Summary

Cost-utility analyses (CUA) are increasingly common in Australia. The EQ-5D is one of the most widely used generic preference-based instruments for measuring health-related quality of life for the estimation of quality-adjusted life years within CUA. There is evidence that valuations of health states vary across countries, but Australian weights have not previously been developed. Conventionally, weights are derived by applying the Time Trade-Off elicitation method to a sub-set of the EQ-5D health states. Using a larger set of directly valued health states than in previous studies, Time Trade-Off valuations were collected from a representative sample of the Australian general population (n=417). A range of models were estimated and compared as a basis for generating an Australian algorithm. The Australia-specific EQ-5D values generated were similar to those previously produced for a range of other countries, but the number of directly valued states allowed inclusion of more interaction effects which increased the divergence between Australia's algorithm and other algorithms in the literature. This new algorithm will enable Australian community values to be reflected in future economic evaluations.

1 Introduction

Economic evaluation of health interventions is integral to the decision-making process in many countries, particularly for government reimbursement decisions. The tools used in the construction of such analyses are, therefore, of increasing importance. Cost-utility analysis (CUA) is the preferred approach in many countries, including Australia. An increasing focus on health-related quality of life has seen the development of standardised descriptive quality of life instruments that allow for direct measurement of the quality of life of patients in clinical settings, trials and observational studies, and valuation via a single index derived from a population-based preference elicitation study. These instruments (termed multi-attribute utility instruments) describe health in terms of a set of dimensions and items, and include an algorithm that assigns an index number to each health state (defined as a specific profile of attribute items representing alternative levels of the different dimensions) represented by the instrument space on a scale with 1 representing full health and 0 representing death. Attaching a value greater than 0 to a health state implies it is better than dead, while a negative value represents a state worse than dead. Existing instruments include the EQ-5D [1], the SF-6D [2], the Health Utilities Index [3, 4] and AQoL [5].

Australia is an unusual case. While CUA has become the preferred approach for the evaluation of pharmaceuticals [6], Australian general population specific weights exist for only one of the more common multi-attribute utility instruments (the AQoL). Therefore, Australian CUAs performed using EQ-5D or SF-6D data have relied on weights from other countries, particularly those from the United Kingdom [1, 2].

Multi-attribute utility instruments have been compared and their role in the economic evaluation of health technologies has been discussed widely in the literature [5, 7]. In this paper, the focus is on the EQ-5D, as it represents the most commonly used generic quality of life descriptive system. The primary aim of this study was to develop Australian based weights for the EQ-5D descriptive system, based on data collected from a sample representative of the Australian general population and using methods that are largely comparable to those used previously to develop weights for other countries.

A secondary aim was to explore methodological issues in the derivation of weights for the EQ-5D, particularly in relation to the choice of health states to be directly valued, and the impact of this choice on the weights derived. In this study the choice of health states was informed by undertaking a simulation study. Several different methods were used to define subsets of health states to be directly valued, and simulation data were generated. The results from each of these subsets were analysed separately and the resulting utility weights were compared for all health states defined by the EQ-5D descriptive system to determine a preferred set of health states to be directly valued. This set was then used for the data collection for the Australian valuation study.

Section 2 of the paper briefly describes the EQ-5D and its development, including the methods that underlie the existing algorithms, and in particular the selection of health states for direct valuation. This section motivates the simulation approach used in this study, and provides a rationale for it. Section 3 describes the methods for the simulation study, and for the data collection and analysis for development of the Australian algorithm. Section 4 presents the results, and Section 5 discusses the choice of algorithm.

2 Overview of the EQ-5D and valuation studies to date

The EQ-5D was originally developed by a European team of researchers. The Measurement and Valuation of Health (MVH) study based at the University of York produced the United Kingdom algorithm [8]. The EQ-5D has five dimensions (Mobility, Self Care, Usual Activities, Pain/Discomfort, and Anxiety/Depression). Each dimension has three levels corresponding to no problems, some problems and severe problems. Consequently there are 243 (3^5) possible health states.

Valuation algorithms exist for a variety of countries, including Spain, the United Kingdom, Zimbabwe, the Netherlands, the USA, Japan, Denmark, and New Zealand [1, 9-14]. Most of these used the Time Trade-Off (TTO) to value individual states, although a Visual Analogue Scale was used in New Zealand [15]. Further, all used direct valuation of a sample of health states, with regression analysis used to develop a linear additive model to predict the values of all other health states. One advantage of the use of a common method for development of algorithms across countries is that it allows comparison between national attitudes to ill-health [9, 16]. Recent evidence suggests that characteristics of the population may drive health state valuations, and that differences in valuations between countries are due to differences in national attitudes to ill health, rather than being artefacts of variations in valuation methods [17].

Relative to other generic health-related quality of life tools, the number of health states in the EQ-5D is small. The SF-6D contains 18,000 health states, and the AQoL allows for more than one billion. Tsuchiya et al. [13] discuss issues relating to the number of unique health states. First, there is a trade-off between the richer descriptive system permissible under those instruments with more health states, and the ease of use associated with tools such as the EQ-5D. In trial-based evaluation of health technologies, it is preferable to not overburden patients with self-complete questionnaires. The brevity of the EQ-5D is an advantage in this regard. However, the simplicity of the descriptive system may make it insensitive to changes in health status, and, therefore, to the relative impact of different interventions on health-related quality of life. Further, in valuation tasks, there may be considerable variability among respondents in their interpretation of a particular descriptor (particularly the distinction between moderate and severe levels). A five level descriptive system has been introduced and is likely to represent an improvement in terms of descriptive ability [18, 19], but no scoring algorithm has yet been developed and the three level descriptive system remains widely used [20].

