Discrete choice experiments and best worst scaling: compliments or substitutes?

Mary Kilonzo, Health Economics Research Unit, University of Aberdeen
Jen Burr, Health Services Research Unit, University of Aberdeen
Mandy Ryan, Health Economics Research Unit, University of Aberdeen
Luke Vale, Health Economics Research Unit & Health Services Research Unit, Aberdeen

Author for Correspondence & Contact Details
Mary Kilonzo
Health Economics Research Unit,
Polwarth Building, Foresterhill,
Aberdeen AB25 2ZD
Telephone: +44 (0) 1224 555181
Fax: +44 (0) 1224 550926

Abstract

Aim: A limitation of discrete choice experiments (DCEs) is that different attribute may be measured on a different underlying scale because they only give information on the weight but not on the scale value. Thus, it is difficult to distinguish the importance of the overall weight of an attribute from the importance of its levels. Best worst scaling (BWS) can be used to separate attribute from scale values and the aim of this paper is to conduct and compare a BWS exercise and a DCE.

Methods: Using a case study in the area of glaucoma a BWS exercise and DCE were conducted. In the DCE choice sets are presented and respondents are asked to choose the most/least preferred. With a BWS exercise a series of profiles are presented and for each profile the attribute level that exhibits the highest (best) utility and the attribute level that exhibits the lowest (worst) utility are picked. A conditional logistic regression model was used to analyse the DCE data and the BWS data were analysed using weighted least squares. Both models were used to generate utility scores for use within a QALY framework.

Results: In general the weightings obtained from the BWS and the DCE were consistent. For both methods better clinical outcomes are preferred to worse clinical outcomes as would be expected. The magnitude of weights was also consistent. However, with the BWS the differences in utility for changes away from the best situation were larger than those estimated by the DCE.

Conclusions: The results of the best worst scaling exercise were consistent with those of the DCE. However, their use in evaluations may result in different conclusions being drawn and further research to explore the implications of this is warranted.

1. Introduction

The last 15 years have seen an increasing use of the discrete choice experiments (DCEs) in health economics[26]. DCEs can be used to compare the relative impact of attributes of the service or product under investigation. Most studies compare relative impacts of attributes by comparing the size and significance of estimated parameters for attributes of interest and obtain valuations using one attribute, normally a cost attribute, as a common numeraire. A limited number of published studies have also used the DCE methodology to estimate values for different health state profiles[8, 12, 25, 26]. To date however only two have used the technique to estimate a utility index which might be used within a QALY framework[21, 24].

One limitation of the DCE approach, that may limit its usefulness for eliciting health state utilities, is that attribute utility estimates are confounded with the underlying subjective utility scale[14]. This is noted by a number of researchers[13, 16, 18, 20]. This confounding of weight and scale make it difficult for individual utility weight comparisons to be made within any DCE on a common scale, unless transformed by a common numeraire such as money

One method recently applied in health economics, to overcome this problem is best worst scaling (BWS)[4]. BWS is attractive because the individual utility weights obtained can be placed on their underlying latent utility scale and an overall utility score obtained that has a non arbitrary value. As with the results of a DCE the higher and lower limits to the utility score enables utilities to be rescaled to a 0-1 scale. The advantages of BWS according to Marley and Louviere[19] include:

- A single pair of best worst choice contains a great deal of information about a persons ranking of options;
- The tasks take advantage of peoples' propensity to identify and respond more consistently to extreme options; and
- It seems to be an easy task for people to perform.

The purpose of this paper is to estimate and compare preference based quality weights using these two methods: DCEs and BWS. Hitherto, BWS has not been used to develop utility weights so as part of our estimation we present a potential method to estimate utility weight that might be used within a QALY framework. This work was conducted as part of a study carried out to develop a utility-based measure of outcome and to estimate health state values in the area of glaucoma. In the following section we describe the experiment. The results from both the DCE and BWS are then presented and discussed. Finally consideration is given to future areas for research.

2. Experiment

The area of application is the development of a preference based utility measure to the area of glaucoma. Glaucoma is a chronic eye disease characterised by progressive damage to the optic nerve

and consequent restriction of the field of vision. The condition does not affect length of life but is associated with impaired health status and health related quality of life. Traditionally the outcome of glaucoma care has been judged on the reduction in intraocular pressure and measures of visual function, mainly an assessment of the visual field. While clinically useful for monitoring progression, such outcomes do not capture the impact on a patients reported health status.

