DCE design the following way:

- (1) Use a “starting” design to produce (for example) the “left-hand side profile” in every choice set, then
- (2) Use some statistical procedure to produce the other profiles in each choice set, from the “left-hand side ones.”

In particular, Case 2 design involves only step (1): there are no “other” profiles in each choice set since the choice set is the profile. Whilst this makes Case 2 design easy in some respects, it has potential problems, which can be serious depending on the issue of interest. These generally arise when attributes all have ordered levels. We use a profile from the EQ-5D health state classification system (Figure 3) to illustrate this.

---

Insert Figure 3 about here

---

All five attributes of the EQ-5D have ordered levels, generally representing “no problems” through to “severe problems”. Presenting this particular profile to respondents would be unwise since the best and worst choices are obvious (to anyone who isn’t a masochist). Conceptually, the random utility term is likely identically equal to zero (not a sampling zero), violating random utility theory and thereby biasing regression estimates (if they are estimable at all). It is the researcher’s responsibility to code with care, so as to minimise the number of choice sets with this property. Unfortunately, as the design becomes larger (so as to estimate interaction terms), this becomes impossible. The first author recently advised a company that at least 75% of its Case 2 data were worthless, thanks to a major choice modelling company providing it with a design that was almost the full factorial. Thus, for many Case 2 applications it may be difficult, or impossible, to estimate interaction terms.



Louviere originally anticipated that Case 2 BWS would allow decomposition of attribute weight and level scale values (McIntosh & Louviere, 2002), a 40 year old problem in mathematical psychology (Anderson, 1970). That is, it is assumed that there is a multiplicative relationship between the importance of an attribute per se – which might vary depending on the context of the choice task – and the level scale of an attribute level – which, conceptually, should be fixed in value no matter what the context of the choice. McIntosh and Louviere were partly right: Marley et al. (2008) proved that although Case 2 BWS does not enable estimation of attribute importance, it does enable the direct estimation of attribute impact, a weaker concept that represents the average utility of an attribute across all its levels. Also, as shown in Section 3.1.1, a case 2 study, in combination with a case 3 study on the same (or suitably related) profiles, in principle may allow the separate measurement of attribute weight and level scale value; should such a study be successful, it would solve this classic measurement problem.

### 2.3 Case 3 (Multi-profile case)

Case 3 BWS is perhaps the most accessible (conceptually at least) to DCE practitioners. It “merely” requires respondents to choose the worst (least attractive) profile/alternative as well as the best (most attractive) one in a DCE. Therefore, virtually all (non pairwise) DCEs administered by CenSoC at UTS are now Case 3 BWS studies.

---

Insert Figure 4 about here

---

Figure 4 provides an example task from a mobile phone study. The increasing use of web-based administration makes expansion of DCEs into Case 3 BWS studies easy and cost-efficient. The additional data provided are valuable in many marketing applications: the additional information obtained about the consumer’s utility function is valuable both for its own sake and in identifying attribute levels that make a good “unacceptable”. The consumer who trades off attributes in a manner predicted by traditional economic theory when choosing most preferred might demonstrate lexicographic preferences when answering least preferred. Such information is valuable to the marketer who wishes to ensure that a product passes an initial “consideration” test by consumers: having an attractive price and a set of desirably valued attributes is of no use if a level on another attribute rules it out of consideration. Availability of data and such “real life” marketing issues caused Case 3 BWS studies to be the primary vehicle for empirical investigations of choice process issues to date. However, process issues apply to all three cases of BWS and these issues are dealt with next.

### 3 Multinomial Logit Models of Best and/or Worst Choice

We begin with notation that applies to all Cases (1, 2, and 3) and talk of “choice options” (or “options”) without distinguishing between objects (Case 1), attribute-levels of a profile (Case 2), and profiles (Case 3). For later results, we need additional notation for Case 2 and Case 3. We also present the results in terms of a numeric “utility” value associated with each choice option (and, as relevant, with each of its attribute-levels), rather than in terms of the utility coefficients (“beta weights”) that are standard in the discrete choice literature; we do this because various theoretical results on BWS can only be stated and proved in the former notation (for example, those in Marley & Louviere, 2005; Marley et al., 2008; Marley & Pihlens, 2012). However, for the reader’s benefit, we do introduce the utility coefficient notation when discussing Case 3.

Let  $S$  with  $|S| \geq 2$  denote the finite set of potentially available choice options, and let  $D(S)$  denote the *design*, i.e., the set of (sub)sets of choice alternatives that occur in the study. For example, participants might be asked about their preferences for mobile phones by repeatedly asking them for choices amongst a sets of four different options:  $S$  represents the collection of mobile phones in the study, and each element of the set  $D(S)$  represents the set of options provided on one particular choice occasion. For any  $Y \in D(S)$ , with  $|Y| \geq 2$ ,  $B_Y(x)$  denotes the probability that alternative  $x$  is chosen as best in  $Y$ ,  $W_Y(y)$  the probability that alternative  $y$  is chosen as worst in  $Y$ , and  $BW_Y(x, y)$  the probability that alternative  $x$  is chosen as best in  $Y$  and the alternative  $y \neq x$  is chosen as worst in  $Y$ . Most previous work using similar mathematical notation has used  $P_Y(x)$  or  $P(x|Y)$  where we use  $B_Y(y)$ . We use the latter for best, and  $W_Y(y)$  for worst, to distinguish clearly between such *Best* and *Worst* choice probabilities.