A second issue arises from the number of states that require direct valuation. For any given descriptive system, the higher the proportion of states that are directly valued, the less restrictions are placed on the functional form of the algorithm (for example, allowing estimation of interactions between dimensions).With instruments such as the AQoL, HUI3 and SF-6D that incorporate several thousand separate health states, any valuation study is necessarily limited in the proportion of health states that can be directly valued. In the case of the AQoL and HUI3, the developers of the instrument assume a priori a multiplicative functional form, which then limits the number of distinct states that need to be valued. In the case of the SF-6D, the functional form is assumed to be additive, but even a relatively large valuation study can only include a small proportion of the total number of health states, which in effect limits the investigation to a linear additive functional form without interactions.

Tsuchiya notes that a larger number of directly valued states makes the evaluation exercise more onerous. Dolan [8], in the original valuation study included 43 states in the valuation sample, but Tsuchiya et al. have argued that a subset of 17 states is appropriate. [13]. While there is a potential trade-off between valuation of a larger proportion of states and the burden of data collection, electronic methods of data collection can reduce the marginal cost of data collection, thus allowing a larger number of respondents for the same data collection resources. For example, use of an on-line panel or computer-assisted telephone interview (CATI) techniques can reduce recruitment and interview costs. Electronic methods for data collection allow more respondents to be questioned, thus either reducing the number of states each respondent has to face, and/or increasing the proportion of states that can be directly valued (rather than estimated through the subsequent algorithm).

A third issue, related to but distinct from the number of health states that need to be directly valued, is the selection of the particular health states that are to be valued. For example, for the EQ-5D it is unreasonable to ask a single respondent to value more than a small subset of the total 243 health states within a TTO framework. Two major approaches have been taken to constructing this subset; the 43 state approach used in the UK valuation survey (of which a subset of 13 of the 43 health states was valued by each respondent), and the 17 state approach (of which all 17 were valued by each respondent) used in the Japanese valuation survey [1, 13]. The approach to selection of health states used by Dolan et al. was based on classifying health states as very mild, mild, moderate and severe (based on the levels of each dimension) and then selecting a subset (n=43) that included full health, the worst health state in the EQ-5D, and health states from each of these severity groups. While the basis of selection is not described in the papers reporting the study, the approach ensures that each dimension is represented at the no problems, some problems and severe problems levels. It also excluded ‘implausible’ health states, defined as combinations of level 1 on usual activities (“No problems with performing one’s usual activities”) with level 3 on mobility (“Confined to bed”) or level 3 on self-care (“Unable to wash or dress oneself”) [8].

Tsuchiya et al. use a subset of 17 of the original Dolan set of 43, described as ‘the minimum set of health states required to estimate the value set’, although it is not clear from the paper on which criteria this statement is made. In neither case is it clear that experimental design principles underlie the choice of health states to be valued. It is noteworthy that the states selected under both the Dolan approach and the Tsuchiya approach have a relatively higher proportion of dimensions at Level 1 (i.e. No Problems) and a high co-occurrence of Level 1 in multiple dimensions. The implication of this is that the point precision will differ between directly valued states, and the uncertainty around the extrapolated values will be greater in those health states with relatively more Level 2 and 3 attributes. A recent study has directly valued 101 of the states, but this has not yet been replicated elsewhere [21].

Given the relatively small number of health states in the EQ-5D, it is feasible to value all states directly. This has the advantages of reducing the need to extrapolate between directly valued states, and allowing for estimation of a wider range of interaction effects. Given that some health states are implausible, it may not be appropriate to value all states, because the cognitive task of requiring respondents to imagine an implausible health state may be unreasonable in a valuation task. Overall there has been relatively little empirical exploration of the impact of selection of health states on the valuation algorithm. Tsuchiya et al. found that the performance of the 17 state approach was very similar to the 43 state approach used by Dolan et al. In both cases, there were no significant interaction terms apart from the N3 term (a dummy variable defined as equal to one when any dimension is at the worst level which, although not an interaction term in a statistical sense, functions like one in the algorithm). However, it is not clear which interaction terms can be modelled using direct valuation of either the 43 or 17 states given above because of the lack of information about the orthogonality of domains in the subset of health states included. In this study, we used a Monte Carlo simulation study to investigate whether two different assumed underlying utility functions can be recovered given direct evaluations of specific subsets of EQ-5D health states. We then used the results of the simulation study to inform the selection of the health states for our main data collection; the approach to this is described below.

3 Methods

Monte Carlo simulation study

The aim of the Monte Carlo study was to test whether the selection of health states included in a valuation study impacts on the extent to which the parameters of an underlying utility function can be investigated. Clearly, this is not a question suited to investigation through empirical means; rather, the use of simulated data is necessary to explore these issues.

The broad approach was to assume a specific functional form and set of coefficients to represent the systematic component of the utility function defined over EQ-5D space, generate simulated data based on each of these functional forms, and then test whether the parameters of the utility function could be estimated from these simulated data. Two different underlying models of utility were specified - a main effects only model and a model with main effects and interactions. The two models are described in more detail below We used five different design approaches to select the health states for which data would be simulated, and generated the simulated data for each design approach and each assumed utility function (thus, ten simulation valuation sets for each combination of design strategy and underlying models of utility). The design approaches are described in detail below. To generate the data for each health state included in the simulation valuation sets, we calculated the systematic component of the utility function based on the assumed coefficients (see Table 1 for details), and added a standard normal error term (zero mean and variance of one)

, to give the total random utility for each simulated observation. We did this for a simulated sample of 300 respondents each valuing 15 health states (therefore the total number of observations in each simulation was the same). We repeated this process 100 times, thus the simulation valuation data sets comprised 100 independent simulated samples of 300 respondents for each of the five designs paired with each of the two functional forms. The designs provided the X-matrix of the simulated samples. These simulated data were then used to estimate the parameters for different models to determine if the original utility function from which the data were generated could be recovered given the design approach and the selection of health states directly valued.

