SEPARATING WEIGHT AND SCALE VALUE: AN EXPLORATION OF BEST-ATTRIBUTE SCALING IN HEALTH ECONOMICS

Emma McIntosh[1] and Jordan Louviere[2]

[1] **Research fellow, HERC, University of Oxford, Old Road, Oxford, UK OX3 7LF**
emma.mcintosh@ihs.ox.ac.uk
Tel: 01865 – 226634 Fax: 01865 – 226842

[2] **Professor, CHERE, University of Technology, Sydney, 1-59 Quay St, Building 5, Level 2, Block C, Room C202, Haymarket NSW 2007, Australia**

Paper Presented to the Health Economists Study Group Meeting

Brunel University, July 3rd-5th 2002

WORK IN PROGRESS: Please do not quote without permission from the authors

Acknowledgement: This study obtained funding from the UK NHS R&D programme. The views in this paper are those of the authors.

ABSTRACT

Rationale

One topic often overlooked by discrete choice practitioners is the inability of traditional conjoint methods to separate weight (the impact of information) and scale value (the location of information on the underlying utility scale). As a consequence, the units of measurement and the scale origins of part worth utility estimates for each attribute level *differ* for each attribute, so they cannot be compared without transforming to common units (e.g. calculating the compensating variation with respect to a cost attribute). Traditional conjoint tasks require subjects to evaluate product/service profiles holistically; hence yield no information on choice *among* attributes, only information on choice of levels *within* attributes. Thus, one needs additional information provided by choices *among* attributes to separate attributes weights from attribute scale values and produce a common underlying scale.

Best attribute scaling (BSc) is a relatively new approach for separating attribute weights and scale values. The key behavioural assumption of the BSc approach is that attribute *levels* chosen by subjects reveal their respective weights and positions on the underlying (latent) subjective scale. The resulting latent scale is an absolute probability scale, which provides a know scale origin (zero) and unit of measurement (one) in contrast to current DCE methods that produce relatively arbitrary scales. Although adherents of conjoint methods claim (with little empirical support) that rating-based conjoint methods produce interval scaled measures (arbitrary zero, meaningful differences like the centigrade temperature scale), this "claim" rests on very strong assumptions about the ability of humans to produce cardinal measures that is at best controversial. All else equal, methods that make less demanding measurement assumptions should be preferred to those requiring more demanding ones, which is a clear advantage of BSc because it relies only on the ability of subjects to make discrete choices. Another advantage is that BSc is consistent with random utility theory, and hence can provide valuation measures that are consistent with economic theory.

Empirical study

The empirical study involves a random sample of members of the general public in Scotland, UK who were surveyed on their preferences for dental treatment for third molar complaints. Advantages of revealing the underlying latent scale of preferences elicited through discrete choice experiments are explored within this paper. Further, more general advantages of the use of this approach in health economics will also be discussed. The extent to which the interpretation of choice experiment results is enhanced by BSc will also be discussed.

BEST-ATTRIBUTE SCALING ~ AN INTRODUCTION

An interesting and important, but little appreciated aspect of all conjoint, choice experiment and axiomatic utility theory related methods for eliciting preferences and tradeoffs and estimating willingness to pay is that the measurement scales derived from applying these techniques are incommensurate. That is, it is not possible to carry out inter-dimensional 'utility' comparisons between attributes because the measurement scales that one derives are unique to each attribute. More specifically, the utility estimates derived from conjoint and choice experiments lie on interval scales in which differences between levels within an attribute are meaningful, but differences between utility estimates of different attributes are not meaningful. The basic reason for this is that the utility estimates for each attribute have a unique scale origin and unit of measurement.

Consequently, one must use a numeraire like cost and derive marginal rates of substitution to transfer all attributes onto a common numeraire. Only this way is it possible to make legitimate inter-dimensional utility comparisons. Without going into the theoretical complexities of this issue, it is possible to illustrate the problem with reference to the following examples:

1. A subject is asked to evaluate 8 health-related product or service descriptions derived from a factorial design in which two levels (each) of risks, benefits and costs are systematically combined/varied, as illustrated below. The subject is asked to "state" whether they would switch to each of the 8 options (one-at-a-time), or stay with what is presently available, see Table 1.

Table 1 Yes/No frequency count

Option/Scenario	Levels of Benefits	Levels of Risks	Levels of Costs	Response (yes/no)
1	1	1	1	Yes
2	1	1	2	No
3	1	2	1	No
4	1	2	2	No
5	2	1	1	Yes
6	2	1	2	Yes
7	2	2	1	No
8	2	2	2	No

The subject's responses in this example refer to an entire combination of attribute levels, and their preferences are expressed relative to the status quo or what they currently have available. The subject does not discriminate, compare or choose among the attribute levels themselves, only combinations of levels. If the utility function is additive (or any multi-linear form), and we calculate the average utility for each attribute level, holding all else constant, it is easy to show that we obtain the outcome that the utility of the levels of a particular attribute is equal to the true but unobserved utility up to a positive linear transformation. A different positive linear transformation applies (in principle) to each attribute; hence, the estimated utility scales all differ.

2. A subject is asked to evaluate 8 pairs of health-related product or service descriptions derived from a factorial design in which two levels (each) of risks, benefits and costs are systematically combined/varied for both pair options, as illustrated in Table 2 below. The subject is asked to "state" which one of the two options they would choose in each of the 8 choice sets or scenarios (one-choice set-at-a-time); optionally, they also could be given a third option of choosing to stay with what is presently available, which is a constant in each

choice set. We also ignore design issues associated with dominated/dominated options in choice sets for the sake of example simplicity.