Many models of choice, especially those involving best-worst scaling, are based on extensions of the *multinomial logit* (MNL) model. The best choice MNL model assumes there is a scale  $u$  such that for all  $y \in Y \in D(S)$ ,

$$B_Y(y) = \frac{e^{u(y)}}{\sum_{z \in Y} e^{u(z)}}. \quad (1)$$

The value  $u(y)$  for an option  $y$  is interpreted as the utility for that option. The representation restricted to  $Y \subseteq S$ ,  $|Y| = 2$ , is the *binary MNL* model. Various results show that  $u$  can be assumed to be a *difference scale* - that is one, that is unique up to an origin. For instance, the *Luce choice model* corresponds to the MNL model when the latter is written in terms of  $b = e^u$ ;  $b$  is shown to be a ratio scale in Luce (1959), which implies that  $u$  is a difference scale. Parallel observations hold for the representations throughout this paper that are written in terms of  $u$ , where the quoted theoretical results were obtained using  $b = e^u$ .

The parallel MNL model for worst choices assumes there is a scale  $v$  such

that for all  $y \in Y \in D(S)$ ,

$$W_Y(y) = \frac{e^{v(y)}}{\sum_{z \in Y} e^{v(z)}}.$$

Marley and Louviere (2005) present a theoretical argument for the case where  $v = -u$ , i.e., we have

$$W_Y(y) = \frac{e^{-u(y)}}{\sum_{z \in Y} e^{-u(z)}}. \quad (2)$$

Assuming  $v = -u$  in the MNL models for best and worst implies that the probability that  $y \in Y$  is selected as *best* in a set  $Y$  with scale values  $u(z)$ ,  $z \in Y$ , is equal to the probability that  $y \in Y$  is selected as *worst* in a set  $Y$  with scale values  $-u(z)$ ,  $z \in Y$ . In particular, when (1) and (2) both hold, we have that for all  $x, y \in X, x \neq y$ ,  $B_{\{x,y\}}(x) = W_{\{x,y\}}(y)$ , and write the common value as  $p(x, y)$ .

We now present three MNL-based models for best-worst choice.

Perhaps the most natural generalization of the above MNL models to best-worst choice is the *maxdiff model*. This model makes the strong assumption that the utility of a choice alternative in the selection of a best option is the negative of the utility of that option in the selection of a worst option, and this utility scale  $u$  is such that for all  $x, y \in Y \in D(S)$ ,  $x \neq y$ ,

$$BW_Y(x, y) = \frac{e^{[u(x)-u(y)]}}{\sum_{\substack{\{p,q\} \in Y \\ p \neq q}} e^{[u(p)-u(q)]}}. \quad (3)$$

It is known that the three models (1), (2), and (3), with a common scale  $u$ , satisfy a common random utility model, based on the extreme value distribution<sup>1</sup> - it is the *inverse extreme value maximum<sup>2</sup> random utility model* (Marley & Louviere, 2005, Def. 11, and Appendix A of this chapter). An alternate description that leads to the maxdiff model is given by the following process description (Marley & Louviere, 2005): Assume that the best (respectively, worst) choice probabilities satisfy (1) (respectively, (2)); equivalently, each of these choice probabilities is given by the relevant component of the above random utility model. A person chooses a best and a worst option, independently, according to the above best (respectively, worst) random utility process; if the resultant best and worst options are different, these form the best-worst choice pair; otherwise, the person resamples both the best and the worst option until the selected options are different. Marley and Louviere (2005) show that, under the above assumptions, this process also gives the maxdiff representation, (3).

Once one begins thinking of best and worst choices as possibly being made sequentially, there are several other plausible models forms for the combined

<sup>1</sup>This means that: for  $-\infty < t < \infty$   $\Pr(\varepsilon_z \leq t) = \exp -e^{-t}$  and  $\Pr(\varepsilon_{p,q} \leq t) = \exp -e^{-t}$ .

<sup>2</sup>We have added *maximum* to Marley & Louviere's definition to emphasize that the random utility models of *choice* are written in terms of maxima, whereas the equivalent ("horse race", accumulator) models of response time are written in terms of minima. See Hawkins et al. (2012) for such response time models for best-worst choice.

best-worst choices. Assuming the best choices satisfy (1) and the worst choices satisfy (2), the *best then worst* MNL model states: for all  $x, y \in Y \in D(S)$ ,  $x \neq y$ ,

$$BW_Y(x, y) = B_Y(x)W_{Y-\{x\}}(y),$$

Similarly the *worst then best* MNL model states: for all  $x, y \in Y \in D(S)$ ,  $x \neq y$ ,

$$BW_Y(x, y) = W_Y(y)B_{Y-\{y\}}(x).$$

Repeated best-worst choices satisfying a common one of the above models lead naturally to models of rank order data. Various authors are exploring such models, and their generalizations to include heterogeneity - see Collins and Rose (2011), Scarpa and Marley (2011), Scarpa, Notaro, Raffelli, Pihlens, and Louviere (2011), Marley and Pihlens (2012), and Marley and Islam (2012).

For the additional material on Case 2 and Case 3, we limit consideration to the maxdiff model (for best-worst choice). Parallel notation applies to models based on repeated best and/or worst choices.

The notation already introduced, where  $x, y$ , etc., denoted generic objects, is all we need to state later theoretical results for Case 1. However, for Case 2 and Case 3 we need the following additional notation.

There are  $m$  attributes, usually with  $m \geq 2$ , and we let  $M = \{1, \dots, m\}$ . Attribute  $i$ ,  $i = 1, \dots, m$ , has  $q(i)$  levels; we call these *attribute-levels* and sometimes let  $p, q$  denote typical attribute-levels, with the context making clear which attribute is involved. A *profile* (traditionally called a *multiattribute option*) is an  $m$ -component vector with each component  $i$  taking on one of the  $q(i)$  levels for that component. Given a set  $P$  of such profiles, let  $D(P)$  denote the *design*, i.e., the set of (sub)sets of profiles that occur in the study. We denote a typical profile by

$$\mathbf{z} = (z_1, \dots, z_m), \tag{4}$$

where  $z_i$ ,  $i = 1, \dots, m$ , denotes the level of attribute  $i$  in profile  $\mathbf{z}$ . For Case 1, we assume that each object  $x$  has a scale value  $u(x)$ , so the representation of the maxdiff model is as in (3); it follows from the results in Marley and Louviere (2005) that  $u$  is a difference scale, i.e., unique up to an origin<sup>3</sup>. For Case 2, we assume that attribute-level  $p$  has a scale value  $u(p)$ , and the representation of the maxdiff model is as in (3); it follows from the results in Marley and Louviere (2005) that  $u$  is a difference scale, i.e., unique up to an origin. For Case 3, we assume that each profile  $\mathbf{z}$  has a scale value  $u^{(m)}(\mathbf{z})$  with  $u^{(m)}$  a difference scale. We also assume the additive representation

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

where each  $u_i$  is a separate (different) difference scale.

