CHERE
CENTRE FOR HEALTH ECONOMICS RESEARCH AND EVALUATION

UNIVERSITY OF TECHNOLOGY SYDNEY

UTS:CHERE

The Centre for Health Economics Research and Evaluation (CHERE) was established in 1991. CHERE is a centre of excellence in health economics and health services research. It is a joint Centre of the Faculties of Business and Nursing, Midwifery and Health at the University of Technology, Sydney, in collaboration with Central Sydney Area Health Service. It was established as a UTS Centre in February, 2002. The Centre aims to contribute to the development and application of health economics and health services research through research, teaching and policy support. CHERE's research program encompasses both the theory and application of health economics. The main theoretical research theme pursues valuing benefits, including understanding what individuals value from health and health care, how such values should be measured, and exploring the social values attached to these benefits. The applied research focuses on economic and the appraisal of new programs or new ways of delivering and/or funding services. CHERE's teaching includes introducing clinicians, health services managers, public health professionals and others to health economic principles. Training programs aim to develop practical skills in health economics and health services research. Policy support is provided at all levels of the health care system by undertaking commissioned projects, through the provision of formal and informal advice as well as participation in working parties and committees.

University of Technology, Sydney
City campus, Haymarket
PO Box 123 Broadway NSW 2007
Tel: +61 2 9514 4720
Fax: + 61 2 9514 4730
Email: mail@chere.uts.edu.au
www.chere.uts.edu.au

Valuing EQ-5D Health States: A Review and Analysis

Presented at AHES 2007

Richard Norman[1], Paula Cronin[1], Rosalie Viney[1], Madeleine King[1], Deborah Street[1,2], John Brazier[3], Julie Ratcliffe[3]

CHERE WORKING PAPER 2007/9

1. Centre for Health Economics Research and Evaluation
Faculty of Business
University of Technology, Sydney

2. Centre for the Study of Choice (CenSoc)
Faculty of Business
University of Technology, Sydney

3. Health Economics and Decision Science, ScHARR
University of Sheffield

First Version: October 2007
Current Version: October 2007

Abstract

Objective: To identify the key methodological issues in the construction of population-level EQ-5D / Time Trade-Off (TTO) preference elicitation studies.

Study Design: This study involves three components. The first was to identify existing population-level EQ-5D TTO studies. The second was to illustrate and discuss the key areas of divergence between studies, including the international comparison of tariffs. The third was to portray the relative merits of each of the approaches, and to compare the results of studies across countries.

Results: While most papers report use of the protocol developed in the original UK study, we identified three key areas of divergence in the construction and analysis of surveys. These are the number of health states valued in order to determine the algorithm for estimating all health states, the approach to valuing states worse than immediate death, and the choice of algorithm. Finally, the evidence on international comparisons suggests differences between countries, although it is difficult to disentangle differences in cultural attitudes with random error and differences due to methodological divergence.

Conclusion: Differences in methods are likely to obscure true differences in values between countries. However, population-specific valuation sets for countries engaging in economic evaluation would better represent societal attitudes.

Introduction

Health care spending, as a proportion of gross domestic product (GDP) has increased over the past thirty years in developed countries. In the United States, this percentage increased from 7.6% in 1972 to 14.0% in 1992, to around 16.0% in 2004 [1, 2]. This is not confined to health care systems with predominantly private funding. Publicly funded systems, as in the United Kingdom, and social insurance-based systems, as in France, have also witnessed large absolute and proportional increases in expenditures. As society's investment in health care increases, appropriate and transparent decision-making by policy makers becomes increasingly important. As a result, economic evaluation is used increasingly by health system decision-makers in Australia and internationally to determine allocation of health care resources between services, and levels of subsidy. Public funding decisions require assessment of resource allocation across diverse diseases and treatments with varying impacts on health outcomes. Cost-utility analysis (CUA) is the main approach used to measure and value the impacts of treatments: the US Panel on Cost-Effectiveness in Health and Medicine recommends the use of quality adjusted life years (QALYs) [3]; the UK National Institute of Clinical Excellence has most commonly used CUA [4, 5] and has recently recommended that it should be the preferred outcome measure; and CUA is increasingly used in Australia in the evaluation of pharmaceuticals and medical services. In the recently released PBAC guidelines, a preference is expressed for the use of CUA.

The advantage of CUA is that it quantifies and values benefit from interventions across all fields of health and healthcare in a common metric (predominantly the Quality-Adjusted Life Year (QALY)), allowing decision-makers to assess the relative value of many different demands for resource allocation. The time in a particular health state is weighted by a quality weight (utility weight) between zero (death) and one (full health) that reflects society's willingness to trade-off between quality of life and survival. Weights less than zero reflect health states worse than death. QALYs are designed to allow comparisons across interventions with disparate outcomes, across different

health care conditions and population groups, thus providing information required by health policy makers. While CUA is simple in concept, it presents challenges in practice. Eliciting valuations for all health states that may be relevant to a disease or intervention is time-consuming and costly. Further, comparison of valuations across interventions and diseases requires comparability of methods.

The QALY approach requires an accurate description of the health outcomes associated with a chronic condition and a method for eliciting preferences for the health outcomes. A number of approaches for describing and eliciting preferences for health outcomes have been developed, but increasingly it is recognised that there is value in standardised methods for measuring and valuing health states for use in economic evaluation. The measurement of health-related quality of life (HRQoL) has become increasingly important for economic evaluations as cost-utility analysis has become the preferred method of investigating cost-effectiveness. In particular, the potential role for standardised instruments which allow for both description and valuation of quality of life is increasingly recognised. This contrasts with the approach that has been commonly used of development of scenario based vignettes based on clinical and quality of life data, which are then used as a basis for health state valuation. The QALY framework assumes that the value of a chronic health profile can be characterised by the product of life expectancy and HRQoL. Therefore, it is of paramount importance to consider how HRQoL can estimated in a way that represents improvements across the spectrum of potential interventions without bias.

