30/06/2006

Paper presented to the Health Economists Study Group Meeting, University of York, 26-28 July 2006

Exploring the Social Value of Health Interventions: A Stated Preference Discrete Choice Experiment.

Colin Green[1*], Karen Gerard[2]

[1] Wessex Institute for Health Research & Development, University of Southampton (c.green@soton.ac.uk)

[2] Health Services Research Group, School of Nursing & Midwifery, University of Southampton (kareng@soton.ac.uk)

- Work in progress: Not for citation without authors' permission -

Acknowledgements: We would like to thank the MORI Social Research Institute for undertaking data collection, especially Gregor Jackson. We would also like to thank colleagues for helpful comments. We would like to acknowledge financial support from the Wessex Institute for Health Research and Development, University of Southampton, which covered data collection costs.

1. Introduction

The paper is motivated by an interest in how society may wish to allocate health care resources when faced with difficult priority setting decisions over which healthcare interventions should be available within the NHS.

There is a growing body of work, referred to as 'empirical ethics' (Culyer 2001), which examines some common conceptions of equity, examining the equity vs. efficiency trade-off in a resource allocation context. However, a recent review of this literature by one of the authors (CG) finds the empirical evidence in this area to be of a largely methodological nature, and in many instances unhelpful to the decision maker. One prominent finding from this review was that the majority of studies in this area have looked at specific social values in isolation (e.g. severity, age, social class) and there is a need to undertake empirical work looking at the relationship between key social values.[1] The study presented in this paper seeks to do just that.

This paper presents the findings from a discrete choice experiment undertaken to elicit the preferences of a sample of the general public over alternative healthcare interventions. It is "work in progress" and represents a preliminary reporting of study methods, results, and the potential for findings to be used to inform health policy decisions.

2. Outline of the Discrete Choice Framework

Discrete choice experiments are a stated preference technique being increasingly used in health economics to elicit individual preferences (Ryan & Gerard 2003). A discrete choice experiment quantifies preferences by analysing responses individuals provide in surveys about how they would choose between alternatives in hypothetical scenarios. These surveys present respondents with a choice between alternative scenarios, i.e. alternative descriptions of a good or service (e.g. healthcare intervention), using a set of attributes. The attributes used are varied across a specified and plausible range of levels to describe the scenarios in the choice set.

In the health economics literature the discrete choice framework has mainly been used as an alternative approach for health benefit valuation and to value the broader outcomes associated with health care provision (e.g. non-health attributes, satisfaction with services), (Ryan & Gerard 2003). The vast majority of studies have elicited individual values from

[1] The review undertaken (as yet unpublished) builds on earlier reviews in this area, with these earlier reviews also highlighting the need for research to investigate the relative importance of different social values (decision making criteria), (e.g. Sassi et al 2001, Schwappach 2002, Dolan et al 2005).

respondents who have been asked to consider the choice context as it applies to them. However, there have been a small number of discrete choice studies that have explored preferences in a 'social' context, where respondents have been asked for their preferences when considering choices (i.e. resource allocation and priority setting choices) involving other people in society (e.g. Farrar et al 2000, Ratcliffe 2000, Schwappach 2003, Gyrd-Hansen 2004).

In general, the DCE approach involves a blend of economic theory, econometric methods and survey design theory (psychology and statistical design). The approach is broadly consistent with microeconomic theory of consumer behaviour, and it draws upon Lancaster's theory of consumer choice (Lancaster 1966), and random utility theory (RUT) (McFadden 1973, Hanemann 1984). These theoretical underpinnings provide a conceptual basis for the discrete choice experiment.

In discrete choice modelling (i.e. RUT), the utility function is composed of a deterministic component (V) and an unobservable or stochastic (random) component (ε):

$$U_{in} = V_{in} + \varepsilon_{in} \qquad (1)$$

where V_{in} is the deterministic component of the utility for individual n and option i; ε_{in} is the random or unobservable component for individual n and option i. V_{in} is the indirect utility function in which the attributes are arguments. V is therefore characterised as:

$$V_{in} = \sum_i \beta_i X_{in} + \sum_p \psi_p W_{pn} \qquad (2)$$

where V is the utility, X is a matrix of attribute levels, W is a vector of p individual characteristics, β_i are the coefficient estimates for each attribute in the matrix X (marginal utilities), and the ψ_p coefficients represent the extent to which personal characteristics influence choice.

Applying random utility theory it is possible to consider the probability of choosing option i over any other option j belonging to the same choice set (C). Assuming the ε term is independently and identically distributed (using a Gumbel distribution, Louviere et al, 2000), the logistic distribution and the conditional logit model can be used to derive the probability outcomes.

The conditional logit model takes the form[2]:

$$Pr(i) = e^{Vi} / \sum e^{Vj} \quad [C \text{ of } j=1..,n] \qquad (3)$$

3. The Discrete Choice Experiment: Eliciting Social Preferences for Different Healthcare Interventions

3.1 Methods

The aim of the discrete choice experiment was to explore the general relationship between key generic social values expected to inform health care priority setting choices.

The two prime research objectives of the study were:

1. To test the feasibility of undertaking discrete choice experiments to inform on the social preferences of the general public (in this healthcare context)

2. To provide some empirical evidence and insight to the way the general public may trade-off different social values (attributes) in the experiment undertaken.

The methods for the study are described here against the five main stages in the conduct of a discrete choice experiment; establishing the attributes and levels, selecting alternative scenarios to present to respondents (experimental design), questionnaire development (survey format), data collection and data analysis and interpretation.

3.1.1 Attributes and levels

The process of defining attributes and levels was informed by a detailed review of the health care literature (empirical studies on equity), a review of health policy documents (guidance published by NICE health technology appraisal programme), discussions with experts (clinical, public health, decision-maker, methodological), pre-pilot testing and a formal pilot study. The literature review undertaken included studies which had recently reported on focus group studies examining priority setting decision making (e.g. Cookson & Dolan 1999, Dolan & Cookson 2000). Pre-pilot research involved testing of potential attributes and levels. The formal pilot was used to test the level of understanding over the final attributes chosen for the survey.