Designs

Five approaches to selection of health states were considered in the simulation study. The first two replicated the Dolan (43 states) and Tsuchiya (17 states) designs. The third used an orthogonal main effects plan (OMEP) in which each pair of levels of particular dimensions appears with equal frequency allowing independent estimation of the main effects. [22] For a 3^5 design, an 18 state OMEP was identified; this was a fractional factorial which permitted the estimation of all main effects while maintaining orthogonality and (usually) balance. Full health was one of the states within the OMEP; therefore, the design included 17 states and thus was similar to the Tsuchiya approach with the key difference being the use of an OMEP to derive the health states. The fourth design was an exhaustive design in which all states were directly valued (albeit by a smaller number of respondents in order to keep the total sample size constant between designs). This will be called the full factorial (FF). The fifth design was the exhaustive design with implausible states removed, or the plausible full factorial (FFP). Our main concern with implausible states was that respondents were likely to provide unreliable responses to health states which did not correspond to something they could imagine, However, this could not be identified in simulated data but we nevertheless included this design because it allowed us to investigate the size and direction of potential bias that might arise statistically from excluding such a systematic subset of the 243 possible EQ-5D health states.

The definition of implausible states for this study differs slightly from that used by Dolan et al. (1997). A state was excluded as implausible if it combined level 3 on mobility with either level 1 on usual activities or level 1 on self-care (“No problems with self-care”). This removes 45 states from the EQ-5D.

Econometric Models for the simulation study

Two econometric models were estimated for each design. The first was a linear additive main effects model including a coefficient for each level of each dimension plus the N3 term included in most previous EQ-5D algorithms (a dummy equal to one if at least one dimension is at the worst level). This model has been assumed predominantly in the existing country specific algorithms (e.g. [8, 9]). The second was a linear additive main effects model that included a parameter for each main effect as before but replacing the N3 term with every two factor interaction between the two less than full health levels of the five dimensions (e.g. Mobility 3 x Pain/Discomfort 2).

Generalised Least Squares (GLS) was used to estimate the parameters in each of the two models, and for each of the 100 simulated samples for each design. In terms of selecting a preferred design, we were interested in three major outcomes: the ability of the model to recover two-factor interaction terms rather than simply the more blunt N3 term, the precision with which the design could recover a set of assumed coefficients, and finally the plausibility of the valued states. The first two criteria can be tested empirically in the simulation study, the third requires a judgment. The decision regarding design attempted to balance these concerns.

General population valuation task

The TTO task was run through an online interface, and was designed such that each respondent valued 11 randomly selected health states from the selected design, as well as the pits state (33333). For each state, the individual was asked if ten years in that state followed by death was preferable to immediate death. For states considered better than immediate death, a 'ping pong' approach was taken, aiming to identify a period of time *x* such that the respondent was indifferent between *x* years in full health, and ten years in the state being valued , with the smallest gap between observable *x*'s being 0.05. If an individual failed to identify a point of indifference, a score midway between values of *x* was assigned. The score assigned to the state was *x*/10. If immediate death was preferable to ten years in the state followed by death, the task was amended to a choice between a) immediate death, and b) *x* years in the health state, followed by (10-*x*) years in full health, followed by death. As with states better than immediate death, a 'ping pong' approach was used. When the value of *x* was adjusted until the individual was indifferent between the options, the health state was valued as (*x*/10)-1. Thus, the boundaries of valuation are -1 and 1. The reasons for doing so have been widely discussed elsewhere, including a recent review article [23].

Recruitment and Data Collection

Recruitment and data collection for the main study was undertaken by a market research company, who had received training in the administration of the on-line task. The study was approved by the institutional Human Research Ethics Committee. The sample frame comprised individuals who had consented to be on the market research company data base, a large existing panel. Respondents were recruited by telephone, and invited to attend the interview in four locations, specifically metropolitan Sydney, Sydney suburbs (Parramatta), metropolitan Melbourne and rural New South Wales (Orange). Respondents were randomly recruited to defined sample characteristics to match the Australian age and gender split. Respondents attended an organised session, and interviewed in groups of four with a trained interviewer available to assist. After the task began, there was no interaction between the four respondents, and the trained interviewer was instructed to only assist with matters of interpretation of the question, and any IT issues. The reason for using this electronic approach relative to a straightforward online survey was that recent evidence has suggested that results generated using that latter approach may produce large numbers of health state valuations clustered around -1, 0 and 1 [24]. Each respondent was paid $60 for completion of the survey. Data were automatically captured in a computer-based central database of results. After an introduction to the task provided by the interviewer, each respondent completed the EQ-5D to familiarise them with the instrument. They then valued 12 states using a Time Trade-Off (11 and the pits state), assisted through the task by the interviewer.

Analysis

A number of linear additive specifications were proposed in order to test for interactions. The range of utility functions used in the regression analysis is given in Table 1.

Table 1 here

Model 1 consisted of a main effect for each movement away from full health. Therefore, a move from Level 1 to Level 3 in a particular dimension (for example mobility) was represented by the sum of the co-efficient moving from Level 1 to Level 2 (named MO2) and from Level 2 to Level 3 (MO3). Therefore, the value *y* placed on a health state was as follows:

$$y = \alpha + \beta'_{dl}\ x'_{dl} + \varepsilon$$

where β′ is a vector of co-efficients and x′ is a vector of dummy variables for dimension *d* at level *l*. Model 1(b) repeated Model 1, but constrained α to be 1 to represent full health. Model 2 (and 2(b)) repeated these specifications, but included a simple interaction term N3, which is a dummy variable equal to 1 if and only if at least one dimension is at the worst level.

$$y = \alpha + \beta'_{dl}\ x'_{dl} + \gamma N3 + \varepsilon$$

Model 4 (and 4(b)) accounted for the more exhaustive nature of the states directly valued, repeating Models 1 and 1b but considering each pairwise interaction term.

$$y = \alpha + \beta'_{dl}\ x'_{dl} + \beta' x'_{dl} + \varepsilon$$

Finally, Models 3 and 3b replicated Models 4 and 4b, but included only interactions between dimensions at level 3.

To reflect the panel nature of the data, all specifications adopted a random effects Generalised Least Squares model (estimated with xtreg in STATA 10.1). Thus, the error term was decomposed into a conventional error term for each observation (assumed to be normally distributed with mean equal to zero), and an individual-specific error term representing the extent to which the intercept of an individual differs from α .