---

<sup>3</sup>That paper uses a representation in terms of  $b = e^u$ , and  $b$  is shown to be a ratio scale, i.e., unique up to a multiplicative scale factor. This implies that  $u$  is a difference scale, i.e., unique up to an additive constant (or origin). This relation holds for all the results stated in this paper as having been demonstrated in Marley & Louviere (2005), Marley, Flynn & Louviere (2008) or Marley & Pihlens (2012).

For Case 2, we use the following revised notation: for a typical profile  $\mathbf{z} \in P$ , let  $Z = \{z_1, \dots, z_m\}$  and let  $BW_Z(z_i, z_j)$  denote the probability that, jointly, the attribute-level  $z_i$  is chosen as best in  $Z$  and the attribute-level  $z_j$  is chosen as worst in  $Z$ .

**Definition 1** (Adapted from Marley, Flynn, & Louviere, 2008, Def. 12) *A set of best-worst choice probabilities on a finite set of profiles  $P$  satisfies an **attribute-level maxdiff model (on single profiles)** iff there exist a real-valued scale  $u$  on the attributes such that for every profile  $\mathbf{z} \in P$  [equivalently, for every such  $Z = \{z_1, \dots, z_m\}$ ] and  $i, j \in M$ ,  $i \neq j$ ,*

$$BW_Z(z_i, z_j) = \frac{e^{[u(z_i) - u(z_j)]}}{\sum_{\substack{k, l \in M \\ k \neq l}} e^{[u(z_k) - u(z_l)]}} \quad (i \neq j). \quad (5)$$

Marley et al. (2008) show that, under reasonable mathematical assumptions,  $u$  is a difference scale, i.e., unique up to an origin.

For Case 3, we use the following revised notation: for typical profiles  $\mathbf{x}, \mathbf{y} \in P$ ,  $BW_X(\mathbf{x}, \mathbf{y})$  is the probability that, jointly, the profile  $\mathbf{x}$  is chosen as best in  $X$  and the profile  $\mathbf{y}$  is chosen as worst in  $X$ .

For completeness, we first present the maxdiff model on profiles written with the options in boldface to represent those profiles.

**Definition 2** *A set of best-worst choice probabilities for a design  $D(P)$ ,  $P \subseteq Q$ ,  $|P| \geq 2$  satisfies a **maxdiff model on profiles** iff there exist a real-valued scale  $u^{(m)}$  on  $P$  such that for every  $\mathbf{x}, \mathbf{y} \in X \in D(P)$ ,  $\mathbf{x} \neq \mathbf{y}$ ,  $|X| \geq 2$ ,*

$$BW_X(\mathbf{x}, \mathbf{y}) = \frac{e^{[u^{(m)}(\mathbf{x}) - u^{(m)}(\mathbf{y})]}}{\sum_{\substack{\mathbf{r}, \mathbf{s} \in X \\ \mathbf{r} \neq \mathbf{s}}} e^{[u^{(m)}(\mathbf{r}) - u^{(m)}(\mathbf{s})]}} \quad (\mathbf{x} \neq \mathbf{y}). \quad (6)$$

Marley et al. (2008, Theorem. 8) show that, under reasonable mathematical assumptions,  $u^{(m)}$  is a difference scale, i.e., unique up to an origin.

In the case where

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

we have the following representation of the maxdiff model on profiles.

**Definition 3** (Adapted from Marley & Pihlens, 2012, Def. 2). *A set of best-worst choice probabilities for a design  $D(P)$ ,  $P \subseteq Q$ ,  $|P| \geq 2$ , satisfies a **preference independent maxdiff model** iff there exists a separate (different) nonnegative scale  $u_i$  on each attribute  $i$ ,  $i = 1, \dots, m$ , such that for every  $\mathbf{x}, \mathbf{y} \in X \in D(P)$ ,  $\mathbf{x} \neq \mathbf{y}$ ,  $|X| \geq 2$ ,*

$$BW_X(\mathbf{x}, \mathbf{y}) = \frac{e^{\sum_{i=1}^m [u_i(x_i) - u_i(y_i)]}}{\sum_{\substack{\mathbf{r}, \mathbf{s} \in X \\ \mathbf{r} \neq \mathbf{s}}} e^{\sum_{i=1}^m [u_i(r_i) - u_i(s_i)]}} \quad (\mathbf{x} \neq \mathbf{y}). \quad (7)$$

Marley and Pihlens (2012, Theorem 6) show that, under reasonable mathematical assumptions, each  $u_i$  is a difference scale, i.e., unique up to an origin, with different origins for each  $i$ .

The more standard notation in the discrete choice literature for (7) assumes that there is a vector

$$\boldsymbol{\beta} = (\beta_1, \dots, \beta_m),$$

with the  $i^{\text{th}}$  component sometimes called the utility coefficient for attribute  $i$ , such that

$$BW_X(\mathbf{x}, \mathbf{y}) = \frac{\exp^{\boldsymbol{\beta}'(\mathbf{x}-\mathbf{y})}}{\sum_{\substack{\mathbf{r}, \mathbf{s} \in X \\ \mathbf{r} \neq \mathbf{s}}} \exp^{-\boldsymbol{\beta}'(\mathbf{r}-\mathbf{s})}} \quad (\mathbf{x} \neq \mathbf{y}).$$

However, that notation requires the attribute-levels to either be numerical or coded in some numerical fashion (e.g., with dummy codes), and is not suitable for proving results of the kind described next.