---

[2] The conditional logit model is more precisely written as $Pr(i) = e^{\lambda Vi} / \sum e^{\lambda Vj}$, [C of j=1..,n], where λ is a scale parameter which is inversely related to the standard deviation of the error terms. This scale parameter affects the estimates of the β coefficients, and it varies with each model. Louviere et al (2000) and Louviere (2001) have discussed this scale parameter in some detail, highlighting that it is not possible to identify this scale parameter (due to its multiplicative form), and identifying it as a limitation of the assumptions of the econometric techniques used. However, Louviere (2000) has also stated that it is legitimate to set the λ scale parameter equal to one in those cases when single choice experiment models are estimated and reported

The attributes selected were generic attributes relevant in many resource allocation (priority setting) decisions.

Table 1 presents the attributes and levels used in the study. Early DCE design (tested in pre-pilot work) was more complex than that in the final version (Table 1) – e.g. considering attributes for category of health intervention, acute vs. chronic health condition, decision-making perspective – but pre-pilot research suggested that a simple format was preferable, especially as this form of survey was largely untested in a face-to-face interview format with the general public.

Table 1. Discrete choice experiment: Attributes and Levels

Attribute	Description	Level 1	Level 2	Level 3	Level 4
Severity	Whether patients are severely affected by their condition.	Yes	No	-	-
Health Improvement	The average health improvement expected from treatment.	Large	Moderate	Small	Very Small
Value for Money	Cost-effectiveness of treatment - the value for money expected from the treatment.	Very Good	Fairly Good	Fairly Poor	Very Poor
Other Treatment	Whether other effective treatments are available for the patient group.	Yes	No	-	-

Severity: ***Whether patients are severely affected by their condition*** *(Yes or No)*
Severity of health, the pre-treatment health state of patients, is identified in the current literature as a potential priority setting criterion. It is supported by respondents in a number of empirical studies (e.g. Nord 1993, 1995, Dolan 1998, Ubel 1999, Oddsson 2003, Cookson & Dolan 1999), and also in feedback from experts, and within pre-pilot findings.

For the purposes of the DCE survey undertaken here the EuroQol (EQ-5D) health classification system was used to articulate a meaning for severity that could be generalised across health states. The 3 levels of the EuroQol were considered analogous to mild, moderate and severe levels for each dimension, and level 3 descriptions were used as examples of how patients may be severely affected by their health condition.

Improvement in health: ***The average health improvement expected from treatment.***
(Large/Moderate/Small/Very Small)
Health gain is a prime candidate for consideration in priority setting choices. The objective of health care interventions is to deliver clinically effective treatment which provides an

improvement in the health of the patient. It is a commonly used objective in health care i.e. health (gain) maximisation. Although in practice health improvement may be closely linked to efficiency (e.g. in cost effectiveness analysis), it was clear from the literature review undertaken, and from pre-pilot research that the level of health improvement was a key generic issue in its own right.[3]

Value for money: ***The value for money expected from the treatment.*** *(Very Good/Fairly Good/Fairly Poor/Very Poor)*

The use of cost-effectiveness in health care is now widespread and efficiency is a well supported motive in the allocation of health resources. The NHS, through the NHS R&D HTA Programme, and NICE, place the cost effectiveness of treatment as a prime consideration for both the assessment and appraisal of interventions respectively. The use of value for money (cost effectiveness) in the survey was dictated by the policy context, as well as the policy documents in this area.

Following pre-pilot work, it was decided that the use of the language of economic evaluation, cost effectiveness analysis, with related efficiency terms for cost per life-year saved, or cost per QALY, were inappropriate for use in a survey of the general public. The terminology of "value for money" was used to express the notion of cost effectiveness and efficiency. This term was regarded as a commonly understood term, and very much related to efficiency and cost effectiveness. Pre-pilot work, and expert opinion, indicated that value for money was a reasonable and acceptable term to use to elicit preferences surrounding efficiency. The labels for the levels were tested in the pre-pilot work and the formal pilot.[4]

Other treatments: ***Whether other effective treatments are available for the patient group.*** *(Yes or No)*

Unlike the other attributes above, 'other treatments' is more of a contextual consideration, concerning the circumstances of the priority setting dilemma (non-health attribute). When reviewing the published NICE technology appraisal guidance, the availability of other treatments, or not, was often presented. In some specific instances the fact that the intervention under consideration was the only available effective treatment was noted explicitly (e.g. appraisal of riluzole for treatment of motor neurone disease, NICE 2001). Whilst there have been few empirical studies to investigate this issue, Williams et al (2005),

[3] Furthermore, small health gains may represent any of the value for money levels used (e.g. 'very good' to 'very poor'), as can large health gains.

[4] Qualitative levels of 'value for money' and 'health improvement' will be/can be mapped to quantitative levels, as per study by Bryan et al (2002).

in a study undertaken to investigate the use of economic evaluation in the process of health technology appraisal in the UK (within NICE), state that the availability of alternative treatments, or not, was a potential modifying factor in the health technology process witnessed at NICE (Williams et al 2005).

3.1.2 Experimental design

This study uses recent recommendations on experimental design for DCE studies (Street and Burgess, 2003, 2004). These recommendations stress the requirement of orthogonality in the design of DCE studies. The design used has a full factorial of 64 combinations of attributes and levels ($4^2 \times 2^4$). A suitable main effects fractional factorial design was used from a web-based catalogue (Sloan, 2003). This entailed the use of a fractional design with 16 scenarios, which Street and Burgess (2003, 2004) have proved to have the properties of near optimal (94.4%) statistical efficiency for parameter estimation for a model of main effects. This statistical efficiency is maintained with the pairing of scenarios into choice sets, using the 'foldover' design recommendations of Street and Burgess (2004). There was no correlation between attributes in the foldover design used, (and the attribute levels appeared the same number of times in each column). Two versions of the survey instrument were used, each with eight of the 16 pairwise choices. An additional orthogonal column from the statistical design was employed to split the 16 pairs into two blocks of eight questions (parameter estimates are therefore independent of the assigned blocks, Hensher et al 2005).

3.1.3 Questionnaire development

Framing of the Choices

The choices were framed in the context of an NHS priority setting choice/dilemma. Respondents were introduced to the fact that decision-makers in the NHS are often faced with difficult priority setting decisions. They were asked to consider a situation where they were in the position of a health care decision maker and were faced with some of these difficult decisions on how best to use its limited budget for the provision of health care services. In each of the DCE questions respondents were asked (in the context of health care decision-maker) to consider two unlabelled treatment options where there was only funding to support one of the options. This framing of the DCE questions was similar to that used successfully in a sample of the general public by Bryan and colleagues (2002) in their study examining the validity of the QALY maximisation objective. The questions were a 'forced choice' format, but interviewers had a number of response categories where a choice was not made (i.e. don't know/unable to choose, none, refused, other).