Qualitative research methods were used, involving people with glaucoma in two focus groups, to establish the attributes and levels to be included in the utility measure. In the first stage of this qualitative component, potentially relevant items, attributes and levels of difficulty were identified from existing vision and glaucoma specific quality of life instruments and studies reporting on disability and quality of life in glaucoma,[1, 5, 9, 10, 11, 15, 22, 29] and additionally from expert opinion. These were collated and were used to define the framework of the focus group discussions. Subjects were recruited from two ophthalmology centres in the UK (Aberdeen and Leeds).

Thirty people were invited to attend two separate focus groups, 17 agreed to participate, nine in the group in Scotland (Aberdeen), and eight in the group in England (Leeds). An experienced qualitative researcher led both focus groups and explored the participant's views on the areas we had collated concerning the effects of glaucoma on vision, mobility, role performance, mood and any adverse effects of treatment. The discussions were audio taped, and subsequently transcribed and analysed using framework methodology to identify key areas that were meaningful to patients for inclusion in the glaucoma profile measure[23].

The key identified areas reflecting health status were near vision tasks (NV), treatment effects both in and around the eye (EE) and effects on general health (GE), illumination (LG), mobility (M), and visual judgement for activities of daily living (ADL). The levels were based on four levels of difficulty associated with each attribute: no difficulty; some difficulty; quite a lot of difficulty; severe difficulty.

From the six attributes, each with four levels, there were $6^4 = 4096$ possible profiles. A fractional factorial design was used to reduce the 4096 profiles to 32 profiles, while still being able to infer utilities for all possible profiles. For the DCE foldover techniques were used to derive 32 choice sets from these profiles, ensuring orthogonality, minimum overlap and level balance of the design for the DCE[17]. Figure 1 provides an example of 1 of the 32 choices in the DCE.

Figure 1 Example of a DCE choice set

SITUATION A	SITUATION B
No difficulty with: • Central and near vision • Lighting and glare • Mobility **Some difficulty** with: • Activities of daily living • Eye discomfort • Other effects of glaucoma and its treatment	**No difficulty** with: • Central and near vision **Some difficulty** with: • Lighting and glare **Quite a lot of difficulty** with: • Activities of daily living • Other effects of glaucoma and its treatment **Severe difficulty** with: • Mobility • Eye discomfort
(Tick one box only) ☐ **Situation A**	☐ **Situation B**

For the BWS respondents were presented with the 32 scenarios. For each scenario the respondent was asked to select the attribute they most preferred and one they least preferred. Figure 2 provides an example.

Figure 2 Example of BWS scenario

Best aspect	**Aspects of situation A**	***Worst aspect***
√	**No difficulty** with lighting and glare	
	Some difficulty with activities of daily living	
	Quite a lot of difficulty with mobility	
	Quite a lot of difficulty with eye discomfort	
	Severe difficulty with central and near vision	√
	Severe difficulty with the effects of glaucoma and its treatments	

Pilot work compared presenting 8, 16 and 32 choices to respondents. Results indicated that responses were not adversely affected by the number of choices, though several respondents commented on the length of the questionnaire. As a consequence one block was used, with all subjects receiving the 32 choices.

Subjects were selected from attendees at four hospital-based clinics and one community-based glaucoma clinic across two eye centres in the UK. All patients with glaucoma as the main diagnosis, or ocular hypertension (the single most important risk factor for the development of glaucoma[7]) on

treatment, with a reliable visual field test in at least one eye and who were willing to complete the questionnaire were eligible. People suspected of glaucoma but not on treatment were excluded. We also recruited volunteers from the International Glaucoma Association (IGA), a patient organisation, following an advertisement in the patient newsletter and the IGA website, anybody with self reported glaucoma was eligible. Four hundred and seventy three people received the questionnaire, 225 from the clinic-based sample and 248 from the self-selected sample from the IGA. Approval was obtained for each phase of the study from the Central Office of Research Ethics Committees. The research was conducted according to the tenets of the Declaration of Helsinki.

3. Econometric analysis

Econometric techniques were used to analyse the DCE and BWS responses. A conditional logistic regression model was used to analyse the DCE response data with the following equations initially being estimated:

$$QW_{ij} = \sum \alpha_{dl} X_{dl} + e + u \qquad (1)$$

where QW_{ij} is the quality weight for outcome state i as valued by individual j, X_{dl} is a vector of dummy variables where d represents the attribute from the profile measure and l the level of that attribute.

Flynn and colleagues[3] describe four models for analysing best worst scaling choice data. These models illustrate paired and marginal model analyses. 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. Model estimation can be performed at either the individual or sample level. In both cases the estimates are arrived at by summing the number of times that an individual selected the particular attribute as best/worst. The analysis of the data was based on the sample and used both paired and marginal methods. As our study was a balanced study each pair had an equal number of chances of being chosen.