In terms of identifying a preferred algorithm for use in Australian cost-utility analyses, evaluation of models was based on consistency of signs and orderings of co-efficients (as the EQ-5D is monotonic), model fit and logical orderings of predicted health state values. With regard to model fit, we examined the log-likelihoods using the Akaike and Bayesian Information Criteria (AIC and BIC). The advantage of AIC and BIC is that they consider both the number of constraints and the predictive value of the algorithm [25, 26]

Results

Simulation study

Table 2 presents the simulation study results for the main effect model. The second column is the coefficient that was assumed (and from which the simulated data were generated). The remaining columns show the means and standard deviations for the coefficients which were estimated based on data simulated using each of the five design approaches. If the assumed underlying utility function included only main effects, all design approaches performed relatively well. The means and standard deviations across all simulations are shown in Table 1. Under these assumptions, the best performing designs in terms of the size of the standard deviations are the OMEP, the full factorial (FF) and the full factorial with only plausible health states (FFP).

Table 2 here

When two-factor interactions were included in the assumed utility function, the only design approaches which allow estimation of all two-factor interaction terms are the FF and the FFP. The number of two factor interactions which could not be estimated was higher for the OMEP and the Tsuchiya approach than for the Dolan approach. While the FF produces the least bias and the best precision, the FFP approach performs almost as well when interactions are included in the assumed utility function. As the effect of asking respondents to value implausible states cannot be captured in a simulation study, it is not possible to determine the trade-off between error that would be introduced by high variance in valuations of implausible states compared with the error introduced by excluding these states from the design. The results of the simulation study suggest that a less restrictive experimental design such as FF or FFP would allow for the possibility of estimating interaction effects whereas existing experimental design strategies do not. It was decided that the FFP represented the most appropriate design approach, allowing for estimation of interactions, without introducing the possibly unreasonable cognitive task of valuation of implausible states. This design comprises 198 health states, i.e. the entire EQ-5D set of health states minus those combining Mobility 3 with either Self-Care 1 or Usual Activities 1.

Time Trade-Off

417 respondents undertook the task, with 101-108 completing in each location. The demographic characteristics of the sample are compared with those of the Australian population in Table 3. In general, the age and gender distribution of the sample was similar to that of the Australian population, although older Australians were under-represented. As all respondents provided a complete set of valuations, it was not necessary to exclude any from the analysis.

Table 3 here

Five of the responders gave the same value for every health state they saw, of which two were conventional non-traders valuing all health states at one as they were unwilling to sacrifice any life expectancy for improved quality of life.

Table 4 here

The results from the eight specifications are given in Table 4. A variety of models were run in which the effect of interview location was investigated, but these were not generally statistically significant. Therefore, all results are based on the pooled sample.

In the simple main effect models (1 and 1(b)), and in those which include the N3 term (2 and 2(b)), all coefficients are negative, and all level 3 coefficients have a larger absolute value than their respective level 2 coefficient. This is as expected and reflects the monotonic nature of the levels of each dimension in the EQ-5D. All coefficients are highly statistically significant ($p<0.01$). In addition, including a constant term improved model fit across all specifications. Of the two-way interactions included in models 3 and 3b, significance at the 5% level is only met by 4 of the 38 interactions. However, those that remain significant in both models are interactions of the worst levels of the dimensions Mobility, Pain / Discomfort and Anxiety / Depression. This suggests that these interactions are potentially important in obtaining accurate utility estimates for very poor health states.

Model 3 and 3b included only the ten interactions between dimensions at level 3, and retained the four statistically significant coefficients ($p<0.01$) with a further three statistically significant at the 10% level.

Figure 1 here

Comparison of the valuation from different utility model specifications of the 198 plausible health states is given in Figure 1. There is a high degree of agreement between specifications, both in terms of scores and ranking. The minimum pairwise correlation co-efficient was 0.960, and the minimum pairwise Spearman co-efficient was 0.970. However, this high level of agreement across the algorithm ignores an important issue at the better end of the scale representing typically mild health states. For the most commonly observed non-full health states (principally those with 4 dimensions at Level 1, and one at Level 2), the difference in algorithms is markedly dependent on whether the intercept is constrained to unity (which is investigated by some published EQ-5D algorithms, but not used in any of their recommended algorithms). This is of particular importance as these health states are likely to be relatively common when using self-assessed EQ-5D health in economic evaluation of any population other than very ill patient groups.

Discussion

The simulation approach used in this study demonstrates that previous time trade-off studies designed to develop EQ-5D algorithms lack sufficient coverage of the EQ-5D space to allow identification and estimation of interactions that may be present between dimensions and levels. Our data collection and comparison of models suggests that a more complex algorithm may be appropriate. Current models which include only the N3 term are essentially additive. In this study, the model that provides the best fit includes a more complex set of interactions of dimensions at their worst levels. The fact that these interaction terms are generally positive and, therefore, in the opposite direction to the main effects suggests that there is a multiplicative effect, that is the additional decrement in utility associated with a worsening in a second dimension is smaller than the decrement for the first worsening dimension.

In all specifications, the constant term is significantly different from one. This is consistent with the findings of other studies, and suggests that it is appropriate to include an unconstrained constant term in the TTO algorithm. The inclusion of a constant term that is not constrained to unity is typically interpreted as capturing the effect of any move away from full health. However, it does impact on the valuation of the milder health states, an impact that is particularly evident in comparison of models 1 and 1b. Anchoring prevents a ceiling effect which can be seen in the non-anchored algorithms. However, this ceiling may be justifiable in that the constant plays a role relative to dimensions at level 2 or 3 that the N3 term plays relative to dimensions at level 3 only.

The simple main effect models 1 and 1b can be rejected on model fit, with significantly poorer AIC/BIC values than other models. The significant interaction terms present in the other models suggest that neither 1 nor 1b are appropriate. In addition, the inclusion of an anchoring point on these models has the largest impact on health state valuations.