Pre-pilot research involved testing of early formats of background text to be used, the framing of the questions, attributes and levels, and showcards to be used in the main survey. A formal pilot was used to test the final format of the survey.

Appendices 1 and 2 present the format used for the DCE questions. A showcard was used to present the attribute descriptions and levels. The interviewer talked through each question (Appendix 1) using showcards (Appendix 2) for each pairwise choice.

The interview comprised an introduction by the interviewer (context, priority setting decisions, attributes to be used), a warm-up task whereby respondents were asked to rank the attributes in order of importance, a warm-up DCE question, a specific DCE consistency question, followed by eight further DCE questions per respondent. Respondents were asked to indicate how difficult they found it to (a) understand the questions, and (b) to answer the questions (using a Likert scale). At the design stage a mean interview time of 25-30 minutes was the target.

Where respondents answered the consistency question incorrectly the interviewer asked the respondents to explain their choice (in an attempt to investigate such inconsistencies).

3.1.4 Data collection

The main survey instrument was administered on a sample of the adult general population (aged 18 and over) in the Southampton (UK) City Council area. Interviews were carried out in a face-to-face format in-home by the MORI Social Research Institute during September - October 2005.

The adopted approach to sampling was a random location quota sample. This sets fixed quotas of people to be interviewed in a number of randomly selected sampling points. Sampling points were based on 'Output Areas' (OAs) in the Southampton City Council area, the smallest building block of the Census. For the sample used in this survey 32 OAs were randomly selected (by MORI) proportionate to population size, controlling for socio-demographic composition. Quotas were set – individually at each sample point – to reflect the socio-demographic profile of residents, on gender, age and work status, using profile data from the 2001 Census.

There were two versions of the main survey instrument, each contained 8 of the 16 DCE combinations from the efficient design used. Respondents were randomly assigned one of the two main survey instruments

There are no commonly used formal sample size calculation methods for this type of public preference study. Pearmain et al (1991) have reported that for DCE studies samples sizes circa. 100 are able to provide a basis for modelling of preference data. The survey aimed to interview 250 people. This target sample size was determined pragmatically on the basis of available resources. However, from a review of the literature on DCE experiments (and comments from methods experts) a sample size of 250 appears credible for a study of this type.

3.1.5 Preliminary Data analysis & interpretation

The data were modelled using a random utility maximisation framework (Louviere et al 2000, Hensher et al 2005), and STATA 8.1 software. The model set out in the design of the study is of the form:

$V = \beta_0 + \beta_1$ SEVERITY $+ \beta_2$ HEALTH IMPROVEMENT $+ \beta_3$ VALUE FOR MONEY $+\beta_4$ OTHER TREATMENTS

where the deterministic part of utility (V) is a function of the attribute levels between options and where the coefficients (part-worth utilities) β_1 to β_4 and constant β_0 are estimated in the model.

In the design used and the model presented the constant will represent the worst case scenario in the factorial design, across attributes and levels. That is, an intervention for 'non-severely' affected persons, where 'other treatments' are available, where the health improvement is 'very small', and the value for money is 'very poor'. The effect on utility (the desirability) of other scenarios will be relative to this base case option (constant). Dummy variables have been used (see Appendix 3) to account for this approach. This is helpful in interpreting the findings of the study.

A priori the coefficients are expected to have a positive sign, indicating an increase in utility (probability of being chosen for funding) with increases in the level of attributes, from the base case scenario. Where *a priori* relationships between choice and independent variables (attributes) hold, this is a good sign of theoretical validity. If the attributes and the constant

are important the coefficients estimated will show statistical significance, and therefore have an influence on respondents' likelihood of choosing the option. Coefficients are used to consider the relative importance per unit of change in the attributes. It is important that this is done carefully, and in the context of the qualitative changes in the attribute levels used.

In the current analysis the model is used to consider main effects. It is important to consider *a priori* where subgroups, defined according to sample characteristics, are thought to have a potential impact on choice. These subgroup impacts are termed interaction effects, i.e. the characteristics interact with the attribute or main effects. A common consideration in discrete choice models is effect of the income of respondents on the choice made, especially where attributes consider willingness to pay. However, *a priori* income level was not considered to be an interaction term in the current analysis as the choice is unrelated to respondent income. In the analysis here, *a priori* there is a view that the age of respondents, their health status and their experience of illness, may have an impact on the response data (interaction effect). These factors are investigated using sub-group analysis, applying a log likelihood ratio test. Where this indicates a significant interaction further investigation is carried out in the discrete choice model.

In the current analysis a fixed-effects conditional logit model is used (STATA), with some exploration of the panel data using a random effects Probit model. In future analyses a random parameter logit model will be explored as a better way of accounting for random effects. However, in the short term it is accepted that the assumption of independent errors may underestimate the standard errors in the data output, and impact on statistical significance of the beta coefficients.

Summary statistics are presented to provide information on the quality of the estimated model; these statistics represent a standard set of measures to reflect overall model significance (likelihood ratio test), goodness of fit (pseudo R^2 e.g. McFaddens R^2), and predictive capability.

3.2 Results

3.2.1 Response rate

259 of 263 respondents in the survey provided preferences when presented with the DCE survey question. There were 4 respondents who did not provide DCE responses for any of the 8 DCE questions, indicating 'don't know' (or with response not stated) on the questionnaire. These 4 respondents are excluded from analyses.

In response to a 'consistency check' question in the survey, 13 respondents (5%) chose the option that was dominated. Although asked to explain their reason for choosing this option such qualitative data did not prove helpful. However, other than in the four cases above, the response data from these respondents has been included in the analysis.

3.2.2 Characteristics of the sample

Appendix 4 presents a summary of the socio-demographic characteristics of the sample. These data indicate a reasonably representative sample of the Southampton City Council geographic area. The sample did have a larger proportion of retired and home workers, than in the population at large, which may be expected in an 'in-home' interview survey of this type.

