THE UNIVERSITY OF
MELBOURNE

Melbourne Institute Working Paper Series

Working Paper No. 24/06

Designing Choice Experiments with Many Attributes: An Application to Setting Priorities for Orthopaedic Waiting Lists

Julia Witt, Anthony Scott and Richard H. Osborne

MELBOURNE INSTITUTE
of Applied Economic and Social Research

Designing Choice Experiments with Many Attributes: An Application to Setting Priorities for Orthopaedic Waiting Lists*

Julia Witt[†], Anthony Scott[†] and Richard H. Osborne[#]

[†] **Melbourne Institute of Applied Economic and Social Research,**
The University of Melbourne

[#] **Centre for Rheumatic Diseases, Department of Medicine, Royal Melbourne Hospital,**
The University of Melbourne

Melbourne Institute Working Paper No. 24/06

ISSN 1328-4991 (Print)

ISSN 1447-5863 (Online)

ISBN 0 7340 3225 0

October 2006

* This study was funded by the Victorian Department of Human Services. Dr. Osborne was supported in part by the Baker Trust, Buckland Foundation, the Arthritis Foundation of Victoria, and a National Health and Medical Research Council Career Development Award. The authors wish to thank Prof. Stephen Graves, Peta Chubb, Kerry Haynes, Mr Richard de Steiger, Mr Stan O'Loughlin, Mr Chris Haw, Mr Ian Critchley and the Orthopaedic Association of Australia (Victorian branch). Thanks also to John Creedy for comments and to seminar participants at the Universities of Aberdeen, Newcastle and Melbourne, particularly Mandy Ryan, Mabel Amaya-Amaya, and Emily Lancsar.

Melbourne Institute of Applied Economic and Social Research
The University of Melbourne
Victoria 3010 Australia
***Telephone* (03) 8344 2100**
***Fax* (03) 8344 2111**
***Email* melb-inst@unimelb.edu.au**
***WWW Address* http://www.melbourneinstitute.com**

Abstract

Stated preference discrete choice experiments are being increasingly used to value the quality of health care services. To date in the health economics literature, discrete choice experiments have used only a relatively small number of attributes due to concerns about task complexity, non-compensatory decision rules, simplicity of experimental designs, and the costs of surveys. This may lead to omitted variable bias and reduced explanatory power when attributes have been pre-selected from a longer list. There may be situations where it is desirable to include a longer list of attributes, such as attaching weights to quality of life instruments to obtain single index scores. The aim of this paper is to examine the feasibility of using a 'blocked attribute' design in a DCE with 11 attributes. This method allocates the 11 attributes across three separate experimental designs and pools the data for analysis. We examine this issue in the context of attaching weights to a disease specific quality of life instrument used to prioritise orthopaedic waiting lists in Victorian hospitals. We produce a single index measure of utility for health states of patients, bounded between zero and one. The use of such a design seems feasible, although issues remain to be resolved about how the ranking should be used in practice to set priorities for waiting lists.

1. Introduction

Stated preference choice experiments are now regularly used to value the quality of health care services and examine individuals' preferences over different 'packages' of health care (Ryan and Gerard, 2003). Respondents are presented with choice sets in which they choose between different scenarios (alternatives) consisting of several attributes that describe each scenario, with levels of the attributes varying between scenarios. The alternatives that the respondent chooses reveal their preference for the attributes that describe these alternatives through the trade-offs that are being made. This paper applies a stated preference choice experiment to orthopaedic (knee and hip) surgery in order to elicit which attributes are most important in prioritizing patients on a waiting list. The alternatives represent health states and the 11 attributes include level of pain, financial difficulties, and enjoyment of life. Since 11 attributes are generally considered too many, the goal of this paper is to use a "blocked" design in which the attributes are split over three separate choice experiments and the data from these pooled to establish the relative ranking of all attributes. Since separating the attributes and then pooling the data has not been done before, issues surrounding this approach are discussed. However, it is an important new method to consider since it is not always feasible or desirable to reduce the number of attributes used.

In health care, the number of attributes has been kept to relatively small numbers (Ryan and Gerard, 2003), and 11 attributes would generally be considered too many. There are two reasons for this. The first is that with a large number of attributes, individuals may not make trade-offs but instead use other decision heuristics or lexicographic decision rules. This violates a key assumption of economic choice theory that rules out the interpretation of the data as utilities (Scott, 2002). A small number of attributes reduces task complexity for respondents and is more likely to enable compensatory rather than non-compensatory decision rules to be used. However, studies have shown that the presence of non-compensatory decision making is unlikely to influence regression results (Johnson and Meyer, 1984; Ryan and Ameya-Ameya, 2004).

A second reason for having a small number of attributes is more pragmatic in that the fewer permutations of attributes and levels, then the smaller the number of choice sets that need to be presented to respondents, and therefore the smaller the sample size required to complete a given number of choice sets. This again reduces the cognitive burden of the choice tasks and introduces the possibility that interaction and higher-order effects can be estimated. It is also not always possible to generate full or fractional factorial designs with a large number of attributes from existing software packages and design catalogues (Bradley et al., 2001; Hahn and Shapiro, 1966).

