Best-Worst Scaling: What it can do for health care and how to do it

Terry N Flynn[1*], Jordan J Louviere[2], Tim J Peters[3], Joanna Coast[4]

[1]MRC Health Services Research Collaboration. Department of Social Medicine, University of Bristol, Canynge Hall, Whiteladies Road, Bristol BS8 2PR, UK

[2]School of Marketing, University of Technology, Sydney, PO Box 123, Broadway NSW 2007, Australia

[3]Department of Community Based Medicine, University of Bristol, The Grange, 1 Woodland Road, Bristol BS8 1AU, UK

[4]Department of Social Medicine, University of Bristol, Canynge Hall, Whiteladies Road, Bristol BS8 2PR, UK

Abstract

Statements like "quality of care is more highly valued than waiting time" generally can neither be supported nor refuted in traditional Stated Preference Discrete Choice Modelling (SPDCM). This is because in traditional SPDCM the regression constant term represents the total utility of some baseline specification of the good or service under evaluation. The constant cannot be decomposed to elicit the relative contributions of individual attributes to the total utility of this baseline specification. There are a number of published papers that fail to recognise this limitation of traditional SPDCM. This has led authors to draw unwarranted conclusions that may unfairly discredit discrete choice experiments in health care.

Best-worst scaling can overcome the problem of estimating attribute importance because it asks respondents to perform a different choice task to that in traditional SPDCM. However, whilst the nature of the best-worst task is generally understood, there are a number of issues relating to the design and analysis of a best-worst choice experiment that require further exposition. This paper will illustrate how to aggregate and analyse choice data of the type utilised by McIntosh and Louviere (2002) and will demonstrate how richer insights can be drawn using a less aggregated method.

Best-Worst Scaling: What it can do for health care research and how to do it

1. Introduction

Stated Preference Discrete Choice Modelling (SPDCM) involves eliciting people's preferences for goods or services based on their intentions expressed in hypothetical situations (Louviere Hensher and Swait, 2000). This distinguishes it from revealed preference analysis, which utilises people's observed behaviour in real markets. It is increasingly used in health care research and other areas of applied economics, where the production and distribution of goods or services by non-market methods means that revealed preference data are unavailable. Even when a particular good or service is traded in markets, it may be difficult or impossible to estimate the effect that particular attributes have on total utility due to a lack of variability in these attributes across the market. A related problem occurs when two or more attributes tend not to vary independently of each other in market-traded goods, leading to multi-collinearity in estimates (Louviere Hensher and Swait, 2000). SPDCM can vary attributes systematically across hypothetical specifications of the good or service and observe the choices people make in order to estimate the utilities of the attributes.

In some applications, most notably health care, policymakers are interested in comparing the absolute importance (utilities) of attributes. Unwarranted conclusions over attribute importance have been made in traditional SPDCM studies before. For example, in the study by Vick and Scott on preferences for GP consultations, the statement made that "being able to talk to the doctor" was more highly valued than "who chooses your treatment" was not warranted from the study conducted (Vick and Scott, 1998) and similar claims generally can neither be supported nor refuted in traditional SPDCM. The use of mostly two-level attributes in the study, with most of these attempting to capture good/bad extremes, did not enable a meaningful comparison of attribute importance, since *only one* attribute in this study had as its lower level some meaningful measure of 'zero' – the 'being able to talk to the doctor' attribute with its 'the doctor does not listen to you' level. Thus, any attempt to

compare the utility of moving from the lower level to the higher level across attributes is akin to choosing the tallest person from a group where only one is standing up. It is unsurprising, therefore, that this attribute was found to be most important to patients.

However, best-worst scaling (Marley and Louviere, 2004), devised by Finn and Louviere (Finn and Louviere, 1992) and introduced to health care research by McIntosh and Louviere (McIntosh and Louviere, 2002) is a novel method that is capable of addressing such issues. The reason why the best-worst approach can address these issues is that by asking respondents to perform a different choice task to that in traditional SPDCM it elicits additional information. The full nature of the task and the analytical framework are described later, but the next section will explain the limitations of traditional SPDCM in making inferences about attribute importance.