Model 2 is the model that is most consistent with existing studies internationally and provides a point of comparison between the Australian population's preferences and those of other populations in other countries. This is presented in Figure 2. The N3 term is significant, and has a similar effect in the Australian models to that seen in other countries. This comparison also suggests broad consistency between Australian valuations and internationally.

Models 3 and 4 take a more sophisticated approach to interactions, and both represent an improvement in model fit over Model 2. Model 4 includes all interaction terms, whereas Model 3 includes only interactions between level 3 of dimensions. Additional combinations of interactions were considered (such as including only interactions involving at least one level 3 dimension, or limiting interactions to specific dimensions), but did not prove better than those reported here. In terms of AIC and BIC, Model 3 is preferred to Model 4. In both cases some of the interaction terms are not significantly different from zero. This is particularly the case in Model 4. While this may be the effect of sample size given that this model includes a large number of estimated coefficients, the fact that there is not a consistent pattern of interactions also suggests that many of these effects may not impact on the valuation placed on the health state beyond the main effect. As expected given the pattern of non-statistically significant interaction terms in Model 4, it does not provide an improvement over Model 3 when compared using the AIC and BIC.

In Model 3 the interaction terms are more consistently significant, and typically positive. In particular, the interaction terms for mobility with pain/discomfort, self care with pain/discomfort and self care with anxiety/depression are all statistically significant and positive. Comparing Models 2 and 3 it can be seen that the main effect for these terms in Model 4 is much larger. While not all interaction terms are significant, the improvement in fit and the significance of interactions between the mobility, pain/discomfort, self care and anxiety/depression dimensions suggests that this model is to be preferred over Model 2, and provides a more appropriate algorithm for the Australian population. In balancing parsimony with predictive value, we recommend Model 3 as the preferred Australian algorithm although we also recommend that the effect of using alternate specifications be considered as part of sensitivity analysis in economic evaluation.

There were 14 non-monotonic pairwise orderings of health states in the algorithm implied by Model 3. In these pairs, the value placed on the poorer health state exceeded the value placed on the better one by up to 0.079 (mean of 0.028). Because the interaction effects generally offset the main effect (reflecting the fact that the move to a worse level on one dimension depends on the levels of other dimensions, and is generally smaller when other dimensions are already at lower levels), it is possible for non-monotonic effects to occur. Given the means and standard deviations of the estimated coefficients that generate these implausible orderings, it is likely that this is a result of sample size rather than valuations (that is, they are generally very small and may result from random error in the data). These non-monotonic orderings are problematic because if used in economic evaluation they would produce implausible cost-effectiveness results, and, therefore, a method was proposed for removing these from the final algorithm implied by Model 3.

The scores for these health states were then amended by considering each illogically ordered pair and assigning to each the mean value of the two health states under the algorithm. This approach was taken as it minimised the maximum movement of a health state away from the state assigned through the preferred algorithm. Functionally this is equivalent to treating the valuations of the two health states as the same, and treating the non-monotonic effects as random error. In situations in which a health state is in more than one illogically ordered pair, the mean score which does not produce a new illogically ordered pair was selected.

The updated valuation of all EQ-5D health states under Model 3 with the amendment for illogical pairings in given in Appendix 1.

The comparability of the amended Model 3 algorithm to that produced elsewhere can be addressed using the graphical approach taken by Badia et al. [9]. Ranking each of the 243 states using the Australian algorithm, each state is valued under a selection of the pre-existing algorithms and placed on one graph. Figure 2 compares the Australian weights with a selection of other studies (in this case, UK, Spain, Japan).

This study provides the first Australian general population derived TTO EQ-5D weights for use in Australian cost-utility analysis. The broad consistency of the health state values predicted by Model 3 with those from other studies undertaken elsewhere using the same regression gives us confidence that the valuation studies are comparable. However, the more comprehensive approach taken in this study to both the absolute number and descriptive content of health states included for direct valuation within the preference elicitation study, suggests that a more complex scoring algorithm than traditionally applied may be more appropriate. Further research is required to confirm the pattern of interactions in other countries and settings.

1. Dolan P, Gudex C, Kind P, et al. The time trade-off method: results from a general population study. Health Econ 1996;5:141-54.
2. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ 2002;21:271-92.
3. Horsman J, Furlong W, Feeny D, et al. The Health Utilities Index (HUI): concepts, measurement properties and applications. Health Qual Life Outcomes 2003;1:54.
4. Torrance GW, Furlong W, Feeny D, et al. Multi-attribute preference functions. Health Utilities Index. Pharmacoeconomics 1995;7:503-20.
5. Hawthorne G, Richardson J, Day NA. A comparison of the Assessment of Quality of Life (AQoL) with four other generic utility instruments. Ann Med 2001;33:358-70.
6. Department of Health and Ageing. *Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee (Version 4.2) (http://www.health.gov.au/internet/main/publishing.nsf/Content/pbacguidelines-index)*. Canberra, 2007.
7. Brazier J, Deverill M, Green C. A review of the use of health status measures in economic evaluation. J Health Serv Res Policy 1999;4:174-84.
8. Dolan P. Modelling Valuations for EuroQol Health States. Med Care 1997;35:1095-108.
9. Badia X, Roset M, Herdman M, et al. A comparison of United Kingdom and Spanish general population time trade-off values for EQ-5D health states. Med Decis Making 2001;21:7-16.
10. Jelsma J, Hansen K, De Weerdt W, et al. How do Zimbabweans value health states? Popul Health Metr 2003;1:11.