Table 2 Yes/No choice frequency count

Choice Set/Scenario	Option A			Option B			Choice (A or B)
	Benefits Levels	Risks Levels	Costs Levels	Benefits Levels	Risks Levels	Costs Levels	
1	1	1	1	2	2	2	A
2	1	1	2	2	2	1	B
3	1	2	1	2	1	2	A
4	1	2	2	2	1	1	A
5	2	1	1	1	2	2	B
6	2	1	2	1	2	1	B
7	2	2	1	1	1	2	A
8	2	2	2	1	1	1	A

As in the case of one scenario at a time, the choice options in the above example constitute a combination of attribute levels. Thus, subjects choose between combinations of attribute levels, they do not choose attributes or attribute levels per se. In the above example, the utility function is expressed as a function of differences (or contrasts) in the attribute levels:

$$U_{ab} = \Sigma_k \beta_k (X_{ka} - X_{kb}) + \varepsilon_{ab}, \qquad (1)$$

Where U_{ab} is the unobserved ("generic") utility associated with differences in the attributes of options A and B, $(X_{ka}-X_{kb})$ represents differences (or statistical contrasts) in the k-th (K=1, 2, 3) attribute levels of options A and B (the so-called "systematic component" of utility), and ε_{ab}, is the random component of utility associated with differences in options A and B.

As before, if we average the utility expression represented by Equation (1) over each attribute to estimate the utility of each attribute level, we obtain the result that the utility estimates (the measures of utility for each level, also called "part-worths" or "scale values") are related to the true but unknown utilities of each attribute level up to a positive linear transformation, but the linear transformation is different for each attribute. Thus, once again we have the situation that the utility estimates of each attribute are measured on an interval scale in which differences in utility estimates for levels within an attribute are meaningful, but inter-attribute level comparisons of utility estimates are not meaningful without rescaling by a common numeraire like cost.

3. Consider adding the following elicitation task to example 1: As in example 1, a subject is asked to evaluate 8 health-related product or service descriptions derived from a factorial design in which two levels (each) of risks, benefits and costs are systematically

combined/varied, as illustrated below. The subject is asked to "state" whether they would switch to each of the 8 options (one-at-a-time), or stay with what is presently available. Now, however, we *also ask the subject two additional questions, namely which of the attribute levels in the scenario are, respectively, the most and least attractive* (or some other relevant subjective dimension) aspects of the scenario. We add this new feature to the task in Table 3.

Table 3 Most/least frequency count

Option/Scenario	Attributes/Levels			Original Response (yes/no)	Attractiveness	
	Levels of Benefits	Levels of Risks	Levels of Costs		Most	Least
1	1	1	1	Yes	R	B
2	1	1	2	No	R	C
3	1	2	1	No	C	R
4	1	2	2	No	B	R
5	2	1	1	Yes	R	C
6	2	1	2	Yes	R	C
7	2	2	1	No	B	R
8	2	2	2	No	B	R

B=Benefits, R=Risks, C=Costs

In Table 3, not only does the subject evaluate the package or bundle of attributes, but the subject also provides information about preferences for attribute levels because they now choose *among* competing attribute levels as well as competing options. Thus, in contrast to the previous methods, with sufficient response data containing this added information one can estimate the utilities of all attributes and levels *on a common scale*. In fact, one simple way to do that is an absolute or probability scale that can be derived simply by calculating the absolute and relative choice frequencies, as shown in Table 4 below.

Table 4 Absolute and relative frequency counts

Attribute levels	Yes count (No) for Benefits	Yes count (No) for Risks	Yes count (No) for Costs	Most count (Least) for B	Most count (Least) for R	Most count (Least) for C
1	1 (3)	3 (1)	2 (2)	1 (1)	4 (0)	1 (3)
2	2 (2)	0 (4)	1 (3)	2 (0)	0 (4)	0 (0)

Emphasizing that this is illustrative, not definitive, it hopefully is obvious that if the sample size (number of responses/scenarios) were sufficiently large, one can calculate meaningful absolute or relative frequency estimates from the above data. To illustrate, the simple example allows us to calculate the following estimates from the data of this one subject:

From the Yes/No response data:

- Benefits = 1 (1/3 = 0.33), Benefits = 2 (2/2 = 1.0)
- Risks = 1 (3/1 = 3.0); Risks = 2 (0/4 = 0)
- Costs = 1 (1/1 = 1.0); Costs = 2 (1/3 = 0.33)

From the Most response data:

- Benefits = 1 (1/8 = 0.125); Benefits = 2 (2/8 = 0.250)
- Risks = 1 (4/8 = 0.500); Risks = 2 (0/8 = 0.000)
- Costs = 1 (1/8 = 0.125); Costs = 2 (0/8 = 0.000)

From the Least response data:

- Benefits = 1 (1/8 = 0.125); Benefits = 2 (0/8 = 0.000)
- Risks = 1 (0/8 = 0.000); Risks = 2 (4/8 = 0.500)
- Costs = 1 (3/8 = 0.375); Costs = 2 (0/8 = 0.000)

From both the Most and Least response data:

- Benefits = 1 or 2 (4/16 = 0.250)
- Risks = 1 or 2 (8/16 = 0.500)
- Costs = 1 or 2 (6/16 = 0.375)

The above results tell us that Risks were most likely the key attribute because the levels of Risks received the highest number of combined most and least choices, followed by Costs and then Benefits. We also can see that the Most choices suggest a large difference in the utility of the two levels of Risks (0.5 vs 0.0), compared with the levels of Costs and Benefits, which exhibited the same difference in choices (0.125) between levels. The Least choices tell a similar story, although they suggest that there is a larger difference in the levels of Costs than there is for Benefits.

Figures 1 and 2 below are graphs of 'Most versus Least' choices for the six attribute levels for both the above empirical example and the theoretically perfect (ie, deterministic) case.