Variables required for the BWS regression include the best worst count frequencies, and the natural log of best worst counts. The number of unique best worst pairs that could be estimated was:

$$2\sum_{i=1}^{K-1}\left[L_i \sum_{k=i+1}^{K} L_K\right].$$

In our case study this equates to 480 unique best worst observations. Data were analysed using weighted least squares and each possible pair was available to be chosen twice. Weight was the

adjusted choice totals (details below). The least valued impact was that of treatment effects both in and around the eye so it was omitted leaving the equation to be estimated as:

$$\ln(f) = \text{cnst} + \beta_1 \text{NV} + \beta_2 \text{LG} + \beta_3 \text{M} + \beta_4 \text{VJ} + \beta_6 \text{EG} + \beta_{11} \text{NV no} + \beta_{12} \text{NV some} + \beta_{13} \text{NV quite a lot} + \beta_{21} \text{LG no} + \beta_{22} \text{LG some} + \beta_{23} \text{LG quite a lot} + \beta_{31} \text{M no} + \beta_{32} \text{M some} + \beta_{33} \text{M quite a lot} + \beta_{41} \text{VJ living no} + \beta_{42} \text{VJ some} + \beta_{43} \text{VJ quite a lot} + \beta_{51} \text{EE no} + \beta_{52} \text{EE} + \beta_{53} \text{EEquite a lot} + \beta_{61} \text{GE no} + \beta_{62} \text{GE some} + \beta_{63} \text{GE quite a lot} \quad (2)$$

Where: F is the number of times a particular best worst choice was selected across all scenarios and across all respondents, adjusted to eliminate sampling zeros as some of best worst pairs were never chosen. This adjustment was achieved by adding a small number: the reciprocal of the sample size to enable logs to be taken[6]. NV, EE GE, LG M, and VJ are the attributes under consideration. Each attribute appears in the equation four times: once for the impact of the attribute and thrice for the three levels of that attribute. β represents the parameter weights to be estimated for the impacts and levels of all the attributes. Frequencies were then arranged in a dummy matrix that was effects coded such that the best attribute level had a value of +1 and the worst attribute level was -1 and all the rest were 0. Effects coding is an alternative to dummy coding where the effects are uncorrelated[2]. The reference point in effects coding is defined as the negative sum of the estimated coefficients, the utility of the Lth level is $\beta_1 + \beta_2 + \beta_3 + \beta_{L-1} \times (-1)$ which means that the reference point is now internalised in the β estimates and cannot be carried over the β_0 coefficient. Both the DCE and BWS analyses were performed using STATA9.2SE[28].

For both the DCE and the BWS rationality tests were included to ensure respondents were engaging in the exercise and taking it seriously. For the DCE Sen's expansion and contraction rationality tests were employed[27]. To test the expansion property respondents were first asked to choose the worse of 2 profiles (A or B). This choice was then widened to a choice between 3 profiles (A, B, or C) in a non-consecutive question. A person that chose situation B in the first choice question should not chose A in the expanded one to satisfy the expansion property. To tests the contraction property respondents were asked to choose a situation from a set of three alternatives (A, B or C). This choice was then narrowed to a set of two alternatives (A or B). The test was satisfied if a respondent who chose situation A in the first choice did not choose option B in the reduced choice set. Respondents were excluded from the regression analysis if they failed *both* tests. For the BWS a repeated question was included.

4. Estimating utility weights

For the DCE quality weights were calculated by summation of the coefficients associated with the best level for each attribute. The weights for all other levels of each attribute were then estimated as a

proportion of this score, allowing all combinations to be estimated on a 0 to 1 scale while maintaining the ratio between the comparisons.

Whilst it has been suggested that the results of BWS might be used to estimate utilities that could be used within a QALY framework, we are aware of no examples where this has been performed. Whilst output from a BWS experiment can be used to anchor utilities at the worse level of a given attribute[4], these estimates are not suitable for use within a QALY framework simply because an individual's health is not described by a single attribute level but by the level for each attributes. Therefore, a further rescaling was performed in which the worst possible attribute level for each attribute were combined and scaled to 0 and the best attribute level for each attribute were combined and scaled to 1. All other possible combination of attributes and levels would be associated with a utility score between 0 and 1. The details of this are described later in the Results.