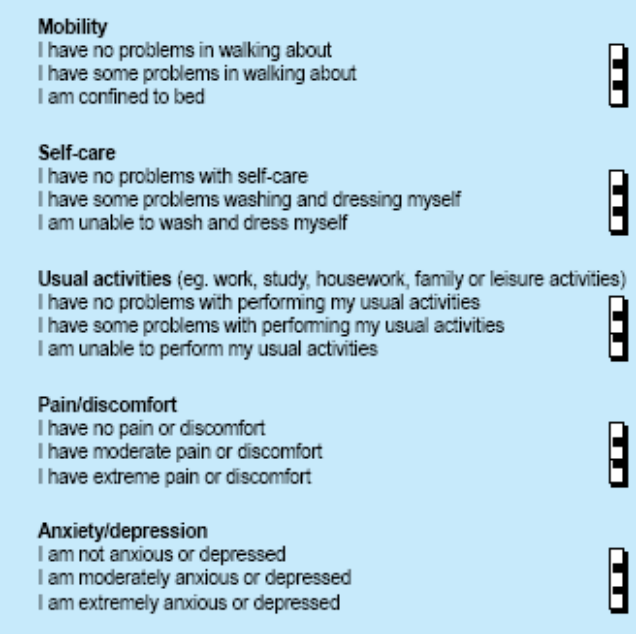
Multi-attribute utility (MAU) instruments such as the EQ-5D, the SF-6D, HUI (Health Utilities Index), and AQoL (Assessment of Quality of Life), have facilitated comparability. [6, 7] These generic instruments comprise a descriptive quality of life instrument (a set of items or statements with multiple response categories that cover a range of dimensions of HRQoL) and a set of utility weights which reflect the strength of a community's preferences for the health states. The utility weights are derived from a scoring algorithm that relates each health state described by the MAU instrument to a single number that

reflects the value of that health state. The scoring algorithm is usually generated by eliciting responses from a population sample using a scaling technique such as time trade-off (TTO), standard gamble (SG) or visual analogue scale (VAS). Statistical methods are used to estimate a model that relates the responses derived from the scaling technique to the dimensions and levels in the descriptive system. This model is the basis of the scoring algorithm for the utility weights. For utility weights to be meaningful for economic evaluation, the scaling technique must reflect trade-offs individuals are willing to make between health outcomes. The numeraire for capturing such trade-offs has typically been time (for example, QALYs), with the time adjustment derived from one of the scaling techniques above. The key advantage of MAUI approach is that it provides community based valuation of health states but direct contemporaneous description of QOL by patients who are experiencing the state.

While the role of MAU instruments in economic evaluation is increasing, there remain a number of methodological issues associated with them. In particular, the applicability of a particular instrument depends on its capacity to reflect accurately the health state valuations associated with particular health states, and to capture and quantify the value of the change in quality of life associated with particular treatments. Recent reviews have noted that there are significant differences in the performance of different MAUIs. [8] While this has been attributed to differences in the dimensions in the instruments and to the preference elicitation techniques, there has been relatively little critical appraisal of the methods of development of MAUI scoring algorithms. In this paper, we examine these issues by considering the EQ-5D [9]. We chose the EQ-5D because it is widely used and there are a number of different studies which have been undertaken to develop country specific scoring algorithms. Because the focus of this review is on one MAUI, we do not consider the psychometric aspects of the instrument, but rather focus on the methods for development of the scoring algorithm. Many of the issue we raise are relevant to other MAUIs.

Overview of the EQ-5D

Figure 1: The EQ-5D (Source: Kind et al, 1998)



The EQ-5D is a tool developed by the Euroqol group (www.euroqol.org). The EQ-5D has five dimensions, intended to represent the major areas in which health changes can manifest. These areas are mobility, self-care, usual activities, pain/discomfort and anxiety/depression (See Figure 1). Each dimension contains three levels, loosely termed 'No Problems', 'Some Problems', and 'Significant Problems'. Thus, there are $3^5 = 243$ potential states in the descriptive system. The TTO approach is used to value a selection of these states, and then to impute values for the remainder using simple regression. The use of TTO for valuing EQ-5D states is well described elsewhere ([9], [10] etc). For states considered to be preferable to immediate death, a respondent is faced with a choice between ten years of a particular chronic health state defined in EQ-5D space with a period of *x* years in full health. The aim of the TTO is to identify a value of *x* for which the individual is indifferent in the choice. The value for the health is defined as *x* / 10.

Our analysis of this EQ-5D / TTO approach involves two strands: Firstly, we look at how to elicit societal valuations for EQ-5D states under the standard TTO protocol. This paper identifies some key themes and issues that run across the population valuation studies. It then identifies some key themes

and issues that run across the literature base. Finally, it looks at international comparisons and discusses whether it is necessary to provide nationality-specific tariffs for the EQ-5D valuation system.

Methods

The initial target of the study was to identify all large general population valuations studies employing the EQ-5D as the tool for describing health. EMBASE and MEDLINE were searched for such papers. To be considered for inclusion, the analysis had to present primary research in English and be published since 1995. Since it was expected that a proportion of good quality reports may be unavailable in peer-reviewed publications, the reference lists of papers identified in the main search were used to identify further studies. Since all of these identified non-peer reviewed publications were available on the EuroQol website (www.euroqol.org), the list of EuroQol Plenary Meeting Proceedings was scanned for further studies relevant to this work. To be included, a study had to attempt to value all 243 states described by the EQ-5D. Beyond this constraint, we chose to be conservative in our approach to exclusion as we were seeking to identify divergence in approach. The one significant exclusion was that we decided not to consider EQ-5D studies that used the visual analogue scale (VAS) as the primary method for valuing states. These were excluded due to the potential for context bias [11], and that states are not valued within a framework of choice.

For each identified study, details most relevant to the analysis of the methods used were identified. Key areas for discussion were selected. These areas were the precise formulation of the algorithm, the number of states directly valued in the survey to generate weights, the method to value states worse than death, the influence of time preferences of results, and international comparisons in predicted values across EQ-5D space.

The algorithms were compared by expanding the approach used by Busschbach et al. who compare the directly valued states in the UK, Germany and Spain.[12] For this, the UK results are used as the benchmark. The

predicted preference scores for the states under the UK algorithm are ranked in descending order. The preference scores under each of the other algorithms are generated, using the same ordering as the UK study. Using this approach, we can identify the tendency for countries to trade-off quantity of life for quality of life, and identify whether countries differ in their relative valuations of the five dimensions.

Results

10 papers were identified ([13, 14, 15, 16 ,9, 17, 18, 19, 10, 20]), of which 8 were published in peer-reviewed journals. It should be noted that there are, at present, no such results for Canada or Australia, two countries strongly supportive of the use of CUA in healthcare decision-making. 2 studies utilised the Visual Analogue Scale (VAS) as the primary method of valuation ([14, 16]) so were excluded. The details of the remaining papers are given in the Appendix.

Three significant methodological differences emerged regarding the survey structure, and the development of the algorithm. The first regarded the number of states that are needed to be directly valued to estimate valuations for the complete EQ-5D space. The second is the approach to valuing states considered to be worse than death. The third is the choice of the algorithm to model those states not directly valued.

The Number of Directly Valued States

Given that the EQ-5D has 243 individual possible states, it is unsurprising that no study has attempted to ask respondents to directly value each of these states. Therefore, the pertinent question becomes how best to form a representative fraction of the entire space which allows a good estimation of the remainder of the EQ-5D. Two approaches have been adopted to form this representative fraction. The original approach [9, 21] valued 43 states, and each respondent directly valued a subset. The alternative approach was developed and uses 17 states, all rated by each respondent [10].