11. Lamers LM, McDonnell J, Stalmeier PF, et al. The Dutch tariff: results and arguments for an effective design for national EQ-5D valuation studies. Health Econ 2006;15:1121-32.
12. Shaw JW, Johnson JA, Coons SJ. US valuation of the EQ-5D health states: development and testing of the D1 valuation model. Med Care 2005;43:203-20.
13. Tsuchiya A, Ikeda S, Ikegami N, et al. Estimating an EQ-5D population value set: the case of Japan. Health Econ 2002;11:341-53.
14. Wittrup-Jensen KU, Lauridsen JT, Gudex C, et al. Estimating Danish EQ-5D tariffs using the time trade-off (TTO) and visual analogue scale (VAS) methods. In: Norinder A, Pedersen KL, Roos P, (eds). Proceedings of the 18th Plenary Meeting of the EuroQol Group. Copenhagen: 2001.
15. Devlin NJ, Hansen P, Kind P, et al. Logical inconsistencies in survey respondents' health state valuations -- a methodological challenge for estimating social tariffs. Health Econ 2003;12:529-44.
16. Norman R, Cronin P, Viney R, et al. International Comparisons in Valuing EQ-5D Health States: A Review And Analysis. Value Health 2009;12:1194-200.
17. Knies S, Evers SM, Candel MJ, et al. Utilities of the EQ-5D: Transferable or Not? Pharmacoeconomics 2009;27:767-79.
18. Janssen MF, Birnie E, Haagsma JA, et al. Comparing the standard EQ-5D three-level system with a five-level version. Value Health 2008;11:275-84.
19. Janssen MF, Birnie E, Bonsel GJ. Quantification of the level descriptors for the standard EQ-5D three-level system and a five-level version according to two methods. Qual Life Res 2008;17:463-73.
20. Brazier J, Ratcliffe J, Salomon JA, et al. Measuring and valuing health benefits for economic evaluation Oxford: Oxford University Press, 2007.
21. Lee YK, Nam HS, Chuang LH, et al. South Korean Time Trade-Off Values for EQ-5D Health States: Modeling with Observed Values for 101 Health States. Value Health 2009.
22. Dey A. Orthogonal Fractional Factorial Designs New York: Wiley, 1985.
23. Tilling C, Devlin N, Tsuchiya A, et al. Protocols for time tradeoff valuations of health states worse than dead: a literature review. Med Decis Making;30:610-9.
24. Norman R, King M, Clarke D, et al. Does mode of administration matter? Comparison of on line and face-to-face administration of a time trade-off task. Qual Life Res 2010;19:499-508.
25. Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control 1974;19:716-23.
26. Schwarz GE. Estimating the dimensions of a model. Annals of Statistics 1978;6:461-64.

Figure 1: Comparisons of utility weights by model using Model 3 as base

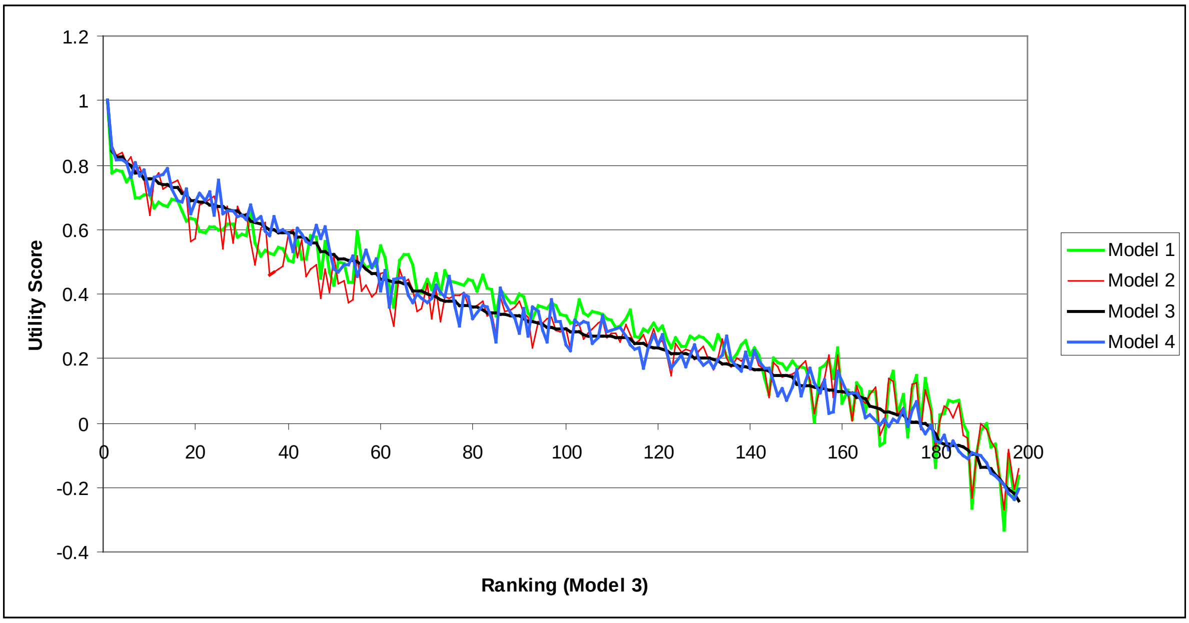

Figure 2: Comparison of utility weights for all 243 health states based on the preferred Australian algorithm (Model 3) with corresponding values from existing UK, Japan and Spain algorithms

