The University Of Sheffield.

UNIVERSITY of York

EEPRU Policy Research Unit in Economic Evaluation of Health & Care Interventions

Quality review of a proposed EQ-5D-5L value set for England

October 2018

Authors: Mónica Hernández-Alava[1], Stephen Pudney[1] & Allan Wailoo[1]

[1]ScHARR, University of Sheffield

ACKNOWLEDGMENTS

This research was funded by the UK Department of Health Policy Research Programme through its Policy Research Unit in Economic Evaluation of Health & Care Interventions (EEPRU). EEPRU is a collaboration between researchers from two institutions (Centre for Health Economics, University of York and School of Health and Related Studies, University of Sheffield). Pudney's work on this review was part-funded by the Economic and Social Research Council through the Research Centre for Micro-Social Change (grant RES-518-28-5001). The views expressed in this report are those of the authors and not necessarily those of the ESRC or Department of Health and Social Care. Any errors are the responsibility of the authors.

We gratefully acknowledge the co-operation of the EQ-5D-5L research team (University of Sheffield, Office of Health Economics and University of Technology Sydney), particularly Nancy Devlin and Koonal Shah, who made available their research materials and provided comments on the factual accuracy of the contents. The data are owned by the Office of Health Economics. We thank Steve Brown and Rebecca Palmer for permission to use statistical information from the HubBLe and Big CACTUS trials and Arjun Bhadhuri and Abualbishr Alshreef, for providing information from those studies. Dr Mark Strong and Professor John Brazier gave helpful comments on the report. Five anonymous referees provided comments.

ABOUT THE AUTHORS

Mónica Hernández-Alava is Reader in Health Econometrics in the Health Economics and Decision Science section of the School of Health and Related Research at the University of Sheffield and Adjunct Associate Professor, Aalborg University, Denmark. She is a specialist in the micro-econometric analysis of health outcome measures and was Principal Investigator of the MRC Methodology Project "*Modelling Generic Preference based Outcome Measures - Development and Comparison of Methods*". She is also Visiting Research Associate and holder of a Biomarker Data Research Fellowship at the Institute for Social and Economic Research, University of Essex.

Steve Pudney is Professor of Health Econometrics in the Health Economics and Decision Science section of the School of Health and Related Research at the University of Sheffield; Adjunct Professor at the Centre for Health Economics, Monash University; and co-Director of the ESRC Research Centre on Micro-Social Change. He was a member of the 2014 Research Excellence Framework sub-Panel for Economics and Econometrics and is a Fellow of the Royal Statistical Society and past member of the Editorial Board of its Journal.

Allan Wailoo is Professor of Health Economics in the Health Economics and Decision Science section of the School of Health and Related Research. He is director of the NICE Decision Support Unit and co-director of the Policy Research Unit in Economic Evaluation of Health and Care Interventions at the University of Sheffield. He is a specialist in the methodology and practice of cost-effectiveness evaluation.

Executive Summary

Background and objectives

- EQ-5D is an instrument for measuring and valuing health related quality of life. It comprises a classification system that allows respondents to describe their health, and a set of associated "utility" values for those health states based on preferences of the general public which allow the calculation of Quality Adjusted Life Years (QALYs). The descriptive system covers five dimensions: mobility, ability to self-care, ability to undertake usual activities, pain and discomfort, and anxiety and depression.

- Until recently, EQ-5D allowed respondents to respond at one of 3 levels (3L) for each dimension (no problems, some problems, extreme problems). The 3L is the most widely used preference based measure used in economic evaluation in England, due in part to the fact that NICE recommends use of 3L health utilities in economic evaluations submitted to its Technology Appraisal Programme (NICE, 2013).

- A new version of the instrument, EQ-5D-5L, aims to improve measurement by using five levels of severity for each dimension (no problems, slight problems, moderate problems, severe problems, and extreme problems). Utilities have been published for the 5L for England (Devlin et al. 2018). It has been shown that economic evaluations undertaken using 5L rather than 3L are likely to generate very different results (Wailoo et al., 2017, Hernandez et al. 2017), so 3L and 5L cannot be used interchangeably if consistent decision making is required.

- Before a decision can be made about 3L versus 5L for economic evaluations that inform UK health policy, there is a requirement to critically assess the data and statistical methods that were used to generate the 5L value set. This reports contains this quality assurance assessment.

The Valuation Methodology

- The EuroQol Group has developed and promulgated a valuation protocol known as the EuroQol Valuation Technology (EQ-VT) for EQ-5D-5L. The EQ-VT has two main elements: a set of ten lead-time time trade-off (TTO) experiments and a set of seven discrete choice (DC) experiments, which require members of the public to compare and evaluate hypothetical health states. EQ-VT specifies the numbers of experimental subjects, the experimental tasks and health states to be valued and provides a digital environment for computer-assisted personal interviewing. The TTO and DC experiments involve only a small proportion of the 3,125

logically possible 5L health states; statistical modelling is used to extrapolate the observed data to the remaining health states.

- The English valuation assessed in this report was based on EQ-VT version 1.0. Serious concerns with data quality were subsequently identified from the five countries that used version 1.0. EQ-VT 1.0 has now been superseded and is no longer in use by EuroQol, so there is good reason to investigate the reliability of the data underlying the published value set.

Data Quality: TTO data

- The TTO experimental design examines 86 specified health states individually – (2.8% of the set of possible health states distinguished by 5L). Each experimental task requires the individual to decide on a trade-off between length of life and state of health, so that 10 years in a given impaired health state can be expressed as equivalent to a shorter life in perfect health.

- There is strong evidence, both direct (self-reported) and indirect (poor data quality) that many participants in the experiments either found this difficult or did not engage effectively with the experimental tasks. Even casual inspection of the individual-level TTO data revealed immediately that much of the data treated as accurate in the published valuation analysis is logically inconsistent or otherwise potentially misleading.

- We defined several different criteria to indicate problem responses (Table 2.5 in the report). Very large proportions of the data were found to suffer these problematic responses. For example, within the set of experimental participants:

    - 92% displayed at least one inconsistency, providing a higher or equal valuation to a health state defined unambiguously worse;
    - 30% were unable to distinguish more than four distinct value levels among the ten states they were asked to evaluate;
    - 29% reported at least one impaired health state as strictly worse than the worst state (55555);
    - over 50% displayed combinations of features in their responses sufficient to cast grave doubts on the validity of their entire TTO evidence.

Data Quality: DC data

- The DC experiment involves 392 distinct specified states (12.5% of the 3125 logically possible EQ-5D-5L states), via 196 pairwise comparisons (0.01% of the logically possible 2-state comparisons). The DC experiments do not explore variations in length of life alongside health states and they only give rankings of states rather than quantitative differences. DC data are therefore less informative than TTO data, and play a more limited role in driving the results of the overall valuation.

- There is little scope to examine validity of the DC data because the experimental design ruled out any combinations of choices capable of displaying logical inconsistencies. The only clear test that we can make on the DC data is a test of the assumption of statistical independence within the sequence of seven tasks undertaken by each participant. Unlike the TTO experiments, there is no evidence of any statistical dependence between the outcomes, nor of any systematic association with the position of the task within the sequence.

Specification and Estimation of the Valuation Model

- There are several potentially serious flaws in the specification of the statistical model used to extrapolate data to health states not covered by the valuation experiments. These are set out in Table 3.1 of the report. In our view, the most worrying are:

    - Apart from a proportionately small number of sample adjustments, the model assumes that all TTO responses are accurate within the resolution of the measurement software. This conflicts with the strong evidence of poor quality TTO data.

    - The specification of the model entails the assumption that valuations exceeding 1.0 (perfect health) are possible but not observed because of a censoring process. In fact valuations for impaired health states above 1.0 are ruled out theoretically, and the upper bound should be modelled as an inherent limit, not as censored observation. This misspecification means that health states with only mild impairments will be systematically over-valued.

    - Although a joint TTO-DC model is estimated, there are conflicts between the utility assumptions made in constructing the TTO and DC components of the joint model.

- Estimation is based on a Bayesian approach implemented using WinBUGS software. The Bayesian approach combines sample TTO and DC data with prior information to form a posterior distribution for the parameters which determine valuations. The estimation software uses an iterative simulation approach (called MCMC) which, after an appropriately long sequence of iterations will simulate random draws from the posterior distribution. We find several major problems with the application:
    - The model is unidentified, so statistical inference based on data alone is impossible.
    - Priors on key parameters are informative, but there is no justification for the priors used, nor is any sensitivity analysis presented.
    - Some elements of the model are specified in a manner that is likely to cause problems for the MCMC algorithm used in WinBUGS. In particular, the method used to impose increasing values for improving health states and the latent class specification of unobserved behavioural heterogeneity cause the MCMC algorithm to converge slowly or cycle rather than converge.
    - We used standard software (CODA) to check on convergence of the estimation algorithm. There is a clear failure to achieve convergence, and the published estimation results are based on an inadequate number of iterations to ensure convergence to the stationary distribution. Moreover, possibly as a result of inappropriate model specification or parametrisation, it may be difficult in practice ever to achieve convergence.

Conclusions and Recommendations

Our examination revealed serious deficiencies in the TTO data. The same may be true of the DC experiments but the experimental design precludes any detailed assessment of data quality.
There are numerous, serious concerns with the specification and estimation of the statistical model.

On the basis of these findings, and the lessons that can be learnt from the substantial programme of work on EQ-5D-5L, we make the following recommendations:

R1 A 5L value set for use in policy applications must be based on good quality data. A new programme of further development, including a new data collection initiative, should be considered to put EQ-5D-5L on a sufficiently firm evidential basis.

R2 A value set that is to be used in decision making must be based on a statistical modelling process that is robust and fit for purpose. If new data is collected, the statistical analysis should not simply replicate the analysis that has been reviewed here.

R3 This review has demonstrated the value of in-depth assessment for research findings which are critical to policy. The academic peer review system cannot provide the depth of review that is required to comply with the recommendations of MacPherson, since journal referees do not have access to underlying data or computer codes, nor do they have the time or professional incentive to review in full detail.

R4 It is good practice in scientific research for all relevant evidence to be made freely available to facilitate replication and secondary studies. This is even more important when the evidence underpins critical policy decisions. Full datasets, coding scripts and statistical analyses should be open to scrutiny to the fullest extent possible, within the bounds set by the need to protect personal information and respect research ethics.