In practice, the number of attributes included in most choice experiments are usually pre-selected from a larger list and are judged to be the most 'salient' or 'important' using evidence from the literature, from focus groups, or by including attributes relevant to changes in policies that have not yet been introduced. Some evidence has shown that results can be sensitive to the inclusion or exclusion of attributes, so that pre-selection of a sub-set of attributes may cause bias in the estimates, particularly if an excluded attribute is correlated with an included attribute or makes a significant contribution to the explanatory power of the model (Louviere and Islam, 2004). Often it is only through careful piloting and focus group work that such bias can be minimised, although there is little guidance on how such piloting should be conducted. A further method that has been used to reduce such bias is to ask respondents to assume that all other attributes are the same across choices, i.e. that their difference is zero.

The aim of this paper is to undertake a choice experiment in the situation where there are 11 attributes, ten of which have four levels, and one has three levels. Instead of presenting each respondent with all 11 attributes, the attributes are 'blocked' or allocated across three experimental designs. Blocking is usually used to reduce the number of choice sets that each respondent has to answer whereas we use it to reduce the number of attributes. The use of a 'blocked attribute' design raises a number of methodological issues that will be examined in this paper.

The context of this study is in attaching weights to a disease specific quality of life instrument which is used to prioritise individuals on a waiting list for orthopaedic surgery (hip or knee replacement). There is a great need for evidence-based prioritisation of individuals on elective waiting lists as in many countries waiting lists exist and progression along the list is somewhat haphazard and not based on clinical or social need (Dreinhoefer et al., 2006; Quintana et al., 2000; Dolin et al., 2003). Prioritisation systems have been developed in some countries, notably New Zealand, Canada and the UK and were developed using clinical consensus and Delphi techniques (Hadorn and Holmes, 1997; Arnett and Hadorn, 2003; Woolhead et al., 2002). They have resulted in modest improvements in access and equity for people requiring joint replacement probably due to limited concordance between actual clinical need and queue positioning (Lack et al., 2000; Coleman et al, 2005).

In this paper the prioritisation is based on a disease-specific quality of life tool containing 11 quality of life questions (Multi-attribute Arthritis Priority Tool – MAPT). This was used to assess quality of life and therefore the need for joint replacement surgery for six hospitals in Victoria, Australia. The 11 dimensions (see appendix) were organised in Guttman scales with three to four levels of clinically defined health states with increasing severity. Having a large number of quality of life dimensions is typical in many generic and disease specific quality of life instruments. There is a small but growing literature on the use of choice experiments to generate utilities across different descriptions of health related quality of life (McKenzie et al, 2001; Ryan et al., 2006). Choice experiments provide a potentially more efficient method of producing utilities compared to other methods such as time-trade-off (TTO) and standard gamble (SG). Choice experiments are more firmly grounded in economic theory (random utility theory), are less costly to administer since a postal survey can be used, and are arguably easier to understand and complete. However, choice experiments to date have typically been based on a relatively small number of attributes and so their suitability to providing utility weights when there are many dimensions is an important issue to explore.

The rest of the paper is organized as follows. Section 2 describes the questionnaire and how the three designs were generated from it. Section 3 outlines the separate regression models of the three designs and the pooled model and deals with issues that may arise from pooling the data. Section 4 explains the empirical approach. Section 5 presents the results of the application to the orthopaedic waiting list. Section 6 discusses and Section 7 concludes.

2. Questionnaire design

The 11 quality-of-life questions (shown in the Appendix) were allocated across three different experimental designs. Some of the questions captured different components of the same underlying concept, which may result in dimensions being dependent on one another. This is another common feature of quality of life instruments. This issue was partly overcome through the allocation of questions to the three experimental designs. For example, questions one to three are all about pain, and so one question was allocated to each of the three designs. Similarly, questions four and five are both about 'looking after yourself', so these were also included in different designs. Finally, questions eight and nine are about financial difficulties and so these were included in separate designs. This ensured there was orthogonality between otherwise similar questions.

Five questions (out of the eleven) were used in each design with two questions included in all three designs (questions six and ten). Within each design, trade-offs will be made relative to these two attributes. It was important to include two common attributes rather than one to ensure that a common trade-off was being made in each design. In the analysis, this will help to standardise the regression coefficients against the average effect of the two common attributes. This therefore reduces any bias caused by the inclusion of different sets of attributes in each experimental design.

Each design had a possible 4^5 = 1,024 different combinations of attribute levels (5 attributes with 4 levels each). Note that one design that includes eleven questions with four levels each would have resulted in a possible 4^{11} = 4,194,304 combinations, considerably larger than the full factorial with just five attributes. Since it is not possible to present all 1,024 scenarios to each respondent, a fractional factorial design was produced generating 16 scenarios for each design (Bradley et al., 1991). This orthogonal design assumed that there were no interaction effects between the attributes.

