Best Worst Scaling: Theory and Practice

A. A. J. Marley*1,2 and T. N. Flynn2

* (corresponding author)

Department of Psychology
University of Victoria
PO Box 3050 STN CSC
Victoria
BC V8W 3P5
Canada
email: ajmarley@uvic.ca

2T. N. Flynn
Centre for the Study of Choice
University of Technology Sydney
PO Box 123 Broadway
Sydney, NSW 2007
Australia
email: Terry.Flynn@uts.edu.au

International Encyclopedia of the Social and Behavioral Sciences
2nd Edition
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. The three “cases” of BWS are described, followed by a summary of the main models and related theoretical results, including an exposition of possible theoretical relationships between estimates from two of the cases. This is followed by the theoretical and empirical properties of “best minus worst scores.” The entry ends with some directions for future research.

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

Louviere and Woodworth (1990) and Finn and Louviere (1992) developed a discrete choice task in which a person is asked to indicate the “least preferred” item in a choice set, in addition to indicating the (traditional) “most preferred” item; the approach that Louviere pioneered is now called best worst scaling (BWS). Louviere’s initial focus was on “objects”, such as attitudes, general public policy goals, brands, or anything that did not require a detailed description, such as would be required for a consumer product such as a soft drink or a car. Thus, Finn & Louviere (1992) used BWS to examine 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. Figure 1 contains a BWS question similar to ones used in that study.

Almost immediately, Louviere began applying BWS to more complex items such as attribute-levels describing a single alternative (profile), or complete alternatives (profiles) of the type familiar to choice modelers. The former case, requiring respondents to identify the best attribute-level and worst attribute-level within an alternative, was relatively unfamiliar to choice modelers, whereas the latter case is a “natural” extension of a standard discrete choice experiment (DCE) in which the decision maker selects the most preferred item (Louviere, Hensher & Swait, 2000; Hensher, Rose & Greene, 2005).

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). After a 10-year delay, Marley & Louviere (2005) completed an extensive analysis of BWS as a theory, explaining the processes that individuals might follow in providing best and worst data, and presenting plausible mathematical forms for such processes; summaries of those, and more recent, results are presented in Section 3. Louviere has recently returned to the use of BWS as a method of data collection (Louviere et al., 2008), in which case its main purpose, via repeated rounds of best-worst choices, is to obtain a full ranking of items in a manner that is “easy” for respondents and can then be analyzed in various ways.

The remainder of the entry is as follows. Section 2 illustrates the three types (“cases”) of BWS that Louviere developed. Section 3 summarizes the basic models, with Section 3.3 summarizing models of ranking by repeated best and/or worst choices; Section 4 presents properties of simple “scores”; and Section 5 summarizes and discusses future research.
2 BWS: The Three Cases

2.1 Case 1 (Object case)
Case 1 BWS (as already illustrated in Fig. 1) 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. 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 discrete choice experiments (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.

2.2 Case 2 (Profile case)
Case 2 BWS, which was introduced by McIntosh and Louviere (2002), is illustrated in Figure 2 using a question from a dermatology study (Coast et al., 2006; Flynn et al., 2008).

![Insert Figure 2 about here]

The choice set has the structure of a single profile (multiattribute alternative, see Case 3). However, the respondent is not required to consider the value of the profile as a whole, but has to consider the attribute-levels that describe it, and choose the one that is best (most attractive) and the one that is worst (least attractive). Case 2 BWS is popular in health because 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. Also, recent theoretical results show that a Case 2 study, in combination with a Case 3 study of choices between the same (or suitably related) profiles, may allow the separate measurement of attribute weight and level scale value (Flynn & Marley, 2012); should such a study be successful, it would solve this classic measurement problem (Anderson, 1970; Marley, Flynn, & Louviere, 2008).

2.3 Case 3 (Multi-profile case)
Case 3 BWS requires a respondent to choose the worst (least attractive) profile as well as the best (most attractive) one in the current choice set. Figure 3 shows an example question from a mobile phone study (Marley & Pihlens, 2012). The additional data provided by worst choices are valuable in many marketing applications, both in terms of better estimation of parameters and in identifying attribute levels that make a product “unacceptable”.