### 3.1 Properties of scores for the maxdiff model

We now present theoretical results for best minus worst scores (defined below) for the maxdiff model of best-worst choice. Some of these results were proved in Marley and Pihlens (2012), the rest were proved in Marley and Islam (2012). The proofs in Marley and Islam were for (partial or full) ranking probabilities that belong to the class of *weighted utility ranking models*. This class includes the maxdiff model of best-worst choice as a special case, along with the MNL model for best choice and the MNL model for worst choice; however, it also includes many interesting ranking models, such as the *reversible ranking model* (Marley, 1968). For simplicity and relevance, we state the results for the maxdiff model - that is, for Case 1 we have (3); for Case 2, we have (5); and for Case 3 we have (6) - or (7) when we assume additivity. Nonetheless, we know that, empirically, the score measures used in these results are useful for preliminary analyses of the data, independent of the model that is eventually fit to the data - see Section 4.2.

We first state results that hold in common for Case 1, 2, and 3; these are results that do not depend on the actual structure (type) of the “choice options” - where a (choice) option is to be interpreted as an object (Case 1), an attribute-level (Case 2), or a profile (Case 3).

Using notation paralleling that in Marley & Louviere (2005) for Case 1, for each option  $x$  in the design, let  $\hat{b}(x) - \hat{w}(x)$  denote the number of times option  $x$  is chosen as best in the study minus the number of times option  $x$  is chosen as worst in the study. We call this the *score* for  $x$  (in this particular design) and refer to “the scores” for these values across the options in the design.

#### Scores: Property 1 (for Case 1, 2, and 3)

Using general language (with undefined terms in quotation marks), the following states a result due to Huber (1963) in such a way that it applies to the maxdiff model for options; Marley and Islam (2012) state the terms and results exactly. Assume that one is interested in the rank order, only, of the (utility)

scale values in the maxdiff model. An *acceptable loss function* is a “penalty” function with a value that remains constant under a common permutation of the scores and the scale values, and that increases if the ranking is made worse by misordering a pair of scale values. Let  $S$  be a master set with  $n \geq 2$  elements and assume that, for some  $k$  with  $n \geq k \geq 2$ , every subset of  $S$  with exactly  $k$  elements appears in the design<sup>4</sup>  $D(S)$ . Then, given the maxdiff model, ranking the scale values in descending order of the (best minus worst) scores, breaking ties at random, has “minimal average loss” amongst all (“permutation invariant”) ranking procedures that depend on the data only through the set of scores.

*Comment 1* The above result actually holds for the class of *weighted utility ranking models*, a class that includes the MNL for best; MNL for worst; and the maxdiff model for best-worst choice (Marley & Islam, 2012).

*Comment 2* Given the above property of the scores, they are likely useful starting values in estimating the maximum likelihood values of the utilities  $u(x)$ ,  $x \in S$ . In fact, various empirical work on the maxdiff model gives a linear relation between the (best minus worst) scores and the (maximum likelihood) estimates of the utilities<sup>5</sup> (Louviere et al., 2008, 2012). Also, Marley and Islam (2012) show similar results for weighted utility ranking models applied to the ranking data of a case 3 study of attitudes toward the microgeneration of electricity.

**Scores: Property 2 (for Case 1, 2 and 3)**

The set of (best minus worst) scores is a sufficient statistic.

*Comment 1* The above result actually holds for the class of *weighted utility ranking models*, a class that includes the MNL for best; MNL for worst; and the maxdiff model for best-worst choice (Marley & Islam, 2012, Theorem 3).

The following result shows that, in the sense stated, the best minus worst scores reproduce the difference between the best and the worst choice probabilities given by the maximum likelihood estimates of the maxdiff model.

**Scores: Property 3 (for Case 1, 2 and 3)**

For the maxdiff model, the (best minus worst) score for an option  $x$  equals the sum over  $X \in D(P)$  of the weighted difference between the (marginal) best and the (marginal) worst probability of choosing option  $x$  in set  $X$ , with those probabilities evaluated at the values of the maximum likelihood parameter estimates; the weight for a set  $X$  is the number of times  $X$  occurs in the design  $D(P)$  (Marley & Pihlens, 2012, prove this for case 3; the result for case 1 and case 2 is then immediate).

The following result applies to the attribute levels in case 3; there is no equivalent result in case 1 or case 2 as there are no “implied” choices of attribute-levels in those cases.

<sup>4</sup>Further work is needed to extend the theoretical result to, say, balanced incomplete block (BIBD) designs. See Marley & Pihlens (2012) for related discussions of *connected designs*.

<sup>5</sup>Assume that the the maxdiff model holds, and a balanced incomplete design (BIBD) is used for the survey. If the utility (scale) values are in a small range - say, [-1,1] - then a linear relation holds under a first-order Taylor expansion of the maxdiff choice probabilities (Marley & Schlereth, unpublished).

**Scores: Property 4 (for Case 3, with the preference independent maxdiff model, Definition 3)**

Using notation from Marley & Pihlens (2012) for choice among profiles (Case 3),  $\widehat{b}_i(p) - \widehat{w}_i(p)$ ,  $p \in Q(i)$ ,  $i \in M$ , denotes the number of times attribute-level  $p$  is ‘chosen’<sup>6</sup> as best minus the number of times  $p$  is ‘chosen’ as worst. Then Marley and Pihlens show that, for the preference independent maxdiff model (Def. 3),

- i) The set of values  $\widehat{b}_i(p) - \widehat{w}_i(p)$  is a sufficient statistic.
- ii)  $\widehat{b}_i(p) - \widehat{w}_i(p)$  equals the sum over all  $X \in D(P)$  of the weighted difference between the (marginal) best and (marginal) worst probability of ‘choosing’ attribute-level  $p$  in set  $X$ , with those probabilities evaluated at the values of the maximum likelihood parameter estimates; the weight for a set  $X$  is the number of times  $X$  occurs in the design  $D(P)$ .