Figures

Tables

SECTION 1: BACKGROUND

1.1 Introduction

The EQ-5D descriptive system covers five dimensions of health: mobility, ability to self-care, ability to undertake usual activities, pain and discomfort, and anxiety and depression. The original version of EQ-5D (the EQ-5D-3L, 3L from here) allows respondents to indicate the degree of impairment on each dimension according to three levels (no problems, some problems, extreme problems). 3L has an associated set of utility values ("value set" in EuroQoL Group nomenclature) based on estimates of the preferences of the general population in the UK (Dolan 1997), as well as for many other countries.

The 3L is the most widely used preference based measure used in economic evaluation in England. This is due, in part, to the fact that NICE recommends health utilities from the 3L be used in reference case economic evaluations submitted to its Technology Appraisal Programme (NICE, 2013).

A new version of the instrument, EQ-5D-5L, includes five levels of severity for each dimension (no problems, slight problems, moderate problems, severe problems, and extreme problems) with the intention of improving the instrument's sensitivity and reducing ceiling effects.

Utility values for this 5L version are now available for England (Devlin *et al.* 2018). The NICE Methods Guide expected to allow the 5L to be used once tariffs became available. Economic evaluations undertaken for other purposes have started using the 5L.

However, work undertaken for NICE via its Decision Support Unit (Wailoo *et al.*, 2017, Hernandez *et al.* 2017) demonstrated that economic evaluations undertaken using 5L rather than 3L are likely to generate very different results. These differences stem both from the responses individuals give to the descriptive system and the valuation systems. 5L utilities are shifted further up the distribution towards full health and compressed into a smaller space, than the 3L. These differences are large, having profound effects on the estimates of cost-effectiveness. New technologies that improve quality of life alone appear less cost effective if health gain is valued using 5L rather than 3L. Technologies that improve length of life only can appear more cost-effective (Pennington *et al.* 2018).

3L and 5L cannot therefore be used interchangeably if consistent decision making is required. NICE, and other decision makers, will need to recommend one version of the EQ-5D as their preferred option. Public sector decision makers must take steps to quality assure models that

underpin those decisions, appropriate to the level of risk and consistent with the Macpherson report (Macpherson 2013).

Given the widespread use of EQ-5D in economic evaluation in UK health care, the number and magnitude of decisions that could be influenced by the 5L valuation work is substantial. There is therefore a high level of risk associated with the move to 5L. Accordingly, NICE and the Department of Health and Social Care commissioned the Policy Research Unit in Economic Evaluation in Health and Care Interventions (EEPRU) to undertake quality assurance.

1.2 An outline of the valuation methodology

EuroQol has developed and promulgated a valuation protocol known as the EuroQol Valuation Technology (EQ-VT) for EQ-5D-5L. The EQ-VT comprises prescribed numbers of experimental subjects, experimental tasks and health states to be compared. It also provides a digital environment for computer-assisted personal interviewing. The EQ-VT has two main elements: a set of ten lead-time time trade-off (TTO) experiments and a set of seven discrete choice (DC) experiments. The same group of individual participants is required to undertake both the TTO and DC tasks. The basic details of EQ-VT and the way it was developed are set out by Oppe *et al.* (2014).

Each TTO task evaluates a specified EQ-5D-defined impaired health state against a state of full health, in two stages. First, the participant attempts to choose a trade-off point at which full health with reduced survival is judged to be as good as the impaired health state lasting for 10 years. Figure 1.1(a) shows the EQ-VT screen where this is done, using an iterative procedure starting from the mid-point of the 10-year window.[1] If the participant feels that the specified state is worse than death, then no trade-off can be made within the 10-year window, and the procedure switches to stage 2. This implements an extended lead-time TTO with a total period of 20 years (See Figure 1.1(b)). Equivalence points are determined approximately, in steps of 0.5 years.

The set of health states evaluated by TTO is allocated using a mixture of hand selected states and a simulation procedure described by Oppe and van Hout (2017) which, in this application gives a sparse but broad coverage of the 3,125 possible health states defined by the EQ-5D-5L health description. The potential weakness of this type of generic experimental design is that the choice of health states to be assessed experimentally through TTO and DC tasks is not necessarily

[1] Figure 1.1 and Figure 1.2 are reproduced with thanks from the interviewer briefing provided to us by OHE.

aligned with the configuration of health states found in real-life cost-effectiveness studies. The large and growing published literature on the EQ-5D-5L health description and value set says very little about the implications for actual cost-effectiveness work. We make a start on investigation of this issue in section 2.2.4, but much remains to be done.

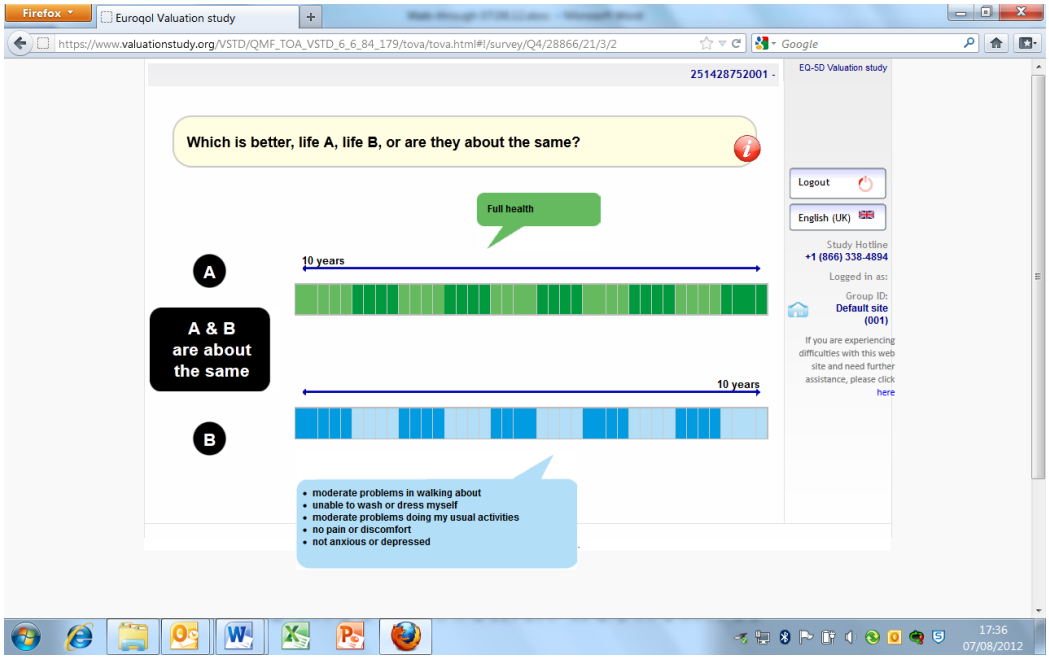
The lead-time phase was introduced to deal with problems encountered in the TTO experiments used to value the older 3L version of EQ-5D (Dolan 1997), which led to large numbers of TTO tasks which returned negative valuations for which an arbitrary rescaling was employed. The choice of a 10-year lead-time corresponds to the reasonable view that $v = -1$ is a realistic *a priori* lower bound for any health state valuation.

The equivalence point T reached by the participant (measured from the start of the extended 20-year window) ranges from $T = 0$ if a decade spent in the impaired state is perceived to be as bad as the loss of two decades of full health, to $T = 20$ if the state is perceived to be equivalent to full health. The equivalent value of the health state is then defined as:

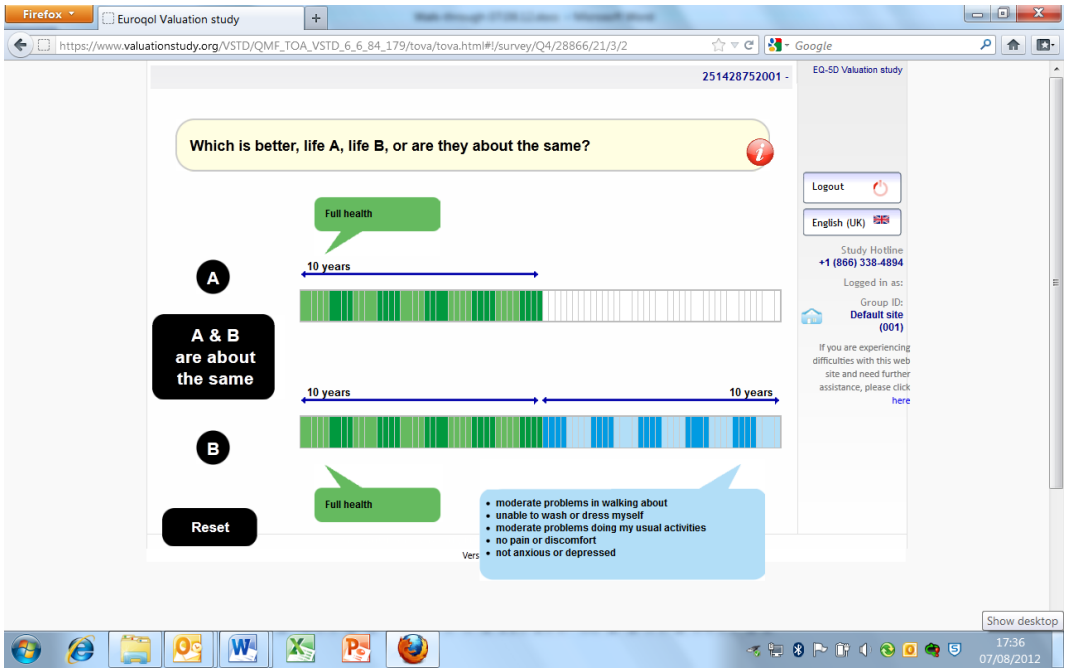
$$v = \frac{T - 10}{10}, \tag{1}$$

which ranges from -1 for the maximum sacrifice that can be measured, to +1 for equivalence to full health.

We draw attention to three potentially problematic outcomes for T. At $T = 0$, no trade-off takes place, implying a valuation v below -1; at $T = 10$, the TTO outcome is exactly at the seam between the primary TTO time frame and the secondary lead-time; at $T = 20$, the participant is unable to distinguish the specified state from full health. If the experiments work well and participants are able to make judgements matching the finer 5L classification, we would expect there to be few outcomes at these levels.



(a) The initial screen (states better than dead)



(b) Secondary screen (states worse than dead)

Figure 1.1: EQ-VT screens showing the two components of the composite TTO task

For the DC component, seven pairwise comparisons were allocated to each participant. For each DC task, the participant is presented with two EQ-5D health states (on a screen shown in Figure 1.2) and asked to rank them.