2. The constant term in traditional discrete choice modelling

A traditional SPDCM exercise involves choosing the most preferred specification of a good ('alternative' or 'scenario') from a choice set of competing scenarios (Louviere and Timmermans, 1990). The size of the choice set often depends on the nature of the problem and/or the discipline in which the study is being conducted. In many health care studies it has mostly been of size two (a pairwise comparison) (Farrar et al., 2000) whereas in marketing studies it has more often been of varied size (Louviere and Woodworth, 1983). When respondents choose their preferred scenario, they are effectively providing information about their preferences compared with a particular scenario for a set of attribute differences – for instance in the case of a car this might be the set comprising the additional utility of blue over red, the additional disutility of an extra cost of £100 and the additional utility of four doors compared with two. The constituent utility part-worths (utilities of individual attribute levels, as opposed to total utilities of entire scenarios) of that particular scenario are not estimable. In other words the constant term in the estimated regression represents some 'bundle' of attribute levels that cannot be decomposed into its constituent parts. The exact representation of the constant term depends upon how the attribute levels were coded in the main regression. If they were coded as dummy variables, then an attribute with n levels associated with it will have (n-1) degrees of freedom; hence there are (n-1)

dummy variables, each one representing the additional utility of that level from the reference/benchmark level. Since one dummy variable is omitted from every attribute the regression constant term represents the total utility of that scenario defined by the omitted (benchmark) level on each attribute.

If the attribute levels are coded using effect coding, again one level per attribute is omitted but the constant now represents the grand mean over all observations. The (n-1) estimated independent effect codes represent the additional utility of that attribute level from the mean utility (over all levels), permitting estimation of respondents' average propensity to choose particular attribute levels. The additional utility of the omitted level is equal to minus the sum of the other estimated level utilities. Effect coded variables are correlated within attributes but are uncorrelated with the grand mean, unlike dummy variables (Louviere Hensher and Swait, 2000).

Some artificial results from a traditional SPDCM exercise performed using dummy variables are illustrated in Figure 1. This example evaluated a good with five attributes, each with three categorical (ordered) levels, 'low', 'medium' and 'high'. The explanatory variables for the regression comprised five multiplied by two (three minus one), i.e. ten dummy variables. The constant term (not presented) represents the utility of the specification defined by level one on all attributes and the utilities of the remaining ten attribute levels are estimated relative to this scenario.

<Figure 1>

Whilst attribute 2 looks to be important, this is misleading – it is merely the case that levels 2 and 3 of this attribute have a large additional utility relative to level 1. These results tell us nothing about how large each of the levels is in relation to each other in the benchmark specification of the good. Best-worst scaling can achieve this, however, by asking respondents to perform a different task which elicits additional information that can be exploited in the analysis of choice data.

3. Best-Worst Scaling – The choice task

Unlike traditional SPDCM, best-worst presents the respondent with each scenario one at a time – in other words the choice set is of size one. Thus, rather than (internally) evaluating and comparing the utility of entire scenarios, respondents evaluate and compare the utilities of the attributes on offer (or, rather, the particular attribute levels on offer), picking that pair of attributes that maximises the difference in utility between them. Thus, variations on best-worst scaling have appeared before, in the guise of "maximum difference analysis" (Szeinbach et al., 1999). Ideally respondents are also asked to accept or reject the scenario on offer (with rejection implying acceptance of whatever the status quo is) in order to perform a traditional SPDCM analysis for comparison.

Thus, when a respondent has to make a best-worst choice in a scenario, he/she chooses that attribute (based on the observed level) that exhibits the highest utility and that attribute (based on the observed level) that exhibits the lowest utility.
Statistically, this process is represented by:

- Identifying every possible pair of attributes available
- Calculating the difference in utility between the best and worst attribute in every pair. This consists of a fixed component (equal to the unobserved but fixed utilities of each attribute level) plus a random component
- Choosing that pair that gives the largest difference in utility

Thus Figure 2 shows an example scenario from a best-worst task.