There were no differences in socio-demographic characteristics across the two versions of the questionnaire used (Chi^2 tests).

3.2.3 Data analysis

Results from the main effects conditional logit model are presented in Table 2. Model A reports findings for a main effects only model, whilst Model B includes one interaction variable.

As expected all main effects are positive, showing an increase in utility where there is a change in the attribute level, from the base case. Where dummy variables have been used for health improvement and value for money the staging of the impact is as expected over the incremental changes in the attribute (e.g. $\beta_3 < \beta_4$). The main effects are also all statistically significant (at the p=0.05 level or below). These findings present some indication of the theoretical validity of the model.

Sub-group analysis for age and experience of illness showed no significant difference between findings. For self-reported health status there was a significant difference (in sub-group) and those respondents who stated a bad health status (either bad or very bad) appeared to place less weight on the 'other treatments' attribute, (the attribute indicating that no other effective treatments were available). The number of respondents with a self-reported health status of bad/very bad was small, only 9.3%. This reflected a similar proportion to that reported in the Southampton City Council area in the 2001 Census (8.6%). There was no significant interaction effect with health status for other attributes in the model.

The introduction of the interaction effect (bad health status *x* other treatment) marginally improved the robustness of the model. On this basis we suggest Model B is the most

appropriate (at the present time) and is the one used in the remainder of this paper. The findings show that statistical significance of the coefficient for the 'other treatment' attribute improved and increased the effect of that attribute when taking into account the proportion of persons in the sample/population with self-reported 'bad' health (i.e. coefficient *x* proportion).

Table 2. Discrete Choice Model Results

Input / Attribute (level)	**MODEL A Main Effects** Coefficient	**MODEL B Main effects with interactions** Coefficient (OR)
Constant (β_0)	0.3783 *	0.3783 * (1.46)
Severely affected (β_1)	0.5690 *	0.5581 * (1.75)
No other treatments (β_2)	0.1015 **	0.1482 * (1.16)
Small health improvement (β_3)	0.3759 *	0.3838 * (1.47)
Moderate health improvement (β_4)	0.8561 *	0.8619 * (2.37)
Large health improvement (β_5)	1.4775 *	1.4713 * (4.35)
Fairly poor vfm (β_6)	0.1823 **	0.2130 ** (1.24)
Fairly good vfm (β_7)	0.9359 *	0.9543 * (2.60)
Very good vfm (β_8)	1.2709 *	1.2756 * (3.58)
Interaction: Health status (bad) x No other treatments (β_9)	N/A	-0.4180 ** (0.66)
Summary Statistics:		
Log-likelihood	1127.74	1107.48
Model Chi^2 (df)	554 (9)	550 (10)
Pseudo R^2	0.197	0.199
% correct predictions	74%	74%

* p=0.01, ** p=0.05

The most important attribute is that describing the level of health improvement, followed by value for money, with severity the next important and other treatments being the least important of the attributes. The interpretation of the importance of each attribute by level is important relative to the base case (constant), or a more appropriate comparator. For example severity (treating a severely affected patient group) is more important than a change from very poor value for money to fairly poor value for money, or from a very small health improvement to a small health improvement. It can also be interpreted as being more important that a change from a small to a moderate health improvement, and from fairly good to very good value for money.

The most important single increment in utility is that from (fairly) poor value for money to (fairly) good value for money, with an increase in utility of 0.74. Interpreting the coefficients as odds ratios gives a clear view of the relative importance of incremental changes from the base scenario (worst case scenario), see Figure 1.

3.3.3 Using the results to predict utility levels and the probability of success

The utility function for each of the 64 scenarios (full factorial) can be estimated. Results (Table 3) showed utility scores ranged from a high of 3.79 for the best case scenario to a low of 0.34 for the worst case scenario. This data provides a relative measure of the desirability of the alternative scenarios, with the utility scale used to consider interval scale properties. Furthermore, using the conditional logit model (equation 3) it is possible to estimate a probability of success (i.e. being selected for funding) for each of the scenarios, relative to one another (i.e. probability of being selected from the total group of 64 options). The probabilities for the 64 scenarios sum to 1.00. In this context the probability scale is arguably more suitable for use as an absolute measure of desirability across the alternatives, and we suggest it can be used as a more appropriate measure of the strength of preference across scenarios. Table 3 presents the estimates of probability, interpreted as a measure of strength of preference (to be easier on the eye probabilities have been multiplied by 1,000). Table 3 presents the 64 scenarios in a ranking according to the measure of preference (probability of being funded). It shows that in over 70% of cases the top 25 scenarios would be chosen. The ten most preferred scenarios account for 42% of the probability of being funded. When considering the 25 'least preferred' scenarios there is an aggregate probability of 13% that one of those scenarios would be funded (given the context of the current study).

Figure 1. Model results: Odds ratios

(impact on the odds of success – being funded – relative to the worst case scenario)

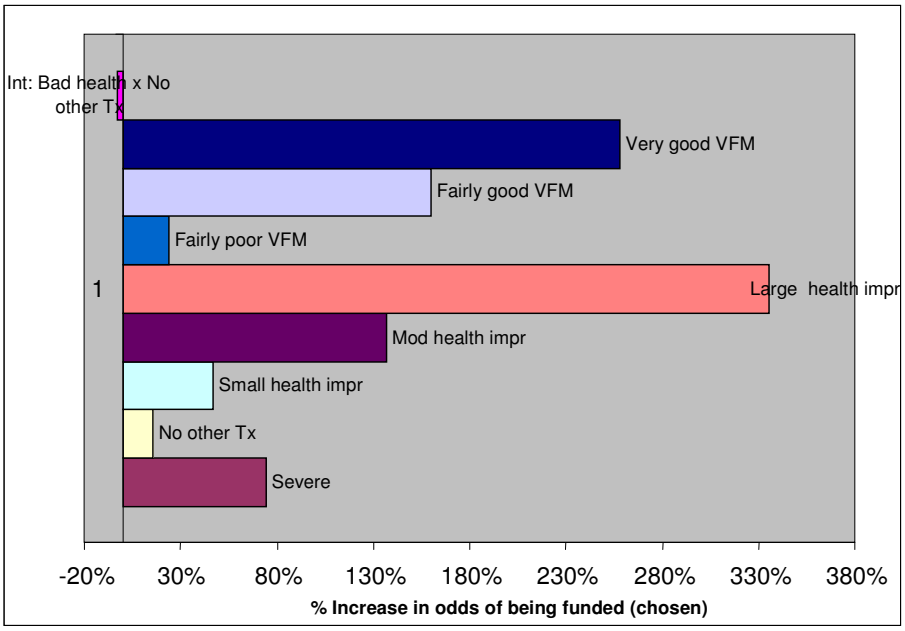

Table 3. Estimated utility score and preference weight by scenario (full factorial)