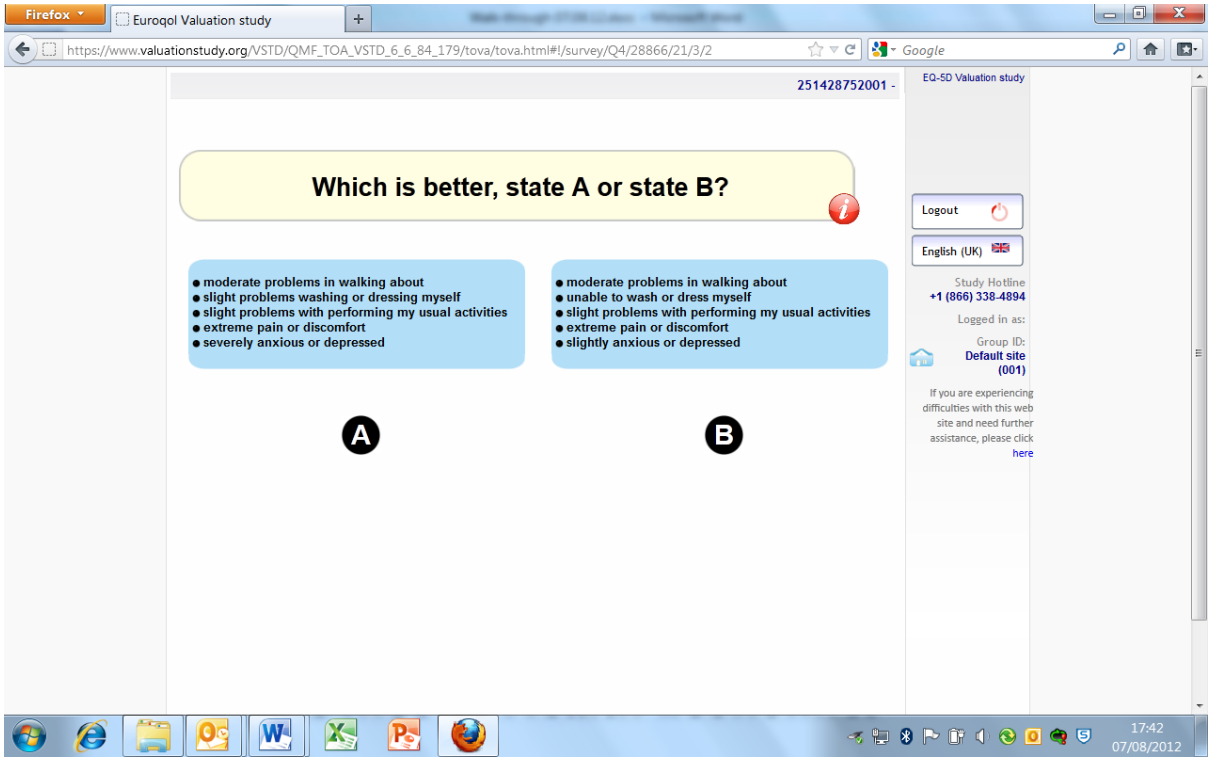


Figure 1.2: EQ-VT screen showing a discrete choice task

DC experiments are much less informative than TTO experiments for two reasons: they give no indication of the trade-off between health state and length of life; and they do not give any quantitative information on the margin by which one state is preferred to another. As a consequence, we would normally expect TTO data to be the dominant source of information in this type of valuation study.

The results of the TTO and DC experiments provide the input into a model-based statistical analysis that aims to do two things:

- It "averages out" random differences between individuals' valuations of the same experimentally-specified health states, using a conditional expectation predictor. If successful, this means that the resulting valuations represent the population as a whole rather than the particular randomly-selected individuals involved in the experiments.
- It gives a basis for extrapolating from the small set of health states covered by the experiments to the much larger set of health states that might be observed in real-life studies.

These two aspects of the valuation analysis are quite distinct and they require judgements about sample size and experimental coverage. These issues are examined in more detail in Section 2.

1.3 The review process

The objective of this review is to give an assessment of the quality of the EQ-5D-5L valuation tariff published by Devlin *et al.* (2018) that was funded by the EuroQoL Research Foundation and the UK Department of Health. We were supplied with a range of datasets and supporting documentation by the authors. These materials are listed in Appendix A1.

The assessment is in two parts:

- Section 2 of the report gives an assessment of the quality of experimental data on which the valuation model is based.
- Section 3 assesses the specification and implementation of the statistical model used to construct the valuations

Section 4 summarises the findings and gives recommendations of two kinds:

- Recommendations for NICE and other decision makers in the English health care system on the proposal to adopt the 5L value set for England
- General recommendations for future research in this area

SECTION 2: DATA QUALITY

In this section, we investigate the reliability of the experimental TTO and DC data on which the Devlin *et al.* (2018) valuation study is based.

2.1 The research protocol

The experiments were conducted in accordance with the standardised EQ-VT protocol which has been used internationally by a number of EuroQol affiliated groups. There are both advantages and disadvantages of following a standardised research protocol. On one hand, it promotes international comparability and can – if it is well designed – promote the spread of good practice. On the other hand, it may fail to accommodate adequately the differences (in target populations, technologies and geographical areas) between applications, and it may discourage critical thinking about individual applications.

A further difficulty is that the protocol has changed over time. The TTO and DC experiments reviewed here were conducted under EQ-VT version 1.0, which has been superseded successively by two revised versions. Version 1.0 is reported to have given rise to problems: "*In the first wave of valuation studies applying the first version of the protocol (EQ-VT Version 1.0), major data issues were observed leading to EQ-VT Version 1.1, a comprehensive research program and finally to the improved EQ-VT Version 2.0.*" (Ludwig *et al.* 2018). An unpublished research paper presented at the 2014 Scientific Meeting of the EuroQol Group and accessible through the EuroQol website (Shah *et al.* 2014) documents some of these data problems. Neither of the key published sources (Devlin *et al.* 2018, Feng *et al.* 2018) is specific about the use of a problematic early version of EQ-VT, nor do they reference Shah *et al.* (2014).

International experience underlines the problems with EQ-VT 1.0 and suggests that major improvements are achievable. Five countries used EQ-VT v1.0 (England, Netherlands, Spain, Canada and China). The associated published papers give little information on data quality, but the following relevant issues were reported.

Canada: only TTO data were used and it was decided to exclude 11.2% of the participants. The authors report excluding data from participants under either of two circumstances: "*(a) giving the same or a lower score for the very mild health state compared with 55555; and (b) giving the same or a lower score for the very mild health state compared with the majority of the health states that are dominated by the very mild health state within the same block. The definition of "majority" used here was 3 of 5/6 health states or 4 of 7/8 health states that are dominated by the very mild health state in the same block. As a result, a total of 136 of 1209 participants met the exclusion criteria.*" (Xie et al 2016)

Netherlands: 87 respondents (8.8%) valued state 55555 at least 0.5 higher than at least one other state. (Versteegh et al 2016)

Spain: the authors discussed general concerns about data quality and attempted to address these in a subsequent paper. They reported "*protocol violations*", that "*for most respondents in the valuation study (76.1%), interviewers did not show and explain the iterative procedure allowing for WTD responses.*" The authors refer to "*satisficing*" as a lack of engagement with the exercise: 61.3% of the cTTO responses exhibited this behaviour. (Ramos Goni et al in press)

It is our understanding that the major issues with the data quality from these first three studies were identified and acknowledged by the EuroQoL group which led to a one-year moratorium on 5L valuation studies and the development of EQ-VT v1.1. Version 1.0 is no longer in use. Version 1.1 introduced a quality control procedure based on four criteria, three of which relate to the process of the interview (e.g. the time to explain the preliminary wheelchair example) and one relates to extreme inconsistencies in valuations relative to 55555. Interviews failing these criteria are flagged as being of suspect quality, allowing feedback to interviewers, retraining and deletion of suspect data.

Korea: EQ-VT 1.1 was used and 34% of respondents provided at least one inconsistent TTO response (Kim et al 2016). Consistency means that if one health state is better than another in at least one dimension and not worse in any other, then the valuation of the former state should also be higher in order to be consistent. Kim et al used a definition of "weak" inconsistency in this study. Here, if the better health state is valued equal to the unambiguously worse health state, this is not considered an inconsistent response.

Uruguay: EQ-VT 1.1 was used and data from 10 interviewers (220 interviews) were excluded as a result of feedback from the Quality Control process. (Augustovski et al 2015)

Version 2.0 built on v1.1 by the addition of a feedback module to respondents as an internal check on their responses.

Germany EQ-VT 2.0 was used; 17.69% of respondents had at least one inconsistency (i.e., health state A defined as better than health state B but A having a lower cTTO value) prior to feedback, which reduced to 12.6% after feedback. (Ludwig et al 2018)

It seems that the changes to the VT have led to improvements in data quality. The methods used in the English valuation study are no longer in use because of this; data for countries such as Germany are substantially better quality.

2.2 Sample size and sample coverage

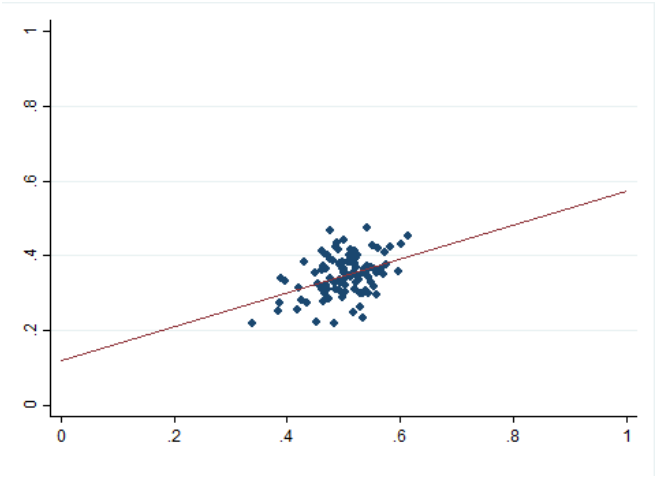
Before examining the data in detail, we first consider the important distinction between sample size and sample coverage for the purposes of valuation. It is important to distinguish sampling error from specification error. Sampling error arises from the random selection of the sample and can be reduced by increasing the sample size. However, if the statistical model is incorrectly specified, an increased sample size does not solve the problem – it merely gives more precise estimates of the wrong thing. Coverage by the experimental sample of the population of adults and the range of possible health states is key to our ability to produce robust valuation results.

Even in a small sample, model misspecification can be more easily detected – and its consequences ameliorated – if the sample design achieves good coverage of the range of values in the population.