Figure 1 Most v's least choices - empirical case

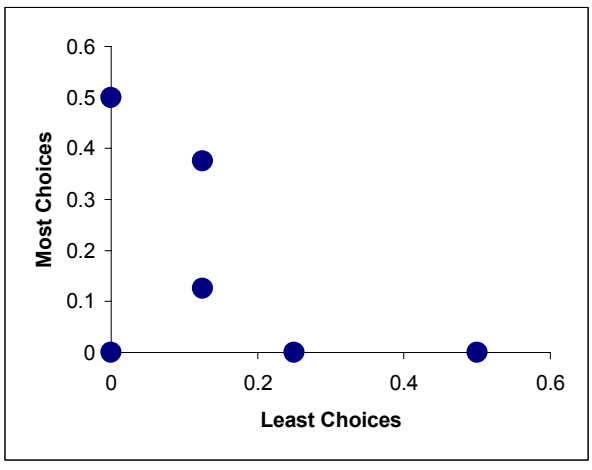
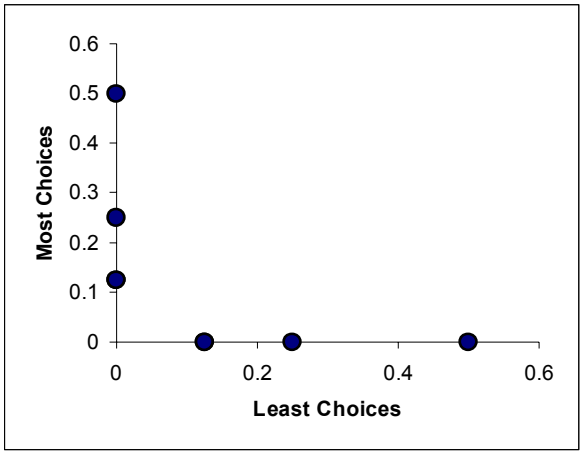


Figure 2 Most v's least choices - theoretical case



Theoretically, if the subject uses a fixed ranking of the attribute levels and always chooses consistently in accordance with that ranking, we would observe the graph labeled "Most VS Least Choices – Theoretical". That is, most and least choices are mirror images. Note how the empirical illustration seems to be a reasonable first approximation to this expected relation. If this expected relationship between most and least choices holds true to a reasonable first approximation, then we can combine both sets of choices to obtain an improved set of utility estimates {Luce RD 1959 225 /id}. We can then combine these choices by stacking the choice frequencies and coding the design matrix such that the attribute levels have opposite signs for most and least choices. The necessary assumptions for this approach to be consistent with random utility theory are as follows:

1. Each of the levels of the K attribute are discrete choice options;

2. The utility of any particular attribute level cannot be directly observed, but can be inferred from the choices of the attribute levels in a most-least (best-worst) task under the assumption that

   $U_i = V_i + \varepsilon_i$,

   where I is the total number of levels observed for all attributes; U_i is the inherently unobservable utility of the i-th (i = 1,…, I) attribute level; V_i is the systematic or observable utility component of the i-th attribute level; and ε_i is the random utility component associated with the i-th attribute level.

3. ε_i has some distribution, which we initially assume is extreme value type I.

4. Most and Least choices are rank-order consistent, mirror images, and hence can be stacked as above to "double" the amount of useful information for estimation purposes.

The above assumptions lead to formulating the problem as a simple multinomial logit (MNL) model with two sources of data that can be combined as indicated above. (Louviere J, Hensher DA, & Swait J 2000) (chapters 8 and 13) discuss how to combine sources of preference data and estimate

utility parameters from the pooled data. That is, we wish to estimate the probability that a particular attribute level is the most attractive (preferred, important, whatever) given the data, or

$P(i|C) = P[(V_i + \varepsilon_i) > (V_j + \varepsilon_j)]$, for all j other levels competing with i in scenario C.

Where all terms were previously defined, except for C that represents choice sets, which in this case are particular attribute combinations or scenarios. If the random components are distributed as extreme value type I, it is well-known that the resulting model will be MNL:

$$P(i|C) = \exp(V_i) / \sum_j \exp(V_j) \quad (2)$$

In the present case, the choice options are attribute levels. It is important to note that the parameters of this model can be estimated from relative or absolute frequency count data in a variety of ways as discussed by Louviere and Woodworth (Louviere J J & Woodworth G 1983) and many subsequent authors. Louviere and Woodworth (Louviere J J & Woodworth G1983) discuss the use of weighted least-squares, and we illustrate the use of this approach in this paper.

As Louviere and Woodworth (Louviere J J & Woodworth G1983) note, equation (2) can be linearised by noting that the denominator is a constant for all choice options in any particular choice set. Thus, we can rewrite equation (2) as follows:

$$P(i|C) = \exp(V_i) / \exp(Q_{1c}), \quad (2a)$$

where Q_{1c} is the denominator in choice set C. Taking natural logs of both sides yields

$$\text{Log}_e[P(i|C)] = V_i - Q_{1c} \quad (2b)$$

Equation (2b) provides a simple proof that the natural logarithms of the probabilities, or equivalently, their empirical realizations, the absolute or relative choice frequencies in each choice set, are estimates of the desired utilities of each choice option up to a linear transformation. Thus, the natural logarithms of the choice frequencies of each option (ie, each attribute level) estimate the desired utilities on an interval scale.

A fixed utility form of the model based on Luce's (1959) Choice Axiom (Luce RD 1959) also can be derived using only the basic frequency counts. Luce's model is

$$P(i|C) = V_i / \sum_j V_j, \quad (3)$$

Fixed utility models assume that the probabilities are random, not the utilities. However, it is hopefully obvious that in the MNL case, these two models are structurally equivalent. As above, we note that the denominator of equation (3) is constant in each choice set, C. Thus, we can rewrite equation (3) as follows:

$$P(i|C) = V_i / Q_{2c}, \quad (3a)$$

where Q_{2c} is the denominator in choice set C. In this case, we can immediately see without further proof that the choice frequencies of each attribute level provide estimates of the utilities of each attribute on a ratio scale. Specifically, we can simply sum the total frequency of choices of each attribute level across all choice sets, adjust for differences in the appearance/availability of the levels due to the design, and these empirical frequencies provide direct estimates of the utilities on a ratio scale if the Luce model is approximately correct. By "availability" we mean that if the number of attribute levels differ for different attributes each level will occur (be available to be chosen) a different number of times. Thus, one has to adjust for difference in availability between attributes, which in the case of 2, 4 and 8 level attributes would simply be to adjust the resulting choice frequencies by multiplication or division by a constant ratio. For example, to keep the number of observations constant, choices of two-level attributes could be divided by two, choices of four level attributes would be unchanged, and choices of eight level attributes would be multiplied by two.

EMPIRICAL APPLICATION OF BSC ~ PREFERENCES FOR THIRD MOLAR MANAGEMENT

Background

This paper reports the results of a BSc stated choice survey to the general population to elicit preferences/values for attributes of third molar care. The attributes and levels are concerned with both extraction of third molar teeth and conservative management.