Table 1: The States Selected by Dolan (1996) and Tsuchiya (2002)

States used by Tsuchiya et al, 2002	States used by Dolan et al, 1996			
	Very Mild (2 of 5)	Mild (3 of 12)	Moderate (3 of 12)	Severe (3 of 12)
11112	11112	11122	13212	33232
11113	11121	11131	32331	23232
11121	11211	11113	13311	23321
11131	12111	21133	22122	13332
11133	21111	21222	12222	22233
11211		21312	21323	22323
11312		12211	32211	32223
12111		11133	12223	32232
13311		22121	22331	33321
21111		12121	21232	33323
22222		22112	32313	23313
23232		11312	22222	33212
32211				
32223				
32313				
33323				

Lamers et al. investigate these alternative approaches. [18] The authors re-analysed data from Dolan et al., [9] simulating valuation studies in which 12, 17, 22, 27, 32, 37 and 42 of the 42 states[1] are directly valued in samples of 50, 100, 200, 300, 400, 600 and 800 per state valued. The outcome for each of these combinations is the mean absolute error (MAE) between the predicted values from the subsequent algorithm and the values observed in the data set.[2]

As expected, the MAE is negatively associated with both the sample size and the number of health states directly valued. Additionally, they contrast these data with the results of Dolan et al. [22] which suggests that not only does the 17-state approach used by Tsuchiya et al. [10] lead to a lower MAE than that of

[1] Note that this deals with 42 since it is assumed that all respondents value 11111.
[2] MAE is a useful tool for estimating appropriateness as it shows the fit of the model to the data. However, other diagnostics might also be of value, for example out-of-sample or split-sample prediction (of directly valued states or otherwise).

Dolan et al. (1996), it leads to a lower MAE than if each respondent valued 17 (or even 22) randomly assigned states from the 42. The mean correlation for the predicted and actual values if 22 states from 42 are randomly selected is 0.986 (SD = 0.006) whereas the figures for the 17 states used by Tsuchiya et al. was 0.989 (SD = 0.002). [10] Thus, the direct valuation of 17 states leads to more accurate prediction within a main effects model.

A related question concerns whether the 17 and 43 state approaches are optimal in terms of study design. To allow equal precision in each of the effect estimates, it is necessary to have equal frequency of appearance for each of the levels. A disproportionate number of relatively better health states will lead to better precision at that end of the scale, and to less at the lower end. In addition, estimating interaction terms between dimensions becomes difficult as certain combinations of states rarely (or never) occur in directly valued states. In the 17 states directly valued by Tsuchiya et al,[10] and the 43 valued by Dolan et al. [9], there is a disproportionate number of relatively better states i.e. level one attributes are over-represented.

Valuing States Worse Than Death

While it is plausible that the poorer states in the EQ5D might be considered worse than immediate death, certain methodological issues arise from generating an algorithm with a subset of states that include states worse than death. While anchoring death at zero and full health at one gives meaning to states that lie in that range, it is difficult to interpret different values below zero. The lack of a tool which is well-suited to this task means that existing papers have taken a range of approaches to valuing these states, some of which raise further problems.

All papers begin from the same starting point, by asking respondents to choose between immediate death and a period of ten years of life, some of which is spent in the state worse than death, and some in full health. In the majority of papers, if the individual is indifferent between immediate death and

x years of the bad state followed by (10-*x*) years of full health, the score for the state worse than death is then calculated[3] in the following way[10 13 15 18 20 21]

Preference score (State worse than death) = (*x* / 10) – 1 (1)

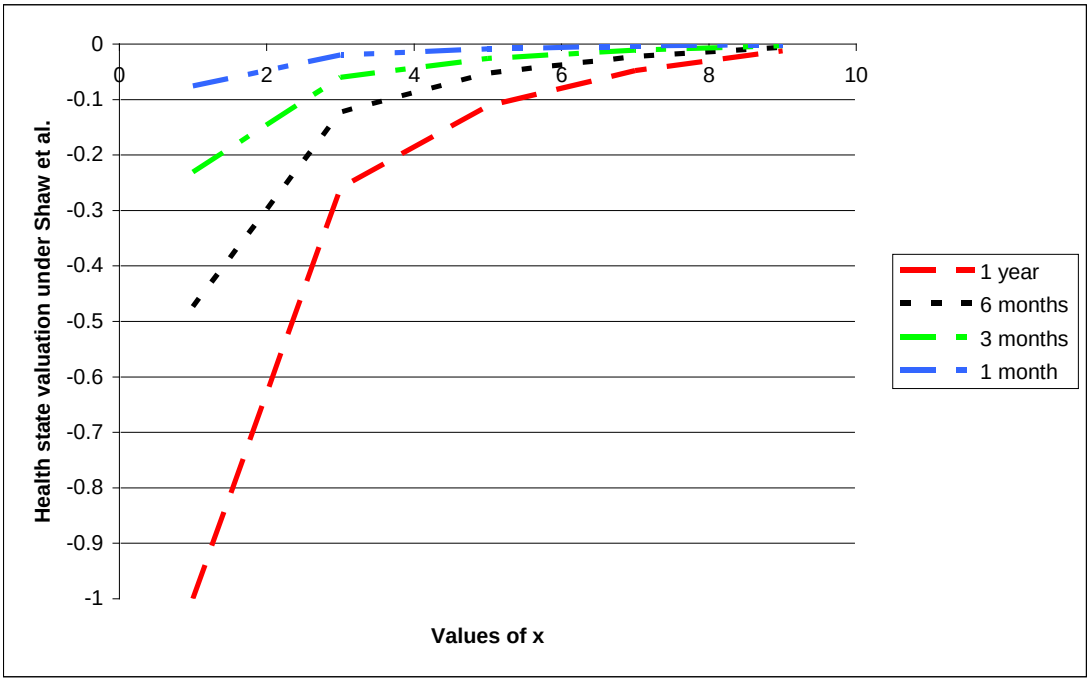
Since *x* is bounded by zero and ten, the preference score for states worse than death are bounded by 0 and -1. The one divergence from this orthodoxy is found in Shaw et al, 2005, for whom

Preference score (State worse than death) = *x* / (10-*x*) (2)

They allowed the value for *x* to be between 0.25 and 9.75 years, meaning that the preference score is initially as low as -39. This leads to a dramatic asymmetry between states better than immediate death and those worse. This is important as it means that the impact of a brief period in the severest health state is of the same magnitude as a much longer period in full health. While a poor state such as this might be plausible, it could be argued that the uncertainty surrounding interpretation of states worse than death means that the value we place on these states should not have excessive influence on the final algorithm. Shaw et al suggested that states worse than death should be bounded by negative one so they then applied a linear transformation to the raw scores, constraining all scores to be in this range. [19] The major problem with this linear transformation is that the valuations in this range are dependent on the minimum length of time the respondent is allowed to endure in the bad health state. If the minimum period allowable in the poor health state increases to, for example, one year, all negative values would be divided by nine. The effect of dividing the different health valuations by different factors (defined by the shortest allowable period in the poor health state) is illustrated in Figure 2.