**3.1.1 Relations between the scales in Case 2 and Case 3**

Note that in Case 2, we have the *same* difference scale  $u$  on each attribute, whereas in Case 3, we have a *different* difference scale  $u_i$  on each attribute  $i$ . Since these scales are derived in different (types of) experiments, there is no necessary empirical relation between them. However, one would hope that such (Case 2 and Case 3) experiments are not measuring totally different properties of the attributes. One theoretical (and, potentially, empirical) property one could hope for is that for each attribute  $i$ , separately, the scale  $u_i$  of Case 3 is strictly monotonic increasing with respect to the scale  $u$  of Case 2. But we are also assuming that the  $u_i$ ,  $i = 1, \dots, m$ , and  $u$ , are separate (different) difference scales. Then, under weak mathematical conditions<sup>7</sup>, the relation between  $u_i$  and  $u$  has to be linear: i.e., there are constants  $\alpha_i > 0$  and  $\beta_i$  such that, for each attribute-level  $z_i$ , we have  $u_i(z_i) = \alpha_i u(z_i) + \beta_i$  (see Aczél, Roberts, & Rosenbaum, 1986; and Marley & Pihlens, 2012, for related results<sup>8</sup>). The interested reader can get a feeling for why this result is true by trying to make a strictly monotonic increasing function, other than a linear one, “work” in preserving the difference scale properties of the  $u_i$  and  $u$ .

Thus, if have data from both a Case 2 and a Case 3 experiment where the above results hold, then substituting the expressions  $u_i(z_i) = \alpha_i u(z_i) + \beta_i$  in the Case 3 representation of the preference independent maxdiff model, (7), gives: for  $\mathbf{x}, \mathbf{y} \in X$ ,  $\mathbf{x} \neq \mathbf{y}$ ,

$$BW_X(\mathbf{x}, \mathbf{y}) = \frac{e^{\sum_{i=1}^m \alpha_i [u(x_i) - u(y_i)]}}{\sum_{\substack{\mathbf{r}, \mathbf{s} \in X \\ \mathbf{r} \neq \mathbf{s}}} e^{\sum_{i=1}^m \alpha_i [u(r_i) - u(s_i)]}} \quad (\mathbf{x} \neq \mathbf{y}).$$

<sup>6</sup>Of course, in this Case 3, attribute-level  $p$  is (only) ‘chosen’ as a consequence of its being a component of the profile chosen on a given choice opportunity.

<sup>7</sup>The function is continuous; alternatively, it is bounded from above on an interval.

<sup>8</sup>In both cases, the results were developed in terms of ratio scales  $b = e^u$ ,  $b_i = e^{u_i}$ , rather than the equivalent difference scales  $u$ ,  $u_i$ .



In this expression, the “importance weights”  $\alpha_i$  are known (from the relation between the data in the two experiments). However, from a Case 3 (or Case 2) experiment, alone, such weights are not identifiable (see Marley et al., 2008).

The above results relating the Case 2 scale to the Case 3 scales were based on the assumptions that the preference independent maxdiff model (Def. 3) holds and for each attribute  $i$ , separately, the scale  $u_i$  of Case 3 is strictly monotonic increasing with respect to the scale  $u$  of Case 2. Although this is an interesting, and possibly desirable result, the data may be otherwise. Weaker relations between the scales in Case 2 and Case 3 are obtained under weaker assumptions (Appendix B).

## 4 Data analysis

Analysis of best-worst data can be conducted in a variety of ways (Louviere et al., 2008; Flynn et al., 2007; Flynn 2010). Two broad methods will be explored here: maximum-likelihood-based methods that are familiar to DCE practitioners, including conditional logistic regression and its extensions, and the best-minus-worst scores introduced in Section 3. It will be demonstrated that although the scores are not part of the typical academic “toolbox” of methods, they have advantages in understanding preference heterogeneity.

### 4.1 Maximum Likelihood (ML) based methods

There is increasing support in statistical packages for “partial ranking” models, for use when the researcher only has data from certain ranking depths, e.g. the top two ranks, or top and bottom ranks. These can be used to analyse best-worst data, but researchers are strongly recommended first to conduct simpler analyses on best and worst data separately to understand whether assumptions made in these models are satisfied. Basic ranking models assume the data at various ranking depths (1) come from a common underlying utility function and (2) have the same variance scale factor. These assumptions can be tested by plotting conditional logit (that is, multinomial logit) estimates of the best data against those for the worst data: a negative linear relationship indicates assumption (1) holds, and a slope with absolute magnitude of 1 indicates (2) holds. Indeed, multiplying all independent variables for the worst data by -1 is a “trick” which allows the researcher to stack the “worst choice sets” below the “best choice sets” and estimate a conditional logistic regression which treats these as “just more best choice sets”. The intuition behind this is as follows: for plausible models (including the conditional logit for best, conditional logit for worst, and maxdiff for best worst) choosing worst from four items with latent utilities of 2, 4, 7, 8 is observationally equivalent to choosing best from four items with latent utilities of  $-2, -4, -7, -8$ . A final adjustment to worst choice data is to delete the item already chosen as best from each choice set. This assumes a sequential model of best, then worst. Knowledge of how the respondent answered each choice set allows the analyst to make the appropriate

deletion (whether best or worst) and web-surveys can force the respondent to provide a best item from  $J$ , followed by a worst from  $J - 1$  (or vice versa).

Increasing evidence suggests the variance scale factor for best and worst data is often not equal to one and future studies should estimate it, using (minus) the estimated value as the factor used to multiply independent variables by for worst choice sets in order that they can be treated as best sets. This reflects practice in data pooling techniques proposed by Swait and Louviere (1993) and formal testing of data pooling should ideally be conducted using methods such as those they propose.

