Available online at www.sciencedirect.com
ScienceDirect
Journal of Health Economics 26 (2007) 171–189
JOURNAL OF HEALTH ECONOMICS
www.elsevier.com/locate/econbase

Best–worst scaling: What it can do for health care research and how to do it

Terry N. Flynn [a,*], Jordan J. Louviere [b], Tim J. Peters [c], Joanna Coast [d]

[a] *MRC Health Services Research Collaboration, Department of Social Medicine, University of Bristol, UK*
[b] *Centre for the Study of Choice, University of Technology, Sydney, Australia*
[c] *Department of Community Based Medicine, University of Bristol, UK*
[d] *Health Economics Facility, University of Birmingham, UK*

Received 9 February 2005; received in revised form 27 March 2006; accepted 19 April 2006
Available online 16 May 2006

Abstract

Statements like "quality of care is more highly valued than waiting time" can neither be supported nor refuted by comparisons of utility parameters from a traditional discrete choice experiment (DCE). Best–worst scaling can overcome this problem because it asks respondents to perform a different choice task. 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 illustrates how to aggregate and analyse such data and using a quality of life pilot study demonstrates how richer insights can be drawn by the use of best–worst tasks.

JEL classification: C35; C91 C92; I39

Keywords: Discrete choice experiments; Utility; Attribute importance

1. Introduction

In a discrete choice experiment (DCE) people's preferences for goods or services are elicited based on their intentions expressed in hypothetical situations (Louviere et al., 2000). This stated preference analysis distinguishes it from revealed preference analysis, which utilises people's

* Corresponding author at: Canynge Hall, Whiteladies Road, Bristol BS8 2PR, UK. Tel.: +44 117 928 7375; fax: +44 117 928 7236.
E-mail address: terry.flynn@bristol.ac.uk (T.N. Flynn).

0167-6296/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhealeco.2006.04.002

observed behaviour in real markets. DCEs are increasingly used in health services research (HSR) 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. In a DCE the researcher 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 various attribute levels (often referred to as part-worth utilities).

In some applications, most notably health care, policymakers are interested in comparing the absolute impact (utilities) of attributes. An example would be testing the hypothesis that "waiting time for an appointment is more important to patients than continuity of care" — a statement purely about attributes, with no reference to the associated levels. Some studies have referred to this as the issue of separating attribute weights and scales — estimating the utility associated with a particular attribute per se (its impact weight in a utility function) separately from the additional utility gained/taken away by that attribute exhibiting an attractive/unattractive level (the scale value).

Unfortunately traditional DCEs cannot address attribute impact issues by comparing utility estimates and unwarranted conclusions about attribute impact have been drawn in previous studies. For example, in the study by Vick and Scott on preferences for GP consultations, the statement "the most important attribute was 'being able to talk to the doctor', whilst 'who chooses your treatment' was least important" was not warranted from the study conducted (Vick and Scott, 1998). Claims such as this are difficult to support or refute in traditional DCEs without careful consideration of design issues (in order to allow comparisons in log-likelihoods or willingness to pay estimates) and are impossible to test at the respondent level. The use of mostly two-level attributes in the study by Vick and Scott, with most of these attempting to capture good/bad extremes, did not enable a meaningful comparison of attribute impact, 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. Whilst such limitations of traditional DCEs are more likely to be acknowledged now, many applied practitioners remain unaware of these issues.

However, best–worst scaling (Marley and Louviere, 2005), devised by Finn and Louviere (1992) and introduced to health care research by 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 DCEs, it elicits additional information. The full nature of the choice task and theoretical model are set out in Section 3 whilst Section 4 summarises a pilot best–worst study that was conducted in the field of quality of life valuation. Section 5 describes the analytical framework, using the quality of life study to illustrate practical and theoretical issues. Section 6 sets out some more advanced issues in best–worst and Section 7 concludes. First of all, the limitations of traditional DCEs will be described, together with the justification for using best–worst scaling.

2. Limitations of traditional discrete choice experiments

There are two principal limitations of traditional DCEs that have prompted research into best–worst scaling. First, the 'pick one' nature of the task is a relatively inefficient way to elicit preference information. Second, the nature of the regression constant term means that attribute impacts are confounded with level scale values. These issues will be explained in turn.

2.1. Efficiency of choice tasks in DCEs

A traditional DCE involves choosing the most preferred specification of a good ('alternative' or 'scenario') from a choice set of competing scenarios, repeated over a number of such choice sets so as to observe trade-offs (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 HSR 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). 'Pick one' tasks require few statistical assumptions but many researchers, when faced by design and budgetary constraints, have utilised rating and ranking tasks to elicit additional preference information (Hausman and Ruud, 1987).

The problem with ranking and rating tasks is that they may induce behaviour in respondents that violates the statistical assumptions inherent in these models. Any rating or ranking exercise can be 'exploded' into a series of pairwise comparisons which can be analysed (Ben-Akiva et al., 1991). A pairwise comparison task is the least demanding for respondents cognitively and if the results from the ranking or rating exercise model do not agree with the exploded pairwise comparison estimates (after adjusting for any differences in random variation inherent in the tasks) this inconsistency indicates a violation of the assumptions of the rating/ranking model. This has been observed in a variety of studies and across several disciplines, with the violations commonly occurring in the results from the 'middle' rankings (Ben-Akiva et al., 1991; Bradley and Daly, 1994). Best–worst tasks provide a middle ground: they provide more information than a 'pick one' task, but by forcing respondents to consider only the extremes of the utility space they provide richer insights while minimising the chances of introducing assumptions about human decision-making that are not satisfied in practice.