Ranked by pref weight	Attributes (scenarios): Value for Money	Health Improvement	Severity	Other Treatments	Utility (logit)	Pref Weight*	*Cumm. Pref. weight*	Public Preference (tentative judgement)
1	very good	large	severe	No Tx	3.79	68.87	*6.89%*	Very Strong
2	very good	large	severe	Other Tx	3.64	59.38	*12.83%*	Very Strong
3	fairly good	large	severe	No Tx	3.47	49.94	*17.82%*	Very Strong
4	fairly good	large	severe	Other Tx	3.32	43.06	*22.13%*	Very Strong
5	very good	large	not severe	No Tx	3.23	39.41	*26.07%*	Very Strong
6	very good	moderate	severe	No Tx	3.18	37.44	*29.81%*	Very Strong
7	very good	large	not severe	Other Tx	3.09	33.98	*33.21%*	Very Strong
8	very good	moderate	severe	Other Tx	3.04	32.28	*36.44%*	Very Strong
9	fairly good	large	not severe	No Tx	2.91	28.58	*39.29%*	Very Strong
10	fairly good	moderate	severe	No Tx	2.86	27.15	*42.01%*	Very Strong
11	fairly good	large	not severe	Other Tx	2.77	24.64	*44.47%*	Very Strong
12	fairly poor	large	severe	No Tx	2.73	23.80	*46.85%*	Very Strong
13	fairly good	moderate	severe	Other Tx	2.71	23.41	*49.20%*	Very Strong
14	very good	small	severe	No Tx	2.71	23.21	*51.52%*	Fairly Strong
15	very good	moderate	severe	Other Tx	2.63	21.43	*53.66%*	Fairly Strong
16	fairly poor	large	not severe	No Tx	2.58	20.52	*55.71%*	Fairly Strong
17	very good	small	severe	Other Tx	2.56	20.02	*57.71%*	Fairly Strong
18	very poor	large	severe	No Tx	2.52	19.23	*59.64%*	Fairly Strong
19	very good	moderate	not severe	Other Tx	2.48	18.47	*61.48%*	Fairly Strong
20	fairly good	small	severe	No Tx	2.38	16.83	*63.17%*	Fairly Strong
21	very poor	large	severe	Other Tx	2.37	16.58	*64.82%*	Fairly Strong
22	very good	very small	severe	No Tx	2.32	15.81	*66.41%*	Fairly Strong
23	fairly good	moderate	not severe	No Tx	2.30	15.54	*67.96%*	Fairly Strong
24	fairly good	small	severe	Other Tx	2.24	14.51	*69.41%*	Fairly Weak
25	very good	very small	severe	Other Tx	2.17	13.64	*70.77%*	Fairly Weak
26	fairly poor	large	not severe	No Tx	2.17	13.62	*72.14%*	Fairly Weak
27	fairly good	moderate	not severe	Other Tx	2.16	13.40	*73.48%*	Fairly Weak
28	very good	small	not severe	No Tx	2.15	13.28	*74.80%*	Fairly Weak
29	fairly poor	moderate	severe	No Tx	2.12	12.94	*76.10%*	Fairly Weak
30	fairly poor	large	not severe	Other Tx	2.02	11.74	*77.27%*	Fairly Weak
31	fairly good	very small	severe	No Tx	2.00	11.47	*78.42%*	Fairly Weak
32	very good	small	not severe	Other Tx	2.00	11.45	*79.57%*	Fairly Weak
33	fairly poor	moderate	severe	Other Tx	1.97	11.16	*80.68%*	Fairly Weak
34	very poor	large	not severe	No Tx	1.96	11.01	*81.78%*	Fairly Weak
35	very poor	moderate	severe	No Tx	1.91	10.46	*82.83%*	Fairly Weak
36	fairly good	very small	severe	Other Tx	1.85	9.89	*83.82%*	Fairly Weak
37	fairly good	small	not severe	No Tx	1.83	9.63	*84.78%*	Fairly Weak
38	very poor	large	not severe	Other Tx	1.81	9.49	*85.73%*	Fairly Weak
39	very good	very small	not severe	No Tx	1.76	9.05	*86.63%*	Fairly Weak
40	very poor	moderate	severe	Other Tx	1.76	9.02	*87.53%*	Fairly Weak
41	fairly good	small	not severe	Other Tx	1.68	8.31	*88.36%*	Fairly Weak
42	fairly poor	small	severe	No Tx	1.64	8.02	*89.17%*	Fairly Weak
43	very good	very small	not severe	Other Tx	1.62	7.80	*89.95%*	Fairly Weak
44	fairly poor	moderate	not severe	No Tx	1.56	7.40	*90.69%*	Very Weak
45	fairly poor	small	severe	Other Tx	1.49	6.92	*91.38%*	Very Weak
46	fairly good	very small	not severe	No Tx	1.44	6.56	*92.04%*	Very Weak
47	very poor	small	severe	No Tx	1.43	6.48	*92.68%*	Very Weak
48	fairly poor	moderate	not severe	Other Tx	1.41	6.38	*93.32%*	Very Weak
49	very poor	moderate	not severe	No Tx	1.35	5.98	*93.92%*	Very Weak
50	fairly good	very small	not severe	Other Tx	1.29	5.66	*94.49%*	Very Weak
51	very poor	small	severe	Other Tx	1.28	5.59	*95.05%*	Very Weak
52	fairly poor	very small	severe	No Tx	1.26	5.46	*95.59%*	Very Weak
53	very poor	moderate	not severe	Other Tx	1.20	5.16	*96.11%*	Very Weak
54	fairly poor	very small	severe	Other Tx	1.11	4.71	*96.58%*	Very Weak
55	fairly poor	small	not severe	No Tx	1.08	4.59	*97.04%*	Very Weak
56	very poor	very small	severe	No Tx	1.05	4.42	*97.48%*	Very Weak
57	fairly poor	small	not severe	Other Tx	0.94	3.96	*97.88%*	Very Weak
58	very poor	very small	severe	Other Tx	0.90	3.81	*98.26%*	Very Weak
59	very poor	small	not severe	No Tx	0.87	3.71	*98.63%*	Very Weak
60	very poor	small	not severe	Other Tx	0.72	3.20	*98.95%*	Very Weak
61	fairly poor	very small	not severe	No Tx	0.70	3.13	*99.26%*	Very Weak
62	fairly poor	very small	not severe	Other Tx	0.55	2.70	*99.53%*	Very Weak
63	very poor	very small	not severe	No Tx	0.49	2.53	*99.78%*	Very Weak
64	very poor	very small	not severe	Other Tx	0.34	2.18	*100.00%*	Very Weak