Study Design

The following outlines the main stages in the BSc study.

Stage 1 Establishing the attributes

The attributes were established using the most recent literature on the management of third molars. This literature includes a report of a Working Party convened by the Faculty of Dental Surgery (The Royal College of Surgeons of England1997). Brickley et al. (1995) (Brickley M, Armstrong R, Shepherd J, & Kay E1995) also provides information on the probabilities of complications when the conservative approach is taken.

Stage 2 Assigning levels to the attributes

Levels of the attributes are given in terms of the range of probabilities associated with attributes of third molar care, these are also provided in the literature.

Stage 3 Devising a statistically efficient design

Following stages 1 and 2 above, respondents are then presented with hypothetical scenarios which combine different levels of attributes. The number of possible scenarios was reduced to a manageable level using an orthogonal main effects design, thus ensuring the absence of multicollinearity. In BSc designs the full orthogonal matrix is presented to individuals hence criteria such as minimal overlap and utility balance are not a consideration.

Stage 4 Presenting scenarios using the best-worse format

Once the design criteria from stage 3 has been fulfilled the scenarios are then replicated within a questionnaire format with an introduction and example. In the BSc format, each individual scenario from the orthogonal matrix is presented to individuals requesting them to state whether they would accept the scenario as well as to identify the best and worse levels within each scenario.

Stage 5 Eliciting preferences & best/worse levels for each scenario

Respondents are asked to evaluate one replication of a regular, orthogonal, main effects fraction of a Q^M factorial in which each of the L attribute levels appears exactly 1/Q of the time. With the BSc approach respondents are then asked to choose one attribute level in each of the profiles that is the best (most attractive, salient, desirable e.t.c.) and one which is the worst (least attractive, least desirable e.t.c.) of all the profile attribute levels. Thus the BSc approach requires one extra decision to be made about each task than a standard discrete choice approach. 400 questionnaires were sent to a random sample of the general population in Dundee, Scotland.

Stage 6 Test Luce's underlying model assumption re.mirror image of most/least choices

Theoretically, if a subject uses a fixed ranking of the attribute levels and always chooses consistently in accordance with that ranking, 'most' and 'least' choices would graph as mirror images. If this expected relationship between most and least choices holds true to a reasonable first approximation, then we can combine both sets of choices to obtain a set of utility estimates.

Stage 7 Calculate the observed choice frequencies for most and least choices for each attribute level

Each attribute is broken down into its individual levels and dummy variables created for each level. Frequencies of most and least attractive for each attribute levels are then estimated and this creates

two new variables per attribute level 'most attractive' and 'least attractive'. It is these new variables which become the dependent variables for the count model (unlike yes/no responses in probit/logit estimation of traditional discrete choice models).

Stage 8 Based on outcome of Stage 6, stack the data for estimation purposes (adjust for 'availability' bias by re-weighting choices as required)

Once the most and least frequencies have been graphed to show that they satisfy Luce's assumption about mirror relations to a close first approximation {Luce RD 1959 225 /id}, the data then have a theoretical justification for being stacked estimation purposes. Both sets of choices are then combined to obtain an improved set of utility estimates. The choices are combined by stacking the choice frequencies and coding the design matrix such that the attribute levels have opposite signs for most and least choices.

Stage 9 Analyse the data and identify scale and weight values

BSc response data can be used to estimate non-parametric and parametric individual or aggregate-level weights and scale values. In order to test hypotheses about weights and scale values an error theory is required, therefore the parametric approach is preferred. With an error theory, the multinomial logit model is traditionally used to analyse the BSc data however with small samples sizes weighted least squares is the most appropriate method. The parameters of this model can be estimated from relative or absolute frequency count data in a variety of ways as discussed by Louviere and Woodworth (Louviere J J & Woodworth G1983). Weight and scale values then are estimated and placed on their underlying latent scale for comparability of all attributes and levels on a common scale.

Results

The attributes and levels were obtained using the most recent literature on the outcomes of third molar extraction and conservative management of third molars, see Table 5.

Table 5 – Attributes and levels for discrete choice experiment

Attribute	Levels	Reference
Days of severe pain where painkillers are taken	0, 1,3 (days)	(The Royal College of Surgeons of England1997)
Bouts of mild dental pain lasting up to 2 days	Never Once a week Once a month	(Brickley M, Armstrong R, Shepherd J, & Kay E1995)
Prolonged bleeding	0, 2.5%, 5%	(The Royal College of Surgeons of England1997)
Sensory nerve damage	0, 5%, 10%	Effectiveness matters
Probability of crowding	0%, 15%, 30%	(Brickley M, Armstrong R, Shepherd J, & Kay E1995)
Pericoronitis episodes (Painful inflammation of the gum)	0,2,5	(Brickley M, Armstrong R, Shepherd J, & Kay E1995)
Cost (£)	£0, £5, £15, £20, £25, £30	Current NHS charges

The experimental design was based on a well-known orthogonal main effects design, which is a subset of the 6 x 3^5 factorial. That is, all main effects can be estimated independently of one another under the assumption that the underlying utility process is strictly additive. In other words, if one uses this design, one must know that the underlying utility process is strictly additive or be willing to assume that it is, which is equivalent to assuming that all attribute interactions are statistically insignificant. The attribute levels in Table 5 above were presented using 'user friendly' terms which people could understand. The choice scenarios were preceded by detailed descriptions of the attributes and levels as well as an example An example of the resulting choice sets presented to respondents within the postal questionnaire based upon the attributes and levels in Table 5 is given in Figure 3.