2.2. The constant term in traditional DCEs

There are a variety of best–worst techniques that can be applied in the context of choice experiments and to understand what best–worst *scaling* does it is useful first to consider the DCE of the type most often applied to date in HSR, the pairwise comparison. In a pairwise comparison DCE respondents are asked to choose between two scenarios, A and B. In many HSR studies no opt-out (i.e. reject both) option has been presented. As such, the statistical information obtained has represented conditional demand; inference being conditional on respondents 'buying' (or more accurately in the UK NHS simply choosing) health care. In a pairwise comparison when respondents choose their preferred scenario they are effectively providing information about their preferences for a set of attribute differences — for instance in the case of an appointment this might be the set comprising the additional utility of a long appointment over a short one and the additional disutility of having no choice over the doctor seen. Regression estimates from probit (or logit) models represent the additional utility/disutility of moving between levels of the attributes. The model is, in essence, a difference model and the constant (representing any systematic propensity to choose A rather than B, across all choice sets and respondents) would not be expected to be significantly different from zero. Being a conditional demand model, the total utility of any scenario relative to 'not buying/choosing' is not recoverable.

Introduction of an opt-out option (reject both A and B) enables estimation of unconditional demand: it allows respondents to stick with some status quo and 'not demand'. However, whilst this solves the issue of estimating a full demand function (and thereby generally, though not necessarily,

introducing a significant constant term), there is an issue with what the constant represents. In such a model the constant term represents some 'bundle' of attribute levels that cannot be decomposed into its constituent parts. In an epidemiological setting this conceptualisation of the constant is entirely natural. For example, consider death rates among smokers/non-smokers and males/females. The odds ratio of smokers to non-smokers can be estimated, as can that of males relative to females. The 'reference case' (and hence constant term) is the odds of death among non-smoking females (for instance): in a female non-smoker the contributions of being female and being a non-smoker to death cannot be separated (and it is nonsensical to try). However, in a study investigating preferences, it is both sensible and desirable to estimate the contributions of attribute levels to the reference case: sensible because people can make choices about the relative desirability of the various components of the reference case; desirable because an attribute's 'impact' – defined as the average utility of an attribute across all of its levels – cannot be estimated when one of its levels is not estimated separately from the other attributes comprising the reference case.

More formally, consider a traditional DCE incorporating an opt-out option using a design with K attributes where L_k represents the number of levels of attribute k. Each attribute will have $(L_k - 1)$ degrees of freedom: $L_k - 1$ dummy/indicator variables or effects codes can be estimated for attribute k, making $\sum_{k=1}^{K}(L_k - 1)$ in total. 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 since one dummy variable is omitted from every attribute the regression constant term represents the total utility of the reference case — that scenario defined by the omitted (benchmark or reference) level on each attribute. Depending upon the coding of the dummy variables, this reference case can be the worst possible specification of the good or service, from which all estimates represent contrasts, or it can be a specification that in fact does not/cannot exist. In other words for attribute k the $L_k - 1$ dummy variables represent the additional utility of each of the remaining $L_k - 1$ levels of attribute k from the reference level.

If the attribute levels are coded using effects coding, again one level per attribute is omitted but the constant now represents the grand mean over all observations — the average utility across the sample, which permits estimation of respondents' average propensity to choose particular attribute levels. The $L_k - 1$ estimated independent effects codes represent the additional utility of that attribute level from the mean utility (over all levels and respondents). The additional utility of the omitted level is equal to minus the sum of the other estimated level utilities. Effects coded variables are correlated within attributes but are uncorrelated with the grand mean, unlike dummy variables (Louviere et al., 2000) and there has been a recent warning in the health economics literature against the use of dummy variables (Bech and Gyrd-Hansen, 2005).

It should be clear that both dummy variables and effects code estimates from a traditional DCE are simply deviations from a total utility, whether of some reference case of a good or the mean utility. An attribute with two levels will have one dummy variable or effects code associated with it. If its coefficient is not statistically significant then it is incorrect to conclude that the attribute is unimportant to respondents: they merely do not perceive a detectable difference in the two level scale values. The impact that any attribute as a whole has to them (defined as the average utility across the L_k levels on the latent, or unobserved, scale) cannot be estimated because one of the levels is not estimated. Returning to the epidemiological example, whilst eliciting the contributions of 'female' and 'non-smoker' to the mortality of a female non-smoker is nonsensical, eliciting a respondent's relative preferences for 'a short appointment' and 'being seen by any doctor' in

'a short appointment with any doctor' is not: by asking respondents to make an explicit choice between attribute levels in a given appointment, best–worst scaling can make inferences about their relative contributions to the utility of a short appointment with any doctor.

More formally, best–worst scaling can provide a more useful reference case than unconditional demand models from traditional DCEs, that is, the reference case is a single attribute level, not an entire scenario. By estimating all of an attribute's levels on the same scale, it allows the researcher to estimate the average utility for attribute k across all its L_k levels — the attribute impact. Only one attribute has a level 'missing' from estimation; this is the only attribute for which the impact cannot be estimated and it acts as the reference case. The level scale values can be calculated for all K attributes. There are advantages to traditional DCEs as well as disadvantages and issues in drawing on the strengths of both methods will be discussed in Section 6 but first the theory and properties of best–worst scaling will be outlined.

3. Best–worst scaling — theory and properties

3.1. The choice task

Consider, as before, a choice experiment with K attributes where L_k represents the number of levels of attribute k. Each scenario (doctor's appointment etc.) is described by each of these K attributes taking a particular level. However, unlike a traditional DCE the scenarios are presented one at a time to respondents. Thus, rather than (internally) evaluating and comparing the utility of entire scenarios, respondents evaluate and compare the utilities of all the attribute levels on offer, picking that attribute level that exhibits the highest utility and that attribute level that exhibits the lowest utility. In effect, the respondent has provided that pair of attribute levels that he/she considers to be furthest apart on the latent utility scale.