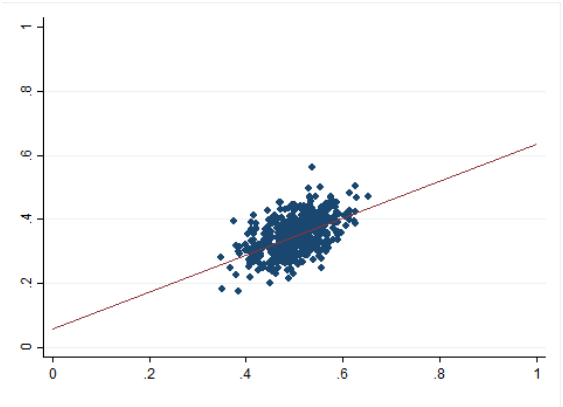
2.2.1 An illustrative example

Figure 2.1 illustrates this with an artificial example of straight line regression, generated through Monte Carlo simulation. Panel (a) shows the results of a regression of y on x with a sample of n = 100 observations, and the covariate x randomly generated with little dispersion. The slope coefficient is estimated as 0.45 with a standard error of 0.10. Panel (b) shows the result of generating a sample five times larger: the estimate is now 0.48 with a reduced standard error of 0.04. The improved precision comes purely from the ability of the regression to "average out" random variation more effectively. Panel (c) goes back to the original sample of n = 100, but rescales the x-observations to cover a wider range. The estimate is again 0.48 with a standard error of 0.04.

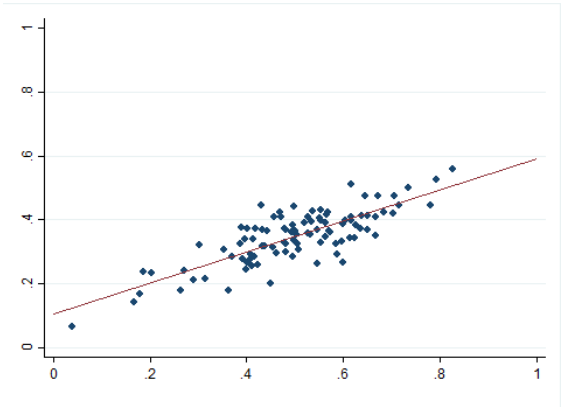
So, in this example, increased sample size and improved coverage are equally effective in reducing sampling error. However, the estimate in panel (c) is obviously far more robust than the estimate of panel (b), despite its much smaller sample size. The regression fit in panel (b) tells us nothing about individuals for whom x is below 0.3 or above 0.7, so if the true relationship departs from linearity in the extremes of the range, there is no possibility of detecting it or correcting the bias it causes. On the other hand, the smaller sample of panel (c) covers the wider range well and allows a much more robust analysis to be carried out.



(a) Small sample size, poor covariate coverage



(b) Large sample size, poor covariate coverage



(c) Small sample size, good covariate coverage

Figure 2.1: A linear regression example of sample size vs sample coverage: improved coverage of population by the covariate gives a more robust fit than increasing sample size five-fold

Although the TTO- and DC-based valuation models of Devlin *et al.* (2018) are more complex than linear regression, the general principle illustrated in Figure 2.1 remains important. For a statistical analysis to be robust, it is important to achieve good coverage of the range of relevant factors in the population; otherwise, extrapolation to cases not adequately represented in the sample is dangerous. In the case of the TTO and DC experimental data, the key issue is whether the specified experimental tasks cover a sufficiently wide range to represent adequately the range of health states likely to be encountered in practical cost-effectiveness applications.

2.2.2 Coverage of logically possible states

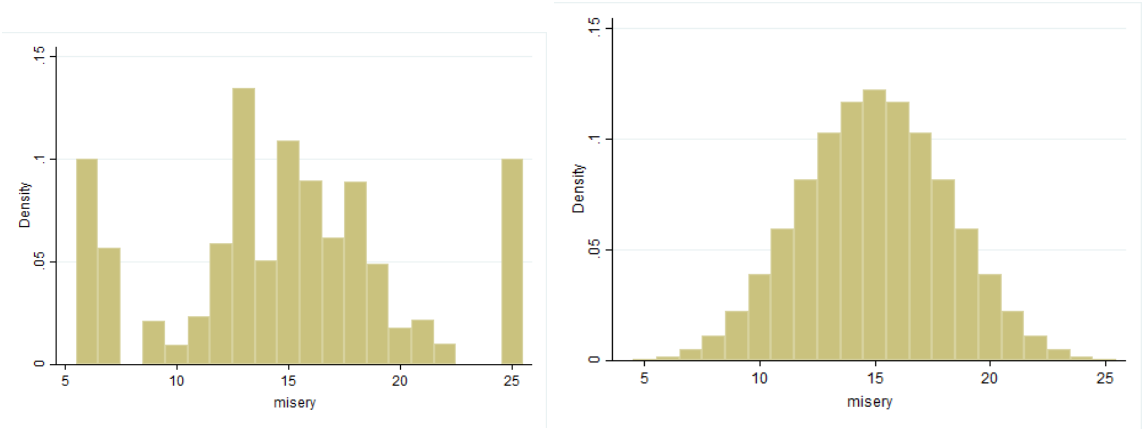
Table 2.1 summarises the degree to which the possible health states are covered by the design of TTO and DC tasks. There are 5^5 = 3,125 logically possible states defined by the EQ-5D-5L health description, and the TTO experimental design examines 86 specified health states

individually – under 3% of the full set of health states. The DC experiment involves 392 distinct specified states, 12.5% of the full set. However, DC experiments are not very informative – we only observe rankings within specified pairs of states. Thus it could be argued that the relevant indicator of coverage is the number of DC pairwise comparisons as a proportion of the 3,125×3,124/2 = 4.88 million possible comparisons.

Table 2.1: Coverage of health states

	Number of distinct states	% coverage of possible states
Logically possible health states in EQ-5D-5L descriptive system	3,125	-
Own health states reported by participants	180	5.8%
States covered by TTO experimental design	86	2.8%
Separate states covered by DC experimental design	392	12.5%
Logically possible 2-state comparisons	4,881,250	-
Pairwise choice situations covered by DC tasks	196	0.01%

An alternative summary of coverage can be made using a quantitative health index. The simplest construction is the "misery" index, defined as the sum of the five EQ-5D-5L items. Figure 2.2 compares the misery distributions of health states involved in the TTO experiments and the full set of 3,125 possible states. The TTO experimental design uses heavy coverage of the two extremes of the distribution (misery = 5 and 25) as a strategy to anchor the scale, and it also covers states very close to full health (misery = 6). The mean misery level across all 10,000 TTO tasks is 14.8, which is close to the average of 15.0 in an unweighted list of all EQ-5D-5L states. But there are obvious gaps in coverage close to full health (misery = 7) and at very poor health states (misery = 23 and 24). These gaps in coverage might be important for certain cost-effectiveness applications involving treatments for people in good and very poor health states.



(a) Distribution of experimental states (b) Distribution in full list of 3,125 states

Figure 2.2: Coverage of the TTO experiments compared in terms of the additive "misery" index with the full list of EQ-5D-5L states

2.2.3 Coverage of empirically relevant states

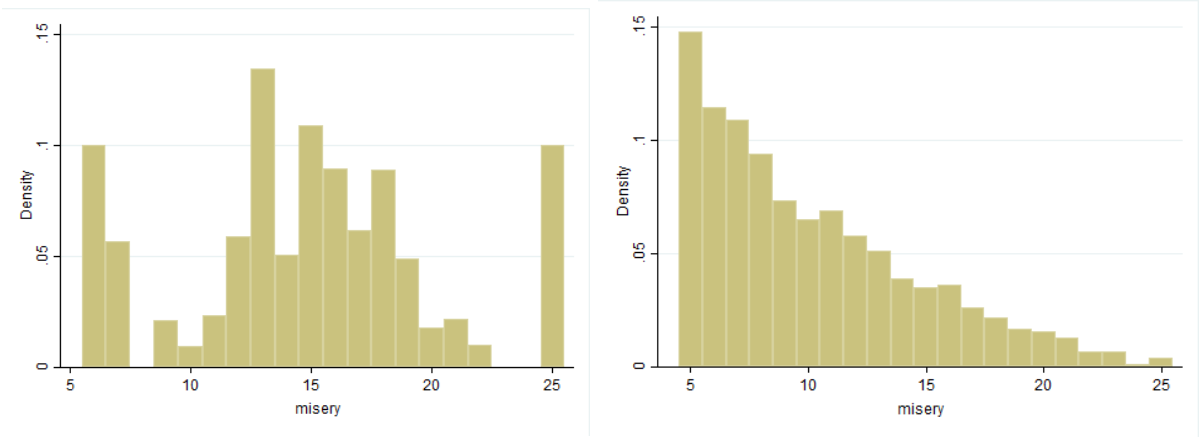
The coverage rates in Table 2.1 are very small but not necessarily a major problem.[2] Figure 2.2 compares the distributions of the misery index in two datasets: (*i*) the set of 10,000 TTO tasks, covering 86 distinct states; and (*ii*) the data from the EuroQol Group's (EQG) reference survey (van Hout *et al.* 2012) containing the responses from 3,637 individuals who report 676 (21.6% of the total) distinct current EQ-5D-5L states. The EQG sample contains a small group of healthy individuals but is otherwise representative of several important disease-specific groups.

The mean of the misery index in the EQG sample is only 10.1, despite that survey's clear focus on disease groups. Figure 2.2(a) shows that coverage of the TTO experimental design in the neighbourhood of that mean level is relatively thin. For example, only 16.9% of the TTO tasks involve health states with misery scores in the range 7-12, whereas almost half (46.8%) of the individuals in the EQG dataset report health states in that range. Note that conventional statistical power calculations (as made by the 5L research team and reported in Oppe and van Hout 2017) only take into account sampling error and not the robustness issue linked to coverage.

We have used the misery index here because it is independent of any value modelling. However, if we use the published Devlin *et al.* (2018) value set to construct an alternative health measure,

[2] For example, in statistical analysis of continuous variables, sample coverage rates are essentially zero, since a finite sample must necessarily miss almost every point on a continuum.

the picture of coverage does not change markedly – for example, 12.7% of the TTO states lie in the utility range 0.6-0.9, whereas 44.8% of EQG respondents report states in that range.



(a) Distribution of experimental states (b) Distribution in EQG reference sample

Figure 2.3: Coverage of the TTO experiments compared in terms of the additive "misery" index in relation to the EQG reference sample

2.2.4 Coverage of states important to cost-effectiveness studies

Although utility scoring systems have been developed specifically for the purposes of cost-effectiveness analysis (CEA), the design of the TTO experiments appears to be largely independent of evidence from real-life CEA studies. As a consequence, we have very little idea of the extent to which the coverage of health states involved in the TTO experiments matches the states which are most influential in determining the outcomes of CEA. Given the wide range of disease areas covered by CEA studies in practice, this is perhaps inevitable for generic (rather than disease-specific) valuation systems like those produced under the EQ-VT protocol.