<Figure 2>

4. Decomposing the constant term

By deciding that the most attractive feature of the good described in this scenario is attribute two (which takes level 'high') and the least attractive feature is attribute one (which takes level 'low'), the respondent has in effect decided that, of all the possible pairs he/she could have chosen from this scenario, this pair exhibits the largest

difference in utility between the two attribute levels (level 'high' of attribute 2 and level 'low' of attribute one).

By asking respondents to make an *explicit* choice between attribute levels in this way, a best-worst regression estimates a difference between these two attribute levels, although one of the *absolute* utilities in the pair remains undefined. Given a sufficient number of scenarios valued by people (with 'sufficient' defined later), the undefined attribute level utility can be estimated relative to another attribute level, and so on. Thus, in effect by 'chaining together' all the differences between all the (in this case fifteen) attribute levels, the best-worst method effectively estimates all attribute level utilities relative to a single attribute level (which remains undefined). Thus it is no longer an entire scenario whose utility acts as benchmark, it is just a single level of a single attribute. Figure 3 illustrates the results of best-worst scaling using the same data as those used in Figure 1, adding the best-worst regression constant term into all 15 attribute level utilities to give total utilities.

<Figure 3>

This shows the crucial benefit of best-worst over traditional SPDCM – the decomposition of the constant term in the regression analysis. The large effect on utility by moving from 'low' to 'medium' and from 'medium' to 'high' in attribute two is seen, as before, but there is more information now available. Unlike the conclusion from Figure 1, attribute three is seen to be more important, on average, than attribute two in terms of overall utility.

This section has indicated the relative advantages of pursuing best-worst scaling rather than traditional SPDCM methods. However, analysing choice data from a best-worst exercise for analysis is less straightforward than that in traditional SPDCM. The theoretical issues and methods of analysis are set out in detail below.

5. Theoretical issues in analysis

For ease of exposition, consider a good or service described by only three attributes (K=3), each with two levels (N=2). In such a situation, when every attribute has the same number of levels, the design is said to be balanced. Unbalanced designs are dealt with in section 8. Assume these attributes are categorical in nature, necessitating the use of dummy variables. (They can be made quantitative with no loss of generality.) If U_*i*_*j* represents the utility of level *j* of attribute *i*, then the best-worst method aims to estimate the part-worth utilities of the six attribute levels (U_1_1 through to U_3_2) on the same scale.

In best-worst, as in other SPDCM, the maximum possible number of scenarios that can result is given by the product of the number of levels (across all attributes). In a balanced design like this, the total number of scenarios is N^K or in this case 2^3=8. The attribute levels available in each of the eight scenarios here are displayed in Table 1.

However, the choices that respondents make are between attribute levels observed *within a given scenario*. Thus, in scenario one the set of three attribute levels available to be picked for 'best' and 'worst' is (U_1_1, U_2_1, U_3_1). The set of six possible best-worst pairs that could be chosen from scenario one is therefore as given in Table 2. More generally, the total number of possible best-worst pairs available to be chosen is given by pairing attribute one with each of the remaining K-1 attributes on offer in a profile, then pairing attribute two with each of the remaining K-2 attributes (attribute one has already been paired with it) etc. The last pairing is that of attribute K-1 with attribute K. The order can then be reversed on all these pairings to give worst-best (W,B) combinations as opposed to the (B,W) that we have so far. Thus the number of pairings in a given scenario is given by:

$2\{ (K-1) + (K-2) + (K-3) + \ldots + (2) + (1) \}$

Or, algebraically, $2\sum_{j=1}^{K-1} j = (K-1)K$ possible best-worst combinations, which in this case is six, agreeing with Table 2.

This is the number of pairs available in a given scenario. However, there are other scenarios to consider: none of them will contain all of the same K(K-1) pairs but, as in

table 1, some of the three attribute levels in scenario one will reappear, and new ones will appear. How many *distinct* pairs are there across all N^K scenarios?

For a given pair of attributes, when the possible levels that could appear across all scenarios are considered, there are N possible best levels and N possible worst levels. Thus there are N^2 distinct pairs for those two attributes and across all attributes there are $N^2 \times K(K-1)$ pairings. In this simple case, there are a total of 24 distinct best-worst pairs that are on offer, with a subset of six of them available to be chosen in any one of the eight scenarios. The design needs to ensure that each of the $N^2 \times K(K-1)$ pairings appears sufficiently often for the required utilities to be estimated. A main effects design ensures that across the scenarios in the best-worst exercise every one of these pairings can be estimated. If interactions are considered important, the appropriate design matrix must be utilised – for instance, a 'resolution 5' design permits all main effects and two-way interactions to be estimated independently of one another.