Subjects are asked to repeat this task for a number of best–worst scenarios (with 'a number' defined later). Additional questions can be asked for each scenario, such as "Would you accept/buy this scenario/good/service", which allows researchers:

1. to estimate a traditional DCE that can be compared with the best–worst estimates, and
2. to use the conditional demand information from the best–worst task to be used in an unconditional demand function. This will be described in Section 6.

In order to understand the best–worst scaling model and how to analyse the data it is necessary to set out the model formally.

3.2. Statistical theory

Best–worst scaling is rooted in Random Utility Theory, a well-tested theory of human decision-making hypothesised by Thurstone and generalised by McFadden (McFadden, 1974; Thurstone, 1927). Although there have been published papers utilising best–worst scaling over the past 15 years (Finn and Louviere, 1992; Szeinbach et al., 1999), the formal statistical and measurement properties were proved only recently (Marley and Louviere, 2005). The statistical model underlying best–worst scaling assumes that the relative choice probability of a given pair is proportional to the distance between the two attribute levels on the latent utility scale. The cognitive process undertaken by individuals is statistically equivalent to:

- Identifying every possible pair of attributes available.
- Calculating the difference in utility between the two attribute levels 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 maximises the difference in utility between them.

Thus, the pair of attribute levels chosen maximises the difference in the part-worth utilities on offer in a scenario. These distances between attribute levels are modelled as a difference model, with variations on best–worst scaling sometimes called "maximum difference scaling" (Szeinbach et al., 1999). Best–worst analysis 'chains together' all estimated differences with the result that the part-worth utilities are estimated relative to a single attribute level, rather than relative to an entire scenario (or the sample mean). More formally, $\sum_{k=1}^{K} L_k - 1$ attribute levels are estimated relative to the remaining level (and hence on a common scale), rather than the $\sum_{k=1}^{K}(L_k - 1)$ in a traditional DCE.

To model the pairs, one first considers how many unique best–worst pairs are available to be chosen in each scenario. Pairs are constructed by pairing attribute one with each of the remaining $K-1$ attributes on offer in a scenario, 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, which produces all possible (B,W) pairs across the scenarios. The order is then be reversed for these pairs to give all possible worst–best (W,B) pairs. Thus, the number of pairs in a given scenario is given by

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

Or, algebraically, there are $2\sum_{j=1}^{K-1} j = (2(K-1)K)/2 = K(K-1)$ possible best–worst combinations, which is the number of pairs in a given scenario, say scenario s. However, there are $S-1$ other scenarios (where S is the total number of scenarios) to consider: none of them will contain the same $K(K-1)$ pairs but some attribute levels in scenario s will reappear, in different combinations with other levels.

The number of *distinct* pairs when all $\sum_{k=1}^{K} L_k$ attribute levels are considered is $2\sum_{i=1}^{K-1}\left[L_i \sum_{k=i+1}^{K} L_k\right]$ (Marley and Louviere, 2005), whilst the number available to be chosen in any given scenario is $K(K-1)$. Typically, small fractional designs like orthogonal main effects plans (OMEPs) do not allow all pairs to be observed independently, hence they cannot all be estimated independently. Thus, the existence of unobserved interactions between attributes may have more serious consequences for parameter estimates than those from a traditional DCE utilising an OMEP. Interactions in best–worst models are the subject of current research.

3.3. Advantages and disadvantages over traditional DCEs

It has been shown that best–worst scaling estimates the utilities of all but one of the $\sum_{k=i}^{K} L_k$ attribute levels in a best–worst choice experiment. This enables the impact of all but one attribute to be estimated, where impact of an attribute is the average across all its levels, which traditional 'pick one' DCEs cannot do. However, like pairwise comparison DCEs, the estimates represent conditional demand: respondents are assumed to choose/purchase. However, by asking the best–worst question alongside a traditional accept/reject question, the strengths of both choice experiments can be drawn upon. This will be explored in Section 6. Analysing choice data from a best–worst exercise for analysis is less straightforward than that in traditional DCEs. The theoretical issues

and methods of analysis will be described with reference to an empirical example, which is described below.

4. Empirical example — quality of life pilot study

A pilot best–worst study was conducted in summer 2005 among 30 people aged 65 and over with the aim of informing a larger quality of life valuation exercise. The study was interviewer-administered, and respondents were presented with hypothetical quality of life states, one at a time. Each state was described by five attributes – attachment, security, role, enjoyment and control – each of which took one of four categorical levels indicating the amount of the attribute that the respondent was capable of experiencing. These attributes and levels were elicited from extensive rounds of qualitative work (Grewal et al., 2006). In such a situation, when every attribute has the same number of levels, the design is said to be balanced. In best–worst, as in other DCEs, the maximum possible number of scenarios, given the design, is given by the product of the number of levels (across all attributes). Thus five attributes ($K = 5$), each with four levels ($L = 4$) means the total number of scenarios is L^K or in this case $4^5 = 1024$. Practical constraints meant that the main survey could not recruit enough respondents in order to estimate interactions. So we decided to administer two versions of an orthogonal main effects plan (OMEP), which were obtained from the following website (http://www.research.att.com/~njas/oadir/) suggested by Street et al. (2005). The original OMEP created survey version 'A', whilst the foldover of this (where levels one and four are switched, as are levels two and three) was used to make survey version 'B'. Respondents were randomised to version A ($n = 12$) or version B ($n = 18$), and each version contained 16 quality of life scenarios. For each best–worst question, respondents were asked to imagine being in that scenario and to choose the best and worst feature. Fig. 1 shows a completed quality of life scenario from the study.