Figure 2: The effect of changing minimum time duration on valuations of states worse than death

[3] Lamers et al appear to use a different algorithm but the reason for divergence is that they treat the time in full health as x, rather than the time in the worse than death state.



As health moves away from zero towards negative one, the effect of this procedural variable becomes increasingly large, and suggests that this divergence from the orthodox position is not justified.

The Construction of the Algorithm

The benchmark UK study [21] prefers a N1 model, in which the algorithm is a main effects model using dummy variables for levels in each dimension worse than 'No Problems', plus the N1 dummy variable, defined as 1 when any of the dimensions are at level 3 (the worst level). Thus,

Valuation = 1 – (constant + $\sum$(dummy$_{l,d}$ * co-efficient$_{l,d}$) + (dummy$_{N1}$ *co-efficient$_{N1}$))

Aside from increased predictive value of the model with this interaction term [9], the intuition behind using such a value is slightly uncertain. The first dimension to move to level 3 will have significant spill-over effects, perhaps not captured by the other dimensions. The need to adapt to a life with a severe impediment has a disutility which is a one-off. Thus, the second

dimension to move to level 3 will have a disutility (illustrated by the co-efficient associated with the respective dummy variable), but may have a lesser impact than if the move had occurred from a state with no pre-existing level 3 problems. The reverse argument, claiming that the N1 term has no intuitive appeal, might argue that the extra predictive value is a remnant of the correction methods used to adjust states worse than death to constrain them between zero and minus one. Since these states are considerably more likely to have level 3 dimensions than the general set of states, it is arguable that applying an erroneous transformation, compressing negative values into too small a range, might be identified through lower co-efficients being applied to level 3 parameters beyond the first.

Other than the N1 variable, most studies do not utilise interaction terms in their final algorithms. However, the intuitive argument in support of interactions can be illustrated using a number of examples (for example, the disutility of not being able to do usual activities may vary, depending on whether the person is mobile as this defines what usual activities consist of). A number investigate alternative model specifications containing interactions [18], but generally (and perhaps surprisingly) find they do not improve the fit of the model ([20, 21, 10, 13]).

The final issue regarding the algorithm is the use and interpretation of the constant term. Conventionally, the intercept reflects the value of the function when all explanatory variables are zero (level 1 in the N1 model). However, in this case, this interpretation does not hold as 11111 is axiomatically described as full health and is anchored at 1. In the identified papers, there are two approaches in the discussion of the intercept. In the majority of studies, the intercept is allowed to vary from zero, and is interpreted as the disutility associated with not being at perfect health, independent of the disutility associated with the movement within the dimension per se (Dolan, 1996). This could be justified in the same way as the N1 variable was justified above. An alternative approach is taken in a recent US study (Shaw et al, 2005). The full algorithm used in this study in illustrated here:

Valuation = 1 – (constant + $\sum$(dummy$_{l,d}$ * co-efficient$_{l,d}$) + β_1D1 + β_2I2-squared + β_3I3 + β_4I3-squared)

where D1 is the number of dimensions not at level one beyond the first, I2 is the number of dimensions at level two beyond the first and I3 is the number of dimensions at level three beyond the first. The differences between this approach and the more commonly utilised N1 approach is that Shaw et al do not allow a constant term (since full health is anchored at 1) and they identified a broader group of statistically significant interaction terms, albeit specified in a different way. One criticism of this approach, and of the N1 approach also, is that it is relatively blunt in its approach to interactions. For example, if we consider the interactions concerning dimensions being at level three, the effect of there being a number of dimensions at level three is independent of the specific dimensions at that level.

International Comparisons

The final question this paper looks at it is the extent to which the use of these different papers affects the preference scores associated with the 243 states in EQ-5D space, and thus whether the choice of algorithm is likely to alter resource allocation decisions. The respective betas are provided in the Appendix. Our results, comparing the wider range of countries using all states defined by EQ-5D space are shown in Figure 3 and Figure 4.