To make a small start in this under-researched area, we have looked at two actual CEA studies, covering a range from mild to very serious: the HubBLe trial for surgical treatment of haemorrhoids (Alshreef *et al.* 2017); and the Big CACTUS study of aphasia treatment for people with stroke (https://www.sheffield.ac.uk/scharr/sections/dts/ctru/bigcactus). Both trials observed EQ-5D-5L health states directly at baseline and 12 months, and Big CACTUS additionally at 6 and 9 months.

The coverage rates are given in Table 2.2. If we take the list of distinct health states specified (as alternatives to full health) in the TTO experiments, they cover only 6.7% or 3.9% respectively of the distinct states reported by the full group of HubBLe or Big CACTUS participants. When we

take into account the frequency with which trial participants report health states by using frequency-weighted coverage rates, the results change in an interesting way. In the HubBLe trial, severity is low and many reports depart from full health 11111 only in a single digit. Such states are heavily represented in the TTO experimental design, so coverage of reports (rather than distinct reported states) increases to over 50% when we weight by frequency. A smaller increase of almost 9 percentage points occurs in the Big CACTUS trial, where health states are typically much poorer than in HubBLe.

For the purposes of cost-effectiveness research, there is a case for judging coverage in relation to the health states of trial subjects who are pivotal in determining the outcome of the cost-effectiveness study – those with the largest absolute magnitude of utility change. For each trial sample, we select the 20% of trial participants showing the largest absolute change over a year and repeat the calculation of weighted and unweighted coverage within that smaller group of trial participants. The coverage rates rise substantially for HubBLe, where coverage of over 70% for the most pivotal reported states seems very good. The more severe states reported by Big CACTUS participants are much less well covered, with only one in eight of the most influential reported states being directly involved in the TTO experiments.

This analysis raises the possibility that the coverage of TTO and DC experiments may be better suited to applications involving mild rather than severe states of ill-health. However, that conclusion is speculative at present because it is not based on large representative range of cost-effectiveness studies. Further research is needed to link the design of valuation methods to actual technology appraisals.

Table 2.2 Coverage of reported health states in two representative cost-effectiveness studies

		Cost-effectiveness study			
		HubBLe		Big CACTUS	
	Experiment type	All cases [1]	Influential cases [2]	All cases [1]	Influential cases [2]
Coverage of (unweighted) list of health states reported by trial subjects	TTO	6.7%	15.1%	3.9%	7.1%
	DC	20.2%	28.3%	15.7%	16.7%
Number of states in trial		*89*	*53*	*363*	*42*
Coverage of health states reported by trial subjects weighted by reporting frequency	TTO	50.3%	71.9%	12.8%	12.4%
	DC	56.4%	73.3%	28.8%	34.7%
Number of states in trial		*346*	*637*	*895*	*193*

[1] All health state reports by trial participants, excluding reports of full health. [2] All health states reported by the 20% of trial participants reporting the largest absolute utility change from baseline to 12 months.

2.3 Sampling of participants

The EQ-VT protocol recommends a sample size of 1,000 individuals, with 10 TTO tasks assigned to each individual. The sample size is small by the standards of major health and social surveys – for instance, the 2016 Health Survey for England (HSE) had over 8,000 respondents. Nevertheless, the sample size should be adequate for representation of major population groups. We are not aware of any formal analysis of sampling error and specification robustness that led to this recommendation. There are two published accounts of the process of data collection (Devlin *et al.* 2018 and Feng *et al.* 2018), two accompanying online appendices, and two pre-publication Discussion Paper versions (Appendix A1.3). However, those sources are not consistent in all details and there are some conflicts in reported sample numbers[3].

Fieldwork was carried out by a highly-regarded survey agency, Ipsos MORI, using a well-established approach. The initial recruitment phase uses two-stage random sampling from the Postcode Address File based on 66 primary sampling units, with random selection of one adult (aged 18+) from each household. This led to slightly over 2,000 potential subjects.

Participants in the study were required to: (*i*) be successfully located by the interviewer; (*ii*) provide personal data on age, gender, health state, education, employment, etc; (*iii*) undertake ten TTO tasks; and (*iv*) undertake seven DC tasks. Non-response could arise through non-contact, outright refusal to participate, refusal or inability to provide all required personal information, or through unwillingness to complete all TTO and DC tasks.

The non-response rate of over 50% is high by the standards of social surveys (*c.f.* 41% for the 2016 HSE). But bias caused by non-response depends on the pattern of non-response rather than its level. In this case, we know little about the way that non-response is related to personal characteristics of the non-respondents. A decision was made by the designers of the experiment to discard all information about individuals who refused participation or gave partial responses, so we have no direct evidence on the personal characteristics related to a high risk of non-response. This contrasts with practice in many health and social surveys, which is to retain partial responses and record as much information as possible on non-contact and refusal cases, to make

[3] The pool of potential subjects is given variously as 2,020 and 2,220 by Feng *et al.* (2018), page 24 and Devlin *et al.* (2018), page 12, respectively. A response rate of 47.7% is reported by Devlin *et al.* (2018), page 12. Depending on the definition used, the final sample size could be 1,000, 999, 996 or 912. The cited 47.7% response rate would imply an initial pool of size 2096, 2094, 2088 or 1912, none of which is consistent with the figures quoted by Feng *et al.* (2018) and Devlin *et al.* (2018), even allowing for rounding error.

possible statistical modelling of the response processes and adjustment for non-response bias. Although there are ethical issues involved in handling non-response, they are not an insuperable barrier, as evidenced by the long record of success for health and social surveys which do collect extensive response data.

In the absence of detailed information on the response process, we can compare the sample structure with available information on the structure of the England and Wales population. We have used the supplied data to reproduce Table 1 from Devlin *et al.* (2018). For a set of 12 participants, most of the personal characteristics are coded as "N/A". It is unclear what this means or why the information is missing. The published Table does not make clear the numerical base on which each sample percentage is calculated. Our Table 2.3 clarifies this and, like the published version, summarises two slightly different samples: one covering the 996 individuals who provided personal data and completed all TTO and DC tasks; the other covering the subset who were also not excluded from the Devlin *et al.* (2018) TTO component of the valuation model on grounds of inconsistent TTO responses.

The statistical modelling reported by Devlin *et al.* (2018) uses sample weights to improve the alignment of the sample and population with respect to age, where age is categorised in four broad groups.[4] The final column of Table 2.3 shows the reweighted sample composition for the set of participants who completed all TTO and DC experiments and provided age information.

The age weighting strategy is not very successful in aligning the sample with the population. The composition of the reweighted sample remains biased towards: women rather than men; retirees rather than employees and students; divorced and widowed rather than never-married; ethnic majority rather than minority groups; and those with health limitations rather the healthy.

[4] The dataset supplied to us contains another variable named "weight" which has a much more complex structure. This does not appear to have been used in the published work and the origin of that weight variable is unclear.

Table 2.3 Background characteristics of the sample

	All participants[1]	After exclusions[2]	General population	Age-weighted[1]
	N **(%)**	***N*** **(%)**	**%**	**%**
Partial sample	***n*** **= 996**	***n*** **= 912**		***n*** **= 996**
Age 18–29	113 (11.3)	105 (11.5)	20.7	20.1
Age 30–44	298 (29.9)	270 (29.6)	26.3	25.9
Age 45–59	250 (25.1)	227 (24.9)	24.7	24.5
Age 60–74	207 (20.8)	191 (20.9)	18.5	18.8
Age 75+	128 (12.9)	119 (13.0)	9.9	10.7
Male	405 (40.7)	372 (40.8)	49.2	40.5
Female	591 (59.3)	540 (59.2)	50.8	59.5
Partial sample	***n*** **= 984**	***n*** **= 900**		***n*** **= 984**
Employed/self-employed	504 (51.2)	466 (51.8)	59.4	52.5
Retired	***277*** (28.2)	256 (28.4)	13.1	24.5
Student	20 (0.2)	19 (2.1)	8.8	3.1
Looking after home/family	83 (8.3)	73 (8.1)	4.2	9.1
Long-term sick/disabled	48 (4.9)	42 (4.7)	3.9	4.7
Other activity[1]	52 (5.3)	47 (4.9)	10.6	6.1
Never married	***237*** (24.1)	213 (23.7)	34.6	28.9
Married	466 (47.4)	434 (48.2)	46.6	45.3
Civil partnership	2 (0.2)	2 (0.2)	0.2	0.2
Separated	37 (3.8)	32 (3.6)	2.7	3.5
Divorced	131 (13.3)	119 (13.2)	9.0	12.4
Widowed	107 (10.9)	99 (11.0)	6.9	9.4
Prefer not to say	4 (0.4)	1 (0.1)		0.4
Christian	636 (64.6	575 (63.9)	59.4	62.8
Other religion	60 (6.1)	53 (5.9)	8.7	6.2
No religion	***280*** (28.5)	266 (29.6)	24.7	30.3
Religion not stated	8 (0.8)	6 (0.7)	7.2	0.7
White	***899*** (91.4)	832 (92.4)	85.4	91.0
Other ethnic group	82 (8.3)	67 (7.4)	14.6	8.7
Prefer not to say	3 (0.3)	1 (0.1)		0.3
Health limited a lot	111 (11.3)	95 (10.6)	5.6	10.4
Health limited a little	***157*** (16.0)	144 (16.0)	7.1	15.4
Not health limited	716 (72.8)	661 (73.4)	87.3	73.2
Degree	211 (21.4)	201 (22.3)		21.7
Nodegree	774 (78.6)	699 (77.7)		78.3
English	919 (93.4)	847 (94.1)		92.6
Any other language	65 (6.6)	53 (5.9)		7.4
Responsible for children	350 (35.6)	314 (34.9)		35.7
Not responsible for children	634 (64.4)	586 (65.1)		64.3
Experienced illness: self	330 (33.1)	297(32.6)		31.5
Experienced illness: family	692 (69.5)	636 (69.7)		68.0
Care for others	416 (41.8)	385 (42.2)		40.3
Own EQ5D =11111	474 (47.6)	437 (47.9)		50.4
Other EQ5D health state	522 (52.4)	475 (52.1)		49.6
EQ-VAS < 80	335 (33.6)	298 (32.7)		32.3
EQ-VAS 80–89	255 (25.6)	241 (26.4)		25.6
EQ-VAS 90–99	337 (33.8)	306 (33.6)		34.9
EQ-VAS 100	69 (6.9)	67 (7.3)		7.3

2.4 Participants' experience of ill-health

A striking feature of the data – which is inevitable in a general-population sample – is the limited direct experience that respondents have of significant health problems. Almost half (47.4%) of the 999 respondents providing personal data reported their current state as full health (EQ-5D-5L description 11111). A third (33.2%) reported some past experience of their own unspecified "serious illness"; over two-thirds (69.4%) reported experience of other family members' illness; and two-fifths (41.8%) reported having cared for others in ill-health. 20% of the sample reported no previous experience of illness in themselves or others, and of that 20%, over two-thirds reported their own current state as full health (11111). Although there is a strong argument in favour of basing utility scores on views held in the general population, the argument is weakened if those views are not well-grounded, so experience of illness is a potentially important characteristic.