If the number of scenarios used in a best-worst exercise is too small, some of the $N^2 \times K(K-1)$ pairings may not appear at all and so cannot be estimated. This might not be such a serious problem because, say, estimates of (B-A) and (C-A) allow (C-B) to be estimated in a design that did not allow direct estimation of the latter, if there are no interactions. Nevertheless it is the need for assumptions about interactions that make reliance on such overly restricted designs unwise.

Best-worst data can be analysed in either an aggregated or disaggregated format. When analysing the results in an aggregated format using weighted least squares (WLS) there are two ways that the choice data can be aggregated and analysed – the 'full' method or the 'restricted' method (Marley and Louviere, 2004). These will be set out first, using a balanced design. Section 8 will describe the adjustment that needs to be made to both of these methods to allow unbalanced designs. Section 9 will set out how to analyse the data when it is fully disaggregated (in other words, respondent level data).

6. Analysis method 1 – Full method

The full method treats each possible best-worst pair as a unique datapoint. Thus the number of observations is equal to the number of unique best-worst pairs that can be estimated, given the design of the exercise, $N^2 \times K(K-1)$ being the maximum number possible in a balanced design. In a main effects design every one of these $N^2 \times K(K-1)$ pairs will have been available to be chosen at least once. In a balanced design, where every attribute has the same number of levels, each possible pair will have been available to be chosen the same number of times. However, in an unbalanced design the number of levels per attribute varies and so, for example, the levels of a four-level attribute will only appear half as often as the levels of a two-level one. This has implications for the analysis that will be described in Section 8.

The data are analysed using weighted least squares with KN dummy variables, D_1_1 through to D_3_2, using the example given above. The equation estimated is given by:

$$\ln(f) = \alpha + \beta_{11} D\_1\_1 + \beta_{12} D\_1\_2 + \beta_{21} D\_2\_1 + \beta_{22} D\_2\_2 + \beta_{31} D\_3\_1 + \beta_{32} D\_3\_2$$

where f is the total number of times a particular best-worst pair was picked across all scenarios and across all respondents, adjusted to eliminate zeros. The natural log of the total number of times each pair was chosen is a linear function of the difference in utility (McIntosh and Louviere, 2002). Any pairs with zero choices must be adjusted slightly to enable logs to be taken. This is achieved by adding the reciprocal of the total sample size to the choice total, as recommended by Ben-Akiwa and Lerman (Ben-Akiva and Lerman, 1985). Table 3 sets out the data used in the final regression. The first three columns are for information only and show the 24 best-worst pairs to be estimated in this example. Column four contains the weight variable (the choice totals f adjusted to eliminate zeros) whilst column five contains the response variable (the natural log of column four). Columns six through 11 comprise the dummy variables that form the explanatory variables in the final regression to estimate the utility part-worths. It can be seen that these are not like traditional dummy variables – a particular dummy variable takes a value of one if that attribute level is best, minus one if that attribute level is worst, and zero otherwise.

7. Analysis method 2 – Restricted method

The restricted method aggregates the data further to estimate the *NK* attribute level utilities using a model that while simpler, requires an adjustment to the choice data in order to eliminate bias in the frequencies. All frequencies are reduced by 1/(|X|-1) where |X| is the number of best-worst/worst-best pairs available in a scenario (equal to K(K-1) in a balanced design). There are a total of *2NK* observations – each of the attribute levels contributes two observations, a best and a worst total. The data are again analysed by weighted least squares but with the equation given by:

$$\ln(g) = \alpha + \beta_{11} D\_1\_1 + \beta_{12} D\_1\_2 + \beta_{21} D\_2\_1 + \beta_{22} D\_2\_2 + \beta_{31} D\_3\_1 + \beta_{32} D\_3\_2$$