Figure 3 Example of a BSc question

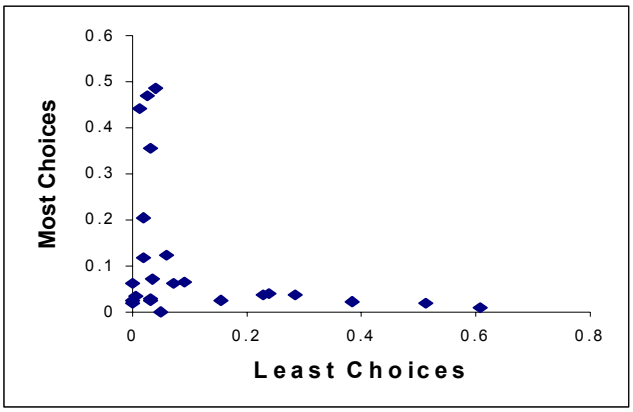
	Description	**Which aspects are the most (M) and least (L) attractive?**
Number of days of severe pain & swelling	**1 day**	
Episodes of mild pain	**Once a week**	
Chance of prolonged bleeding	**5%**	
Chance of nerve damage	**0%**	
Chance of crowding of teeth	**15%**	
Number of episodes of painful inflammation of the gums	**2 episodes**	
Total cost to you (£)	**£20**	

Would you consider this treatment option?

Yes ❒ No ❒

The response rate for the general population survey was 54/400 (14%). The observed choice frequencies for most and least choices for each attribute level were obtained and the data stacked for estimation purposes in the manner illustrated in the preceding section. Figure 4 below graphs the most and least frequencies to show that they satisfy the assumption about mirror relations to a close first approximation, justifying stacking the data for estimation purposes.

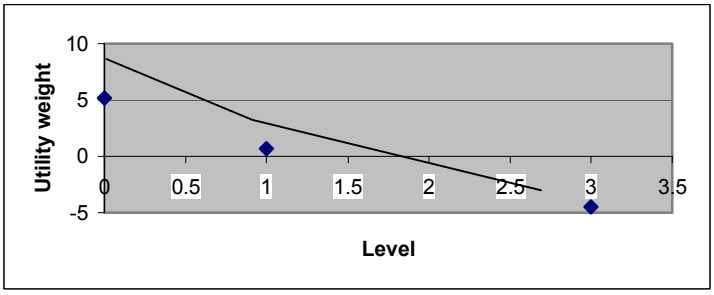
Figure 4 Most VS Least Choices - Dental Study



One feature of the present study is that the 6-level attribute, which is cost, while balanced in the design, results in levels that only occur half as often as the 3-level attributes. That is, each level of cost is only available to be chosen half as often as the levels of the other attributes. To compensate for this "availability bias" one must reweight the choices, which can be done simply by either dividing the choices of all the 3-level attributes by two, or multiplying the choices of the 6-level attribute by two. We did the latter. It also is worth noting that because our purposes in this exercise are illustrative and not substantive, we made no attempt to reweight the data so as to preserve the original degrees of freedom (54 subjects x 18 scenarios x most/least = 1944 possible choices). We also eliminated all non-responses, which were a consistent percentage (approximately 9.5%) of the total choices for both most and least choices.

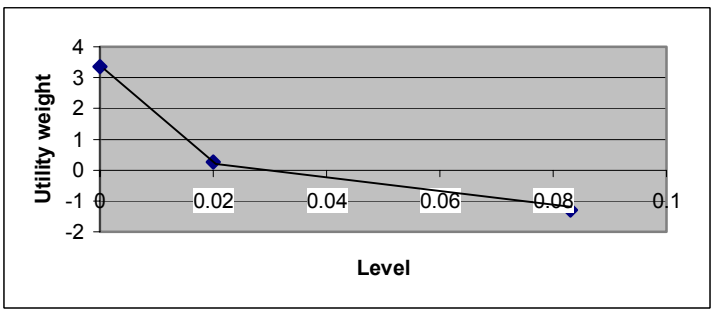
Because the sample size is small there are several zero choice frequencies. According to Ben-Akiva and Lerman (1985) (Ben-Akiva ME & Lerman S 1985) the addition of a small proportion to the zero frequencies results in consistent but inefficient estimates. Thus, we assumed that the zeros were sampling zeros and not structural zeros and added a small proportion to each zero equal to the proportion of choices that would be expected if a level had been chosen once in the total possible number of choices (54 subjects x 18 choices =1/972). We then proceeded to estimate the utilities associated with each attribute level using weighted least-squares. The results from this estimation are shown in Figures 5-10 and their corresponding tables.

Figure 5 Severe pain and swelling (0 days, 1 day, 3 days)



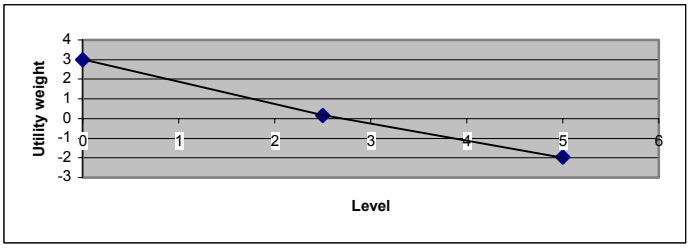
Attribute level	Utility Weight	WTP
0 Days	5.19	/
1 Day	0.66	/
3 Days	-4.48	£38.70

Figure 6 Episodes of mild pain (never, weekly, monthly)



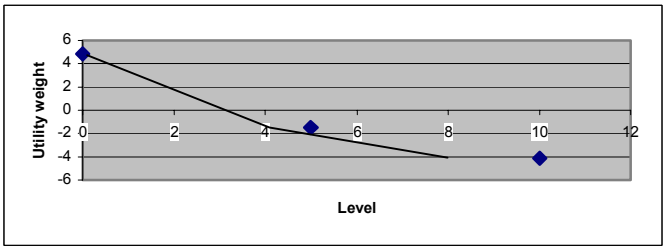
Attribute level	Utility Weight	WTP
0 Episodes	3.36	/
Weekly	0.27	/
Monthly	-1.29	£18.60

Figure 7 Chance of bleeding (%)



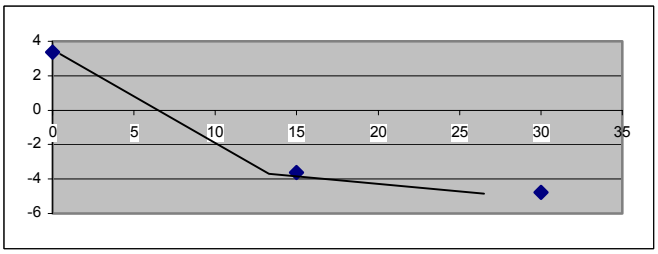
Attribute level	Utility Weight	WTP
0 %	2.99	/
2.5%	0.16	/
5%	-1.96	£19.80