* preference weight = probability (equation [3]) multiplied by 1,000

Table 4 presents the finding from the ranking of attributes by respondents (ranked in order of importance 1st to 4th). The ranking task was primarily a warm-up task to ensure respondents had familiarity with the attributes/levels used, and to offer an opportunity for respondents to ask questions on the attributes where there may have been any misunderstanding. There were no statistically significant differences (*Chi*^{2} test) in ranking by questionnaire version (1 & 2), or by respondent characteristic (e.g. social class, work status, household income), other than for ethnic status. With ethnic status there was significant difference in mean rank for severity of health (*Chi*^{2} *p*-value=0.01), however there was a large difference in numbers in each of these groups (n=216 vs n=40).

Table 4. Percentages of respondents ranking attributes 1st to 4th

	Rank			
Attribute *(mean rank)*	1st	2nd	3rd	4th
Severity *(1.87)*	52.7	18.4	18	10.9
Health Improvement *(1.83)*	40.3	41.1	14	4.7
Value for Money *(3.22)*	9.8	13.7	21.1	55.5
Other Treatments Available *(2.78)*	8.6	24.7	46.7	20

Table 5 presents the data on difficulty respondents had in understanding and answering the DCE questions posed. The majority of respondents reported that they found the questions not very or not at all difficult to understand. Interestingly, and perhaps reassuringly, the majority of respondents did report that the questions were very or fairly difficult to answer.

Table 5. Difficulty understanding and answering DCE questions

Level of Difficulty with Questions	Difficulty "Understanding"	Difficulty "Answering"
Very difficult	10.8%	20.8%
Fairly difficult	29.7%	47.5%
Not very difficult	31.7%	23.6%
Not at all difficult	26.6%	6.6%
Not stated/missing	1.2%	0.8%
Total	100% (n=259)	100% (n=259)

There were no statistically significant differences (*Chi*^{2} test) in responses to the difficulty questions by questionnaire version (1 & 2). Generally there were no significant differences in response to difficulty questions by respondent characteristic. Although there were some

significant differences in understanding the questions by age, with the elderly likely to find questions more difficult to understand, and differences by self reported health status. For health status groups those in bad or very bad health were more likely (16 of 24) to have difficulty understanding the questions (p-value=0.011).[5] There were no statistically significant differences (nor a tendency towards a difference) between subgroups in terms of difficulty answering the questions.

4. Discussion and future work

This study has both methodological and empirical objectives. It tests the DCE methodology in the general public when priority setting choices (including value for money) are presented, and presents findings from a sample of the general public on how they trade-off attributes in the context of the priority setting choices presented.

The study suggests (demonstrates) that it is feasible to present priority setting choices using the DCE framework to samples of the general public. Whilst there is no objective data on the numbers who refused to participate in the study, given the door-to-door nature of the survey, the feedback from interviewers in the pilot study and the main survey was that respondents were keen to participate, seemed engaged in the survey, and generally had few problems completing the survey (accepting the fact that choices were difficult ones). There were only a small number of respondents who 'failed' the consistency question, and the majority of respondents reported little or no difficulty in understanding the questions posed (although circa. 30% did report it was fairly difficult to understand the questions). Whilst respondents indicated that the questions were difficult to answer, it would have been more of a concern if they had not indicated such difficulty.

The study provides data on the relative value of the different attributes, when considering changes across the different healthcare scenarios presented. All attributes used in the study significantly influenced respondents' preferences over health care scenarios. The most important attribute/level changes involved level of health improvement and value for money. The findings show that the general public hold 'severity of health condition' and 'the availability of other treatments' as important social values, not to be ignored, and this provides some support to those arguing against an often dominant 'health maximising' decision maker perspective. However, the findings also indicate that in many instances the level of health improvement and value for money arguments provide a strong indication of the social value (preference) associated with a health care intervention.

[5] This may, in part, explain the interaction effect noted in Table 2.

Of some interest is the difference we see in the ranking of the value for money attribute, by level of importance, and the impact of the attribute on the discrete choices made. Whilst the ranking task was primarily a warm-up task, the difference we see in this attribute particularly, and to some extent in the severity attribute, suggest some thought should be given to the relative merits of the two different tasks i.e. ranking versus attribute based choices. Other studies have also reported such differences. For example, Schwappach (2003) reports differences between ratings of importance and the actual impact of attributes on choices. Schwappach reports that the majority of respondents in his study stated that social class was of "no or little" importance, but this attribute contributed significantly to the allocation decisions in the discrete choice study reported. Such evidence suggests that where respondents are asked direct questions, such as 'is value for money important in priority setting decisions?' they may give the answer they feel is expected (i.e. 'no' or 'of course not'). Yet, when faced with choices, which involve a sacrifice (and may force some deliberation), they do indeed see value for money as important. It is clear that this may be the case with an issue like social class in some instances (political correctness). It may also be an explanation of the conflicting findings in some of the 'empirical ethics' literature around issues such as age and desert. Further consideration of these issues is needed, in the context of the current study.

One of the most interesting and evolving aspects of the current paper (we feel) is the attempt to make a preference judgement across all of the different scenarios in the full factorial, and to use the probability of success (choice) from the complete choice set (n=64) as a measure of strength of preference. Assuming the probability scale has ratio-scale measurement properties, it affords an opportunity for a true strength of preference measure. This area of the paper is the one where we would especially request some comment from the HESG participants.