where *g* is the total number of times a particular attribute level was picked across all scenarios and across all respondents (as opposed to a particular best-worst pair in the full method), with two adjustments: the first adjustment eliminate zeros, as in Section 6 whilst the second adjustment (see above) eliminates the bias in choice frequencies. Thus there are *NK* best totals and *NK* worst totals. Table 4 sets out the data used in the final regression. The first three columns are for information only and show the 12 best-worst totals used in this example. Column four contains the weight variable (the choice totals adjusted to eliminate zeros) whilst column five contains the dependent variable (the natural log of column four). Columns six through 11 comprise the dummy variables that form the explanatory variables in the final regression to estimate the utility part-worths.

8. Unbalanced designs

In an unbalanced design not all best-worst pairs are equally available for selection – the levels of attributes with fewer levels appear more often than those from attributes with more levels. The WLS analysis requires a response variable which represents the number of times each best-worst pair (or attribute level if the restricted method is used) *would have been chosen* if every pair (or attribute level) had been available to be picked the same number of times. More formally, the probability of choosing a particular level is not independent of the probability that it is available to be chosen; hence, one must condition on the probability of being available through the use of an adjustment. The adjustment is performed by:

1. Dividing the number of times each best-worst pair under method 1 or attribute level under method 2 was chosen by the number of times it was available to be chosen across all scenarios and individuals (the availability totals variable). This produces a variable containing the $N^2 \times K(K-1)$ pair (or 2KN attribute level) choice frequencies.
2. Multiplying all the choice frequencies by *one* value from the availability totals variable. In a design where all but one attributes had the same number of levels there will be two availability totals to choose from, that of the K-1 attributes and that of the Kth attribute with a different number of levels. It does not matter which availability total is used, but it is more logical to pick the one that appears most often (that of the K-1 attributes in this example) so that as many as possible of the original choice totals are restored.

Methods 1 and 2 use aggregated data and the effects of individual-level factors upon preferences cannot be estimated. To do so requires fully disaggregated (respondent level) data that must be analysed using multinomial logistic (MNL) or multinomial probit (MNP) regression.

9. Fully disaggregated data

To estimate the effect of respondent characteristics (such as age and sex) upon utilities, covariates must be introduced. These must take the form of respondent-dummy interaction terms and require the analyst to use disaggregated data.

This is because limited dependent variable models require differences in the probabilities of choice for the various outcomes in a choice set to be associated with differences in the explanatory variables. Since respondent characteristics, such as age, do not vary for potential best-worst pairs in a choice set they cannot affect choice probabilities and cannot be separated out from the overall regression constant term. In addition to this practical concern, there is a conceptual one: the main effect of, for example, age upon utility has no meaning – it is only the effect that age has upon the utility gained for a particular attribute that has meaning. Either of the regression models above could be analysed using multinomial regression. The number of

observations would be multiplied by the sample size (assuming no missing data) but the dummy variables above are unchanged. However, it should be noted that method 2 is more amenable to such analysis than method 1 – for instance, Stata (Stata Corporation, 2003) allows a maximum of 50 possible outcomes in MNL/MNP regression and the number of potential pairs under method 1 quickly becomes large as the design gets more complex (that is, as N or K increases).

10. Other methodological issues

Best-worst scaling can give additional information to that obtained in traditional SPDCM. In particular, unlike traditional SPDCM, it is the utility of a single level of one attribute that acts as a benchmark, not an entire scenario. However, there are estimation issues that are not necessarily solved by the best-worst scaling procedures detailed above but which are common to all SPDCM methods. It is by drawing attention to these issues within a best-worst framework that greater awareness of them may be fostered among practitioners of discrete choice modelling in general.