Figure 3 Comparing the United Kingdom to Northern European Countries

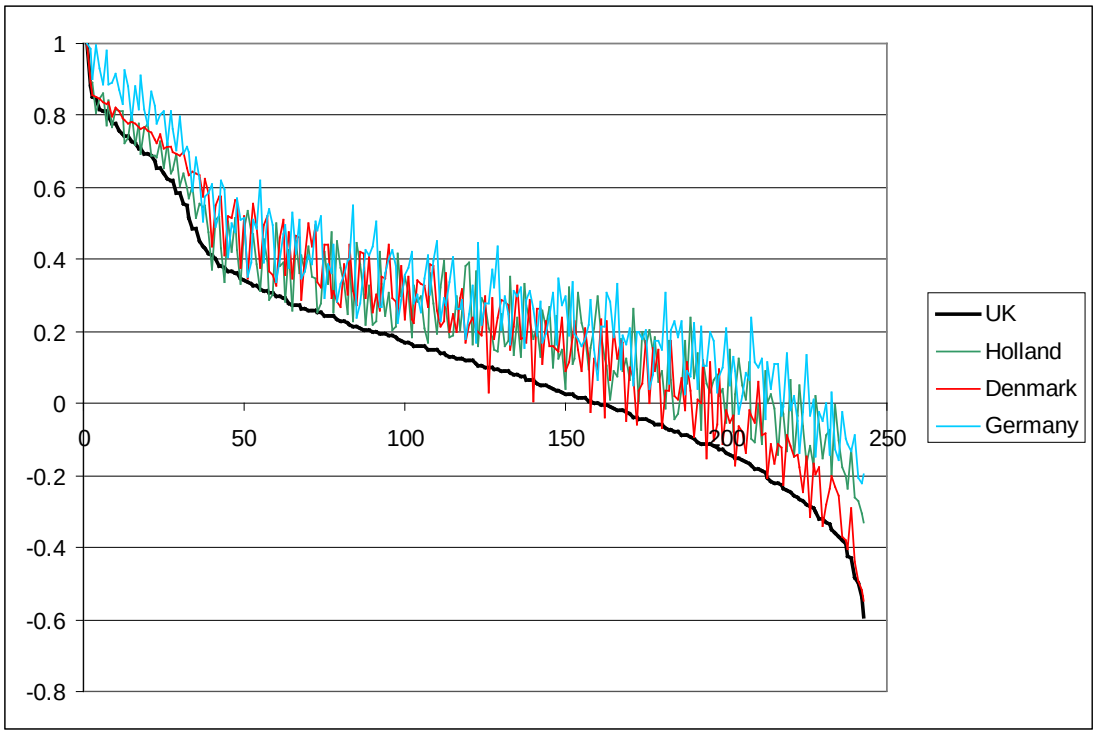
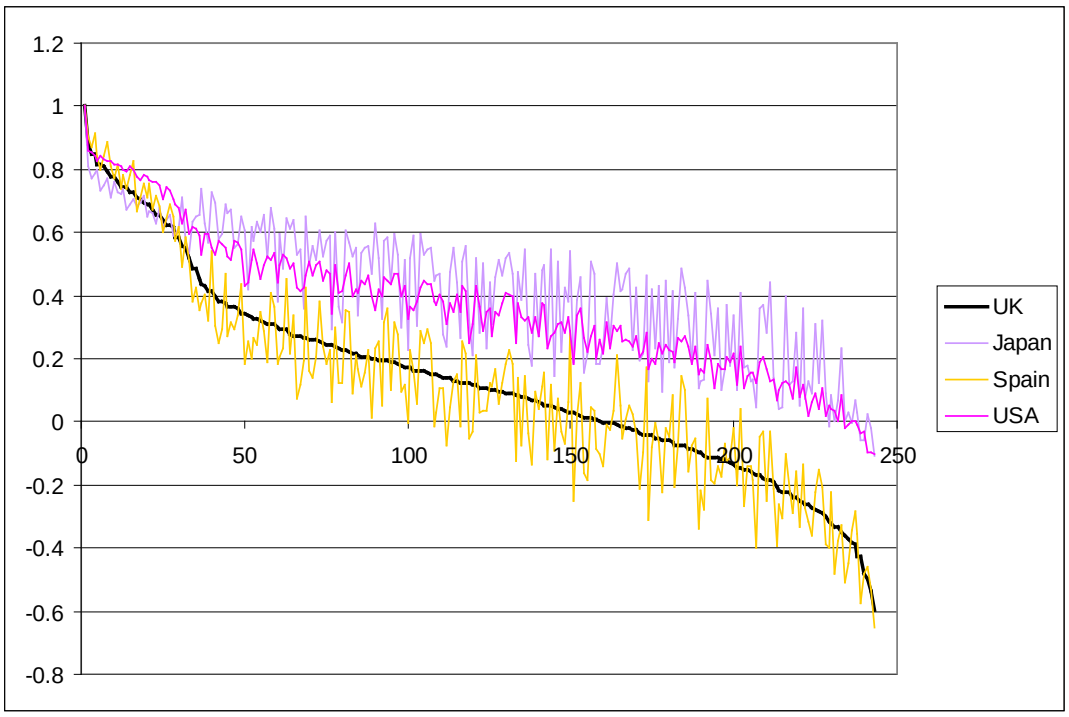


Figure 4 Comparing the United Kingdom to Non-Northern European Countries

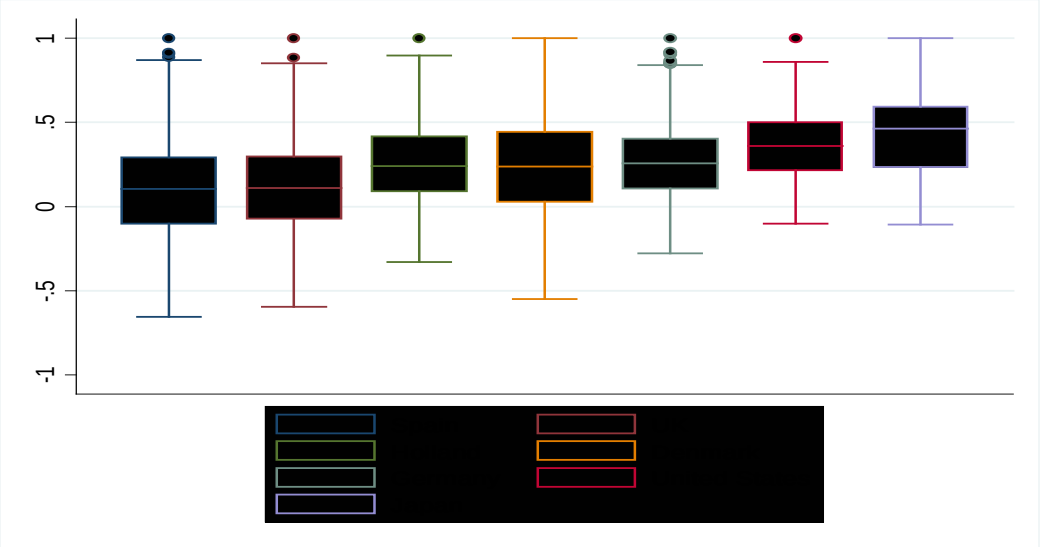


We have compared the algorithms to the benchmark in groups of three. When algorithms from Denmark, Germany, and the Netherlands are compared to

the UK study, they generate similar preference scores across the range of health states. Generally, these lie above the UK figures but follow the same trend. This suggests the various dimensions of the EQ-5D have the same approximate relative importance in these countries, but the absolute disutility attached to worsening in the health state in general is estimated to be lower.

Divergence from this trend can be seen in the countries shown in Figure 4. The Spanish algorithm does not appear to systematically differ from the UK algorithm, but displays more variance from the UK algorithm than the Northern European results (suggesting different emphasis between dimensions). The Japanese results are below those of all other algorithms for mild health states (due to a large constant term in the N1 algorithm), but, for worse states, lie above all other algorithms. Under the Japanese algorithm, there are very few states considered worse than death. Additionally, the Japanese results show considerable variance relative to the UK figures. In comparing the Japanese results to the UK, this seems to be the result of a relatively high importance being associated with mobility, and a relatively low importance being associated with pain and discomfort, and anxiety and depression. The US study follows a similar pattern to the Japanese results but displays less variability relative to the UK. This unwillingness to trade off quantity of life for quality of life in Japan and the US means the spread of HRQoL scores is lower in these countries. As noted by Luo et al. and Noyes et al., this will lead to interventions being less cost-effective in CUA as the quality of life gain is likely to be relatively smaller. [23,24] This is illustrated in Figure 5.