In each design, the levels of one scenario (Scenario A) remained fixed and the remaining 15 scenarios were compared to it, giving 15 discrete choices. The fixed Scenario A was chosen out of the 16 that were generated for the orthogonal design, because it typically had two attributes at each of the extremes (i.e., one attribute level was 1, another attribute level was 4), and the remaining 3 attributes were in between. The use of a fixed scenario makes the choice task easier to complete and assists with meeting some of the key design principles discussed below.

The pairing of the 15 scenarios with the fixed scenario has to satisfy a number of design properties to ensure the choice design is efficient (Zwerina et al., 1996; Carlsson and Martinsson, 2003). Orthogonality (when attribute levels vary independently of each other) is satisfied when one attribute set at a certain level is independent of the levels of other attributes. In fact, each attribute in the fixed scenario is always compared to all three remaining attribute levels within each design. Level balance (when the levels of each attribute appear with equal frequency) is satisfied by using a fractional factorial design that generated equal frequencies of each attribute. Minimal overlap (when the alternatives within each choice set have non-overlapping attribute levels) is satisfied because each of the levels appear with equal frequency and the fixed scenario is always compared to all three remaining attribute levels. This results in minimal overlap because it minimizes the number of times that the difference between the fixed scenario and all other scenarios in each design is zero.

Respondents were asked to choose which of the two patients (represented by the two scenarios) should be given priority for a hip/knee joint replacement. The respondents were 96 of the approximately 150 arthroplasty surgeons in the state of Victoria. 65 surgeons filled out the questionnaire at a dedicated session at the Victorian branch

Australian Orthopaedic Association annual meeting, and the remaining 36 responded by mail. Thirty respondents completed all three designs, i.e. all 45 choices.

3. Regression models

The 'blocked attribute' design assumes that individuals will have the same preferences over the 11 attributes if they are presented together compared to if they are presented in blocks. For each block of attributes (i.e. each experimental design), respondents were asked to assume that all other dimensions of quality of life were the same for the two patients. This is a standard although untested method to attempt to control for omitted variable bias. This is equivalent to assuming that the differences between the levels of the omitted attributes are zero. Each experimental design can be used to estimate a separate regression model:

$$U_{ij}^{1} = \alpha + x_{ij}\beta + z_{ij}\omega + v_{j} + \varepsilon_{ij}^{1} \qquad (1)$$

$$U_{ij}^{2} = \alpha + x_{ij}\beta + w_{ij}\delta + v_{j} + \varepsilon_{ij}^{2} \qquad (2)$$

$$U_{ij}^{3} = \alpha + x_{ij}\beta + q_{ij}\theta + v_{j} + \varepsilon_{ij}^{3} \qquad (3)$$

where U^{1}_{ij} is the utility from alternative *i* chosen by individual *j* in experimental design 1. Since the utility difference between the two alternatives is not observed, *U* is defined as a binary variable (0,1). x_{ij} are the two attributes that are common across the three experimental designs, z_{ij}, w_{ij} and q_{ij} are the attributes that are specific to designs 1, 2 and 3 respectively. $\alpha, \beta, \delta, \omega$ and θ are coefficients to be estimated. v_j are random effects and ε_{ij}^{t} (t = 1, 2, 3, for each design) are the independent and identically-distributed (i.i.d.) error terms.

Alternatively, the data from each design can be pooled:

$$U_{ij} = \alpha^{k} + x_{ij}\beta + z_{ij}\omega + w_{ij}\delta + q_{ij}\theta + v_{j} + \varepsilon_{ij} \qquad (4)$$

where *z*, *w* and *q* will be zero (i.e. a zero attribute difference) in the designs in which they are omitted, in line with our assumption that all other attributes are the same across the two alternatives. α^k are now design-specific constants for each of the *k*=3 designs, and can be represented by dummy variables for each design. This controls for unobserved differences in the average probability of choosing scenario A or scenario B in each design. This includes any 'left-right' bias where respondents may be more or less inclined to choose the constant comparator (Scenario A). It also controls for any differences in respondents' characteristics between designs that were not accounted for in the random allocation of designs across respondents. The pooled model has random effects, v_j, and the error term ε_{ij}

$$\varepsilon_{ij} = \begin{pmatrix} \varepsilon_{ij}^1 & \varepsilon_{ij}^2 & \varepsilon_{ij}^3 \end{pmatrix}$$

where

$$E(\varepsilon_{ij}^t, \varepsilon_{ij}^s) = 0, \forall t \neq s$$

In other words, ε_{ij} is joint normal distributed.

The following probit model is estimated:

$$\Pr(y = 1 \mid x) = \Phi(\alpha^k + x_{ij}\beta + z_{ij}\omega + w_{ij}\delta + q_{ij}\theta + v_j)$$

$$\Pr(y = 0 \mid x) = 1 - \Phi(\alpha^k + x_{ij}\beta + z_{ij}\omega + w_{ij}\delta + q_{ij}\theta + v_j)$$

where $\Pr(y = 1 \mid x)$ is the probability that the respondent chose scenario A.

When combining different datasets, the discrete choice literature focuses on the combination of revealed and stated preference data. These datasets are likely to vary in many ways, including the independent variables used and the context of the study (see, for example, Brownstone et al., 2000; Bhat and Castelar, 2003; and Louviere et al., 1999). It is therefore important to control for heterogeneity that may influence the responses in each dataset. It is also important for the marginal utility of any common attributes to be equal, though there are two reasons why the coefficients on attributes common to each dataset may not be equal.