5. Comparisons between BWS and the DCE

Comparison of the DCE and BWS was performed by comparing:

i) the response rates and results of the rationality tests to the different questions;

ii) the scaled weights generated from each model. (Although the measures are not the same it is possible to compare consistency of the results by looking at the size/magnitude of the estimates and the external and internal validity of the results.)

6. Results

Response rates and results of the consistency test

289 subjects responded to the DCE component of the questionnaire. Three respondents failed both consistency tests in the DCE and 25 chose a different BW pair when presented with exactly the same scenario. The results reported in this paper are based on 286 respondents that passed the DCE consistency test to ensure that the analysis was performed on the same people. Further work to compare results from respondents who passed both the DCE and BWS consistency tests is planned.

Results of the DCE

Table 1 presents the final regression model obtained from the DCE, merging attribute levels where there was no evidence that the levels within attributes were different[1]. These results confirm the theoretical validity of the model**,** with coefficients increasing as the level of the attribute moves from severe to 'better' levels. Moving from 'no difficulty' with near vision tasks to 'severe difficulty'

[1] Where the conditional logistic regression model provided no evidence of a significant difference between levels for a given attribute, using the Wald test, levels were combined and the model re-estimated. Utility scores were estimated using this final model.

results in the most loss of utility followed by activities of daily living and mobility, systemic ('other effects') and eye discomfort were considered to be the least important.

Table 1 Results of the DCE

Attributes and levels of difficulty	Coefficient	Std. Err.	z	P>z	[95% Conf. Interval]
Near vision tasks					
No difficulty	1.254	0.046	27.000	0.000	1.163 1.345
Some difficulty	0.852	0.050	16.980	0.000	0.753 0.950
Quite a lot of difficulty	0.526	0.041	12.720	0.000	0.445 0.608
Illumination					
No difficulty	0.272	0.036	7.500	0.000	0.201 0.343
Mobility					
No difficulty	0.921	0.045	20.380	0.000	0.832 - 1.010
Some difficulty	0.577	0.050	11.530	0.000	0.479 - 0.675
Quite a lot of difficulty	0.349	0.044	7.860	0.000	0.262 - 0.436
Visual judgement for activities of daily living					
No difficulty	0.999	0.045	22.090	0.000	0.910- 1.087
Some difficulty	0.720	0.050	14.420	0.000	0.622 - 0.817
Quite a lot of difficulty	0.431	0.043	10.100	0.000	0.347 - 0.515
Eye Effects					
No difficulty	0.241	0.042	5.720	0.000	0.158 - 0.323
Some difficulty	0.134	0.040	3.380	0.001	0.057 - 0.212
General effects					
No difficulty	0.202	0.040	5.040	0.000	0.123 - 0.281
Some difficulty	0.169	0.042	4.060	0.000	0.087 - 0.250

Results of the BWS experiment

The results of the weighted least squares analysis are reported in Table 2. The upper part of the table shows the importance of attributes overall. Near vision tasks has the largest impact relative to local eye discomfort (the omitted impact). The next three attributes are activities of daily living and mobility and other effects of glaucoma treatment. The values for these four attributes are similar and are statistically significant. The attribute with the least impact is lighting and glare and it is not statistically significantly different from local eye discomfort.

The lower part of Table 2 illustrates the additional importance for each level of each attribute. The majority of attributes adhere to an approximate linear specification as the values decrease as the level of difficulty increases. The utility scores exhibit face validity in that the lower levels of difficulty for each attribute have higher utilities compared with the utility scores for higher levels of difficulty. The greater the level of difficulty the lower utility value, implying that the higher the level of difficulty the less likely it is for respondents to choose that option.