2.5 Participants' self-assessment of difficulties

The designers of the TTO and DC experiments wisely included questions asking subjects to self-assess the degree of difficulty they encountered in completing the TTO tasks. Three questions were asked, about the difficulty of: understanding the TTO questions; distinguishing between the hypothetical lives they were asked to compare; and deciding on the appropriate trade-off point. Each self-assessment was given on a 5-point Likert scale.

Figure 2.3 shows the distribution across the 1,000 TTO participants of these self-assessments. Just over half found it very easy to understand the questions, and just under half found it very easy to distinguish between the different hypothetical lives. In both cases, only around 10% of participants reported serious difficulty (points 4 and 5 on the scale). In contrast, over half the participants agreed with the proposition that they had difficulty in deciding on the equivalence point.[5] These responses give grounds for caution in using the TTO responses and they underline the need for careful assessment of the quality of the TTO data.

[5] Some of the difference between the first two measures and the third may be due to question design – the first two involved respondents agreeing that something was easy, whereas the other involved agreeing that something was difficult. The phenomenon of acquiescence bias may have contributed to a more positive response for the first two questions, and there may be a case for redesigning these questions.

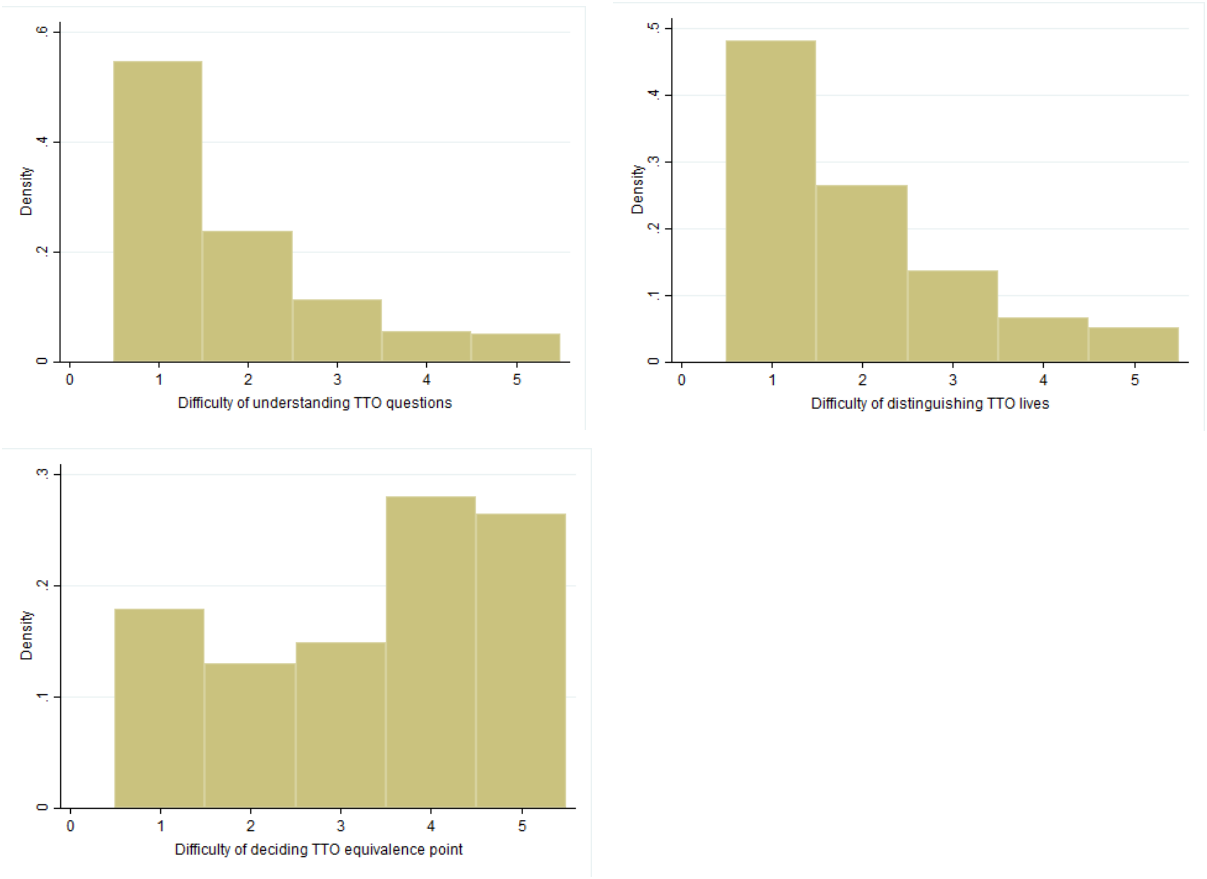


Figure 2.4: Distribution of indicators of perceived difficulty of TTO tasks

DC tasks are potentially simpler than the TTO, since they require only the ability to rank two states in terms of quality of life. Nevertheless, there were only slightly fewer participants reporting the two highest levels of difficulty (49% rather than 54%). Figure 2.4 shows the distributions of the three difficulty indicators.

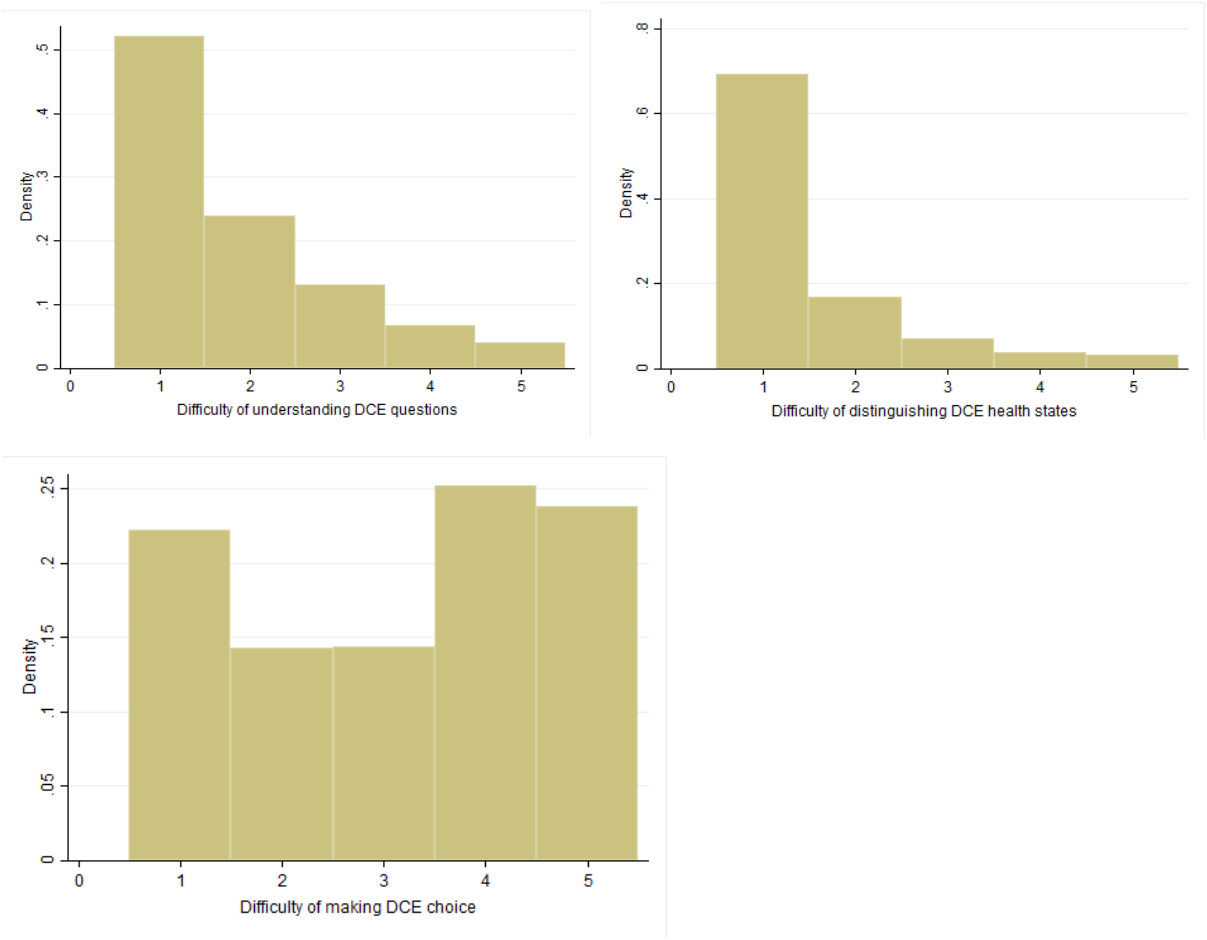


Figure 2.5: Distribution of indicators of perceived difficulty of DC tasks

We estimated logistic regression models of the probabilities of reporting either of the highest two levels of difficulty in making the TTO and DC choices. Table 2.4 shows the marginal effects, defined as the sample mean predicted change in probability as each personal characteristic changes in turn by one unit. The TTO and DC experiments clearly differ in terms of the characteristics of people who report difficulties. For TTO, the results are surprising. As might be expected, respondents whose main language was not English perceived more difficulty, by an average margin of 16 percentage points. But gender and ethnicity are also highly statistically significant, with men and members of ethnic minorities less likely (by 8 and 24 percentage points respectively) to report difficulty in deciding on the equalising point. Since there is no reason to expect any real cognitive advantages of being male or a member of a minority ethnic group, this may capture some difference in their willingness to reveal problems with a difficult task.

No significant influences were found for the DC experiments; the overall *P* value is 0.0049 for TTO and 0.3452 for DC.