By deciding that the most attractive feature of the state described in this scenario is love and friendship (attachment), which takes the level 'a lot' (level three of four), and the least attractive

Quality of life state #3		
Best		**Worst**
✓	You can have a lot of the love and friendship that you want	
	You can think about the future with only a little concern	
	You are unable to do any of the things that make you feel valued	✓
	You can have a little of the enjoyment and pleasure that you want	
	You are able to be completely independent	

Fig. 1. Best–worst scaling scenario from quality of life pilot study.

Table 1
Best–worst pairs available in quality of life scenario; pair two is referred to in the text

Pair	Best attribute level	Worst attribute level
1	You can have a lot of the love and friendship that you want	You can think about the future with only a little concern
2	You can have a lot of the love and friendship that you want	You are unable to do any of the things that make you feel valued
3	You can have a lot of the love and friendship that you want	You can have a little of the enjoyment and pleasure that you want
4	You can have a lot of the love and friendship that you want	You are able to be completely independent
5	You can think about the future with only a little concern	You are unable to do any of the things that make you feel valued
6	You can think about the future with only a little concern	You can have a little of the enjoyment and pleasure that you want
7	You can think about the future with only a little concern	You are able to be completely independent
8	You are unable to do any of the things that make you feel valued	You can have a little of the enjoyment and pleasure that you want
9	You are unable to do any of the things that make you feel valued	You are able to be completely independent
10	You can have a little of the enjoyment and pleasure that you want	You are able to be completely independent
11	You can think about the future with only a little concern	You can have a lot of the love and friendship that you want
12	You are unable to do any of the things that make you feel valued	You can have a lot of the love and friendship that you want
13	You can have a little of the enjoyment and pleasure that you want	You can have a lot of the love and friendship that you want
14	You are able to be completely independent	You can have a lot of the love and friendship that you want
15	You are unable to do any of the things that make you feel valued	You can think about the future with only a little concern
16	You can have a little of the enjoyment and pleasure that you want	You can think about the future with only a little concern
17	You are able to be completely independent	You can think about the future with only a little concern
18	You can have a little of the enjoyment and pleasure that you want	You are unable to do any of the things that make you feel valued
19	You are able to be completely independent	You are unable to do any of the things that make you feel valued
20	You are able to be completely independent	You can have a little of the enjoyment and pleasure that you want

feature is doing things that make you feel valued (role), which takes the level 'unable to do any' (level one of four), 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. Thus, in this quality of life scenario the full set of $K(K-1)=20$ possible best–worst pairs that could have been chosen is given in Table 1.

Respondents repeated this task for the remaining 15 quality of life scenarios. The total number of unique best–worst pairs possible for this design is $2\sum_{i=1}^{4}\left[L_i\sum_{k=i+1}^{5}L_k\right]=320$. The OMEP design ensured that each of these pairings appeared at least once.

Health economists are becoming increasingly familiar with best–worst tasks. However, there is little understanding of how to analyse the data, and because common statistical packages like Stata currently do not have standard commands for best–worst data, there is a need for exposition.

5. Modelling and analysis of choice data

Marley and Louviere (2005) show that best–worst choice data can be modelled in a variety of ways. 'Paired' models use the $2\sum_{i=1}^{K-1}\left[L_i\sum_{k=i+1}^{K}L_k\right]$ best–worst pairs to make inferences about the latent utility scale, whilst 'marginal' models use the $\sum_{k=1}^{K}L_k$ attribute levels. Marginal models are so-called because in a $\sum_{k=1}^{K}L_k$ by $\sum_{k=1}^{K}L_k$ table containing the choice frequencies of all the $2\sum_{i=1}^{K-1}\left[L_i\sum_{k=i+1}^{K}L_k\right]$ best–worst pairs, by aggregating choice frequencies to the margins of the table we obtain frequencies for the attribute levels. That is, the process aggregates over all pairs that include a given attribute level. The choice between paired or marginal method is largely an empirical matter because the measurement properties are the same. However, using the latter may lead to larger standard errors in estimated utility parameters due to fewer observations.

Paired and marginal models can each be analysed at a respondent level or at a sample level. Respondent level analyses require the use of maximum likelihood estimation, whilst sample level analysis can be done with simpler procedures like weighted least squares (WLS). Which estimation procedure one chooses is determined according to the need of the researcher. For example, the average utilities from the sample mean WLS may be satisfactory for some purposes, but taking differences in individuals into account would require respondent-level analysis. Therefore, there are $2 \times 2 = 4$ principal ways of analysing best–worst choice data (see Table 2).

The following sections will set out the methods and relative advantages in more detail. Results from the quality of life pilot study will be used to illustrate the methods, where relevant.

5.1. Paired model analyses (cells 1 and 2)

Paired methods of analysis model the possible best–worst pairs that can be chosen. For sample-level analysis (cell 1) the number of observations is equal to the number of unique best–worst pairs that can be estimated, which is $2\sum_{i=1}^{K-1}\left[L_i\sum_{k=i+1}^{K}L_k\right]$ in a main effects design; every one of these $2\sum_{i=1}^{K-1}\left[L_i\sum_{k=i+1}^{K}L_k\right]$ 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 5.3.