Table 2 Best-worst utilities (paired WLS) method

	Coefficient	SE	T ratio	P>(T)	95% Confidence interval
_cons	2.017	0.041	48.76	0.000	1.936 2.098
Attribute impacts					
Near vision tasks	-0.303	0.047	-6.48	0.000	-0.395 -0.211
Lighting and glare	-0.056	0.051	-1.11	0.269	-0.155 0.043
Mobility	-0.240	0.048	-5.01	0.000	-0.335 -0.146
Activities of daily living	-0.295	0.047	-6.24	0.000	-0.388 -0.202
Eye discomfort	(dropped)				
Other effects treatment	-0.214	0.051	-4.16	0.000	-0.315 -0.113
Level scale values					
Near vision tasks					
No difficulty	1.980	0.049	40.5	0.000	1.884 2.076
Some difficulty	0.451	0.055	8.19	0.000	0.343 0.559
Quite a lot of difficulty	-0.913	0.053	-17.07	0.000	-1.018 -0.808
Severe difficulty	-1.519*				
Illumination					
No difficulty	0.626	0.060	10.47	0.000	0.509 0.744
Some difficulty	0.485	0.053	9.1	0.000	0.380 0.590
Quite a lot of difficulty	-0.516	0.065	-8	0.000	-0.643 -0.389
Severe difficulty	-0.595*				
Mobility					
No difficulty	1.850	0.052	35.64	0.000	1.748 1.952
Some difficulty	0.472	0.054	8.74	0.000	0.366 0.578
Quite a lot of difficulty	-0.930	0.057	-16.25	0.000	-1.042 -0.818
Severe difficulty	-1.393*				
Visual judgment for activities of daily living					
No difficulty	2.036	0.049	41.56	0.000	1.940 2.133
Some difficulty	0.559	0.053	10.47	0.000	0.455 0.664
Quite a lot of difficulty	-0.940	0.055	-17.16	0.000	-1.048 -0.832
Severe difficulty	-1.656*				
Eye effects					
No difficulty	0.898	0.066	13.69	0.000	0.769 1.027
Some difficulty	0.085	0.078	1.09	0.276	-0.068 0.239
Quite a lot of difficulty	-0.142	0.085	-1.67	0.095	-0.308 0.025
Severe difficulty	-0.842*				
General effects					
No difficulty	1.120	0.060	18.79	0.000	1.003 1.236
Some difficulty	0.342	0.070	4.85	0.000	0.204 0.480
Quite a lot of difficulty	-0.619	0.064	-9.64	0.026	-0.745 -0.493
Severe difficulty	-0.842*				

* computed from model results

The coefficients for visual judgement for activities of daily living have the highest and lowest with values of 2.036 for no difficulty and -1.656 for the worst level of the attribute. The common scale shows a pattern of utility weights which adheres to consistent choice behaviour, namely the most preferred attribute level for all attributes is no difficulty and the least preferred attribute level for all

attributes is severe. This common utility scale ranges from to 2.036 to -1.656, approximately 4 utility units. Table 3 describes the results of the method that we have adopted to convert the coefficients obtained from the BWS regression into utility scores scaled between 0 and 1. The first stages of this calculation follow the method used by Flynn and colleagues[3] and we propose a method for discussion to extend this calculations to provide utility scores suitable for use within a QALY framework.

The second column of the table reports the coefficients obtained from the BWS regression. By combining the level scale value with the average attribute impact the value, reported in the second column, the values reported in the third column are estimated. For example, the value for 'no difficulty with near vision tasks' is 1.677, which is equal to the coefficient for this level (1.980) plus the average impact of the attribute (0.303). The next stage of the calculation is to estimate the additional value above the lowest scale value (column 4). The lowest scale value is severe difficulty with visual judgement for activities of daily living which is given the value of 0. The values of all other scale values are rescored relative to this.

Rescaling between the best and worst scale values as is illustrated in column 4 is as far as other analysts have currently reported their results. However, these data are not, without further manipulation, suitable for use within a QALY framework because an individual's health is not described by a single attribute level but by the level for each attributes. To accomplish this, the worst combinations of each attribute levels are all given a 0 score. All other scores for an attribute are re-estimated relative to the 0 score for that attribute level (column 5). For example, the score for general effects – quite a lot of difficulty becomes 0.223, which is the value for quite a lot of difficulty (1.118) less the value for severe difficulty for this attribute (0.895). From the data reported in column 5 it can be seen that the worst possible combination of attribute levels is given the value of 0 and the best possible combination at attribute levels is equal to 15.358 (i.e. 3.500 + 1.222 + 3.243 + 3.692 + 1.740 + 1.962). These data are then rescored between 0 and 1, where the worst possible combination of attribute levels is given the value of 0 and the best possible combination at attribute levels is given the value of 1.