10.1. Dealing adequately with the random component of utility

When designing a discrete choice experiment, attention must be paid to the random component of utility. This can be conceptualised as any combination of a number of sources of variation – respondent inability to fully recognise systematic differences in utility and measurement error being two examples. Recognition of the issues surrounding variation in the random component of utility has been limited in many areas of applied economics and a comprehensive account of developments in the area only appeared in 1999 (Hensher Louviere and Swait, 1999). A full exposition of the issues can be found in Swait and Louviere (Swait and Louviere, 1993), but in short all parameter estimates from choice models based on random utility theory are confounded with an unknown scale factor. This scale factor is inversely related to the variance of the random component of the utilities underlying people's choice behaviour. For example, suppose we wish to estimate utilities from two groups of people. Both groups exhibit the same underlying fixed utilities but the people in group one make more mistakes evaluating the scenarios than those in group two – in other

words the variance of the random utility component is larger in group one than that in group two. Utility estimates from a model using group one will appear to be closer to zero than those in group two, despite the fact that the underlying fixed components of utilities in the two groups are identical. One way of dealing with this in some empirical studies is the introduction of a cost attribute which facilitates estimation of willingness to pay, thereby cancelling the unknown scale factor from numerator and denominator. The assumption of constant marginal utility of money across studies must still be invoked, however, if cross-study comparisons are to be made. For some purposes and in some contexts, including a cost attribute is unlikely to be feasible.

In health economics, the treatment of the variation in utility between respondents within discrete choice experiments has largely been restricted to the use of random effects to model respondent heterogeneity in (usually probit) regression models. It is certainly the case that large SPDCM studies might necessitate blocked designs and the need for distributional assumptions in making inference. However, not only does this particular focus on preference heterogeneity ignore the other factors that might lead to variation in choice behaviour, it is conceptually equivalent to allowing for variation in the fixed component of individuals' utilities (the mean) but not in the variance of the random component. As such it is a partial solution at best and there is evidence to suggest that this simplistic treatment of heterogeneity is not supported empirically (Louviere, 2001). Furthermore, failure to allow for variation in the mean by the inclusion of random effects only leads to incorrect standard errors; failure to recognise variation in the random component of utility leads to incorrect point estimates.

Given recent advances in the estimation and treatment of the scale factor it would seem more logical to exploit the power of best-worst to make individual-level inference than to attempt to introduce random effects into the models detailed above. Indeed, work has begun to utilise the power of best-worst scaling to model individual-level utility functions that require no statistically questionable distributional assumptions surrounding preferences (Louviere and Marley, 2004).

10.2. Sample size issues

Given that best-worst scaling represents a different choice task with different outcomes to that in traditional SPDCM, sample sizes for estimating individual-level utilities are unknown. However, if the analyst is interested in the differences between the proportions of respondents choosing the various attribute levels, equations for confidence intervals can be used to estimate required sample sizes. Such equations would utilise knowledge of factors such as the number of times best-worst pairs were available to be chosen, which is available from the statistical design.

In the current absence of guidance for defining sample sizes for best-worst studies where heterogeneity in respondent preferences is expected, one way forward would be the use of simulation studies. These would vary the size of the random component relative to the fixed component of utility to determine the influence of particular sample sizes upon the reliability of estimates and thus provide guidance for various designs.

11. Conclusion

Best-worst scaling asks respondents to perform a different task from that in most SPDCM exercises performed to data. In so doing it provides additional insights over those from traditional SPDCM studies that should prove attractive in health care research. Its ability to decompose the constant term and thereby allow the absolute importance of attributes to be compared will be valuable in evaluating many aspects of service provision. In such situations the pairwise comparison method is unsuitable and the results from previous studies are to be treated with caution until and unless they can be corroborated with those from best-worst scaling exercises. When researchers are interested in comparisons of marginal changes in attributes pairwise comparisons can be statistically valid but given the potentially easier choice task inherent in a best-worst exercise there may be a case for reconsidering traditional methods here too. This paper provides guidance in analysing best-worst data and future work will provide further evidence to inform sample size calculations and other design issues.

References

Ben-Akiva, M. and Lerman, S. R., 1985. Discrete choice analysis: theory and application to travel demand. MIT Press, Cambridge, MA.

Farrar, S., Ryan, M., Ross, D., Ludbrook, A., 2000. Using discrete choice modelling in priority setting: an application to clinical service developments. Soc. Sci. Med. 50, 63-75.

Finn, A. and Louviere, J. J., 1992. Determining the Appropriate Response to Evidence of Public Concern: The Case of Food Safety. J. Public Policy Mark. 11, 12-25.