Flynn et al. (2008) compared sequential models of the type just described with the maxdiff model. The maxdiff model requires all possible pairs to be modelled in each choice set, making datasets very large. The dermatology Case 2 study (with four attributes) required 12 possible pairs to be set up per choice set. The maxdiff model's advantages are largely theoretical: the trick used in sequential models to turn worst into best choices introduces a very small error in the likelihood function due to the asymmetry in the EV1 distribution (Marley and Louviere, 2005; Flynn et al., 2008). The maxdiff model has no such error, and the properties of the scores presented in Section 3.1 are based on that model. However, in practice, estimates from maxdiff models are generally indistinguishable from those from sequential models and data parsimony means the latter are generally preferred. Finally, the reader should ask themselves whether, when providing best and worst data, they would really consider all possible pairs before giving their answer.

## 4.2 The best-minus-worst scores.

Marley and Louviere (2005) show that the best minus worst scores are not unbiased estimates of the true utilities when the maxdiff model holds. However, they have been found to be linearly related to the ML estimates of the conditional logit model in virtually every empirical study to date. This is probably a manifestation of the linear portion of the logistic (cumulative distribution) function; thus, a non-linear relationship is likely only when the researcher is plotting the scores for a single highly consistent respondent, or for a sample of respondents each of whom is highly consistent and the choices are highly consistent across the sample. In other words, whilst the analyst should be wary of inferring cardinality in the scores for a given respondent, (s)he does not have to aggregate many respondents to obtain scores that are highly linearly related to the conditional logit estimates. Thus, researchers who are not confident of implementing limited dependent variable models such as logit and probit regression can obtain good estimates using a spreadsheet.

The scores also enable considerable insights to be drawn at the level of the individual respondent. For example taxonomic (clustering) methods of analysis have been applied to the scores (Auger et al., 2007). Since the scores are a function of choice frequencies there is no need for any prior rescaling of the data in attempts to eliminate or standardise any respondent-level response styles: two people who agree on the rank order of a list of items but who use different parts of

a rating scale will provide identical best-worst data. Flynn and colleagues have also used the scores to help evaluate solutions from latent class analyses, which can give spurious solutions (Flynn et al., 2010). This use of the scores to judge and guide analyses of the choice (0,1) data that choice modellers traditionally use represents an important aid in decomposing mean and variance heterogeneity. It is well-known that the perfect confound between means and variances on the latent scale holds for all limited dependent variable (including logit and probit) models (Yatchew, & Griliches, 1985), which means that technically there are an infinite number of solutions to any DCE. Judicious use of the scores can help rule out many outcomes.

## 5 Summary and Future Research

Best-Worst Scaling offers a cost-efficient way of obtaining more information from a respondent, and/or a way of evaluating models of choice processes. It is important to note that it is a theory of how an individual makes choices; aggregating across individuals requires assumptions to be made (and tested, where possible). It is for this reason that the BIBDs proposed here are attractive: a single design means there is no potential for any confounding of properties of the design with any given individual’s preferences.

There are a number of fruitful areas for future research. Researchers interested in keeping within a traditional choice modelling paradigm would welcome work to further understand how and to what extent best-worst data can be pooled. Issues such as the size of the variance scale factor at ranking depths and the conditions under which worst models have a different functional form to best are important. Indeed work to further understand process issues generally is welcome: early work suggests that the class of models with natural process interpretations is different from the class of models with useful score properties. Yet use of the scores may give the average applied researchers more confidence, not least in terms of better understanding heterogeneity in preferences and/or scale in their data. Data pooling has normative issues that are pertinent to health economists in particular: for best-worst choice, it is only if all individuals satisfy the maxdiff model that the average of their utility estimates represents the preferences of the “representative individual” used in economic evaluation.

### Appendix A

#### A maximum random utility model for best, worst, and best-worst choices satisfying a “common” MNL model

When treated as a single model, the three models (1), (2), and (3), satisfy an *inverse extreme value maximum<sup>2</sup> random utility model* (Marley & Louviere, 2005, Def. 11). That is, for  $z \in S$  and  $p, q \in S$ ,  $p \neq q$ , there are independent random variables  $\epsilon_z, \epsilon_{p,q}$  with the extreme value distribution<sup>1</sup> such that for all  $y \in Y \in D(S)$ ,

$$B_Y(y) = \Pr \left( u(y) + \epsilon_y = \max_{z \in Y} [u(z) + \epsilon_z] \right), \quad (8)$$

$$W_Y(y) = \Pr \left( -u(y) + \epsilon_y = \max_{z \in Y} [-u(z) + \epsilon_z] \right), \quad (9)$$

and for all  $x, y \in Y \in D(S)$ ,  $x \neq y$ ,

$$BW_Y(x, y) = \Pr \left( u(x) - u(y) + \epsilon_{x,y} = \max_{\substack{p, q \in Y \\ p \neq q}} [u(p) - u(q) + \epsilon_{p,q}] \right). \quad (10)$$

Standard results (summarized by Marley & Louviere, 2005) show that the expression for the choice probabilities given by (8) (respectively, (9), (10)) agrees with that given by (1) (respectively, (2), (3)).

These maximum random utility models are particularly interesting because they can be rewritten in such a way as to also predict response time. The key step is to convert the above maximum random utility model of best and/or worst choice into an equivalent “horse race” *minimum* random utility of best and/or worst choice *and* response time (Marley & Colonius, 1992). See Hawkins et al. (2012) for such extension

## Appendix B

### General relations between the scale values in Case 2 and Case 3

The result in the text relating the Case 2 scale to the Case 3 scales was based on the assumption that for each attribute  $i$ , the scale  $u_i$  of Case 3 is, separately, strictly monotonic increasing with respect to the scale  $u$  of Case 2. Although this is an interesting, and possibly desirable result, the data may be otherwise. A weaker relation is obtained with the following weaker assumption.