Table 2
Models for analysing best–worst scaling choice data

	Level of analysis	
	Sample level (weighted least squares)	Respondent level (multinomial logit)
Best–worst model		
Paired	1	2
Marginal	3	4

Table 3
Section of quality of life dataset for paired method of WLS analysis

Best–worst pair number (1–320) and attribute-level description			Adjusted counts and logged values		Attachment impact weight and levels				Security impact weight and levels				...
	Best	Worst	WLS weight		Love and friendship	None	Little	Lot	Concern	Lot	Some	Little	...
1	Love, f'ship_none	Concern_lot	0.058824	−2.8332	1	1	0	0	−1	−1	0	0	...
2	Love, f'ship_none	Concern_some	0.058824	−2.8332	1	1	0	0	−1	0	−1	0	...
3	Love, f'ship_none	Concern_little	0.058824	−2.8332	1	1	0	0	−1	0	0	−1	...
4	Love, f'ship_none	Concern_none	0.058824	−2.8332	1	1	0	0	−1	1	1	1	...
5	Love, f'ship_little	Concern_lot	0.058824	−2.8332	1	0	1	0	−1	−1	0	0	...
6	Love, f'ship_little	Concern_some	1.058824	0.0572	1	0	1	0	−1	0	−1	0	...
7	Love, f'ship_little	Concern_little	0.058824	−2.8332	1	0	1	0	−1	0	0	−1	...
8	Love, f'ship_little	Concern_none	0.058824	−2.8332	1	0	1	0	−1	1	1	1	...
9	Love, f'ship_lot	Concern_lot	1.058824	0.0572	1	0	0	1	−1	−1	0	0	...
10	Love, f'ship_lot	Concern_some	1.058824	0.0572	1	0	0	1	−1	0	−1	0	...
11	Love, f'ship_lot	Concern_little	0.058824	−2.8332	1	0	0	1	−1	0	0	−1	...
12	Love, f'ship_lot	Concern_none	0.058824	−2.8332	1	0	0	1	−1	1	1	1	...
13	Love, f'ship_all	Concern_lot	0.058824	−2.8332	1	−1	−1	−1	−1	−1	0	0	...
14	Love, f'ship_all	Concern_some	1.058824	0.0572	1	−1	−1	−1	−1	0	−1	0	...
15	Love, f'ship_all	Concern_little	1.058824	0.0572	1	−1	−1	−1	−1	0	0	−1	...
16	Love, f'ship_all	Concern_none	0.058824	−2.8332	1	−1	−1	−1	−1	1	1	1	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
161	Concern_lot	Love, f'ship_none	0.058824	−2.8332	−1	−1	0	0	1	1	0	0	...
162	Concern_some	Love, f'ship_none	0.058824	−2.8332	−1	−1	0	0	1	0	1	0	...
163	Concern_little	Love, f'ship_none	4.058824	1.4009	−1	−1	0	0	1	0	0	1	...
164	Concern_none	Love, f'ship_none	2.058824	0.7221	−1	−1	0	0	1	−1	−1	−1	...
165	Concern_lot	Love, f'ship_little	0.058824	−2.8332	−1	0	−1	0	1	1	0	0	...
166	Concern_some	Love, f'ship_little	0.058824	−2.8332	−1	0	−1	0	1	0	−1	0	...

The data from the pilot study were analysed using weighted least squares with each survey version (A and B) analysed separately. For each version there were 320 observations, representing the unique best–worst pairs, of which Table 3 sets out 22 of these for survey version B: observations 161–166 are included to show the sign reversals required when the best–worst ordering in the 160 pairs is reversed.

The first three columns are for information only. Column four contains the weight variable (the choice totals f adjusted to eliminate zeros — see below) whilst column five contains the dependent variable (the natural log of column four). The remaining columns comprise a subset of the variables that form the explanatory variables in the final regression to estimate the utility part-worths. The impact weights are coded differently from the level scale values. Impact weight for attribute k takes a value one for all pairs in which attribute k was picked as best. Conversely impact weight for attribute k takes a value of minus one for all pairs in which attribute k was picked as worst. For the levels for attribute k, effects coding is used with level four, 'all', being the omitted level. Effects coding makes the levels mean-centred (with the mean being the attribute impact) so the value of level four is simply minus one times the sum of the lower three level estimates. There were $\sum_{k=1}^{K} L_k - 1$ independent variables (in addition to the constant): each attribute except one has an impact variable, whilst all K attributes ($k = 1,\ldots,K$) have $L_k - 1$ effects coded scale level variables. Initially, all five impacts were included in the regression: Stata arbitrarily omits one impact variable to avoid a saturated model. The least valued impact was that of security, so in order to make inference easier it was omitted, leaving the final equation to be estimated as

$$\begin{aligned}\ln(f) = {} & \text{cnst} + \beta_1\,\text{attachment} + \beta_3\,\text{enjoyment} + \beta_4\,\text{role} + \beta_5\,\text{control} + \beta_{11}\,\text{attach\_none} \\ & + \beta_{12}\,\text{attach\_little} + \beta_{13}\,\text{attach\_lot} + \beta_{21}\,\text{secure\_none} + \beta_{22}\,\text{secure\_little} \\ & + \beta_{23}\,\text{secure\_lot} + \beta_{31}\,\text{enjoy\_none} + \beta_{32}\,\text{enjoy\_little} + \beta_{33}\,\text{enjoy\_lot} \\ & + \beta_{41}\,\text{role\_none} + \beta_{42}\,\text{role\_few} + \beta_{43}\,\text{role\_many} + \beta_{51}\,\text{control\_none} \\ & + \beta_{52}\,\text{control\_few} + \beta_{53}\,\text{control\_many} \qquad (1)\end{aligned}$$

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 sampling zeros. This adjustment is achieved by adding a small number (in this case the reciprocal of the sample size), as recommended by Goodman (1968) to enable logs to be taken. 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).

Table 4 gives the regression output when the data for questionnaire B are analysed (one respondent was unable to perform the exercise and provided no usable data).

Attachment has the largest impact, relative to security (the omitted impact). The remaining three attributes have similar impacts and of these, one is statistically significant and one approaches significance. Attachment also has a large range of level scale values, compared to other attributes.