Figure 5: Willingness to Trade: The Range of Scores in Seven Countries



The uncertain element in interpreting these results is to identify whether the differences in algorithms are a result of genuine differences in national attitudes towards ill health, or whether they are the product of different study designs. In support of the former is the fact that Figure 3 suggests convergence between countries in a geographical locality (Northern Europe). However, we believe that, to firmly identify a trend in algorithms between countries, we would require a greater number of studies than currently exist

Discussion

This paper identifies a number of key methodological questions in the construction of population level EQ-5D TTO questionnaires. The number of states that need to be directly valued is considered, and the best solution may depend on whether it is worthwhile to look for interaction terms. We identified study design issues with the sets of states most commonly selected to be directly valued. The decision regarding number of states leads into a number of questions regarding the choice of algorithm. Then, we identified competing approaches for the valuation of states considered to be worse than death and identify that the approach used by Shaw et al. makes valuations heavily dependent on a parameter of model design (specifically the minimum period of the state considered in the TTO) which should have no effect on the valuation.

Whether country-specific algorithms are necessary is a difficult question which we have only partly addressed. There is clear divergence between countries in their valuations, both in terms of their willingness to trade quantity of life for quality, and their relative importance of the five dimensions of the EQ-5D. Our results suggest that a proportion of the divergence in algorithms is attributable to genuine cultural differences, which suggests that country-specific algorithms are of importance. This is particularly true in countries such as Canada and Australia which engage in substantial economic evaluation.

References

1. McPake B, Normand C, Kumaranayake L. *Health Economics: An International Perspective.* . London: Routledge, 2002.
2. Smith C, Cowan C, Heffler S, Caitlin A. National health spending in 2004: recent slowdown led by prescription drug spending. *Health Affairs (Millwood)* 2006;25(1):186-196.
3. Gold M. *Cost-effectiveness in health and medicine*. New York: OUP, 1996.
4. Claxton K, Sculpher M, Drummond M. A rational framework for decision making by the National Institute For Clinical Excellence (NICE). *Lancet* 2002;360(9334):711-5.
5. Taylor R. Using health outcomes data to inform decision-making: government agency perspective. *Pharmacoeconomics* 2001;19 Suppl 2:33-8.
6. Brazier J, Deverill M, Green C, Harper R, Booth A. A review of the use of health status measures in economic evaluation. *Health Technol Assess* 1999;3(9):i-iv, 1-164.
7. Spilker B. *Quality of life and pharmacoeconomics in clinical trials*. Philadelphia: Lippencott-Raven, 1996.
8. 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(5):358-70.
9. Dolan P, Gudex C, Kind P, Williams A. The time trade-off method: results from a general population study. *Health Econ* 1996;5(2):141-54.
10. Tsuchiya A, Ikeda S, Ikegami N, Nishimura S, Sakai I, Fukuda T, et al. Estimating an EQ-5D population value set: the case of Japan. *Health Economics* 2002;11(4):341-53.
11. Krabbe PF, Stalmeier PF, Lamers LM, Busschbach JJ. Testing the interval-level measurement property of multi-item visual analogue scales. *Qual Life Res* 2006;15(10):1651-61.
12. Busschbach JJ, Weijnen T, Nieuwenhuizen M, Oppe S, Badia X, Dolan P, et al. A comparison of EQ-5D time trade-off values obtained in Germany, The United Kingdom and Spain. In: Brooks R, Rabin R, De Charro F, editors. *The measurement and valuation of health status using EQ-5D: A European Perspective*. The Netherlands: Kluwer, 2003.

13. Badia X, Roset M, Herdman M, Kind P. A comparison of United Kingdom and Spanish general population time trade-off values for EQ-5D health states. *Medical Decision Making* 2001;21(1):7-16.
14. Bjork S, Norinder A. The weighting exercise for the Swedish version of the EuroQol. *Health Econ* 1999;8(2):117-26.
15. Greiner W, Claes C, Busschbach JJ, Graf von der Schulenburg JM. Validating the EQ-5D with time trade off for the German population. *Eur J Health Econ* 2004.
16. Devlin NJ, Hansen P, Kind P, Williams A. Logical inconsistencies in survey respondents' health state valuations -- a methodological challenge for estimating social tariffs. *Health Econ* 2003;12(7):529-44.
17. Jelsma J, Hansen K, De Weerdt W, De Cock P, Kind P. How do Zimbabweans value health states? *Popul Health Metr* 2003;1(1):11.
18. Lamers LM, McDonnell J, Stalmeier PF, Krabbe PF, Busschbach JJ. The Dutch tariff: results and arguments for an effective design for national EQ-5D valuation studies. *Health Econ* 2006;15(10):1121-32.
19. 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(3):203-20.
20. Wittrup-Jensen KU, Lauridsen JT, Gudex C, Brooks R, Pedersen KM. Estimating Danish EQ-5D tariffs using the time trade-off (TTO) and visual analogue scale (VAS) methods. In: Norinder A, Pedersen KL, Roos P, editors. *Proceedings of the 18th Plenary Meeting of the EuroQol Group*. Copenhagen, 2001.
21. Dolan P. Modelling Valuations for EuroQol Health States. *Medical Care* 1997;35(11):1095-1108.
22. Dolan P, Gudex C, Kind P, Williams A. A social tariff for EuroQol: Results from a UK general population study. *Centre for Health Economics York Discussion Paper Series* 1995;138.
23. Noyes K, Dick AW, Holloway RG. The implications of using US-specific EQ-5D preference weights for cost-effectiveness evaluation. *Med Decis Making* 2007;27(3):327-34.
24. Luo N, Johnson JA, Shaw JW, Feeny D, Coons SJ. Self-reported health status of the general adult U.S. population as assessed by the EQ-5D and Health Utilities Index. *Med Care* 2005;43(11):1078-86

Appendix

The algorithms across the literature base

	Con	MO2	SC2	UA2	PD2	AD2	MO3	SC3	UA3	PD3	AD3	N1
UK	0.081	0.069	0.104	0.036	0.123	0.071	0.314	0.214	0.094	0.386	0.236	0.269
Spain	0.024	0.106	0.134	0.071	0.089	0.062	0.430	0.309	0.195	0.261	0.144	0.291
Japan	0.148	0.078	0.053	0.040	0.083	0.062	0.418	0.101	0.128	0.189	0.108	0.014
Germany	0.071	0.082	0.063	0.010	0.104	0.017	0.303	0.176	0.058	0.289	0.095	0.285
Holland	0.071	0.036	0.082	0.032	0.086	0.124	0.161	0.152	0.057	0.329	0.325	0.234
Denmark	0.088	0.055	0.066	0.022	0.076	0.059	0.405	0.179	0.055	0.345	0.319	0.159