Table 3 Derivation of utility scores from the results of the BWS

	Coefficient	Value*	Additional value above lowest scale value	Value above worst possible combination of attribute levels	Rescaled between 0 & 1
_cons	2.017				
Attribute impacts					
Near vision tasks	-0.303				
Lighting and glare	-0.056				
Mobility	-0.240				
Activities of daily living	-0.295				
Eye discomfort	0				
Other effects treatment	-0.214				
Level scale values					
Near vision tasks					
No difficulty	1.980	1.677	3.628	3.500	0.2279
Some difficulty	0.451	0.148	2.099	1.970	0.1283
Quite a lot of difficulty	-0.913	-1.216	0.735	0.606	0.0395
Severe difficulty	-1.519*	-1.822	0.128	0.000	0
Illumination					
No difficulty	0.626	0.571	2.521	1.222	0.0796
Some difficulty	0.485	0.429	2.380	1.080	0.0703
Quite a lot of difficulty	-0.516	-0.572	1.379	0.079	0.0052
Severe difficulty	-0.595*	-0.651	1.299	0.000	0
Mobility					
No difficulty	1.850	1.610	3.561	3.243	0.2112
Some difficulty	0.472	0.232	2.182	1.864	0.1214
Quite a lot of difficulty	-0.930	-1.170	0.780	0.463	0.0301
Severe difficulty	-1.393*	-1.633	0.318	0.000	0
Visual judgment for activities of daily living					
No difficulty	2.036	1.741	3.692	3.692	0.2404
Some difficulty	0.559	0.264	2.215	2.215	0.1443
Quite a lot of difficulty	-0.940	-1.235	0.716	0.716	0.0466
Severe difficulty	-1.656*	-1.951	0.000	0.000	0
Eye effects					
No difficulty	0.898	0.898	2.849	1.740	0.1134
Some difficulty	0.085	0.085	2.036	0.927	0.0604
Quite a lot of difficulty	-0.142	-0.142	1.809	0.700	0.0456
Severe difficulty	-0.842*	-0.842	1.109	0.000	0
General effects					
No difficulty	1.120	0.906	2.856	1.962	0.1277
Some difficulty	0.342	0.128	2.079	1.184	0.0771
Quite a lot of difficulty	-0.619	-0.833	1.118	0.223	0.0145
Severe difficulty	-0.842*	-1.056	0.895	0.000	0

Comparison of results

The results, recalibrated on a 0 to 1 scale, for the DCE for each level of each attribute are detailed in columns 2 and 5 of Table 4 and those for the BWS are reported in columns 3 and 6 of Table 4. For example for the DCE no difficulty with central and near vision gives a quality of weight of 0.322, moving to a situation where one has some difficulty will reduce the weight to 0.219 and moving to a situation where one has quite a lot of difficulty reduces the weight to 0.135. These quality weights can be summed to establish a quality weight or score for all profiles.

Table 4 Comparison of the recalibrated scores for each attribute level for the DCE and BWS

Attribute	DCE	BWS	Attribute	DCE	BWS
Near Vision tasks			**Visual judgement with activities of daily living**		
No difficulty	0.322	0.228	No difficulty	0.257	0.240
Some difficulty	0.219	0.128	Some difficulty	0.185	0.144
Quite a lot of difficulty	0.135	0.039	Quite a lot of difficulty	0.111	0.047
Severe difficulty	0	0	Severe difficulty	0	0
Illumination			**Eye discomfort**		
No difficulty	0.070	0.080	No difficulty	0.062	0.113
Some difficulty	0	0.070	Some difficulty	0.035	0.060
Quite a lot of difficulty	0	0.005	Quite a lot of difficulty	0.035	0.046
Severe difficulty	0	0	Severe difficulty	0	0
Mobility			**Other effects**		
No difficulty	0.237	0.211	No difficulty	0.052	0.128
Some difficulty	0.148	0.121	Some difficulty	0.043	0.077
Quite a lot of difficulty	0.090	0.030	Quite a lot of difficulty	0.043	0.015
Severe difficulty	0	0	Severe difficulty	0	0

Table 5 provides examples of how utility scores are calculated for four different situations using the weights produced by the two different methods.

Table 5 Utility scores for different health state profiles

Situation description	Quality weights		Utility Score	
	DCE	***BWS***	***DCE***	***BWS***
You have no difficulty with near vision tasks, illumination, mobility, visual judgment for activities of daily living, eye discomfort and other effects glaucoma and its treatment	0.322 0.070 0.237 0.257 0.062 0.052	0.228 0.080 0.211 0.240 0.113 0.128	1	1
You have some difficulty with near vision tasks and no difficulty with illumination, mobility, visual judgment for activities of daily living, eye discomfort and other effects glaucoma and its treatment	0.219 0.070 0.237 0.257 0.062 0.052	0.128 0.080 0.211 0.240 0.113 0.128	0.897	0.900
You have some difficulty with near vision tasks, some difficulty with illumination, mobility, and no difficulty with visual judgment for activities of daily living, eye discomfort and other effects glaucoma and its treatment	0.219 0.000 0.149 0.257 0.062 0.052	0.128 0.070 0.121 0.240 0.113 0.128	0.738	0.800
You have quite a lot of difficulty with near vision tasks, illumination, mobility and visual judgement for activities of daily living, and some difficulty with eye discomfort and other effects glaucoma and its treatment	0.135 0.000 0.090 0.185 0.035 0.043	0.039 0.005 0.030 0.047 0.060 0.077	0.414	0.258
You have quite a lot of difficulty with near vision tasks, illumination, mobility, and severe difficulty with visual judgement for activities of daily living, eye discomfort and other effects glaucoma and its treatment	0.135 0.000 0.090 0.000 0.000 0.000	0.039 0.005 0.030 0.047 0.046 0.015	0.225	0.182
You have severe difficulty with near vision tasks, illumination, mobility, visual judgement for activities of daily living, eye discomfort and other effects glaucoma and its treatment	0.000 0.000 0.000 0.000 0.000 0.000	0.000 0.000 0.000 0.000 0.000 0.000	0	0