4
3 Models of Best and/or Worst Choice

We use 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). 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 (Louviere, Hensher, & Swait, 2000); 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).

3.1 Notation

Let \( P \) with \(|P| \geq 2\) denote the finite set of potentially available choice options, and let \( D(P) \) 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 cars by repeatedly asking them for choices amongst sets of four different options: \( P \) represents the collection of car types in the study, and each element of the set \( D(P) \) represents the set of options provided on one particular choice occasion. For any \( Y \in D(P) \), with \(|Y| \geq 2\), \( B_Y(y) \) denotes the probability that alternative \( y \) 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 \( Y \).

3.2 Multinomial logit models of best and/or worst choice

Many models of choice are based on extensions of the multinomial logit (MNL) model. The best choice MNL model assumes there is a difference scale \( u \) such that for all \( y \in Y \in D(P) \),

\[
B_Y(y) = \frac{e^{u(y)}}{\sum_{z \in Y} e^{u(z)}}.
\]

(1)

The value \( u(y) \) for an option \( y \) is interpreted as the utility for that option.

The parallel worst choice MNL model assumes there is a difference scale \( v \) such that for all \( y \in Y \in D(P) \),

\[
W_Y(y) = \frac{e^{v(y)}}{\sum_{z \in Y} e^{v(z)}}.
\]
Marley and Louviere (2005) present a theoretical argument for the case where \( v = -u \), i.e., we have
\[
W_Y(y) = \frac{e^{-u(y)}}{\sum_{z \in Y} e^{-u(z)}}. \tag{2}
\]

Note that (1) with (2) implies that the probability that \( y \in Y \) is selected as \textit{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 \textit{worst} in a set \( Y \) with scale values \( -u(z) \), \( z \in Y \).

The MNL models for best choices and for worst choices can be combined in three simple ways to model best-worst choice:

First, assuming the best choices satisfy (1) and the worst choices satisfy (2), the \textit{best, then worst, MNL model} assumes that for all \( x, y \in Y \), \( x \neq y \),
\[
BW_Y(x, y) = B_Y(x)W_{Y_{\{-y\}}}(y),
\]

Second, the \textit{worst, then best, MNL model} assumes that for all \( x, y \in Y \), \( x \neq y \),
\[
BW_Y(x, y) = W_Y(y)B_{Y_{\{-y\}}}(x).
\]

Repeated choices satisfying a common one of the above models lead naturally to models of rank order data - see Section 3.3.

Third, the \textit{maxdiff model for best-worst choice} assumes that the value of a choice alternative in the selection of a best option \( u \) is the negative of the valence of that option in the selection of a worst option \( -u \), and that for all \( x, y \in Y \), \( x \neq y \),
\[
BW_Y(x, y) = \frac{e^{[u(x)-u(y)]}}{\sum_{(p,q) \in Y} e^{[u(p)-u(q)]}}. \tag{3}
\]

### 3.3 Models of ranking by repeated best and/or worst choice

Thus far, we have discussed models for the choice of the best and the worst option in a set of available options. To the extent that those choices are reliable, it is of interest to consider the use of repeated choices of the best and/or worst option to generate a rank order of the available options, and to develop models for such rankings. For 4 options, there are 8 distinct patterns of best and/or worst questions, any one of which an experimenter can ask a participant to use for the generation of a rank order on those options; and, also, at least 8 possible (MNL-based) models of best and/or worst choice for the final rank order. Here, we introduce notation for 2 such models, illustrated with 4-element sets.