Hensher, D. A., Louviere, J. J., Swait, J., 1999. Combining sources of preference data. J. Econometrics. 89, 197-221.

Louviere, J. J., 2001. What if consumer experiments impact variances as well as means: Response variability as a behavioural phenomenon. J. Consum. Res. 28, 506-511.

Louviere, J. J., Hensher, D. A., Swait, J., 2000. Stated choice methods: analysis and application. Cambridge University Press, Cambridge.

Louviere, J. J. and Marley, A. A. J., 2004. Modeling the choices of single individuals by combining efficient choice experiment designs with extra preference information. CenSoC working paper series 04-004. Centre for the Study of Choice, Faculty of Business, University of Technology, Sydney.

Louviere, J. J. and Timmermans, H., 1990. Stated Preference and Choice Models Applied to Recreation Research: A Review. Leisure Sci. 12, 9-32.

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

Marley, A. A. J. and Louviere, J. J., 2004. Some probabilistic models of Best, Worst, and Best-Worst choices. CenSoC working paper series 04-005. Centre for the Study of Choice, Faculty of Business, University of Technology, Sydney.

McIntosh, E. and Louviere, J. J. Separating weight and scale value: an exploration of best-attribute scaling in health economics. Health Economics Study Group. January 2002.

Stata Corporation., 2003. Stata Statistical Software. College Station, TX

Swait, J. and Louviere, J. J., 1993. The role of the scale parameter in the estimation and comparison of multinomial logit models. J. Marketing Res. 30, 305-314.

Szeinbach, S. L., Barnes, J. H., McGhan, W. F. et al, 1999. Using conjoint analysis to evaluate health state preferences. Drug Inf. J. 33, 849-858.

Vick, S. and Scott, A., 1998. Agency in health care. Examining patients' preferences for attributes of the doctor-patient relationship. J. Health Econ. 17, 587-605.

Table 1: Scenario specification of 2^3 design

Scenario Number	Attribute 1 Level	Attribute 2 Level	Attribute 3 Level
1	1	1	1
2	1	1	2
3	1	2	1
4	1	2	2
5	2	1	1
6	2	1	2
7	2	2	1
8	2	2	2

Table 2: Best-worst pairs available in scenario one

Best-worst pair number	Best attribute level	Worst Attribute level
1	U_1_1	U_2_1
2	U_1_1	U_3_1
3	U_2_1	U_3_1
4	U_2_1	U_1_1
5	U_3_1	U_1_1
6	U_3_1	U_2_1

Table 3: Example dataset for full method WLS analysis

	Attribute number and level number		Weight	Response variable	Independent (dummy) variables					
BW pair	Best	Worst	Adjusted total	ln(adjusted total)	D_1_1	D_1_2	D_2_1	D_2_2	D_3_1	D_3_2
1	1_1	2_1	#	#	1	0	-1	0	0	0
2	1_1	2_2	#	#	1	0	0	-1	0	0
3	1_1	3_1	#	#	1	0	0	0	-1	0
4	1_1	3_2	#	#	1	0	0	0	0	-1
5	1_2	2_1	#	#	0	1	-1	0	0	0
6	1_2	2_2	#	#	0	1	0	-1	0	0
7	1_2	3_1	#	#	0	1	0	0	-1	0
8	1_2	3_2	#	#	0	1	0	0	0	-1
9	2_1	3_1	#	#	0	0	1	0	-1	0
10	2_1	3_2	#	#	0	0	1	0	0	-1
11	2_2	3_1	#	#	0	0	0	1	-1	0
12	2_2	3_2	#	#	0	0	0	1	0	-1
13	2_1	1_1	#	#	-1	0	1	0	0	0
14	2_2	1_1	#	#	-1	0	0	1	0	0
15	3_1	1_1	#	#	-1	0	0	0	1	0
16	3_2	1_1	#	#	-1	0	0	0	0	1
17	2_1	1_2	#	#	0	-1	1	0	0	0
18	2_2	1_2	#	#	0	-1	0	1	0	0
19	3_1	1_2	#	#	0	-1	0	0	1	0
20	3_2	1_2	#	#	0	-1	0	0	0	1
21	3_1	2_1	#	#	0	0	-1	0	1	0
42	3_2	2_1	#	#	0	0	-1	0	0	1
23	3_1	2_2	#	#	0	0	0	-1	1	0
24	3_2	2_2	#	#	0	0	0	-1	0	1