The results indicate the respondents have less utility as the difficulty level of the attribute increases, but the BWS utility values are lower than those of the DCE. The size of the BWS index is also lower than those of the DCE. The utility scales generated from the two methods appear to be similar although the values seem to differ depending on the level of difficulty experienced for each attribute. The values for DCE are lower for the lower levels of difficulty but higher in higher levels of difficulty.

6 Discussion

Summary of results

A strength of this study is that we adopted a within sample design, with the same subjects completing the DCE and BWS exercise. Completion rates for both tasks were similar, this may be due to the DCE choices being presented jointly with the BWS questions. The results of the BWS exercise were generally consistent with those of the DCE. However, the results of the different approaches do provide different information. Whilst the estimates from the DCE model represent the additional (dis)utility of moving between attributes and levels, the BWS provides additional information as the reference case is a single attribute level. This enabled us to establish, using the BWS, the attribute that respondents consider to be of greatest importance was near vision tasks (the coefficient for this attribute which indicates its average importance was the largest absolute terms). However the attribute with the highest value was no difficulties for visual judgement for activities of daily living.

A key feature of this study presented here that is worthy of some more debate is how we have converted the BWS results into a 0 to 1 scale suitable for use within a QALY framework. Although this has been suggested as a use for BWS, we are not aware of any work where this has been presented. It is unclear whether the method we have adopted is appropriate and debate is required as to whether a better approach can be identified.

Areas for further research

With respect to the econometric/analytical methods used to analyse the response data, further research is planned with respect to:

- Re-estimation of the model using an individual level analysis as opposed to the sample level analysis reported in this paper.

- Explore the impact of excluding respondents who failed either or both simple rationality tests.

- Even though the BWS task has been said to be an easier task than DCE more respondents (25 versus 3) failed the rationality test. Within the current analysis respondents who failed these rationality tests with the BWS were not excluded from the analysis (although respondents who failed both rationality tests in the DCE were). Analytically, further work is needed to explore the impact of including and excluding respondents who failed these tests within the analysis.

- Explore the impact of combining levels in a reduced model where there was no evidence of a statistically significant difference between levels of attributes in the full model.

- For the DCE, attribute levels were combined and the model re-estimated where there was no evidence of a difference between levels. While it is interesting that there is some difference in what levels are statistically significant and, for the BWS, that the average effect on utility of lighting and glare on utility is 0 further analysis is required to explore the effect of combining the non-significant levels.

- Further consider how utility scores suitable for use in a QALY framework might be estimated. As highlighted above the approach adopted in this paper is presented solely as a means of provoking debate and methods to improve on this approach need to be developed.

- At the applied level, the following will be investigated: The policy impact of using the different valuation techniques. Additional work undertaken by ourselves and colleagues has considered the impact of using the different utilities in a model for screening for glaucoma. To date analyses using EQ-5D and DCE valuations have been performed. Further work might be conducted to compare the results when using the DCE and BWS valuations (which of course are both from the same perspective).