Figure 8 Chance of nerve damage (%)



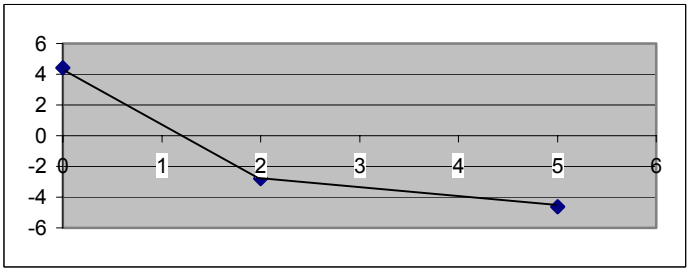
Attribute level	Utility Weight	WTP
0 %	4.83	/
5%	-1.47	/
10%	-4.13	£35.80

Figure 9 Chance of crowding (%)



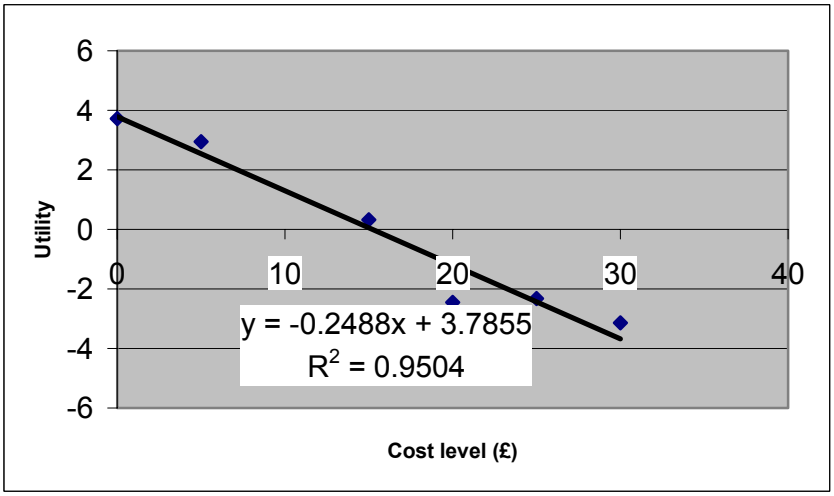
Attribute level	Utility Weight	WTP
0 %	3.36	/
15%	-3.62	/
30%	-4.79	£32.60

Figure 10 Number of episodes of gum inflammation



Attribute level	Utility Weight	WTP
0	4.41	/
2	-2.79	/
5	-4.63	£36.16

Figure 11 Graph of cost results

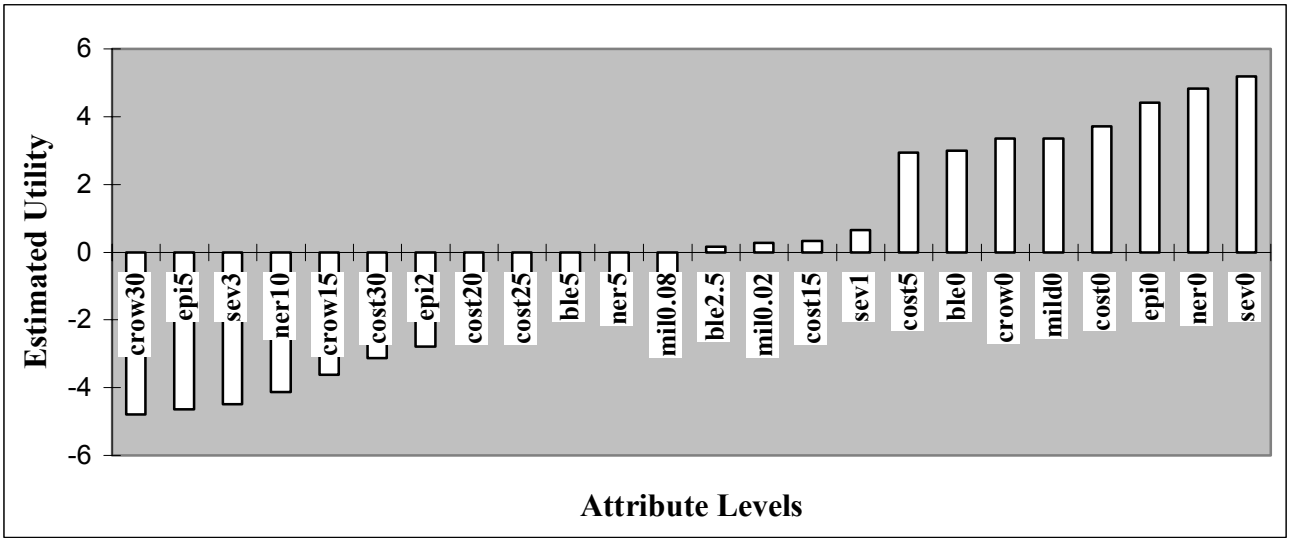


Cost Level Utility Weights

Attribute level	Utility Weight
£0	3.72
£5	2.94
£15	0.33
£20	-2.46
£25	-2.32
£30	-3.13

[Utility = 3.78 – 0.2488cost]

Figure 12 Common Utility scale



Discussion

Utility weights

The results from this BSc experiment show firstly that the majority of attributes adhere to an approximate linear specification, this can be seen in Figures 5-10. The utility weights for each attribute level are provided individually for each attribute as well as being compared on the common utility scale shown in Figure 12. One of the advantages of the common utility scale is the ability to separate the individual levels into a preference ordering distinct from an aggregate 'attribute preference bundle'. The common scale shows a pattern of utility weights which seems to adhere to consistent choice behaviour, namely the most preferred attribute level is 'No days pain and swelling' and the least preferred attribute level is '30% chance of crowding'. This common utility scale ranges from –4.79 to 5.19, approximately 10 utility units. It is feasible to see how this scale could be re-calibrated to produce information for a dental health utility scale.