Table 4: Example dataset for restricted method WLS analysis

	Attribute number and level number		Weight	Response variable	Independent (dummy) variables					
Obs Number	Best	Worst	Adjusted total	ln(adjusted total)	D_1_1	D_1_2	D_2_1	D_2_2	D_3_1	D_3_2
1	1_1		#	#	1	0	0	0	0	0
2	1_2		#	#	0	1	0	0	0	0
3	2_1		#	#	0	0	1	0	0	0
4	2_2		#	#	0	0	0	1	0	0
5	3_1		#	#	0	0	0	0	1	0
6	3_2		#	#	0	0	0	0	0	1
7		1_1	#	#	-1	0	0	0	0	0
8		1_2	#	#	0	-1	0	0	0	0
9		2_1	#	#	0	0	-1	0	0	0
10		2_2	#	#	0	0	0	-1	0	0
11		3_1	#	#	0	0	0	0	-1	0
12		3_2	#	#	0	0	0	0	0	-1

Figure 1: Estimated attribute level utilities from traditional SPDCM exercise

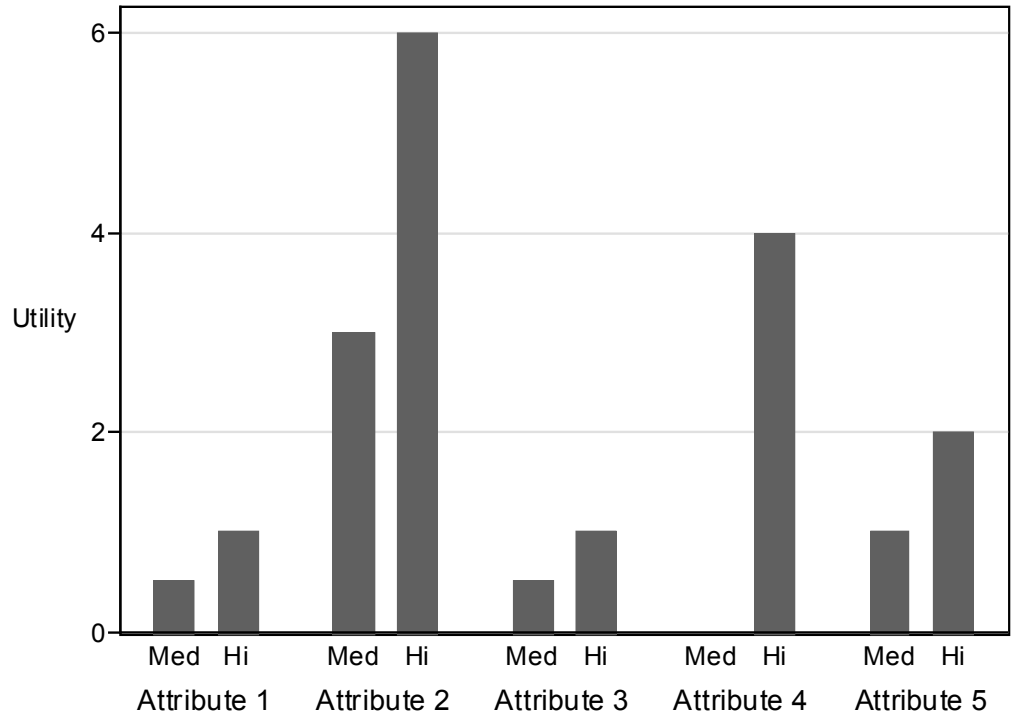

Figure 2: Example of a Best-Worst Scaling exercise scenario

Scenario #3		
Best		Worst
	Level 'Low' of attribute 1	✓
✓	Level 'High' of attribute 2	
	Level 'Medium' of attribute 3	
	Level 'Medium' of attribute 4	
	Level 'High' of attribute 5	

Figure 3: Estimated attribute level utilities from Best-Worst Scaling exercise