Assume that there is a function  $F$  that maps a typical vector of scale values  $(u(r_1), \dots, u(r_m))$  of the attribute-level maxdiff model on single profiles to the overall scale value  $u^{(m)}(r_1, \dots, r_m)$  of the maxdiff model on profiles (Def. 2), i.e.,

$$F(u(r_1), \dots, u(r_m)) = u^{(m)}(r_1, \dots, r_m).$$

Similar to the previous case, we assume that  $u$  and  $u^{(m)}$  are separate (different) difference scales, and we also assume that the mapping  $F$  is *invariant under admissible transformations* (here, changes of origin) in the following sense: There is a function  $G$  on the non-negative real numbers such that for each  $\alpha > 0$ ,

$$F(u(r_1) + \alpha, \dots, u(r_m) + \alpha) = F(u(r_1), \dots, u(r_m)) + G(\alpha),$$

The mathematical question then becomes: what are the possible solutions of the above equation for  $F$  (and  $G$ )? Under quite weak regularity conditions<sup>9</sup>, the general solution has the form (Aczél et al., 1986): for some  $i$ ,  $i \in \{1, \dots, m\}$ ,

---

<sup>9</sup>That certain functions are bounded on an arbitrarily small open  $m$ -dimensional interval.

and a constant  $c_i$ ,

$$F(u(r_1), \dots, u(r_m)) = \left\{ \begin{array}{l} c_1 u(r_m) + f(u(r_1) - u(r_m), \dots, u(r_{m-1}) - u(r_m)) \\ c_i u(r_i) + f(u(r_1) - u(r_i), \dots, u(r_{i-1}) - u(r_i), u(r_{i+1}) - u(r_i), \dots, u(r_{m-1}) - u(r_m)) \\ c_m u(r_m) + f(u(r_1) - u(r_m), \dots, u(r_{m-1}) - u(r_m)) \end{array} \right\}$$

$$if \left\{ \begin{array}{l} i = 1 \\ 1 < i < m, . \\ i = m \end{array} \right.$$

where  $f$  is an arbitrary non-negative function with  $f = const$  if  $m = 1$ . It is important to understand that a specific solution is *one* such function  $f$ , for a particular  $i$ . Also, for the present application, we would expect  $F$  to be strictly increasing in each variable; a sufficient condition for this is that  $c_i > 0$  for each  $i$  and  $f$  is strictly increasing in each variable.

Clearly, the above functional relation between the maxdiff scale  $u^{(m)}$  and the attribute-level scale  $u$  is quite general - for instance, the following is a possible solution, where, for simplicity, we set  $m = 3$ : there are constants  $A, B, C > 0$  such that

$$u^{(m)}(r_1, \dots, r_3) = Cu(r_1) + A(u(r_2) - u(r_1)) + B \log(u(r_3) - u(r_1)).$$

Thus, if the above assumptions hold (and the previous linear ones do not), then it is a challenging empirical task to explore possible relations between the scale  $u$  of Case 2 and the scales  $u_i$ ,  $i = 1, \dots, m$ , of Case 3.

### Acknowledgements

This research has been supported by Natural Science and Engineering Research Council Discovery Grant 8124-98 to the University of Victoria for Marley. The work was carried out, in part, whilst Marley was a Distinguished Research Professor (part-time) at the Centre for the Study of Choice, University of Technology Sydney.

### References

- Aczél, J., Roberts, F. S., & Rosenbaum, Z. (1986). On scientific laws without dimensional constants. *Journal of Mathematical Analysis and Applications*, 9, 389-416.
- Anderson, N. H. (1970). Functional measurement and psychophysical judgement. *Psychological Review*, 77(3), 153-170.
- Auger, P., Devinney, T. M., & Louviere, J. J. (2007). Using best-worst scaling methodology to investigate consumer ethical beliefs across countries. *Journal of Business Ethics*, 70, 299-326.
- Beggs, S., Cardell, S., & Hausman, J. (1981). Assessing the potential demand for electric cars. *Journal of Econometrics*, 16, 1-19.

- Coast, J., Salisbury, C., de Berker, D., Noble, A., Horrocks, S., Peters, T. J., & Flynn, T. N. (2006). Preferences for aspects of a dermatology consultation. *British Journal of Dermatology*, 155, 387-392.
- Cohen, S., & Neira, L. (2004). *Measuring preference for product benefits across countries: Overcoming scale usage bias with maximum difference scaling*. Paper presented at the American Conference of the European Society for Opinion and Marketing Research, Punta del Este, Uruguay.
- Cohen, S. H., & Neira, L. (2003). *Overcoming scale usage bias with maximum difference scaling*. Paper presented at the ESOMAR 2003 Latin America Conference, Punta del Este, Uruguay.
- Collins, A. T., & Rose, J. M. (2011). *Estimation of a stochastic scale with best-worst data*. Paper presented at the Second International Choice Modelling Conference, Leeds, UK.
- 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. (2008). Estimating preferences for a dermatology consultation using Best-Worst Scaling: Comparison of various methods of analysis. *BMC Medical Research Methodology*, 8(76).
- Flynn, T. N., Louviere, J. J., Peters, T. J., & Coast, J. (2010). Using discrete choice experiments to understand preferences for quality of life. Variance scale heterogeneity matters. *Social Science & Medicine*, 70, 1957-1965. doi:10.1016/j.socscimed.2010.03.008
- Hausman, J. A., & Ruud, P. A. (1987). Specifying and testing economic models for rank-ordered data. *Journal of Econometrics*, 34, 83-104.
- Hawkins, G. E., Brown, S. D., Marley, A. A. J., Heathcote, A., Flynn, T. N., & Louviere, J. J. (2012). *Accumulator models for best-worst choices*. Department of Psychology. University of Newcastle, Australia. Newcastle, NSW, Australia.
- Helson, H. (1964). *Adaptation-Level Theory*. New York: Harper & Row.
- Hensher, D. A., Rose, J. M., & Greene, W. H. (2005). *Applied choice analysis: a primer*. Cambridge: Cambridge University Press.
- Huber, P. J. (1963). Pairwise comparison and ranking: optimum properties of the row sum procedure. *Annals of Mathematical Statistics*, 34, 511-520.
- Louviere, J. J., Flynn, T. N., & Carson, R. T. (2010). Discrete choice experiments are not conjoint analysis. *Journal of Choice Modelling*, 3(3), 57-72.
- Louviere, J. J., Flynn, T. N., & Marley, A. A. J. (2012). *Best-Worst Scaling: Theory, methods and Applications*. Manuscript, University of Technology Sydney.