Let \( Y \) denote a typical choice set; \( R(Y) \) the set of rank orders of \( Y \); \( \rho \in R(Y) \) a typical rank order (from best to worst) of \( Y \); and \( p_{R(Y)}(\rho) \) the probability of that rank order occurring. Let \( r, s \) be two indices, each of which can be fixed (throughout a rank order) at the value \( b \) (for \textit{best}) or \( w \) (for \textit{worst}). Let \( P_{R(Y)}^r(\rho_1\rho_2\rho_3\rho_4) \) denote the probability of the rank order \( \rho = \rho_1\rho_2\rho_3\rho_4 \) (from best to worst) when the rank order is obtained by: the first choice is of type \( r \); the second of type \( s \); the third, again, of type \( r \). Then we can have:
i. repeated best:

\[ P_{R(Y)}^{b,b}(\rho) = B_Y(\rho_1)B_{Y-\{\rho_1\}}(\rho_2)B_{\{\rho_3,\rho_4\}}(\rho_3). \] (4)

ii. repeated best, then worst:

\[ P_{R(Y)}^{b,w}(\rho) = B_Y(\rho_1)W_{Y-\{\rho_1\}}(\rho_4)B_{\{\rho_2,\rho_3\}}(\rho_2). \] (5)

The natural first assumption in testing these models is to assume that the best (respectively, worst) choice probabilities satisfy the MNL model (1) (respectively, (2)), and, as needed by data, generalizations of those models that include a scale factor that depends on the current choice set in some identifiable manner; this scale factor relates to the possibly changing variability (“consistency”) of the choices across sets. An example of such a generalization of the MNL model for best choices, (1) is: there is a nonnegative scale factor \( s \) defined for each integer \( 2, 3, \ldots \), and a difference scale \( u \) such that for all \( y \in Y \),

\[ B_Y(y) = \frac{e^{s(|Y|)u(y)}}{\sum_{z \in Y} e^{s(|Y|)u(z)}}. \] (6)

We then define a generalized rank ordered logit model (GROL) as a set of rank orders that satisfy (4) with the best choice probabilities satisfying (6). The form (6) is a special case of what Vermunt and Magidson (2005, Section 2.4) call an MNL with replication-specific scale factor and of Fiebig et al.’s (2010) generalized multinomial logit model (GMNL); the latter model also includes scale (variance) heterogeneity across individuals.

Scarpa, Notaro, Raffelli, Pihlens, and Louviere (2011) collected ranking data by repeated best, then worst, choices and fit that data quite successfully with a model based on repeated best choice, i.e. (4), with those choices satisfying a generalization of the model in (6) that included properties of the scale factor \( s \) on aspects of the design that took account of the difference between the data collection method (repeated best, then worst) and the model (repeated best); it would be interesting to see if their data could be better fit by a model that matched their data collection method, i.e. (5). Collins and Rose (2011) fit related models to stated preference data on dating choices.

3.4 Maximum random utility models of choice and response time

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

\[ B_Y(y) = \Pr\left(u(y) + \epsilon_y = \max_{z \in Y} \{u(z) + \epsilon_z\}\right), \] (7)

\(^1\)This means that: for \(-\infty < t < \infty \) \( \Pr(\epsilon_z \leq t) = \exp -e^{-t} \) and \( \Pr(\epsilon_{p,q} \leq t) = \exp -e^{-t} \).
\[ W_Y(y) = \Pr \left( -u(y) + \epsilon_y = \max_{z \in Y} [-u(z) + \epsilon_z] \right), \] (8)

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

\[ BW_Y(x, y) = \Pr \left( u(x) - u(y) + \epsilon_{x,y} = \max_{p,q \in Y} [u(p) - u(q) + \epsilon_{p,q}] \right). \] (9)

Standard results (summarized by Marley & Louviere, 2005) show that the expression for the choice probabilities given by (7) (respectively, (8), (9)) agrees with that given by (1) (respectively, (2), (3)).

These random utility models are particularly interesting in the present context because they can be rewritten, and extended, in such a way as to also predict response time (Hawkins et al., 2012).

3.5 Extensions of the models to Case 2 and Case 3

The notation, models, and results (such as for scores, Section 4) are easily extended to Case 2 and Case 3 - see Marley, et al. (2008) for Case 2, Marley and Pihlens (2012) for Case 3, and Flynn and Marley (2012) for a summary of all cases.

4 Properties and Uses of Scores for the Maxdiff Model

We now summarize theoretical and empirical results for best minus worst scores (defined below) for the maxdiff model of best-worst choice. Although the theoretical results are not exact for, say, the best, then worst, MNL model, to date no identifiable differences have been found between fits of that model and the maxdiff model (see Flynn et al., 2008, for an example of such fits); hence the score measures are useful for preliminary analyses of the data independent of the model that is eventually fit to the data.