Table 2.4 Marginal effects from logistic regression model of the association between personal characteristics and difficulty in making TTO and DC choices

Characteristic	TTO decision	DC decision
Age	0.002 (0.001)	0.001 (0.001)
Male	-0.079** (0.033)	-0.049 (0.033)
Never married	0.051 (0.043)	0.005 (0.044)
Religious	0.005 (0.037)	0.017 (0.037)
Ethnic minority	-0.238***(0.059)	0.035 (0.063)
Disability	0.039 (0.052)	-0.036 (0.052)
Degree	0.060 (0.039)	0.040 (0.040)
Children	0.073* (0.039)	0.040 (0.040)
Experience of caring	-0.011 (0.033)	0.059* (0.033)
English not main language	0.158** (0.072)	0.117 (0.082)
Empirical prevalence	0.543*** (0.016)	0.490*** (0.016)

Statistical significance: * = 10%; ** = 5%, *** = 1%. Standard errors in parentheses.

As we show statistically in sections 2.6 and 2.7, these indicators of difficulty in deciding on the TTO equivalence point and in making DC choices are strongly associated with the occurrence of problematic outcomes for the TTO and DC experiments.

2.6 The TTO experiments

TTO tasks present a considerable cognitive challenge to participants, since they are required to imagine the impact of being in unfamiliar health states, and deal with the difficulties involved in locating the equivalence point on the trade-off between ill-health and length of life. It would not be surprising to find significant numbers of participants unable to make reliable judgements.

2.6.1 Individual-level analysis of TTO quality

Ideally, participants would be able to discriminate clearly between marginally different health states and order them in a logically consistent way. If that is the case, we would observe substantial variation in the outcome *T* for each TTO task, without large numbers of outcomes piled up at the limits $T = 0$ and 20 or the seam between the two stage of the composite TTO at $T = 10$. We have defined a set of ten indicators that might be useful in identifying individuals generating poor-quality TTO data. They are listed in Table 2.5, together with the proportions of

individuals indicated by each, in the original TTO sample and the subsamples produced by removing individuals discarded or with outcomes modified by Devlin *et al.* (2018).

As discussed in section 2.1 above, some health states in the TTO valuation tasks presented to respondents have a logical ordering. If health state A is better than health state B in at least one dimension, and not worse on any other dimension, then the valuation for state A should also be higher than state B for the response to be considered consistent. This is the definition of inconsistency used in the results below and is used as an indicator of problematic responses. A weaker test classes valuations of A and B that are equal as consistent[6].

Depending on which potential anomalies are regarded as serious, a proportion ranging from 52% to 94% of the individual participants provided at least one outcome which could reasonably be regarded as problematic. The statistical literature on classical and non-classical measurement error would suggest that error rates as high as these are likely to lead to very large biases in inferences drawn from the sample data.

The sample deletions and other special treatment used by Devlin *et al.* (2018) to deal with problematic TTO outcomes make relatively little difference to this picture. Excluding problematic individuals from the data used for analysis only reduces the proportion of problematic respondents in the retained sample by around 4 percentage points. Among respondents who remain in the sample and who have no TTO outcomes given special treatment in the modelling, the proportion who give reason for concern remains high, at 44% to 91%, depending on the criteria used.

[6] "Weak" inconsistency fails to capture some of the problematic responses apparent in Table 2.6 below. See, for example, respondent 1.

Table 2.5 Proportions of individual participants displaying potentially problematic response behaviour

Anomalous outcome type	# individuals	% of sampled individuals		
		Original TTO data (n = 1,000)	After sample deletions[1] (n = 912)	After deletions and special treatment[2] (n = 604)
(1) All individual's TTO trials result in same value of T	23	2.3%	0%	0%
(2) Individual reports at least 1 non-55555 trial with same or lower value of T than trial of 55555	668	66.8%	63.8%	49.8%
(3) Individual reports at least 1 non-55555 trial with strictly lower value of T than trial of 55555	289	28.9%	26.1%	28.6%
(4) Individual reports fewer than 5 distinct values for T	309	30.9%	26.3%	20.9%
(5) Individual reports mild trial (1-point difference from 11111) with same or lower T result as trial of 55555	84	8.4%	0%	0%
(6) Individual reports values T =0, 10 or 20 in every trial	41	4.1%	2.7%	0%
(7) Individual reports all ten trial values T as multiple of 5 years	63	6.3%	4.2%	0.5%
(8) Individual gives only integer values for T	362	36.2%	35.1%	31.1%
(9) 'Seam' outcome of T = 10 in at least two trials with no outcome below 10	164	16.4%	16.4%	0%
(10a) Individual with any inconsistencies between the logical ordering of health states and the TTO valuation	922	92.2%	91.5%	88.4%
(10b) Individual with inconsistencies in more than 20% of tasks between the logical ordering of health states and TTO valuation	518	51.8%	47.4%	39.1%
Individual displays any of anomalies (1), (3), (4) or (5)	520	52.0%	47.6%	44.2%
Individual displays any of anomalies (1), (3), (4), (5), (7), (8) or (9)	711	71.1%	68.4%	60.9%
Individual displays any of anomalies (1), (3), (4), (5), (7), (8), (9) or (10b)	769	76.9%	74.8%	67.4%
Individual displays any of anomalies (1), (3), (4), (5), (7), (8), (9) or (10a)	940	94.0%	93.4%	91.1%

[1] Deletions comprise the 88 individuals excluded completely from Devlin *et al.*'s (2018) analysis on grounds of missing personal characteristics or grossly inconsistent TTO outcomes. [2] "Special treatment" refers to the 308 individuals for whom one or more TTO outcomes are overridden or treated as censored by Devlin *et al.* (2018)

To illustrate the serious concerns about much of the data that are used, unmodified, by Devlin *et al.* (2018), Table 2.6 reproduces the TTO outcomes for the first 30 individuals that remain after excluding individuals who were either excluded from Devlin *et al.*'s (2018) analysis, or whose data were subject to some special treatment (data edits or censoring). Thus all the TTO results in Table 2.6 were treated as fully accurate data. These cases have not been selected in any special way – they are the first 30 listed in order of the identifiers supplied with the datasets. The following list of problematic TTO outcomes illustrates the generally poor quality of the data:

- Individual 1 is undiscriminating, giving a value v= 0.95 for 7 of the 10 tasks.
- Individual 2 inconsistently rates state 33253 better than 23242
- Individual 3 inconsistently rates mild states 12112 and 11212 much worse than states 34244, 43514, 55424 and 44553
- Individual 4 inconsistently rates states 21444 and 53244 worse than 55555
- Individual 8 inconsistently rates mild state 11121 much worse than state 25222
- Individual 10 inconsistently rates four states worse than 55555
- Individual 11 inconsistently rates 44345 better than44125
- Individual 12 rates state 12514 negatively (v = -0.5), while 55555 is rated as v= +0.5
- Individual 13 rates state 14554 more highly than 12344
- Individual 14 rates state 55555 above state 44345
- Individual 15 rates state 35332 above 13122
- Individual 16 is undiscriminating, giving v = 0.05 for 8 of the 10 tasks
- Individual 17 inconsistently rates state 25122 (v= 0) below 45233 and 55233
- Individual 19 rates 6 states below the worst possible (55555)
- Individual 20 is undiscriminating for states with mild difficulties in the first three health domains, rating 12111, 11221 and 11235 as equivalent to full health; and inconsistently rates 34515 above 12514
- Individual 21 inconsistently rates 43514 and 34244 as worse than 55555
- Individual 22 inconsistently rates state 53243 as equivalent to full health
- Individual 23 rates state 55233 well above 45233, and is undiscriminating on poor health states, rating 12244 and 45233 as equivalent to 55555
- Individual 24 rates state 31525 as equivalent to full health and (inconsistently) superior to 21111; also 55233 above 45233, 25122 and 21111
- Individual 25 rates state 24445 as much worse than 55555

- Individual 26 inconsistently rates state 31514 above 11414
- Individual 28 rates 7 of the 10 states (including 55555) at 0.95, above state 42115 at 0.9

Although random response error could account for some of these cases, the large number of them and the egregious nature of some of the anomalies suggest that there might be serious difficulties for participants relating to their engagement with or understanding of the TTO tasks.

Table 2.6 TTO outcomes for the first 30 individuals (ordered by original survey id), excluding individuals excluded from model estimation and individuals with any TTO outcomes overridden or treated as censored in estimation

Individual	EQ5D	*v*	EQ5D	*v*	EQ5D	*v*	EQ5D	*v*	EQ5D	*v*	EQ5D	*v*	EQ5D	*v*	EQ5D	*v*	EQ5D	*v*	EQ5D	*v*
1	21111	1	12112	.95	11212	1	23152	.95	21345	.95	43514	.95	34244	.95	55424	.95	44553	.95	55555	.9
2	11112	.8	12334	.65	32314	0	23242	.45	21334	.55	24342	.45	53412	.4	33253	.6	55225	.05	55555	-.2
3	21111	.95	12112	.25	11212	.2	23152	.85	21345	.85	34244	.85	43514	.4	55424	.5	44553	.8	55555	-.65
4	11121	.9	11414	.3	25222	.8	25331	.4	31514	.4	21444	-.6	35143	.5	53243	-.2	53244	-.4	55555	-.3
5	11121	.95	11414	.7	25222	.8	31514	.5	25331	.7	21444	.5	35143	.5	53243	.7	53244	.5	55555	.2
6	21111	.95	11212	.8	12112	.9	23152	.5	21345	.35	34244	.8	43514	.8	55424	.8	44553	.25	55555	.25
7	12111	1	11122	.95	13224	.8	42321	.85	35311	.9	34232	.6	52335	.75	24445	.5	43555	.2	55555	.2
8	11121	-.2	11414	1	25222	1	25331	-.5	31514	-.5	21444	-.5	35143	-.7	53243	-.5	53244	-.7	55555	-.95
9	11211	.95	12121	.95	23514	.35	52215	.3	12543	.5	32443	.5	45133	.65	34155	.3	43542	.4	55555	-.35
10	11211	1	12121	1	12543	.3	52215	.6	23514	.65	32443	.3	45133	.4	34155	.5	43542	.4	55555	.5
11	11121	1	21112	1	12513	.8	53221	.3	12344	.3	44125	.2	54342	.3	14554	.3	44345	.3	55555	.2
12	12111	1	11221	1	11235	.6	12514	-.5	54231	.8	51451	.6	34515	.6	45144	.5	35245	.8	55555	.5
13	11121	1	21112	1	12513	.8	53221	1	12344	.6	44125	.7	54342	.8	14554	.8	44345	.6	55555	.5
14	11121	1	21112	.9	12513	.8	53221	.9	12344	1	44125	.7	54342	.3	14554	.3	44345	.05	55555	.1
15	11211	.85	13122	.5	42115	.3	11425	.2	51152	.5	22434	.3	35332	.8	45413	.4	24553	.6	55555	.2
16	12111	.1	11122	.05	13224	.05	42321	.05	35311	.05	34232	.05	52335	.5	24445	.05	43555	.05	55555	.05
17	21111	.2	11421	.6	13313	1	25122	0	12244	.6	31525	.3	45233	.2	55233	.5	52455	.5	55555	-.9
18	11112	.8	14113	.7	21315	.3	15151	.2	31524	.2	52431	.8	43315	.3	24443	.2	54153	.1	55555	.1
19	21111	.8	11212	.3	12112	1	23152	.4	21345	0	34244	.3	43514	.2	55424	.1	44553	.2	55555	.5
20	12111	1	11221	1	11235	1	12514	.5	54231	.7	51451	.5	34515	.8	45144	.65	35245	.5	55555	.3
21	21111	1	11212	.8	12112	.9	23152	.5	21345	.5	43514	.2	34244	.2	55424	.6	44553	.1	55555	.25
22	11121	.8	11414	.5	25222	.6	25331	.6	31514	.5	21444	.3	35143	.6	53243	1	53244	.3	55555	.3
23	21111	.95	11421	1	13313	.95	25122	.95	12244	.5	31525	.9	45233	.5	55233	.95	52455	.6	55555	.5
24	21111	.6	11421	1	13313	.7	25122	.6	12244	.8	31525	1	45233	.6	55233	.8	52455	.7	55555	.5
25	12111	.85	11122	.95	42321	.1	13224	.75	35311	.1	34232	.45	52335	.2	24445	-.5	43555	0	55555	0
26	11121	.95	11414	.3	25222	.4	25331	.35	31514	.5	21444	.4	35143	.1	53243	.3	53244	.3	55555	0
27	12111	.95	11221	.95	11235	.9	12514	.55	54231	.8	51451	.95	34515	.45	45144	.3	35245	.5	55555	0
28	11211	1	13122	1	11425	.95	42115	.9	51152	.95	22434	.95	35332	.95	45413	.95	24553	.95	55555	.95
29	11121	.95	21112	.8	12513	.7	53221	.8	12344	.2	44125	.4	54342	.2	14554	.2	44345	.2	55555	.05
30	11211	.95	13122	.95	11425	.95	42115	.8	51152	.8	22434	.9	35332	.95	45413	.95	24553	.8	55555	.5