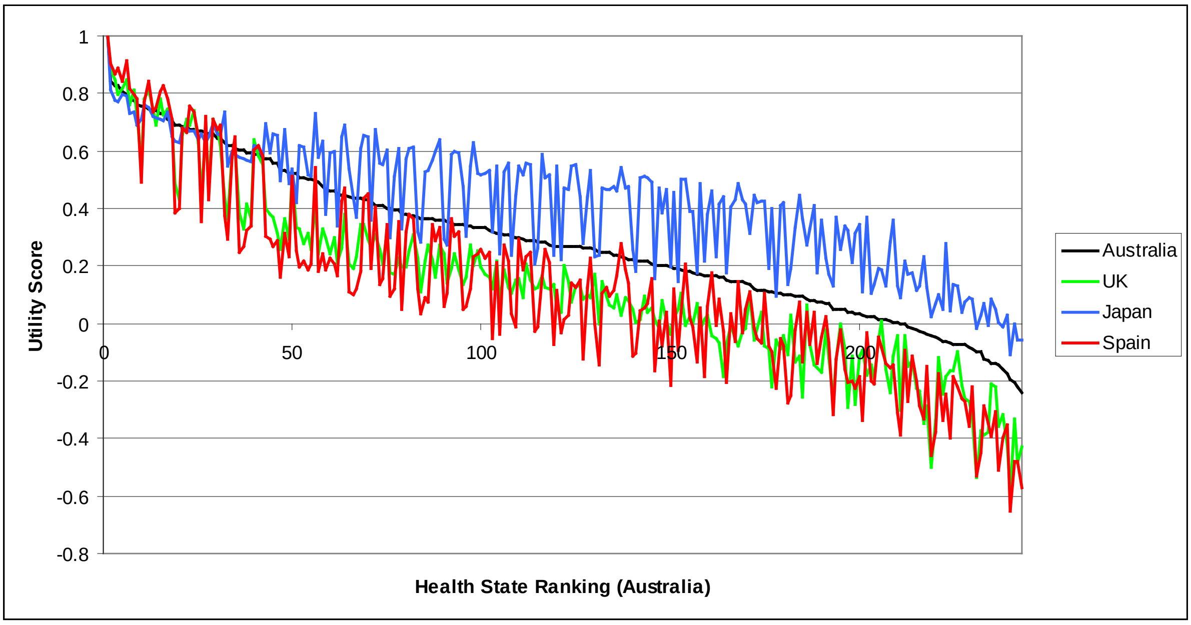

Table 1: Definition of Variables

Variable	Definition	Used in
MO2	1 if mobility is level 2; 0 otherwise	All models
MO3	1 if mobility is level 3; 0 otherwise	All models
SC2	1 if mobility is level 2; 0 otherwise	All models
SC3	1 if mobility is level 3; 0 otherwise	All models
UA2	1 if mobility is level 2; 0 otherwise	All models
UA3	1 if mobility is level 3; 0 otherwise	All models
PD2	1 if mobility is level 2; 0 otherwise	All models
PD3	1 if mobility is level 3; 0 otherwise	All models
AD2	1 if mobility is level 2; 0 otherwise	All models
AD3	1 if mobility is level 3; 0 otherwise	All models
N3	1 if any dimension is level 3; 0 otherwise	Models 2/2b
XXa x YYb	1 if dimension XX is level a (where a≠1) and dimension YY is level b (where b≠1) and XX ≠ YY (in model 3/3b, b=3); 0 otherwise	Models 3/3b/4/4b

Table 2: Simulation Results (Main Effects)

Variable	Coefficient	Dolan (SD)	OMEP (SD)	Tsuchiya (SD)	Full factorial (SD)	Full factorial plausible (SD)
Constant	**0.15**	0.168(0.055)	0.143(0.081)	0.149(0.043)	0.149(0.069)	0.151(0.073)
MO2	**0.1**	0.042(0.052)	0.092(0.048)	0.101(0.082)	0.098(0.049)	0.108(0.041)
MO3	**0.3**	0.276(0.067)	0.302(0.050)	0.318(0.086)	0.300(0.059)	0.312(0.058)
SC2	**0.1**	0.084(0.055)	0.092(0.049)	0.090(0.065)	0.101(0.049)	0.097(0.053)
SC3	**0.2**	0.245(0.067)	0.191(0.054)	0.200(0.080)	0.202(0.056)	0.195(0.056)
UA2	**0.1**	0.113(0.063)	0.099(0.048)	0.098(0.071)	0.101(0.049)	0.098(0.050)
UA3	**0.15**	0.131(0.074)	0.148(0.05)	0.137(0.087)	0.151(0.051)	0.141(0.051)
PD2	**0.2**	0.203(0.052)	0.197(0.049)	0.191(0.067)	0.200(0.047)	0.200(0.047)
PD3	**0.4**	0.522(0.058)	0.404(0.052)	0.394(0.072)	0.396(0.050)	0.394(0.050)
AD2	**0.15**	0.083(0.054)	0.151(0.053)	0.163(0.067)	0.149(0.046)	0.151(0.044)
AD3	**0.3**	0.369(0.060)	0.299(0.056)	0.304(0.072)	0.297(0.050)	0.295(0.049)
N3	**0.2**	0.269(0.073)	0.197(0.075)	0.199(0.071)	0.198(0.075)	0.192(0.075)

Table 3: Sample Characteristics

	TTO Sample (n=417)	Australian Population
Male (%)	50.4	49.3
18-24	9.4	6.6
25-34	9.1	9.1
35-44	8.6	9.4
45-54	10.3	8.9
55-64	8.6	7.3
65+	4.6	7.8
Female (%)	49.6	50.7
18-24	8.2	6.3
25-34	7.9	9.0
35-44	9.6	9.5
45-54	12.0	9.1
55-64	9.1	7.4
65+	2.9	9.4
Australia born (%)	78.4	76.0
Household income (weekly gross) (declined responses excluded) (%)		
Less than $500	21.6	23.5
$500-$999	28.4	24.5
$1,000-$1,999	31.2	33.3
More than $2,000	18.8	18.7
Marital status (declined responses excluded) (%)		
Never married	36.0	34.0
Previously married	12.6	12.7
Married	51.5	53.3
EQ-5D		
Those reporting problems on (%)		
Mobility	12.0	
Self-Care	1.0	
Usual Activities	9.4	
Pain / Discomfort	23.7	
Anxiety / Depression	24.5	
Attitude to task		
Difficult / very difficult	3.4	
Neither easy nor difficult	17.3	
Easy / very easy	79.4	