These values represent the average utilities across the entire sample. The main survey will investigate to what extent there are differences between subgroups in terms of attribute impacts and/or level scale values. Subgroups will be defined by individual-level characteristics, such as socio-economic factors. It is within the context of limited dependent variable models that many researchers will be most familiar with the issues surrounding the effects of individual-level characteristics. 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

Table 4
Best–worst utilities (paired WLS method) for questionnaire B

	Coefficient	S.E.	*T*-ratio	$P>\|T\|$	95% Confidence interval	
_cons	−0.3067	0.0750	−4.09	0	−0.4542	−0.1592
Attribute impacts						
Attachment	0.8105	0.0803	10.09	0	0.6524	0.9685
Security	–	–	–	–	–	–
Enjoyment	0.2632	0.1010	2.61	0.01	0.0645	0.4620
Role	0.1908	0.0974	1.96	0.051	−0.0008	0.3824
Control	0.1076	0.0971	1.11	0.268	−0.0834	0.2986
Level scale values						
Attach_none	−1.9678	0.1129	−17.43	0	−2.1899	−1.7457
Attach_little	0.1694	0.1012	1.67	0.095	−0.0299	0.3686
Attach_lot	0.9053	0.0905	10	0	0.7272	1.0834
Attach_all	0.8932	–	–	–	–	–
Secure_none	−0.6123	0.1180	−5.19	0	−0.8446	−0.3801
Secure_little	−0.3761	0.1302	−2.89	0.004	−0.6324	−0.1199
Secure_lot	0.0373	0.1153	0.32	0.746	−0.1895	0.2642
Secure_all	0.9511	–	–	–	–	–
Enjoy_none	−0.8888	0.1286	−6.91	0	−1.1418	−0.6358
Enjoy_little	−0.3367	0.1632	−2.06	0.04	−0.6578	−0.0155
Enjoy_lot	0.6561	0.1493	4.39	0	0.3622	0.9499
Enjoy_all	0.5694	–	–	–	–	–
Role_none	−0.8956	0.1239	−7.23	0	−1.1394	−0.6518
Role_few	−0.0277	0.1532	−0.18	0.857	−0.3293	0.2738
Role_many	0.4435	0.1363	3.25	0.001	0.1753	0.7117
Role_all	0.4798	–	–	–	–	–
Control_none	−0.8085	0.1122	−7.2	0	−1.0294	−0.5876
Control_few	0.0835	0.1596	0.52	0.601	−0.2307	0.3976
Control_many	0.2780	0.1376	2.02	0.044	0.0071	0.5488
Control_all	0.4471	–	–	–	–	–

Number of observations = 320; adjusted $R^2 = 0.6299$.

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: for example, the main effect of 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.

Maximum likelihood estimation or some other suitable methods (e.g. minimum chi-squared) are required in order to make inference at levels below the sample level. The format of the data may vary according to the statistical package used, but in Stata the clogit command applied to an expanded dataset is the most flexible method. For example, each quality of life scenario for each person would have $K(K-1) = 20$ observations: in terms of independent variables, each of the 20 possible pairs would be coded as for WLS, but the dependent variable would be an indicator variable, taking a value one for the pair chosen and zero otherwise.[1]

[1] The group(*V*_id) tag is added to the regression: *V*_id should be an indicator variable taking a different value for every scenario for every respondent (17 respondents by 16 scenarios gives 272 values in this case).

The effects of respondent-level covariates upon preferences can be estimated easily in such a model. For instance, by effects coding the sex variable, the impact and level variables still can be interpreted as averages across the entire sample. However, the *female_impact* and *female_level* interactions represent the additional utility that women experience for a given attribute and levels; the estimate for males is simply a sign change.

5.2. Marginal model analyses (cells three and four)

Marginal methods of analysis model the possible attribute levels that can be chosen. Thus they aggregate the data over best–worst pairs to estimate the $\sum_{k=1}^{K} L_k$ attribute level utilities using a model that, while more parsimonious (and probably easier to code in standard statistical packages), may suffer from larger standard errors as a result. There are a total of $2\sum_{k=1}^{K} L_k$ 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 this time with $\sum_{k=1}^{K} L_k$ independent variables (in addition to the constant); however, as before, there are $K-1$ impact variables and $L_k - 1$ effects coded scale level variables for each attribute, there is an additional best–worst indicator variable taking the value one for all observations where best–worst pairs are ordered one way, and minus one when they are in reverse order. The best and worst constants will differ only if the distribution of best frequencies differs significantly from the distribution of worst frequencies.

The least valued impact was that of security, which was omitted to make inference easier; hence, the final equation estimated was

$$\begin{aligned}\ln(g) = {} & \text{cnst} + \text{bw\_indic} + \beta_1\,\text{attachment} + \beta_3\,\text{enjoyment} + \beta_4\,\text{role} + \beta_5\,\text{control} \\ & + \beta_{11}\,\text{attach\_none} + \beta_{12}\,\text{attach\_little} + \beta_{13}\,\text{attach\_lot} + \beta_{21}\,\text{secure\_none} \\ & + \beta_{22}\,\text{secure\_little} + \beta_{23}\,\text{secure\_lot} + \beta_{31}\,\text{enjoy\_none} + \beta_{32}\,\text{enjoy\_little} \\ & + \beta_{33}\,\text{enjoy\_lot} + \beta_{41}\,\text{role\_none} + \beta_{42}\,\text{role\_few} + \beta_{43}\,\text{role\_many} \\ & + \beta_{51}\,\text{control\_none} + \beta_{52}\,\text{control\_few} + \beta_{53}\,\text{control\_many} \end{aligned} \tag{2}$$

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 paired method), with an adjustment to eliminate zeroes, as in Section 5.1. Thus there are $\sum_{k=1}^{K} L_k = 20$ best totals and $\sum_{k=1}^{K} L_k = 20$ worst totals. Table 5 sets out a subset of the data used in the final regression.