The first is that the β may vary because of different scale parameters (i.e. different means and variances of the unobserved factors) in each dataset. The scale parameter cannot be estimated directly, but Swait and Louviere suggest a simple likelihood ratio test to test for the equality of β whilst accounting for differences in the value of the scale parameter for each model (Swait and Louviere, 1993). If the data fails this test, it is argued that the RP and SP datasets cannot be combined since β varies across the data sources (Louviere et al., 2000; Train, 2002).

The second reason why β may vary is unobserved heterogeneity, of which there may be several sources. Omitted variables bias may be a problem as each data set might have a different and incomplete set of attributes and so if the omitted attributes are correlated with the included variables, then the coefficients will be biased, with the bias different in each dataset depending on which attributes have been omitted. A particular problem with revealed preference data is the endogeneity of attributes which is notoriously difficult to account for in econometric models without additional information. A further source of unobserved heterogeneity between revealed and stated preference datasets is differences between the characteristics and preferences of the samples. This is related to selection bias which may lead to different estimated marginal utilities from each dataset.

Models (1) to (4) were estimated using a random effects probit model. The random effects were used to account for correlations across observations caused by each respondent filling out at least 15 choices. This captures unobserved factors specific to each respondent. The dependent variable was either 0 or 1, depending on which scenario the respondent chose. The independent variables were the differences between the levels of Scenario A and Scenario B.

4. Prediction of utility scores

All types of discrete choice model are grounded in random utility theory, which is a discrete choice approach to consumer theory. If an individual chooses scenario B over scenario A, it is assumed that the utility of B is greater than A, although the actual utility cannot be observed, only the ordinal ranking. The coefficients from the probit model were used to predict rankings of individual patients who vary in terms of attribute levels. In generating the predictions from the models, the values of Scenario A were set to equal the ‘best’ possible health state. Since scenario B will always be ‘worse’ than A, this anchors the predicted probabilities to the ‘best’ health state. The predicted probabilities can therefore be interpreted as the utility scores, bounded between zero and one, from a comparison of the patient’s actual health state to the best possible health state. Any combination of the attributes of the questionnaire can be used to evaluate the model at specific points, allowing prioritization of a wide spectrum of patients. A higher utility score (predicted probability) means that patients are in a worse health state. In terms of setting priorities for waiting lists, a higher score therefore means that the individual is given a higher priority on the list. The predicted probabilities are obtained from the linear prediction of the probit model evaluated using a standard normal distribution. The predictions therefore depend on the distribution (mean and variance) of the linear prediction from the probit model.[1]

5. Results of the application to orthopaedic waiting lists

A high participation rate by Victorian surgeons was achieved (approximately 64%). While no information was available on the non-respondents, 100% of the surgeons at the scientific meeting completed the questionnaire suggesting the sample contained the majority of surgeons actively engaged in clinical practice. Table 1 shows the results from regression models (1) to (4).

The common parameters β in the three models (look after others and enjoyment of life) are not equal across the models and this is confirmed by the Swait-Louviere test that rejects the null hypothesis of parameter heterogeneity. This test accounts for differences in scale across the models, although this is unlikely to be a major problem since the models we estimate are very similar in terms of the choice task and types of variables included.

[1] The same method was repeated for a logit model to ensure that the probit and logit would give the same rankings. Since there were no outliers in the data that might affect the results, this turned out to be true and so the probit model was chosen.

Table 1. Random effects probit regression results.

	Design 1		Design 2		Design 3		Pooled	
	B	(s.e)	β	(s.e)	β	(s.e)	β	(s.e)
Look after others	0.133**	0.061	0.242**	0.050	0.349**	0.059	0.225**	0.031
Pain stops sleeping	0.699**	0.069	-		-		0.640**	0.064
Enjoyment of life	0.336**	0.068	0.251**	0.049	0.555**	0.059	0.353**	0.032
Enough help	0.116*	0.073	-		-		0.107*	0.063
Difficulties financially	0.387**	0.059	-		-		0.366**	0.056
Pain while resting	-		0.651**	0.052	-		0.649**	0.050
Been in paid work	-		0.135**	0.051	-		0.139*	0.051
Looking after self	-		0.216**	0.055	-		0.236**	0.055
Pain limits walking	-		-		0.753**	0.058	0.691**	0.052
Relationships	-		-		0.021	0.073	0.079	0.068
Change in joint problem	-		-		0.491**	0.055	0.437**	0.049
Dummy for design 2	-		-		-		-0.913**	0.212
Dummy for design 3	-		-		-		-0.357	0.223
Rho	0.266**	0.069	0.203**	0.054	0.195**	0.060	0.160**	0.037
Constant	0.853**	0.236	-0.188	0.162	0.567**	0.174	0.580**	0.166
-2LogL	-272		-389		-310		-1001	
Pseudo R^2	0.34		0.25		0.40		0.36	
Model χ^2 (df)	182(5)**		177(5)**		238(5)**		668 (13)**	
Number of observations	797		760		754		2311	
Number of doctors	54		51		51		96	
Swait-Louviere test of parameter homogeneity (χ^2, df): 58.4(1)**								