For each option \( x \) in the design, the score for \( x \) (in this particular design) is 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 refer to “the scores” for these values across the options in the design.

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

Marley and Islam (2012) state the following 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 \( P \) be a master set with \( n \geq 2 \) elements and assume that, for some \( k \) with \( n \geq k \geq 2 \), every subset of \( P \) with exactly \( k \) elements
appears in the design\(^2\) \(D(P)\). 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.

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

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

The set of (best minus worst) scores is a sufficient statistic. As with the first theoretical property, this one holds for the class of weighted utility ranking models (Marley & Islam, 2012, Theorem 3).

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

The best minus worst scores have been found to be linearly related to the maximum likelihood estimates of the best, then worst, MNL model, and of the maxdiff model, in virtually every empirical study to date. For example, Flynn (2010) obtained this result in a Case 2 quality of life study, and Marley and Islam (2012) present such linear relations for both profiles and attribute-levels in a Case 3 study of consumer attitudes to the installation of solar technology for household electricity production. Such linearity 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.

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) and Flynn and colleagues have used the scores to decide which solutions from latent class analyses "make sense" (Flynn et al., 2010).

## 5 Summary and Future Research

Further research is needed to understand the factors affecting the scale factor at different choice depths and the conditions under which the data require different functional forms (models) for best and worst choices. Also, research to date suggests that the class of models with natural process interpretations is different from the class of models with useful score properties. Nonetheless, the scores may give the average applied researcher more confidence, not least in terms of better understanding heterogeneity in preferences and/or scale factors in their data.

Data pooling has normative issues that are particularly pertinent to health economists: 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

\(^2\)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.
the “representative individual” used in economic evaluation. Also, appropriate data pooling exercises (likely involving Case 2 and Case 3 data) may allow the separation of attribute weight and level scale value (Marley, Flynn, & Louviere 2008; Flynn & Marley, 2012). While the distinction between these two measures is not recognized by traditional economic welfare theory, it offers benefits to health researchers constructing multi-dimensional health outcome instruments from individual symptom scales: for example, rejection of a simple “sum-score” aggregation rule necessitates decisions as to what importance weights are to be applied to the individual symptom scales (Fayers & Machin, 2007, p.218).

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


**Figure 1. A completed example BWS ‘object case’ question**

<table>
<thead>
<tr>
<th>Most</th>
<th>Issue</th>
<th>Least</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pesticides used on crops</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hormones given to livestock</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Irradiation of foods</td>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
<td>Excess salt, fat cholesterol</td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>Antibiotics given to livestock</td>
<td></td>
</tr>
</tbody>
</table>

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

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

Please imagine being offered the appointment described above and tick which feature would be best and which would be worst.
Figure 3. An example BWS ‘multi-profile case’ question

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

Centre for the Study of Choice  http://www.censoc.uts.edu.au/
Society for Judgment and Decision Making http://www.sjdm.org/
Daniel L. McFadden http://emlab.berkeley.edu/~mcfadden/
A. A. J. Marley http://www.uvic.ca/psyc/marley/

Cross References to Related Entries in the Encyclopedia

Note to the Editors: These are only for Section 4. We do not know how to access other sections (such as those on Economics and Medicine)

41028. Experimental Laboratories: Biobehavioral, Neuroscientific, Psychological
41029  Experimental Laboratories: Social and Economic
43031. Decision and Choice: Luce's Choice Axiom
43033. Decision and choice: random utility models of choice and response time
43035. Decision and choice: utility and subjective probability, empirical studies
43060. Measurement Theory: Conjoint Analysis Applications
43063. Measurement theory: probabilistic
43082. Psychometrics: Preference Models with Latent Variables
43094. Stochastic dynamic models (choice, response, time)
43108. Decision and Choice: Risk, Empirical Studies
43109. Decision and Choice: Risk, Theories
44040. Panel Surveys, Uses and Applications