The first three columns are for information only. 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). The remaining columns comprise the independent variables to estimate the utility part-worths. Table 6 displays the regression results, which are consistent with those from the paired method.

Attachment has the largest impact relative to security and the three remaining attributes are of similar relative magnitudes to those in Table 4. All three are insignificant, perhaps reflecting the smaller sample size (40 rather than 320). The main survey will investigate differences (if any) between the methods in more detail in order to provide more guidance to researchers.

Marginal models also can be analysed in a multinomial framework. As before, we use the *clogit* command in Stata to illustrate. Each scenario for each person contributes ten observations: five attribute levels can be picked as best, and five as worst. Coding of the independent variables is the same as for WLS regression, and a new dependent variable should be coded in the same

Table 5
Section of quality of life dataset for marginal method WLS analysis

Best–worst attribute level number (1–40) and attribute-level description			Adjusted counts and logged values		B–W indicator	Attachment impact weight and levels				Security impact weight and levels				...
	Best	Worst	WLS weight			Love and friendship	None	Little	Lot	Concern	Lot	Some	Little	...
1	Love, f'ship_none	–	5.0588	1.6211	1	1	1	0	0	0	0	0	0	...
2	Love, f'ship_little	–	25.0588	3.2212	1	1	0	1	0	0	0	0	0	...
3	Love, f'ship_lot	–	48.0588	3.8724	1	1	0	0	1	0	0	0	0	...
4	Love, f'ship_all	–	53.0588	3.9714	1	1	−1	−1	−1	0	0	0	0	...
5	Concern_lot	–	0.0588	−2.8332	1	0	0	0	0	1	1	0	0	...
6	Concern_some	–	1.0588	0.0572	1	0	0	0	0	1	0	1	0	...
7	Concern_little	–	9.0588	2.2037	1	0	0	0	0	1	0	0	1	...
8	Concern_none	–	19.0588	2.9475	1	0	0	0	0	1	−1	−1	−1	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
21	–	Love, f'ship_none	46.0588	3.8299	−1	−1	−1	0	0	0	0	0	0	...
22	–	Love, f'ship_little	4.0588	1.4009	−1	−1	0	−1	0	0	0	0	0	...
23	–	Love, f'ship_lot	0.0588	−2.8332	−1	−1	0	0	−1	0	0	0	0	...
24	–	Love, f'ship_all	0.0588	−2.8332	−1	−1	1	1	1	0	0	0	0	...
25	–	Concern_lot	28.0588	3.3343	−1	0	0	0	0	−1	−1	0	0	...
26	–	Concern_some	17.0588	2.8367	−1	0	0	0	0	−1	0	−1	0	...

Table 6
Best–worst utilities (marginal WLS method) for questionnaire B

	Coefficient	S.E.	T-ratio	$P>\|T\|$	95% Confidence interval	
_cons	2.1911	0.0913	23.99	0	1.9999	2.3822
bwindic	−0.3728	0.1347	−2.77	0.012	−0.6548	−0.0908
Attribute impacts						
Attachment	1.1060	0.1795	6.16	0	0.7304	1.4816
Security	–	–	–	–	–	–
Enjoyment	0.3432	0.2067	1.66	0.113	−0.0895	0.7759
Role	0.3200	0.1988	1.61	0.124	−0.0960	0.7361
Control	0.1340	0.1989	0.67	0.509	−0.2823	0.5502
Level scale values						
Attach_none	−2.2663	0.1951	−11.61	0	−2.6747	−1.8578
Attach_little	0.2634	0.1909	1.38	0.184	−0.1361	0.6630
Attach_lot	0.9522	0.1661	5.73	0	0.6045	1.2999
Attach_all	1.0507	–	–	–	–	–
Secure_none	−0.7786	0.2189	−3.56	0.002	−1.2368	−0.3203
Secure_little	−0.3598	0.2494	−1.44	0.165	−0.8818	0.1622
Secure_lot	0.0032	0.2194	0.01	0.988	−0.4560	0.4624
Secure_all	1.1351	–	–	–	–	–
Enjoy_none	−1.1218	0.2467	−4.55	0	−1.6382	−0.6053
Enjoy_little	−0.1863	0.3237	−0.58	0.572	−0.8638	0.4912
Enjoy_lot	0.5679	0.2896	1.96	0.065	−0.0382	1.1741
Enjoy_all	0.7401	–	–	–	–	–
Role_none	−1.1992	0.2340	−5.13	0	−1.6889	−0.7095
Role_few	0.0998	0.3027	0.33	0.745	−0.5339	0.7334
Role_many	0.5288	0.2632	2.01	0.059	−0.0221	1.0796
Role_all	0.5707	–	–	–	–	–
Control_none	−1.2768	0.2157	−5.92	0	−1.7284	−0.8253
Control_few	0.1551	0.3152	0.49	0.628	−0.5046	0.8147
Control_many	0.4585	0.2680	1.71	0.103	−0.1025	1.0194
Control_all	0.6633	–	–	–	–	–

Number of observations = 40; adjusted $R^2 = 0.8103$.

way as for the paired analysis, that is, this time it is an indicator variable for which attribute level was chosen as best or worst.[2]

5.3. 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 that represents the number of times each best–worst pair (or attribute level if the marginal 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

[2] The group(V_id) tag is again added to the regression. In Stata a warning message is generated since there are two non-zero observations per choice set (a best and a worst).