Notes: ** $p \leq 0.05$; * $0.05 < p \leq 0.10$

A reason why the coefficients are different might then be the existence of unobserved heterogeneity. However, some sources of bias can be ruled out. First, there will be little difference between the samples as each experiment was randomly allocated to respondents, so selection bias is unlikely to explain the differences in coefficients. Second, there was little difference in the context in which the data were collected. Third, thirty respondents completed all three experimental designs which further provided confidence in the homogeneity of the samples. The inclusion of design-specific constants (dummy variables) in the pooled model will account for unobserved differences between the samples and between the designs. The dummy variable for design 2 was statistically significant, suggesting that respondents in design 2 were more likely to choose scenario A. This is likely to be due either to unobserved differences in respondents or due to differences in the average utility of scenario A over scenario B.

Omitted variables are likely to be the source of the differences in coefficients. A key aspect of the blocked design was the deliberate omission of variables and our request to respondents to assume that all omitted attributes were the same for each choice. However, respondents (surgeons) are familiar with the process of choosing whether or not to admit patients and so will be familiar with all 11 attributes influencing this decision. This is certainly the case for the thirty respondents who filled out all three designs. For example, they may associate a high level of pain (an included attribute) with 'difficulties in looking after yourself' (an omitted attribute) and therefore are more likely to choose the scenario with a high level of pain. The coefficient on pain will therefore be biased upwards.

Omitted variable bias is a more serious problem in probit and logit models than in linear models because even if the bias is independent of the x's, the probit coefficients are inconsistent. However, Woolridge (2002) has shown that the probit of y on $\mathbf{x}$ consistently estimates β/σ rather than β (given normality conditions and the correct probit structural equation), and that for the purpose of obtaining relative effects, this is as good as estimating β, if the magnitudes of the β do not have to be meaningful. Since we estimate marginal effects (on the standard normal distribution) to "rank" the attributes, partial effects are sufficient. Furthermore, while the relative ranking of the two common attributes in relation to the other included attributes changes across our three designs and the pooled model, the relative ranking of only the two common attributes does not. In other words, 'looking after others' always has a smaller coefficient than 'enjoyment of life', and therefore 'enjoyment of life' is always more influential in prioritizing the patient on the waiting list than 'looking after others'. The fact that 'looking after others' is relatively less important in design 2 than in design 1, for example, only shows that in design 1 one of the included attributes (namely 'enough help') is relatively more important than 'looking after others'. This is confirmed in the pooled model. Therefore, omitted variable bias does not pose a problem in our pooled model and a random effects approach will capture all individual-specific heterogeneity.

The issue with omitted variables bias is then whether we can pool the data given the specific context of our analysis. A key issue is that, assuming the 11 attribute utility function is the complete utility function, the pooled model includes all of the omitted variables in the estimation. This is quite different to combining stated and revealed preference data when the researcher does not know what the omitted variables are. Furthermore, since each respondent filled out at least 15 choices, the omitted variables are the same for all 15 choices and so any bias will be constant across respondents and will therefore be captured in the random effect rather than the idiosyncratic error.

A potential problem with the random effects model is that it assumes the random effects are uncorrelated with the included attributes. It therefore assumes that the omitted variables within the random effect are not correlated with the included attributes. To test this assumption, we estimated all four models with random and fixed effects logit, and found that the coefficients were similar when a fixed effects model was used. This was confirmed by Hausman tests. This suggests that any omitted variables in the random effect were not correlated with the included attributes, and is therefore not an explanation for the difference in coefficients across the models.

If we can rule out the main sources of unobserved heterogeneity as causes for the difference in coefficients between the models, the only other reason why the coefficients are different is that the marginal utility and marginal rate of substitution between the common attributes is sensitive to which other attributes are included. Respondents are making trade-offs with a different set of attributes in each design, and so the coefficients and the trade-off between the two common attribute coefficients, β_1/β_2, may be sensitive to which attributes are included. However, even if this is the case, the data can still be pooled because this is a result of the fact that the relative ranking of the two common attributes is relative to all other attributes, as captured by the pooled model. The ordering of the coefficients on the common attributes is not affected by what other attributes are included in each design, as already mentioned. Specifically, ‘enjoyment of life’ is always more important than ‘looking after others’, regardless of where these rank relative to the other attributes that are included in the design. The pooled model uses the relative rankings of all attributes in relation to each other to provide an overall ranking. That the coefficients of the two common attributes change is a reflection of the effect that the inclusion of all attributes has on the relative ranking. This is not an indication that individual tastes are non-transitive or not well formed: if individual tastes were non-transitive, this would be a problem even if the data were not pooled.