REFERENCES

1. Azuara-Blanco A, Aspinall PA, O'Brien C. Quality of life in patients with glaucoma: A conjoint analysis approach. 2003 ARVO Annual Meeting, May 2003; Poster no. 2004.
2. Bech M, Gyrd-Hansen D. Effects coding in discrete choice experiments. Health Economics 2005 14, 1079-1083.
3. Flynn TN, Louviere JJ, Peters TJ, Coast J. Best-worst scaling: What it can do for health care and how to do it. J Health Econ 2007 26, 171-189.
4. Flynn TN, Louviere JL, Peters TJ, Coast J. Estimating preferences for a dermatology consultation using Best-Worst Scaling: Comparison of three methods of analysis. Health Economists' Study Group, City University, London, January 2006.
5. Frost NA, Sparrow JM, Durant JS, Donovan JL, Peters TJ, Brookes ST. Development of a questionnaire for measurement of vision-related quality of life. Opthalmology 1998 5, 185-210.
6. Goodman LA. The analysis of cross classified data: Independence, quasi independence, and interaction s in contingency tables with or without missing entries. Journal of the American Statistical Association 1968 63, 1091-1131.
7. Gordon MO, Beiser JA, Brandt JD, et al. The Ocular Hypertension Treatment Study: Baseline factors that predict the onset of primary open-angle glaucoma. Arch Ophthalmol 2002 Jun 120(6), 714-20; discussion 829-30.
8. Hakim Z, Pathak DS. Modelling the EuroQol data: A comparison of discrete choice conjoint and conditional preference modelling. Health Econ 1999 8(2), 103-16.
9. Jampel HD, Schwartz A, Pollack I, Abrams D, Weiss H, Miller R. Glaucoma patients' assessment of their visual function and quality of life. J Glaucoma 2002 11, 154-63.
10. Jampel HD, Schwartz GF, Robin AL, Abrams DA, Johnson E, Miller RB. Patient preferences for eye drop characteristics: A willingness-to-pay analysis. Arch Ophthalmol 2003 121, 540-6.
11. Janz NK, Wren PA, Lichter PR, Musch DC, Gillespie BW, Guire KE, Mills RP, CIGTS SG. The collaborative initial glaucoma treatment study: Interim quality of life findings after initial medical or surgical treatment of glaucoma. Ophthalmology 2001 108, 1954-65.
12. Johnson R. Willingness to pay, willingness to wait and super QALYs. RTI Health Solutions Monograph. RTI Health Solutions. 2005.
13. Keeney RL, Raifa H. Decisions with multiple objectives: Preferences and value trade offs. New York: Wiley. 1976.
14. Lancsar E, Louviere JJ, Flynn T. Several methods to investigate relative attribute impact in stated preference experiments. Social Science & Medicine, 2007 64(8), 1738-1753.
15. Lee BL, Gutierrez P, Gordon M, Wilson MR, Cioffi GA, Ritch R, Sherwood M, Mangione CM. The glaucoma symptom scale. A brief index of glaucoma-specific symptoms. Arch Ophthalmol 1998 116, 861-6.
16. Louviere J, Swait J. Separating weights and scale values in conjoint tasks using best worst attribute scaling. Unpublished work 2001
17. Louviere J, Hensher D, Swait J. Stated preference modeling: Theory, methods and applications. Cambridge, UK: Cambridge University Press; 2000.

18. Lynch J. Uniqueness issues in the decompositional modelling of multiattribute overall evaluations. Journal of marketing research 1985 22, 1-9
19. Marley AAJ, Louviere JJ. Some probabilistic model of best, worst and best worst choices. Journal of mathematical psychology 2005 49, 464-480
20. McIntosh E. Using discrete choice experiments to value benefits of health care. PHD Thesis, University of Aberdeen 2003
21. McKenzie L, Cairns J, Osman L. Symptom-based outcome measures for asthma: The use of discrete choice methods to assess patient preferences. Health Policy 2001 57(3), 193-204.
22. Nelson P, Aspinall P, Papasouliotis O, Worton B, O'Brien C. Quality of life in glaucoma and its relationship with visual function. J Glaucoma 2003 12, 139-50.
23. Ritchie J, Spence L. Qualitative data analysis for applied policy research. In: Bryman A, Burgess R, eds. Analysing Qualitative Research. London: Routledge 1994, 173-194.
24. Ryan M, Netten A, Skatun D, Smith P. Using discrete choice experiments to estimate a preference-based measure of outcome - an application to social care for older people. J Health Econ 2006 25, 927-44.
25. Ryan M, San Miguel F. Revisiting the axiom of completeness in health care. Health Econ 2003 12(4), 295-307.
26. Ryan M, Gerard, K. Using discrete choice experiments to value health care: current practice and future prospects. Applied Health Economics and Policy Analysis 2003 2, 55-64.
27. San Migeul F, Ryan M, Amaya-Amaya M. 'Irrational' stated preferences: A quantitative and qualitative investigation. Health Economics 2005 14, 307-22.
28. StataCorp. STATA statistical software 9SE College Station, Texas: STATA Corporation; 2005.
29. Zanlonghi X, Arnould B, Bechetoille A, Baudouin C, Bron A, Denis P, Nordmann JP, Renard JP, Rigeade MC, Rouland JF, Sellem E. [Glaucoma and quality of life]. Journal Francais d Opthalmologie 2003 26(Spec 2), S39-44.