Compensating variation

As long as cost is included as an attribute in a BSc study it possible to calculate WTP. Specifically, one first graphs the cost results, which are shown in the Figure 11. As can be seen, the cost levels are linear to a close first approximation, and hence, all that has to be done is to regress the utility estimates against the actual price levels (£'s) to obtain the appropriate unit of measurement to calculate compensating variation. Specifically, using the utility function estimated in Figure 11 ($Y = -0.2488x + 3.7855$) the differences in any utility levels can be easily calculated both within and between attributes. Note, that this makes it computationally straightforward to perform inter-dimensional attribute comparisons, and the implied WTP differences for the lowest and highest valued level of each attribute are shown in the small tables adjoining Figures 5-10. These figures

were obtained by subtracting the lowest utility from the highest utility in any particular attribute to get a difference in utility of highest to lowest, WTP is then obtained by dividing the utility difference (from least to most) by the cost coefficient from the regression. The WTP values for shifts from the least attractive level to the most attractive level are provided alongside each attribute figure (5-10). A reduction in the number of days suffering pain and swelling of three days is valued most highly, at £38.70. This is closely followed by a value of £36.16 for a reduction in number of episodes of inflamed gums from 5 to 0. Individual WTP values for each absolute attribute *level* can also be compared on an absolute WTP scale. This can be compared to the traditional DCE approach which commonly estimates aggregate utility units per *attribute*. The WTP results show that people were WTP £32.60 to move to from 30% chance of crowding to 0% chance. Whilst the cost levels in this study were approximately linear, the BSC results reveal that the majority of preference value from 'crowding' lies in the lower two levels, namely a reduction from 15% to 0%, £27.92, with the marginal value of a reduction in crowding from 30% to 15% being only £4.68. From a policy perspective this reveals that the majority of effort should concentrate on reducing the chance of crowding from 15% to 0%.

Limitations of the data in the empirical study

The response rate for this population sample was 14%, this is a low sample size and cannot be regarded as representative of the general population. It is the case that the BSc questionnaire requires a lengthy questionnaire due to individual scenarios being presented as opposed to the usual paired scenarios, which approximately half the number of choice sets. There were also a relatively large number of attributes in the questionnaire, many of which contained risk probabilities and may have been too cognitively complex for many.

The BSc approach more generally

The BSc approach makes it possible to separate the weight or importance of an attribute from the position of each attribute's levels on the underlying utility scale. Traditional methods for inferring the "importance" of attributes, in fact, confound weight and scale value, and the best that can be said for them is that they estimate "relative weight". Relative weight is the empirical impact of an attribute across all its levels on the dependent variable. Logical ways to estimate this impact include the proportion of variance in the dependent variable explained by each attribute or similar measures for non-linear models, such as the proportion of total log likelihood attributable to each attribute (or effect). Such effects are "relative" because they do not generalize to other contexts in which either new attributes are added, some attributes are subtracted, levels are changed, etc. The key takeaway message however is that these relative weights are a complex combination of both weight and scale position, and these cannot be separated using traditional methods.

In the case of BSc, the weights that can be estimated are also "relative" in the sense that they can change with context, however they can be estimated separately from scale values. That is, one can calculate the overall impact of an attribute, which is the total number of times that it is chosen either best or worst across all scenarios, and one also can calculate the position of each level, which is the total number of times that an attribute level is chosen as best (or estimated from pooled best-worst data). Thus, one can determine whether the effects observed in choice experiments are due to an attribute having a large/small weight or having a large/small difference in scale values. The four conditions, and their implications are shown in Table 6 below. Briefly, different strategies are indicated to change attribute scale positions and/or weights.

Table 6 Conditions and implications of differences in weights and scale values

Implications of differences in weights & scale values		Attribute Scale Values	
		High	Low
Attribute Weights	High	**1.** Very large impacts on choice/behaviour	**2**. Moderate-large impacts on choice/behaviour
	Low	**3**. Low-moderate impacts on choice/behaviour	**4**. Little impact on choice/behaviour

Consider the following types of implications for the four circumstances:

1. Attributes and levels in this condition have very large effects on choices, and if these effects are in the desired direction, that may be sufficient. On the other hand, if the effects are not desirable, then one has two options: try to get consumers to place less weight on the attribute, or try to convince them that particular levels have a different scale position. The former implies a marketing and communications campaign, whereas the latter implies an informational or educational campaign, although changing weights also may require information and education to be fully effective. If we further note, that it is *differences* in attribute levels between competing options that matter in choices, we also can see that competitors (or health care policy makers) probably want to match levels as closely as possible and/or differentiate as much as possible in the case of attributes in this condition.
2. This circumstance is intermediate in the sense that it may be possible to move an attribute in this state to state 1 by a campaign directed at informing or educating consumers that the levels are different than perceived and/or by making actual changes to the levels to increase the differences. Otherwise, the implications are similar to state 1.
3. Attributes in this condition derive their impact from the spread of scale values. Thus, one can try to increase their weights if the attribute per se is desirable, or one can try to decrease the scale value differences if undesirable.

4. Attributes in this condition have virtually no impact, which means they not only are ignored in decision-making, but consumers perceive little difference in their levels. If the attribute is exceptionally desirable from a policy standpoint, one might try to increase its weight or its range of scale values, but it may be that attributes in this condition simply do not matter, which would imply that changes would be very difficult to achieve.

The method of BSc was originally motivated by a large number of practical applications in which conjoint ratings and discrete choice tasks produced results which indicated little statistical support for rejecting the null hypothesis for attributes expected to be significant *a priori* (Louviere J & Swait J 2001). Early work based on the Luce model by Louviere and Woodworth (1983) (Louviere J J & Woodworth G1983) from which the BSc model is based was the first attempt to "derive a single, closed-form model to summarize and aggregate choice response surface for a set of conjoint profiles". In health care however, the primary motivation for such a 'closed-form response surface' is driven by a very different reason – that of using random utility theory to help in the search for good, reliable, theoretically based measures of benefit, preferably with an identifiable underlying scale. With the BSc approach, weights and scale values estimated from 'best' frequencies are ratio scaled because they are estimates of proportions (probabilities), which trivially satisfy ratio scale conditions. In fact, these weights and scale values constitute and *absolute scale with known zero and meaningful unit*. This latter characteristic, whilst less important in marketing where the primary aim is to predict market share, is a key advantage for the use of this methodology in health care. With the framework of economic evaluation driving health care allocation decisions the possibility of a theoretically based methodology to obtain ratio-scaled benefit measures to go alongside the ratio-scaled 'cost' element is an exciting prospect. With the huge number of both generic and disease specific attributes (health related, non-health and process) in health care there is enormous scope for use of this technique within an economic evaluation framework.