Table 3 presents the utility function for each of the 64 scenarios, it estimates the probability of success (using equation 3) and ranks the scenarios from most preferred to least preferred. Reassuringly, the ranking seems to be a fairly sensible one, without any obviously inconsistent rankings (i.e. we do not see any scenarios that appear to be in the 'wrong' place). We accept that there is a fair degree of subjective judgement involved – but life is often not simple. We have even gone as far as suggesting a category of 'strength of preference' for each scenario (from very strong to very weak, see Table 3). This is based on a wish to avoid a specific numeric measure of preference (or giving the impression of being too scientific). The study seeks to give a general feel for the relative value of attributes, and scenarios, and does not seek to be too specific or definitive on the magnitude of differences

between the scenarios described. However, there needs to be some form of cut-off between the preference categories – therefore we have suggested some in the first instance.
In Table 3 the scenarios which account for the top 50% of scenarios likely to be chosen (13 of them) are judged to be very strongly preferred for funding by the general public. Those scenarios with a preference weight greater than the average weight (1.55% [1 ÷ 64), or preference weight of 15.5) are judged to be fairly strongly preferred for funding. The scenarios which are in the bottom half of the ranking (accounting for the bottom 10% of cumulative probability/preference), the least likely to be chosen are judged to be very weakly preferred. Those scenarios below an average probability and outside of the very weak preference category are judged to be fairly weakly preferred. To illustrate how the data may be transferred to a decision making perspective we map some potential health technology appraisal topics to the attributes, levels, and preferences presented in Table 3. Table 6 presents some illustrative (quickly put together) examples of health care interventions. We accept that this area of our work (the interpretation of findings) is a key area for future effort, and deserving of greater attention than given here, but use the present examples for illustration (and to encourage discussion).

Table 6. Preference data applied to a health policy context

Health Technology	Patients severely affected	Other treatments available	Health improve-ment	Value for money	Judgment on public preference*
Drugs for Alzheimer's disease	Yes	No	Small	Very poor *(Fairly poor)*	Very Weak *(Fairly Weak)*
Insulin Pumps for Type 1 diabetes	No	Yes	Small	Fairly poor	Very Weak
Drotrecogin alfa (activated) for severe sepsis	Yes	No	Moderate	Very Good (fairly good)	Very Strong
Riluzole for Motor Neurone Disease	Yes	No	Small	Very Poor *(Fairly Poor)*	Very Weak *(Fairly Weak)*

* Judgment on public preference i.e. *should it be funded?* (Using authors judgment/interpretation of levels/attributes for each of the technologies)

In future work we plan to consider in further detail how the data on general public preference may be used in a policy relevant way, such as one input to the NICE health technology appraisal process. We also plan to re-analyse the discrete choice data using quantitative levels for the health improvement and value for money attributes, transposing quantitative levels on the qualitative levels used in the survey, for purposes of data analysis. This is similar to the approach taken by Bryan et al (2002), in their handling of the quality of life attribute.

There are obvious limitations with the study. Some related to the fact that the paper presents a 'work in progress' view of the data, and data (implications) require further investigation (hopefully assisted with comments from HESG participants). Some limitations are linked to the presentation of the tasks, and the framing of the questions. Such issues are present in all studies of this nature (and will no doubt draw [and deserve] some further comment here). The study design has been deliberately kept to a simple approach, with a set of attributes that are thought to be key generic social values (in priority setting context), with levels that whilst plausible are kept to a minimum. There may be other attributes, and/or other levels, that could be used in similar surveys. At the outset a simple matrix was set out (full factorial of 64) for the attributes and levels, whereby the authors felt it was sufficient for health policy decision-makers to be able to map onto that matrix the scenarios that they faced in a real-time decision-making manner. But importantly we wished to have a full factorial that did not present a vast matrix (of scenarios) when it came to matching health interventions to the range of options in the full factorial. At a practical level it seems intuitively appealing that the full factorial, as presented in this paper, may be viewed on one page of A4 (it is easy in discrete choice designs to get to full factorials in excess of 500 scenarios). We feel that most decision-makers (and others) are able to make a judgement on the attribute levels that are appropriate for the health care interventions that may be under consideration by them for appraisal/funding decisions. In doing so they may get some help from the data presented on the preferences of the general public for such generic scenarios (matched to healthcare interventions). Obviously such data will need to be used in combination with other information available to the decision maker.

Among the many possible further issues we plan to give priority in future work to developing the data analysis and the interpretation of findings in a policy context. As such we particularly welcome your comments on these aspects of the study.

Appendix 1. Example of DCE question

Question:	SHOWCARD A health care decision maker is faced with difficult choices on how to allocate its budget. Imagine a choice where there are two options for the use of available funds. Given that only one of the options can receive funding, which option would you support?
Option K	Patients are severely affected by their condition. With treatment the average patient has a very small improvement in their health. The treatment is regarded as being very poor value for money. There are no other effective treatment options available.
Option L	Patients are not severely affected by their condition. With treatment the average patient has a large improvement in their health. The treatment is regarded as being very good value for money. There are other effective treatment options available.

Appendix 2. Example of Survey Showcard (for the above question)

Health condition: Long-term health condition

	OPTION K	OPTION L
Severity of patients	Severely affected	Not severely affected
Improvement in health	Very small improvement in health	Large improvement in health
Does the treatment offer Value for money	Very poor value for money	Very good value for money
Other effective treatment options available	No	Yes

You are reminded that only one of these options can be provided from the limited funds available.

A decision has to be made by the health care provider.