Finally, we know that the unobserved factors are different in each design and, assuming our 11 attribute model represents the complete utility function, we know what the unobserved variables are. Whether the value of β varies across these models or not, the pooled model (4) will estimate the average value of β across all three experimental designs. Because we know what the unobserved factors are in models (1) to (3) and that they are included in the pooled model, the average value of β in the pooled model should be unbiased as it is estimated in the presence of the ‘unobserved’ factors which are now included in the pooled model. The ‘unobserved’ factors causing the omitted variable bias become ‘observed’. Any remaining source of bias will result from our 11 attributes not representing the full utility function. But this is the situation faced if we had included all 11 attributes in one design.

Pooling the data in this way compromises the design property of minimum overlap, which essentially means we have less information on which to base the estimation of the parameters since one third of the values of each of these variables will be zero. This will influence the efficiency with which the parameters are estimated rather than cause bias.

In the pooled model, the largest coefficients are those for the three pain attributes, followed by the change in hip or knee problem, difficulties financially, enjoyment of life, looking after self, looking after others, and being in paid work. ‘Difficulties with relationships’ is not statistically significant and ‘enough help looking after self’ is only significant at the 10% level. All coefficients are positive, suggesting that more severe levels of each attribute lead to a higher probability of being prioritised. Rho is a test of the significance of the random effects, and assesses whether there is a correlation between the multiple answers of each respondent. It shows that such a correlation did exist.

These coefficients and the constant were used to evaluate the model at different attribute levels. For each attribute, the attribute differences range from -3 to +3, so some of the evaluations will be out-of-sample predictions. We provide an example for 10 individuals who completed the MAPT instrument. The levels of each attribute were compared to the combination of attributes representing the best possible health state (i.e. attribute values are all equal to 1). The differences in levels between the actual and best health state were used to calculate predicted probabilities using the marginal utilities from the pooled regression model. So for each individual, the linear prediction $\hat{y}$ is:

$$\hat{y}_j = \alpha + d_j \beta_m$$

Where d_j is the attribute difference for individual j and β_m is the regression coefficient for the m attributes. For the 10 hypothetical individuals this results in the linear predictions utility scores and ranking shown in Table 2. This shows the predicted probability that patient B will be prioritised over patient A (who is in the best health state). These predicted probabilities are bounded between zero and one and can be interpreted as utility scores given the theoretical basis of the choice experiment. These scores are ranked in descending order to establish who has priority among these ten individuals who are waiting for hip/knee surgery.

Table 2. Utility scores and priority ranking for 10 hypothetical individuals.

	Ind.1	Ind.2	Ind.3	Ind.4	Ind.5	Ind.6	Ind.7	Ind.8	Ind.9	Ind.10
$\hat{y}$	12.35	7.93	6.82	10.64	4.96	4.50	8.42	2.65	8.84	7.64
Predicted probability (utility score)	0.95	0.56	0.41	0.86	0.19	0.15	0.63	0.05	0.68	0.52
Ranking	1	5	7	2	8	9	4	10	3	6

In the regression model, it was assumed that the attribute differences were linear as they were treated as continuous variables. To test this assumption, the probit model was re-estimated using dummy variables replacing the continuous variables one at a time. If the coefficients on the dummy variables increase by approximately equal intervals, and the confidence intervals from one dummy variable to the next overlap, the explanatory variables are monotonically increasing, and it can be concluded that they are approximately linear. This turned out to be the case and so using assuming linearity and using continuous variables were appropriate. Likelihood ratio tests comparing models with dummies to models with continuous variables favoured the models with continuous variables. Furthermore, re-coding the variables into dummies did not influence the rankings of individuals.

6. Discussion

We used a discrete choice experiment to weight a quality of life tool to allocate queue position of patients waiting for hip and knee joint replacement surgery. To avoid respondents having to trade off 11 attributes, we used a 'blocked attribute' design where the 11 attributes were allocated to three separate experimental designs. Two questions were repeated across the questionnaires, the rest differed between them. It was not possible to reduce the number of attributes because these came from an existing questionnaire and each attribute was regarded as clinically essential for prioritizing patients. In fact, because designs may get created from the literature, focus groups or interviews where important attributes are defined, designers may not always have the flexibility to reduce the number of attributes. Our methodology does not require reducing the number of attributes and therefore avoids biases that are likely to exist in other studies. Regression results were used to calculate predicted utility scores, bounded between zero and one, that can be used to rank patients on orthopaedic waiting lists.

Since the MAPT is a disease specific quality of life instrument, there are also clear applications of this methodology to generating single index scores for quality of life instruments with many dimensions, without having to artificially reduce the number of quality of life dimensions.

With respect to the choice experiment, there was no opportunity to compare an 11 attribute design with our 'blocked attribute' design and this is a topic for further research. However, even then it would be difficult to distinguish between non-compensatory decision making and making inherently different trade-offs. We have argued that the failure of the Swait and Louviere test for parameter homogeneity is not an issue in this context, since the unobserved factors are captured in the individual-specific random effects and the pooled model includes the known omitted variables causing the bias and therefore estimates an unbiased parameter. This, however, assumes that the 11 attributes represent the complete utility function and this may not be the case. Nevertheless, this would also be the case if we had conducted the experiment with an 11 attribute design.