Early work in the health care field attempting this approach used interval scaled 'arbitrary' scores. However, such a summation exercise did not take into consideration the importance of the scale factor. As a consequence, such 'utility scores' could have a limited usefulness in a technical or allocative efficiency setting in economic evaluation, the scores acting simply as ordinal indicators of value within the given setting. Hence the importance of the scale factor and the lack of a fixed anchor point make the traditionally derived utility scores arbitrary, unlike the BSc utility weights with known zero and meaningful unit.

In addition to the potential benefits of this method discussed above there are a number of more general advantages of the using BSc method. Firstly, whilst the method can be used as a standalone

technique the interpretation of results from other conjoint techniques is also greatly enhanced by BSc results. In this empirical study, the response functions for each attribute were generally linear however it may be the case in other studies that the BSc approach identifies non-linear response functions thus providing information on the appropriate specification of variable form. For example, BSc studies may reveal a preference for say, the lowest price level, showing that price has little additional impact at higher levels. This advantage can be carried over to the valuation of health states where it is possible to imagine patients having non-linear preferences for levels of particular health states. The BSc approach would help to identify the exact pattern of utility weight per level per health state attribute. Further, these levels could then be equated on a common utility scale with anchor points at 0 with the highest level feasibly re-scaled to 1. Such an approach may provide an insight for exploring weights of health care attributes and levels for use within the QALY paradigm. At the very least, the BSc approach would be a useful complement to existing techniques exploring this issue in health care.

Previous studies using BSc

A small number of studies to date have used the BSc approach, these applications have mainly been in the field of marketing. Finn and Louviere (1992) (Finn A & Louviere J 1992) used BSc to measure consumer opinions about public policy alternatives; Louviere, Finn and Timmermans (1994) (Louviere J, Finn A, & Timmermans HG 1994)used BSc to measure retail images; and Swait (1994) (Swait J 1994)used BSc to measure brand images and classify subjects into latent segments. To date, there have been only two published papers using a variant of BSc approach in health care. Szeinbach et al (1997) (Szeinbach SL, Barnes, & Garner DD 1997) used 'maximum difference' conjoint analysis to determine which value added services offered by pharmaceutical manufacturers are liked the best (worse) as perceived by hospital pharmacy directors. In a small empirical study carried Szeinbach et al (1999) (Szeinbach SL et al. 1999) entitled 'Using conjoint analysis to evaluate health state preferences' maximum difference conjoint analysis (a variant of BSc) was used to elicit utility weights for health states of the EQ-5D instrument. These data were then compared to values obtained from a visual analogue scale. Results from this study revealed that the application of this technique is useful for health related quality of life research and that the maximum-difference results compare favourably with values obtained from visual analogue scales. Whilst this study only had a small sample (n=33), it provides encouraging evidence of the potential use of this new approach in health care.

Conclusions

This paper presents one of the first empirical applications in health care of a the BSc measurement and scaling model which permits separate identification of attribute weights and attribute part-

worths, something not possible with any form of traditional choice experiments. Whilst this technique is not claiming to be a substitute for traditional choice experiments the proposed approach can be viewed as a complement to existing choice experiments, which provides newer and deeper insights to assist interpretation of results, and/or as a stand-alone technique for part-worth estimation. The BSc approach provides new and additional information about utilities that enhances interpretation of traditional and choice-based conjoint. The use of this methodology in health care has much scope and future work in this area could explore the advantages of the methodology to elicit common utility scales for use within the economic evaluation framework.

References

Ahlquist M & Grondal HG 1991, "Prevalence of impacted teeth and associated pathology in middle aged and older swedish women", *Community Dentistry and Oral Epidemiology*, vol. 19, pp. 116-119.

Ben-Akiva ME & Lerman S 1985, *Discrete choice analysis: theory and application to travel demand* MIT Press, Cambridge, Mass.

Brickley M, Armstrong R, Shepherd J, & Kay E 1995, *International Dental Journal*, vol. 45, pp. 124-128.

Finn A & Louviere J 1992, "Determining the appropriate response to evidence of public concern: The case of food safety", *Journal of Public Policy and Marketing*, vol. 11, no. 1, pp. 12-25.

Louviere J, Finn A, & Timmermans HG 1994, "Retail Research Methods," in *Handbook of Marketing Research*, 2 edn, McGraw-Hill, New York.

Louviere J, Hensher DA, & Swait J 2000, *Stated Choice Methods: analysis and application*, 1 edn, Cambridge University Press, Cambridge.

Louviere J J & Woodworth G 1983, "Design and analysis of simulated consumer choice or allocation experiments: an approach based on aggregate data", *J Market Res*, vol. 20, pp. 350-356.

Louviere J & Swait J. Separating weights and scale values in conjoint tasks using best attribute scaling. 2001. Ref Type: Unpublished Work

Luce RD 1959, *Individual choice behaviour: a theoretical analysis*, 1 edn, Wiley, New York.

Swait J 1994, "A structural equation model of latent segmentation and product choice for cross-sectional revealed preference data", *Journal of Retailing and Consumer Services*, vol. 30, pp. 305-314.

Szeinbach SL, Barnes, J. H., & Garner DD 1997, "Use of pharmaceutical manufacturers value added servcies to build customer loyalty", *Journal of Business Research*, vol. 40, pp. 229-236.

Szeinbach SL, Barnes, J. H., McGhan WF, Murawski MM, & Corey, R. 1999, "Using conjoint analysis to evaluate health state preferences", *Drug Information Journal*, vol. 33, pp. 849-858.

The Royal College of Surgeons of England 1997, The management of patients with third molar teeth: Report of a working party convened by the faculty of Dental Surgery, The Royal College of Surgeons of England, England.