Appendix 3. Dummy coding of attributes in data analysis

Attribute	Levels	Variable (1)	Variable (2)	Variable (3)
Severity	No	0	–	–
(severely affected by condition)	Yes	1	–	–
Health Improvement	Very small	0	0	0
	Small	0	0	1
	Moderate	0	1	0
	Large	1	0	0
Value for Money	Very poor	0	0	0
	Fairly poor	0	0	1
	Fairly good	0	1	0
	Very good	1	0	0
Other Treatments Available	Yes	0	–	–
	No	1	–	–
Interaction terms:				
Health status	Good or fair	0	–	–
	Bad (v-bad)	1	–	–

Appendix 4: Socio-demographic characteristics of the sample used

	Number of People	%
Total	259	100
Gender		
Male	117	45.2
Female	142	54.8
Age		
18-34	90	34.7
35-54	89	34.4
55+	80	30.9
Social grade		
AB	36	13.9
C1	68	26.3
C2	65	25.1
DE	84	32.4
Work Status		
Working full time	92	35.5
Not full time		
Working part-time	41	15.8
House person	23	8.9
Retired	60	23.2
Registered unemployed	5	1.9
Unemployed (not registered)	4	1.5
Permanently sick/disabled	14	5.4
Student	20	7.7
Household income		
Below £17,500	19	7.3
£17,500 - £29,999	46	17.8
£30,000 - £49,999	38	14.7
Above £50,000	16	6.2
Refused/not stated	68	26.3
Ethnicity		
White British	218	84.2
BME[1]/other	41	15.8
Home ownership		
Owner occupier	147	56.8
Social renter	60	23.2
Private renter	46	17.8
Illness/disability		
Yes[2]	103	39.8
No	148	57.1
Household composition		
With children	88	34.0
Without children	170	65.6
Health in general		
Good or very good	185	71.4
Fair	46	17.8
Bad or very bad	24	9.3
Health Insurance		
Yes	36	13.9

[1] Black and minority ethnic
[2] Respondent and/or someone else in household

Bibliography

Bryan S, Roberts T, Heginbotham C, McCallum A. QALY-maximisation and public preferences: results from a general population survey. *Health Econ* 2002 Dec;11(8):679-93.

Cookson R, Dolan P. Public views on health care rationing: a group discussion study. *Health Policy* 1999;49(1-2):63-74.

Culyer AJ. Equity - Some theory and its policy implications. *J Med Ethics* 2001;. 27(4):275-83.

Dolan P. The measurement of individual utility and social welfare. *J Health Econ* 1998 Jan 1;17(1):39-52.

Dolan P, Cookson R, Ferguson B. Effect of discussion and deliberation on the public's views of priority setting in health care: focus group study. BMJ 1999 Apr 3;318(7188):916-9.

Dolan P, Cookson R. A qualitative study of the extent to which health gain matters when choosing between groups of patients. *Health Policy* 2000;51(1):19-30.

Dolan P, Shaw R, Tsuchiya A, Williams A. QALY maximisation and peoples preferences: a methodological review of the literature. *Health Econ* 2005;14(2):197-208.

Farrar S, Ryan M, Ross D, Ludbrook A. Using discrete choice modelling in priority setting: an application to clinical service developments. *Soc Sci Med* 2000;50(1):63-75.

Gyrd-Hansen D. Investigating the social value of health changes. Journal of Health Economics 2004;23:1101-1116.

Hanemann WM. Welfare evaluations in contingent valuation experiments with discrete responses. American Journal of Agricultural Economics 1984;66:332-341.

Hensher DA, Rose JM, Greene WH. *Applied Choice Analysis. A Primer.* Cambridge University Press; 2005.

Lancaster K. A new approach to consumer theory. Journal of Political Economy 1966;74:132-157.

Louviere.J., Hensher D, Swait J. *Stated Choice Methods: Analysis and Applications*. Cambridge: Cambridge University Press; 2000.

McFadden D. Conditional logit analysis of qualitative choice behaviour. In P Zarembka (Ed) *Frontiers in Econometrics*. New York: Academic Press. 1974

National Institute for Clinical Excellence. *Guidance on the use of Riluzole (Rilutek) for the treatment of Motor Neurone Disease*. NICE: London. 2001(www.nice.org).

Nord E. The trade-off between severity of illness and treatment effect in cost- value analysis of health care. *Health Policy* 1993;. 24(3):227-38.

Nord E. The Relevance of Health State After Treatment in Prioritizing Between Different Patients. *J Med Ethics* 1993;19(1):37-42.

Nord E. The person-trade-off approach to valuing health care programs. *Med Decis Making* 1995;. 15(3):201-8.

Nord E, Richardson J, Street A, Kuhse H, Singer P. Maximizing health benefits vs egalitarianism: An Australian survey of health issues. *Soc Sci Med* 1995;. 41(10):1429-37.

Nord E. Concerns for the worse off: Fair innings versus severity. *Soc Sci Med* 2005;. 60(2):257-63.

Oddsson K. Assessing attitude towards prioritizing in healthcare in Iceland. *Health Policy* 2003 Nov;66(2):135-46.

Pearmain D, Swanson J, Kroes E, Bradley M. *Stated preference techniques. A guide to practice*. Richmond, Surrey: The Haag, Netherlands: Steer, Davis, Gleave and Hague Consulting Group; 1991.

Ratcliffe J. Public preferences for the allocation of donor liver grafts for transplantation. *Health Econ* 2000 Mar;9(2):137-48.

Roberts T, Bryan S, Heginbotham C, McCallum A. Public involvement in health care priority setting: an economic perspective. *Health Expect* 1999 Dec;2(4):235-44.

Ryan M, Gerard K. Using discrete choice experiments to value health care programmes: current practice and future research reflections. *Applied Health Economics and Health Policy* 2003;2(1):1-10.

Sassi F, Archard L, Le GJ. *Equity and the economic evaluation of healthcare*. Health Technology Assessment (Winchester, England) 2001;. 5(3).

Schwappach DL. Resource allocation, social values and the QALY: a review of the debate and empirical evidence. *Health Expect* 2002 Sep;5(3):210-22.

Schwappach DLB. Does it matter who you are or what you gain? An experimental study of preferences for resource allocation. *Health Econ* 2003;12(4):255-67.

Sloan N. *A Library of Orthogonal Arrays*. http://www research att com/~njas/oadir/ (last accessed January 2006) 2003. Available from: URL: www.research.att.com/~njas/oadir

Street D, Burgess L. *Quick and easy choice sets: using word processing packages to construct near-optimal stated choice experiments*. Sydney: University of Technology; 2003.

Street D, Burgess L. Optimal and near optimal pairs for the estimation of effects in the 2-level choice experiment. *Journal of Statistical Planning and Inference* 2004;118:185-99.

Ubel PA. How stable are people's preferences for giving priority to severely ill patients? *Soc Sci Med* 1999;49(7):895-903.

Williams I, Bryan S, McIver S. The use of cost effectiveness analysis in technology coverage decisions: evidence from the UK. Abstract (oral presentation) at *Society for Medical Decision Making Annual Meeting*, San Francisco, October 2005.