Orthogonal designs have some limitations because they only ensure independence of the effects of different attributes, and not necessarily that all relevant trade-offs (within each of our three designs) are presented to the respondents. Another approach would be to generate an efficient design rather than an orthogonal design. An efficient design would be generated by assuming prior values for the β's, using these to estimate the covariance matrix, and then minimising the errors to make the β's as close to their true values as possible when generating the design. Our orthogonal design does not guarantee the lowest possible errors, and so more accurate coefficients could be obtained if an efficient design had been used. It is difficult to estimate these prior values of β unless a similar study has been conducted before. However, we were able to check, after our DCE was completed, whether we had low or high D-errors, the measure commonly used to minimise errors in efficient designs. Instead of using prior values for the β's, these calculations use the actual values that were calculated once the experiment was complete. The D-error is based on the following:

$$D-error = \det(\Omega^{*}(\hat{\beta} \mid X)^{1/K}$$

where Ω^{*} is the covariance matrix of the β's, excluding constants, and K is the number of parameters to estimate. For design 1, the D-error is 0.2834; for design 2, the D-error is

0..2871; and for design 3, the D-error is 0.2798. These are all fairly reasonable, so our orthogonal design seems to be relatively efficient as well.

There are a number of issues to be resolved to apply this to actual waiting lists. Since the utility score and ranking is based on the mean and standard deviation of the linear prediction, any new additions and patients leaving the list require the whole ranking to be re-calculated. This may mean that some patients never reach the top of the list since their 'place' will change, and so raises issues about how the ranking will be used in practice. It is likely that the scores would need to be weighted by the length of time spent on the list. It may also be necessary to repeat the exercise at regular intervals, as rankings will change as quality of life changes over time. These issues depend of course on how waiting lists are managed now, of which there is very little information.

There are also issues about doctor's thresholds to place people on the list. One can imagine that the decision is taken in two parts, whether to place a person on the list, and then what priority to give them. A choice experiment could be designed along these lines by including a 'do not place on the list' option. Then it would be possible to model doctors' admission thresholds. These thresholds are likely to be influenced by a number of factors, including the doctor's private practice.

A further issue is that although the MAPT levels will be generated by patients, the utility weights were generated by a group of surgeons. Strong arguments in favour of patients determining both the MAPT levels and the weights can be made, given the important role that community preferences should play. However, this was not possible in this study because in order to use the MAPT at all, surgeons insisted they should be involved in the process of setting priorities on the waiting list and in judging the clinical need of patients for treatment. Debates about whose preferences should be included in any type of quality of life measure are abundant, and in this setting, patients' values might be more appropriate given that they are the ones experiencing the symptoms, not the surgeons. On the other hand, surgeons would be thoroughly familiar with all the attributes, while some patients may not have experienced all the symptoms, and so may not be a good judge of them. Further research should compare the preferences of surgeons with those of patients.

Another issue is that patients have an incentive to overstate the seriousness of their condition on the MAPT, rendering them a higher place on the waiting list. The only control for this is that the doctors, who know the seriousness of their patients' condition, check that their patients did not overstate their condition. But then there might be an incentive for the doctors to see that their own patients get higher on the waiting list as well, exacerbating the problem rather than alleviating it.

There is a more fundamental issue in that the priorities are determined solely on current quality of life, and not on the *expected* cost per QALY or expected capacity to benefit from treatment for an individual patient. One could argue that it is the expected gains in quality and length of life per additional dollar spent that should determine priorities and then ensure that health status is maximised from the available resources. Although the ranking generated here may improve on the current situation of waiting list management, it does not necessarily lead to an efficient allocation of scarce resources. Perhaps the routine administration of the MAPT a few months after surgery would enable us to predict changes in utility for patients, so the benefits of treatment could be estimated and used to prioritise patients coming onto the list with similar characteristics in combination with data on treatment costs. This, however, may greatly increase the complexity of placing patients on the waiting list, not least because capacity to benefit from treatment would depend on a number of other factors, such as age and general health, which are largely patient-specific and multi-dimensional.

7. Conclusion

A discrete choice experiment was used to rank 11 attributes of an existing questionnaire, using a new methodology whereby the attributes were split across three separate designs, and the data pooled to obtain the final ranking. Two potentially serious issues, namely different scale parameters and unobserved heterogeneity, were addressed and found not to be of significance in this particular setting. Thus, it seems possible that attributes can be allocated across separate designs, which has the beneficial effect of reducing the complexity of experiments without compromising the content of the study by reducing the number of attributes.

Appendix: The 11 dimensions of the Multi-attribute Arthritis Priority Tool (MAPT)*

1. **Do you have hip or knee pain that does not get better even when you rest (for example, while sitting)?**
   4 levels

2. **Do you have hip or knee pain when you first go to bed at night that stops you going to sleep?**
   4 levels

3. **Do you have hip or knee pain that limits your walking?**
   4 levels

4. **Does your hip or knee make it difficult for you to look after yourself (such as washing yourself, getting dressed, going to the toilet)?**
   4 levels

5. **Do you get enough help with looking after yourself (such as washing yourself, getting dressed, going to the toilet)?**
   4 levels

6. **Does your hip or knee affect your enjoyment of life?**
   4 levels

7. **Does your hip or knee cause difficulties with your relationships with people close to you (such as wife, husband, children and close friends)?**
   3 levels

8. **Does your hip or knee make it difficult for your household (yourself, family and others) to manage financially?**
   4 levels

9. **Have you been in paid work in the last 6 months?**
   4 levels

10. **Do you need to look after people who require your care (such as a sick or disabled partner or family member)?**
    4 levels

11. **Overall, is your hip or knee problem different now compared with how it was 6 months ago?**
    4 levels

* Full version of questionnaire is available from the author.