- Louviere, J. J., & Hensher, D. A. (1982). On the design and analysis of simulated choice or allocation experiments in travel choice modelling. *Transportation Research Record*, 890, 11-17.
- Louviere, J. J., & Street, D. (2000). Stated preference methods. In D. A. Hensher & K. Button (Eds.), *Handbook in Transport I: Transport Modelling*. (pp. 131-144). Amsterdam: Pergamon (Elsevier Science).
- 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. D. (1959). Individual choice behavior. New York: John Wiley & Sons.
- Luce, R. D. & Suppes, P. (1965). Preference, utility, and subjective probability. In R. D. Luce, R. R. Bush, and E. Galanter. (Eds.). *Handbook of Mathematical Psychology, Vol. III*. New York, NY: John Wiley and Sons, pp. 235-406.
- Marley, A. A. J. (1968). Some probabilistic models of simple choice and ranking. *Journal of Mathematical Psychology*, 5, 333-355.
- Marley, A. A. J., Flynn, T. N., & Louviere, J. J. (2008). Probabilistic models of set-dependent and attribute-level best-worst choice. *Journal of Mathematical Psychology*, 52, 281-296.
- Marley, A. A. J., & Islam, T. (submitted). Conceptual relations between expanded rank data and models of the unexpanded rank data. *Journal of Choice Modelling*.
- 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* (pp. 105-142). New York: Academic Press.
- McIntosh, E., & Louviere, J. J. (2002). *Separating weight and scale value: an exploration of best-attribute scaling in health economics*. Health Economists' Study Group, Brunel University.
- Scarpa, R., & Marley, A. A. J. (2011). *Exploring the consistency of alternative best and/or worst ranking procedures*. Paper presented at the Second International Choice Modelling Conference, Leeds, UK.
- Scarpa, R., Notaro, S., Raffelli, R., Pihlens, D., & Louviere, J. J. (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, 813-828.

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

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

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

Yatchew, A., & Griliches, Z. (1985). Specification error in probit models. *Review of Economics and Statistics*, 67(1), 134-139.



**Figure 1. A completed example BWS 'object case' question**

<b>Most</b>	<b>Issue</b>	<b>Least</b>
	Pesticides used on crops	
	Hormones given to livestock	
	Irradiation of foods	✓
	Excess salt, fat cholesterol	
✓	Antibiotics given to livestock	

Please consider the food safety issues in the table above and tick which concerns you most and which concerns you least.

**Figure 2. A completed example BWS 'profile case' question**

<b>Most</b>	<b>Appointment #1</b>	<b>Least</b>
	You will have to wait two months for your appointment	
	The specialist has been treating skin complaints part-time for 1-2 years	✓
	Getting to your appointment will be quick and easy	
✓	The consultation will be as thorough as you would like	





Please imagine being offered the appointment described above and tick which feature would be best and which would be worst.

**Figure 3. A completed example BWS 'profile case' question based on EQ-5D instrument**

Best		Worst
	Some problems walking about	
✓	No problems with self-care	
	Some problems with performing usual activities	
	Extreme pain or discomfort	✓
	Moderately anxious or depressed	

Imagine you were living in the health state described above. Tick which aspect of this would be best to live with and which would be worst to live with.

Figure 4. An example BWS 'multi-profile case' question

	Phone 1	Phone 2	Phone 3	Phone 4
<b>Phone Style</b>	 Clam or flip phone	 Candy Bar or straight phone	 Swivel flip	 PDA phone with touch screen input
<b>Handset Brand</b>	A	B	C	D
<b>Price</b>	\$49.00	\$199.00	\$249.00	\$129.00
<b>Built-in Camera</b>	No camera	5 megapixel camera	2 megapixel camera	3 megapixel camera
<b>Wireless Connectivity</b>	No Bluetooth or WiFi connectivity	Bluetooth and WiFi connectivity	WiFi connectivity	Bluetooth connectivity
<b>Video Capability</b>	No video recording	Video recording (up to 1 hour)	Video recording (more than 1 hour)	Video recording (up to 15 minutes)
<b>Internet Capability</b>	Internet Access	Internet Access	No Internet access	No Internet access
<b>Music Capability</b>	No music capability	MP3 Music Player only	FM Radio only	MP3 Music Player and FM Radio
<b>Handset Memory</b>	64 MB built-in memory	2 GB built-in memory	512 MB built-in memory	4 GB built-in memory

**Table 1. Choice sets from a  $2^4$  expansion for four objects. A ✓ means the object is in the choice set, a ✗ means the object is not in the choice set.**

	Object W	Object X	Object Y	Object Z
Set 1	✓	✓	✓	✓
Set 2	✓	✓	✓	✗
Set 3	✓	✓	✗	✓
Set 4	✓	✗	✓	✓
Set 5	✗	✓	✓	✓
Set 6	✓	✓	✗	✗
Set 7	✓	✗	✓	✗
Set 8	✓	✗	✗	✓
Set 9	✗	✓	✓	✗
Set 10	✗	✓	✗	✓
Set 11	✗	✗	✓	✓
Set 12	✓	✗	✗	✗
Set 13	✗	✓	✗	✗
Set 14	✗	✗	✓	✗
Set 15	✗	✗	✗	✓
Set 16	✗	✗	✗	✗