Published General Population EQ-5D Health State Valuation Papers

First author, year, and country	Sample size (actually receiving questionnaire), age, specified location (if not nationwide)	Selection of health states	Recruitment methods and withdrawal rates at different stages pre-analysis	Valuation method	Analysis method (including data exclusion) N.B. This is limited to the method for calculating quality of life outcomes	Preferred algorithm
Dolan (1997) **UK**	The sample size was 3395. Age greater than 18.	They roll-backed the states by saying that UA1 cannot go with M3 or SC3. 43 were roughly stratified by seriousness and each responder valued a sample of 13 (11 random (2 "very mild", 3 "mild", 3 "moderate" and 3 "severe") plus unconscious and 33333) No description of how they identified the wider set of 43 states.	6080 addresses drawn from postcode address file. After non-residential properties and refusals were accounted for, 3395 interviews were undertaken and 3337 produced data suitable for analysis.	1. Rating of own health state using EQ-5D and VAS. 2. (Presumably socio-economic data collection since these were reported) 3. Ranking of 15 states (13 plus 11111 and dead) 4. TTO ranking of 13 States worse than dead – the choice was between dying immediately and spending a length of time (10-x) years in the health state followed by x years in full health **(This is true in all papers unless stated)**	Data exclusion (after questionnaire) if 1. Insufficient data 2. All states rated worse than death 3. No understanding of task All states better than dead were converted into a quality of life figure by dividing the number of years by 10. States worse than death were transformed using the formula y = -(x/10)-1	Y = constant + β_1MO + β_2SC + β_3UA + β_4PD + β_5AD + β_6M2 + β_7S2 + β_8U2 + β_9P2 + β_{10}A2 + β_{11}N3 This is a main effects model without interactions, but with a term for whether each of the dimensions are at level 3, and one if any of them is at level 3. They tested for other model involving interactions but these did not improve the model significantly, and introduced inconsistencies.
Tsuchiya (2002) **Japan**	The sample size was 543. Age greater than 20. Random selection of areas within Saitama, Hiroshima and Hokkaido and then random selection of individuals from	They used a modified version of the MVH protocol containing 17 health states[4]. All respondents valued all health states. These 17 were described as "the minimum set of health states needed to estimate the value set".	Brief letters sent out to 972 people. Of thee, 617 agreed to take part in the survey. The survey was undertaken through face-to-face interviews and 78 respondents were rejected, leaving 543 people suitable for analysis.	1. Rating of own health state using EQ-5D and VAS. 2. VAS evaluations of 14 hypothetical health states expressed in EQ-5D 3. Socio-economic background questions 4. Ranking of 19 hypothetical health states expressed in EQ-5D 5. TTO of the 17 hypothetical health states The 14 states were independent	Data exclusion (after questionnaire) if: 1. Completely missing TTO data 2. Only 1 or 2 states valued 3. All states given the same value 4. All states worse than death For states better than dead, use the standard transformation. For worse than dead, use y=-(x/10)-1	The base case algorithm was Preference value = PV = 1 – β_0*const – β_1*mob2 – β_2*mob3 – β_3*sc2 – β_4*sc3 – β_5*ua2 – β_6*ua3 – β_7*pain2 –β_8*pain3 – β_9*mood2 – β_{10}*mood3 – β_{11}*N3 – (β_n*interaction terms) - e They investigated using alternative models, looking at the number of dimensions at

[4] These were 11112, 11113, 11121, 11131, 11133, 11211, 11312, 12111, 13311, 21111, 22222, 23232, 32211, 32223, 32313, 33323 and 33333

	the local electoral registry			(but occasionally overlapping) with the 19. The 19 are the 17 given in the footnote, plus 11111 and dead (these two are not part of the subsequent TTO) Six months was the lowest increment allowable. No description of ping-ponging.		level 3 (C3), the square of the number of dimensions at level 3 (C3sq), the existence of a dimension at level 1 (N1), the number of dimensions at level 1 (C1) and the square of the number of dimensions at level 1 (C1sq). With combinations of these 6 interaction terms, the paper looks at 41 models. R^2 did not significantly improve under these alternatives and none removed heteroskedasticity. The best performers were the plain model without interactions, N3, C3sq and N3 + C3sq.
Badia (1999 and 2001) **Spain**	The sample size was 975. Age unspecified 1 primary health care district in Barcelona covering 4 different socioeconomic areas. Respondents were quota sampled.	They used the 43 health states used in the UK General Population Survey (Dolan et al (1996)). Each respondent rated 13 states – not stated how these were selected from the 43. Since their aim was to compare Spanish and UK figures, likely to be the same process as Dolan (1996).	1 930 individuals were contacted by letter, then by follow-up telephone call. 1 000 agreed to participation. Health state valuations were obtained in face-to-face interviews. 975 were identified as suitable for inclusion in data analysis.	1. Rating of own health state using EQ-5D and VAS. 2. Ranking of 13 health states plus unconscious (but not plus dead) 3. They were then asked to put death in the ranking 4. 13 states randomly ordered and respondent valued each (making sure first was not 33333 or unconscious). The ping-ponging went 5 years, 4/6 years, 3/7 years, 2/8 years, 1/9 years)	For states better than dead, the quality of life value was simply x/10. For states worse than dead, 2 different transformations were used. Firstly, y =-x/(10-x) Alternatively, y = (x/10)-1 They chose the second one as more appropriate	Preference value = PV = $1 - \beta_0$*const – β_1*mob2 – β_2*mob3 – β_3*sc2 – β_4*sc3 – β_5*ua2 – β_6*ua3 – β_7*pain2 –β_8*pain3 – β_9*mood2 – β_{10}*mood3 – β_{11}*unconscious – β_{12}*N3 – (β_n*interaction terms) – e The unconscious term might be a misprint since it can't be included in an analysis as a dummy variable since it is incompatible with scores in any of the dimensions. Also, it is not mentioned in the regression comparison between Spain and the UK.
Lamers (2006 **The**	The sample size was 300. Age between 18 and	They used a modified version of the MVH protocol containing 17	Quota sampling was used. The potential participants were	1. Rating of own health state using EQ-5D and VAS. 2. (Presumably socio-economic	They looked at whether this smaller dataset was adequate for producing full quality of	y = $1 - \beta_0$*const – β_1*mob2 – β_2*mob3 – β_3*sc2 – β_4*sc3 – β_5*ua2 – β_6*ua3 – β_7*pain2 –