References

Arnett, G. and Hadorn, D.C. "Developing priority criteria for hip and knee replacement: results from the Western Canada Waiting List Project." *Canadian Journal of Surgery*, 2003; 46: 290 – 296.

Bhat, C. and Castelar, S. "A unified mixed logit framework for modeling revealed and stated preferences: formulation and application to congestion pricing analysis in the San Francisco Bay area." *Research Report SWUTC/03/167220-1*, Southwest Regional University Transportation Center, Center for Transportation Research, University of Texas at Austin, April 2003.

Brownstone, D., Bunch, D. and Train, K. "Joint mixed logit models of stated and revealed preferences for alternative-fuel vehicles." *Transportation Research Part B*, 2000; 34: 315 – 338.

Carlsson F. and Martinsson P. "Design techniques for stated preference methods in health economics." *Health Economics* 2003; 12: 281 – 294.

Coleman, B., McChesney, S., and Twaddle, B. "Does the Priority Scoring System for Joint Replacement really identify those in most need?" *The New Zealand Medical Journal*, 2005; 118: 1 – 6.

Dolin, S.J., Williams, A.C., Ashford, N., George, J., Pereira, L. and Perello, A. "Factors affecting medical decision-making in patients with osteoarthritis of the hip: allocation of surgical priority." *Disability and Rehabilitation*, 2003; 25: 771 – 777.

Dreinhoefer, K., Dieppe, P., Til, S., Grober-Gratz, D., Floren, M., Gunther, K. Puhl, W. and Brenner, H. "Indications for total hip replacement." *Annals of the Rheumatic Diseases*, 2006; 65(10): 1346 – 1350.

Hadorn, D. and Holmes, A. "The New Zealand priority criteria project: Criteria pilot tests." *British Medical Journal*, 1997a; 314(7074): 131.

Johnson E. and Meyer R. "Compensatory choice models of non-compensatory processes: the effect of varying context." *Journal of Consumer Research*, 1984; 11: 528 – 541.

Lack, A., Edwards, R.T. and Boland A. "Weights for waits: lessons from Salisbury." *Journal of Health Services Research & Policy*, 2000; 5(2): 83 – 88.

Louviere J., Hensher D. and Swait J. *Stated choice methods. Analysis and application*. Cambridge University Press, Cambrigde, 2000 (Ch8).

Louviere J. and Islam T. "To include or exclude attributes in choice experiments: a systematic investigation of the empirical consequences." In: Wiley J., Thirkell P. (eds) *Australia and New Zealand Marketing Academy Conference*, Victoria University of Wellington, New Zealand, 2004.

Louviere, J., Meyer, R., Bunch, D., Carson, R., Dellaert, B., Hanemann, W.M., Hensher, D. and Irwin, J. “Combining Sources of Preference Data for Modeling Complex Decision Processes.” *Marketing Letters*, 1999; 10(3): 205 – 217.

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: 193 – 204.

Quintana, J.M., Arostegui, I., Azkarate, J., Goenaga, J.I., Guisasola, I., Alfageme, A. and Diego, A. “Evaluation by explicit criteria of the use of total hip joint replacement.” *Rheumatology*, 2000; 39(11): 1234 – 1241.

Ryan M. and 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.

Ryan M. and Ameya-Ameya M. “‘Threats’ to and hopes for estimating benefits.” *Health Economics*, 2005; 14: 609 – 619.

Ryan M., Netten A., Skatun D. and Smith P. “Using discrete choice experiments to estimate a preference-based measure of outcome – an application to social care for older people.” *Journal of Health Economics*, 2006 (forthcoming).

Scott A. “Identifying and analysing dominant preferences in discrete choice experiments: an application in health care.” *Journal of Economic Psychology*, 2002; 23: 383 – 398.

Swait J. and Louviere J. “The role of the scale parameter in the estimation and use of multinomial logit models.” *Journal of Marketing Research*, 1993; 30: 305 – 314.

Train K. *Discrete choice methods with simulation*. Cambridge University Press, Cambridge, 2002.

Wooldridge, J.M. *Econometric Analysis of Cross-Section and Panel Data*. MIT Press, London, 2002.

Woolhead, G.M., Donova, J.L., Chard, J.A. and Dieppe, P.A. “Who should have priority for a knee joint replacement?” *Rheumatology*, 2002; 41: 390 – 394.

Zwerina K., Huber J. and Kuhfeld W. *A general method for constructing efficient choice designs*. Durham, NC, Fuqua School of Business, Duke University, 1996.