Table 4: Estimated coefficients from the alternative model specifications

Coefficient (SE)	Model[a]							
	1	1b	2	2b	3	3b	4	4b
Constant[b]	0.855(0.022)**		0.910(0.022)**		0.895(0.022)**		0.848(0.038)**	
MO2	-0.076(0.014)**	-0.094(0.014)**	-0.071(0.014)**	-0.081(0.014)**	-0.068(0.014)**	-0.080(0.014)**	-0.033(0.037)	-0.110(0.032)**
MO3	-0.269(0.019)**	-0.266(0.019)**	-0.264(0.019)**	-0.261(0.019)**	-0.374(0.033)**	-0.372(0.033)**	-0.355(0.047)**	-0.341(0.047)**
SC2	-0.106(0.016)**	-0.138(0.015)**	-0.104(0.016)**	-0.122(0.015)**	-0.087(0.016)**	-0.109(0.016)**	-0.040(0.038)	-0.123(0.031)**
SC3	-0.202(0.017)**	-0.228(0.017)**	-0.169(0.017)**	-0.181(0.017)**	-0.267(0.025)**	-0.291(0.025)**	-0.172(0.051)**	-0.236(0.048)**
UA2	-0.082(0.016)**	-0.110(0.016)**	-0.048(0.017)**	-0.060(0.016)**	-0.053(0.017)**	-0.072(0.017)**	0.002(0.041)	-0.089(0.034)**
UA3	-0.149(0.017)**	-0.175(0.016)**	-0.085(0.018)**	-0.093(0.018)**	-0.139(0.024)**	-0.165(0.023)**	-0.139(0.047)**	-0.203(0.044)**
PD2	-0.073(0.015)**	-0.099(0.015)**	-0.082(0.015)**	-0.098(0.014)**	-0.068(0.015)**	-0.085(0.015)**	-0.031(0.040)	-0.118(0.033)**
PD3	-0.308(0.015)**	-0.331(0.015)**	-0.268(0.016)**	-0.277(0.016)**	-0.449(0.022)**	-0.473(0.022)**	-0.437(0.042)**	-0.519(0.037)**
AD2	-0.090(0.015)**	-0.120(0.015)**	-0.086(0.015)**	-0.103(0.014)**	-0.097(0.015)**	-0.118(0.015)**	-0.087(0.039)*	-0.179(0.032)**
AD3	-0.259(0.015)**	-0.285(0.015)**	-0.214(0.016)**	-0.223(0.016)**	-0.397(0.023)**	-0.424(0.023)**	-0.394(0.042)**	-0.484(0.036)**
N3			-0.180(0.020)**	-0.201(0.019)**				
MO2_SC2							0.013(0.036)	0.039(0.035)

MO2_SC3							-0.029(0.038)	-0.010(0.037)
MO2_UA2							-0.052(0.037)	-0.033(0.037)
MO2_UA3							0.010(0.036)	0.036(0.036)
MO2_PD2							0.002(0.036)	0.036(0.035)
MO2_PD3							0.011(0.038)	0.043(0.037)
MO2_AD2							-0.037(0.036)	-0.003(0.035)
MO2_AD3							-0.027(0.038)	0.007(0.037)
MO3_SC3					0.064(0.034)	0.061(0.034)	0.050(0.043)	0.043(0.043)
MO3_UA3					-0.025(0.034)	-0.031(0.034)	0.015(0.044)	0.018(0.044)
MO3_PD2							0.003(0.050)	0.000(0.050)
MO3_PD3					0.092(0.033)**	0.094(0.033)**	0.107(0.048)*	0.106(0.048)*
MO3_AD2							-0.062(0.049)	-0.063(0.050)
MO3_AD3					0.013(0.035)	0.016(0.035)	-0.019(0.048)	-0.015(0.048)
SC2_UA2							-0.049(0.041)	-0.019(0.04)
SC2_UA3							-0.020(0.043)	-0.002(0.043)
SC2_PD2							-0.008(0.039)	0.030(0.038)
SC2_PD3							-0.049(0.043)	-0.018(0.042)
SC2_AD2							-0.035(0.040)	-0.002(0.039)
SC2_AD3							-0.043(0.043)	-0.015(0.043)

SC3_UA2							-0.075(0.048)	-0.047(0.047)
SC3_UA3					-0.055(0.030)	-0.050(0.030)	-0.087(0.046)	-0.071(0.046)
SC3_PD2							-0.090(0.043)*	-0.060(0.043)
SC3_PD3					0.090(0.030)**	0.100(0.030)**	0.008(0.044)	0.025(0.044)
SC3_AD2							0.019(0.044)	0.050(0.044)
SC3_AD3					0.105(0.031)**	0.104(0.031)**	0.093(0.047)*	0.121(0.046)**
UA2_PD2							-0.012(0.042)	0.024(0.041)
UA2_PD3							0.014(0.045)	0.047(0.045)
UA2_AD2							0.001(0.042)	0.046(0.040)
UA2_AD3							0.015(0.047)	0.057(0.046)
UA3_PD2							-0.029(0.041)	-0.005(0.041)
UA3_PD3					0.025(0.030)	0.032(0.030)	0.008(0.044)	0.037(0.043)
UA3_AD2							0.037(0.042)	0.065(0.042)
UA3_AD3					0.043(0.030)	0.060(0.030)*	0.059(0.045)	0.088(0.044)*
PD2_AD2							-0.005(0.038)	0.031(0.037)
PD2_AD3							0.027(0.040)	0.060(0.039)
PD3_AD2							0.042(0.041)	0.078(0.040)
PD3_AD3					0.185(0.029)**	0.186(0.029)**	0.223(0.041)**	0.258(0.040)**
Log	-3070.7	-3092.8	-3029.5	-3037.5	-2987.8	-2999.1	-2975.8	-2983.6

likelihood								
AIC	6167.32	6209.68	6086.96	6101.05	6021.51	6042.21	6053.57	6067.13
BIC	6252.06	6287.90	6178.21	6185.79	6171.43	6185.60	6385.99	6393.03

* Significant at 5% level

** Significant at 1% level

NOTES:

a As the final set of directly valued health states did not include any co-occurrence of MO3 with either SC1 or UA1, no interaction was fitted between MO3 and SC2, or between MO3 and UA2 in Models 3-4b.

b The null hypothesis is that the constant is one rather than zero