Netherlands	75	health states. All respondents valued all health states. These 17 were described by Tsuchiya (2002) as "the minimum set of health states needed to estimate the value set".	approached by telephone and invited to participate. Face-to-face interviews at the office of the research company were undertaken and the participants each received a 20 Euro gift voucher.	data collection since these were collected) 3. Ranking of 17 states (plus 11111 and dead) 4. TTO process However, they used a computer for the TTO component, with the states presented in a random order. They describe the ping-ponging process as 'outward titration'. It is unclear whether this is the same process described in Tsuchiya, Badia etc.	life figures for the entire range of EQ-5D states.	β_8*pain3 – β_9*mood2 – β_{10}*mood3 — β_{11}*N3 – (β_n*interaction terms) + e A second model was constructed without the N3 term.
Griener (2005) **Germany**	The sample size was 339. Age greater than 15. The selection was based on postal zip codes to provide representative urban/rural split. Locations all in county of Lower Saxony and Bremen)	The health states selected followed the British sample analysis (Dolan (1997))	No detail of how many people were initially approached. 380 returned reply card indicating willingness to participate. These were assigned to interviewers, who contacted participants to arrange mutual date for meeting. If no date could be agreed, or if there was no reply in three consecutive attempts to contact them, the respondent was excluded. 339 interviews were undertaken, of which 334 were completed. A small reward was offered (20 Marks)	1. Background information collection. 2. VAS ranking of all selected states 3. TTO assessment 4. Self-assessment The ping-ponging was limited to whole years, other than in the case the indidividual was indifferent between ten years in state x with less than 1 year in full health. If the point of indifference was between 7 and 8, the authors assumed 7.5. **Each interview lasted 43 minutes.**	Data exclusion (after questionnaire) if: They removed extreme values which eliminated 1% of the VAS scores and 1.5% of the TTO scores.	Initially, the authors used an OLS regression (additive combinations) This contained: 1. A variable for each EQ-5D dimension, showing a move from a less severe level to a more severe level. 2. A dummy variable for a shift from level 2 to level 3 3. An N3 dummy (for any level 3) They investigated an alternative specification, using a multiplicative combination of the parameters. However, they find it is no better than the simple, additive model.
Wittrup-Jensen (2002) **Denmark**	The sample size was 1 331 or 1 332. Age 18-91 (range rather than	14 states used per respondent (22222, 33333, 2 mild states, 8 other states, 2 further	4 075 random addresses were contacted, of which 2 653 were at home within three attempts. 1	The authors used face-to-face interview techniques. Complementary to this, computer-assisted interviewing techniques	Data exclusion if: 1. All states given the same value	Random effects model was used on 12 different algorithms.

	limit)	states related to diabetes or heart disease patients), plus death and 11111. The authors used a split-sample approach (4 groups) to directly value as many states as possible.	332 completed face-to-face interviews at the respondents' home.	were used (with the interviewer inputting responses from respondent). 1. Self-assessment using EQ-5D and VAS. 2. Rankings of the states 3. TTO	2. Less than 2 states valued 3. 11111 or death not valued 4. Death valued higher than or equal to 11111 For health states worse than death, the transformation was y = –x(10-x)	$Y_{it} = \text{constant} + \beta_2 x_{2,it} + \beta_3 x_{3,it} + \ldots + \beta_k x_{k,it} + u_i + \text{error}_{it}$ Of the 12, three that had universally consistent parameters.[5] TTO3 (see footnote) had all significant parameters ($p<0.001$). However, since they wanted comparability with the UK-tariffs, they went with TTO4 (where all parameters were significant other than U2) The RE model, using this algorithm, suffered from mis-specification and heteroskedasticity (both $p<0.0001$), but not from multicollinearity.
Jelsma (2003) **Zimbabwe**	The sample size was 2 488. Age greater than 15 with a minimum primary education and necessarily the oldest household member. Based in a high-density suburb of Harare (Glenview).	38 health states were chosen, based on Dolan (1996). Unconscious and dead not included, and two other states excluded due to an administrative error	All residential plots in suburb (Glenview) were identified, and a random sample of 2 500 were chosen. 2 384 face-to-face interviews were conducted in the houses without prior invitation. If no-one was present, the interviewer returned once at a later point.	1. Self-assessment using EQ-5D and VAS. 2. TTO on seven randomly selected states (from the 38) (Background information was also collected)	Data exclusion if all states valued the same, fewer than three states valued, or if there were more than 3 logical inconsistencies. Demographic data investigated for representativeness. Men underrepresented, as were older people. Sample split into thirds, two thirds used as an internal sample. Data on these were used to predict the results of	N3 model rejected as it lead to counter-intuitive results. They considered interaction effects (and found some significant effects). However, they argued that an algorithm with these interactions is unreliable.

[5] TTO3 = f(MO, SC, UA, PD, AD, M2, S2, U2, P2, A2) where x2 means dimension x is at level 3, TTO4 = f (MO, SC, UA, PD, AD, M2, S2, U2, P2, A2, N3) where N3 means any dimension at level 3, and TT10 = f(MO, SC, UA, PD, AD, M2, S2, U2, P2, A2, N2, N3) where N2 means any dimension at level 2.

					the final third.	
Shaw (2005) **USA**	The sample size was 4 048. Age was 18-99.3 (range rather than limit)	The authors used a split-sample approach, with four modelling samples of approximately 900 people and a validation sample of 400. Each group was assigned full health, 'the pits', unconscious and immediate death. Additionally, they were allocated 2 of 5 mild states (i.e. with one dimension at level 2) and 9 of the remaining 36 states. Therefore, all of the 43 states were valued by at least 900 people.	12 000 addresses initially selected. From this list, 5 237 people were selected for interview. 4048 interviews were collected A small reward was offered ($30)	1. Self-assessment using EQ-5D 2. Ranking of health states 3. Rating of these states on VAS 4. Rating own health on VAS 5. TTO on 13 states of the 15 states (i.e. not including 11111 and dead) 6. Demographic questions 7. Self-completion of the 15-item HUI-2/3 and rating own health on a 5-point scale ranging from excellent to poor.	Data was excluded if: 1. The respondent failed to value more than 1 health state. 2. There was incomplete demographic information. 3. If health states valued the same 4. If all health states were valued as being worse than death. 5. Respondent valued a health state using both sides of the board 6. Respondent valued one or more incorrect health states based on their assigned set of health state cards 7. Respondent valued one or more health states more than once. Included population rescaled to provide representativeness of the sample to the whole population in terms of race and sex.	All algorithms tested included main effects. Interactions were considered but these generated models suffered from multicollinearity and logically inconsistent parameter estimates. Rather than use a constant, they employed a D1 variable, representing the number of movements away from perfect health beyond the first. For example, D1(11112)=0 D1(11122)=1 D1(11123)=1 Interactions between dimensions were considered in a similar way. The variable I3 represented the number of dimensions at level 3 beyond the first. Therefore, I3(11113)=0 I3(11123)=0 I3(11133)=1 The square of this, as well as the similar I2 and I2-squared were also included. I2 was subsequently excluded as it caused multicollinearity.