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 $2\sum_{i=1}^{K-1}\left[L_i\sum_{k=i+1}^{K}L_k\right]$ pair (or $2\sum_{k=1}^{K}L_k$ 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 attribute 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 K-th 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 retained.

6. Other methodological issues

Best–worst scaling can give additional information to that obtained in traditional DCEs. In particular, unlike traditional DCEs, it is the utility of a single level of one attribute that acts as a benchmark, not an entire scenario. However, (at least as applied to date), like pairwise comparisons, it is a conditional demand model: the best–worst task itself gives no information on the attractiveness of the scenario relative to the respondent's current position. Administering the best–worst task as part of a wider choice experiment incorporating an opt-out option ("Would you accept this scenario?") therefore may be useful. The task to synthesise the results from the two different discrete choice experiment tasks then becomes one of adjusting for different levels of variability in responses, a more general issue that has not been sufficiently well recognised in applied economics.

6.1. 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 due to various combinations of several sources of variation, such as respondent inability to fully recognise systematic differences in utility and measurement error (to name only two). 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 et al., 1999). An exposition of the issues can be found in Swait and Louviere (1993), but, briefly, all parameter estimates from choice models based on random utility theory are confounded with an unknown scale factor (which is entirely different from the 'scales' referred to in the impact weights and level scales conceptualisation of best–worst). 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 1 make more mistakes evaluating the scenarios than those in group 2 — that is, the variance of the random utility component is larger in group 1 than group 2. Utility estimates from a model estimated from group 1 will appear to be closer to zero than one estimated from group 2, 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 may be infeasible. Hensher et al. and Swait and Louviere have provided guidance on combining data from multiple data sources: the methods utilise the researcher's ability to estimate the size of the random component in one data set relative to another. The synthesis of a binary choice (accept/reject) question with a best–worst scaling question therefore is an extension of this.

A second issue arising from the random utility component concerns the treatment of respondent heterogeneity. 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 DCEs might necessitate blocked designs and the need to make distributional assumptions for inferences. However, not only does this particular focus on preference heterogeneity ignore other factors that might underlie 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. Hence, it is a partial solution at best, with little evidence to suggest that such a simplistic treatment of heterogeneity is true empirically (Louviere, 2001). Furthermore, 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 logical to exploit the power of best–worst to make individual-level inferences than to attempt to introduce more complex random effects models. Indeed, work has begun to utilise the power of best–worst scaling to model individual-level utility functions that do not require distributional assumptions about preference parameters (Louviere et al., 2004).

6.2. Sample size issues

Given that best–worst scaling represents a different choice task with different outcomes to that in traditional DCEs, 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 known from the statistical design.

In the current absence of guidelines 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.

7. Conclusion

Best–worst scaling asks respondents to perform a different task than most DCEs to date. Hence, it provides additional insights over traditional DCEs that may be attractive to health services researchers. For example, it allows estimation of part-worth utilities relative to a single attribute level, and so allows the impacts of attributes to be compared, which can be valuable in evaluating many aspects of service provision. When researchers are interested in comparisons of marginal

changes in attributes, pairwise comparisons can be used to provide statistically valid estimates; however, because the choice tasks inherent in a best–worst exercise are potentially easier, 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.

Acknowledgements

We acknowledge Dr. Jackie Brown and the helpful comments of both the referees and Professor Mandy Ryan who discussed an earlier version of this paper at the Health Economists' Study Group meeting, January 2005, Oxford. This work was supported by the MRC Health Services Research Collaboration.

References

Bech, M., Gyrd-Hansen, D., 2005. Effects coding in discrete choice experiments. Health Economics 14, 1079–1083.

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

Bradley, M., Daly, A., 1994. Use of the logit scaling approach to test for rank-order and fatigue effects in stated preference data. Transportation 21, 167–184.

Farrar, S., Ryan, M., Ross, D., Ludbrook, A., 2000. Using discrete choice modelling in priority setting: an application to clinical service developments. Social Science and Medicine 50, 63–75.

Finn, A., Louviere, J.J., 1992. Determining the appropriate response to evidence of public concern: the case of food safety. Journal of Public Policy and Marketing 11, 12–25.

Goodman, L.A., 1968. The analysis of cross-classified data: Independence, quasi-independence, and interactions in contingency tables with or without missing entries. Journal of the American Statistical Association 63, 1091–1131.

Grewal, I., Lewis, J., Flynn, T.N., et al., 2006. Developing attributes for a generic quality of life measure for older people: Preferences or capabilities? Social Science and Medicine 62, 1891–1901.

Hausman, J.A., Ruud, P.A., 1987. Specifying and testing economic models for rank-ordered data. Journal of Econometrics 34, 83–104.

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

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.

Louviere, J.J., 2001. What if consumer experiments impact variances as well as means: response variability as a behavioural phenomenon. Journal of Consumer Research 28, 506–511.

Louviere, J.J., Burgess, L., Street, D.J., 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 04-005. Centre for the Study of Choice, University of Technology, Sydney.

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

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

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.

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

McIntosh, E., Louviere, J.J., 2002. Separating weight and scale value: an exploration of best-attribute scaling in health economics, Health Economists' Study Group Meeting, Brunel University.

Street, D.J., Burgess, L., Louviere, J.J., 2005. Quick and easy choice sets: constructing optimal and nearly optimal stated choice experiments. International Journal of Research in Marketing 22, 459–470.

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.

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

Thurstone, L.L., 1927. A law of comparative judgment. Psychological Review 34, 273–286.

Vick, S., Scott, A., 1998. Agency in health care. Examining patients' preferences for attributes of the doctor–patient relationship. Journal of Health Economics 17, 587–605.