What are the personal characteristics associated with problematic TTO outcomes? Table 2.7 summarises logistic regression models for the occurrence of each of the patterns of TTO outcomes (1)-(10) listed in Table 2.5. Explanatory covariates are included to capture possible influence of personal characteristics. The covariates were selected by comparing informally the mean characteristics of participants who gave problematic responses with those who did not, and selecting characteristics with substantial between-group mean differences for at least one of the indicators (1)-(10). We also include a binary variable indicating whether or not the individual agrees (at level 1 or 2 on a 5-point scale from 1 = agree to 5 = disagree) with the proposition that "*I found it difficult to decide on the exact points where life A and life B were about the same*".

There are relatively few statistically significant impacts of personal characteristics, except for the existence of a disability, which is associated with reductions of 5-8 percentage points in the prevalence of problems (4), (5) and (9); and minority ethnic identification, which is associated with increases of 9-17 percentage points in the prevalence of problems (2), (4), (7) and (9).

The most consistent effect, statistically significant in seven of the ten problem indicators, is for self-reported difficulty with TTO. This might be expected to act as an indicator of weak cognitive or empathetic ability to carry out TTO tasks, and therefore to be positively associated with problematic TTO outcomes. But, interestingly, the reverse is true – all significant impacts are negative.[7] The only interpretation that we can offer for this is that the covariate is acting instead as an indicator of the degree to which the participant takes the experiment seriously and struggles hard to give a worthwhile response; a participant who does not take the experiment seriously and simply gives an effort-minimising sequence of responses may then accurately report that (s)he did not find the task difficult.

[7] This finding is not an artefact of the multivariate modelling approach – the mean prevalence of all but one of the problem indicators (1)-(10) is lower for those reporting difficulty in deciding TTO equivalence points.

Table 2.7 Marginal effects from individual-level logistic regression models for the probability of generating one or more potentially problematic TTO outcomes

Personal characteristic	**Category of problematic TTO outcome** (defined in Table 2.5)									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Age	0.000	0.004***	0.002*	0.001	0.001	0.001	0.001	0.001	0.000	0.001
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Male	0.001	0.019	-0.023	0.008	-0.005	0.013	0.011	0.072**	0.007	-0.014
	(0.010)	(0.031)	(0.030)	(0.030)	(0.018)	(0.013)	(0.016)	(0.032)	(0.025)	(0.018)
Single	0.018	0.068*	0.116***	0.009	0.056*	-0.006	0.007	0.006	-0.021	0.019
	(0.018)	(0.038)	(0.042)	(0.040)	(0.030)	(0.017)	(0.023)	(0.042)	(0.031)	(0.021)
Religious	0.021**	-0.030	0.009	0.043	0.032*	0.025**	0.025	-0.018	-0.028	-0.033*
	(0.009)	(0.034)	(0.034)	(0.034)	(0.019)	(0.013)	(0.017)	(0.036)	(0.029)	(0.018)
Ethnic minority	0.012	0.100*	-0.000	0.168***	0.060	0.051	0.086**	0.046	0.170***	-0.001
	(0.020)	(0.055)	(0.058)	(0.063)	(0.042)	(0.033)	(0.042)	(0.063)	(0.061)	(0.034)
English not main language	0.056**	-0.008	0.005	0.072	0.068*	0.017	0.033	-0.105**	-0.043	0.025
	(0.027)	(0.049)	(0.047)	(0.049)	(0.036)	(0.022)	(0.028)	(0.047)	(0.034)	(0.024)
Disability	-0.003	-0.041	0.011	-0.080**	-0.050***	0.002	-0.001	-0.021	-0.061**	-0.018
	(0.012)	(0.038)	(0.037)	(0.035)	(0.018)	(0.016)	(0.019)	(0.038)	(0.026)	(0.022)
Degree	0.033*	0.061*	-0.019	0.058	0.060**	0.029	0.012	0.033	-0.027	0.017
	(0.020)	(0.036)	(0.036)	(0.038)	(0.027)	(0.021)	(0.021)	(0.039)	(0.029)	(0.020)
Has children	0.013	-0.077**	-0.083***	-0.020	-0.016	0.010	0.008	0.057*	0.025	-0.029
	(0.010)	(0.031)	(0.030)	(0.030)	(0.018)	(0.013)	(0.016)	(0.032)	(0.025)	(0.018)
Experience of caring	-0.002	0.071	0.066	0.030	0.032	0.020	0.035	0.094	-0.027	0.009
	(0.021)	(0.087)	(0.075)	(0.072)	(0.037)	(0.023)	(0.028)	(0.079)	(0.054)	(0.045)
TTO difficulty	-0.025**	-0.075**	-0.048	-0.105***	-0.053***	-0.031**	-0.040**	-0.018	0.023	-0.048***
	(0.010)	(0.030)	(0.029)	(0.030)	(0.018)	(0.013)	(0.016)	(0.031)	(0.024)	(0.017)
Joint *P*-value	0.0020	0.0009	0.0220	0.0000	0.0000	0.0013	0.0001	0.0686	0.0606	0.1004

Standard errors in parentheses. Statistical significance: * = 10%, ** = 5%, *** = 1%

Table 2.8 shows the proportions of TTO tasks producing problematic outcomes, rather than proportions of individuals concerned.

Table 2.8 Proportions of TTO tasks producing a potentially problematic outcome

Anomalous outcome type	% TTO tasks
(1a) Non-55555 trial with same or lower value of *v* than trial of 55555	26.3%
(1b) Non-55555 trial with strictly lower value of *v* than trial of 55555	7.4%
(2) A valuation $v = 0$ is given at the seam between the primary TTO window and the secondary lead-time window	13.1%
(3) A state with misery index < 15 is valued below -1	1.9%
(4) A state with misery index ≥ 15 is rated equivalent to full health	6.4%

Since the TTO tasks are in sequence for each individual, we can treat them as successive waves of testing within each individual and use panel data modelling methods to analyse the way that potentially problematic responses develop sequentially. This is important, since the modelling methods used by Devlin *et al.* (2018) make little allowance for statistical dependence between TTO results within individuals.[8] There is good reason to question this assumption, since it is likely that there is some learning by participants as the experiments proceed and there may also be fatigue effects as the burden of the tasks accumulates.

To examine the influence on TTO outcomes of the specified health state and the characteristics of the participant and to investigate dependency within the sequence of tasks, we use the following dynamic panel data probit model:

$$v_{it} = \alpha v_{it-1} + \beta_1 Mo_{it} + \beta_2 Sc_{it} + \beta_3 Ua_{it} + \beta_4 Pd_{it} + \beta_5 Ad_{it} + \beta_6 Mis_{it} + \beta_7 Mild_{it} + \gamma X_{it} + u_i + \varepsilon_{it} \qquad \text{for } t = 2 \ldots 10 \qquad (2)$$

where: v_{it} is the TTO outcome for individual *i* at the *t*-th task in value form and v_{it-1} is the outcome from the previous task; $Mo_{it} \ldots Ad_{it}$ are binary indicators taking the value 1 if the

[8] Devlin *et al.* (2018) actually assume the TTO valuations are independent except for a scale factor which has a 3-point discrete probability distribution (see Appendix 2 for details). That specification would not allow for learning or fatigue effects.

relevant health domain (*Mo* = mobility, *Sc* = self-care, *etc*) is specified at level 4 or worse; Mis_{it} is the misery index; $Mild_{it}$ is a binary indicator identifying a mild state (misery = 6); X_{it} represents a set of other covariates, comprising personal characteristics, elapsed time spent on earlier TTO tasks and an indicator of perceived difficulty in deciding on the equivalence point; u_i is an unobserved persistent random effect; and ε_{it} represents trial-specific random response errors.

The parameter α captures the possible dependence between the current TTO task and earlier ones, so between-task independence requires $\alpha = 0$. The parameters $\beta_1 \ldots \beta_7$ represent the influence of the specified health state on the TTO outcome, and γ captures the influence of personal characteristics. The random effect u_i allows for any tendency of individual *i* to give systematically high or low assessments of all health states.

We estimate the model (2) using Wooldridge's (2005) method for dealing with the correlation between the lagged dependent variable v_{it-1} and the random effect u_i. Coefficient estimates and their standard errors are given in Table 2.9.

In all cases, the parameter α is highly significant, so the independence assumption made by Devlin *et al.* (2018) is clearly rejected.[9] The effect is particularly strong for indicators (3) and (4), which pick up any tendency to misclassify health states at the lower and upper extremes respectively. This strong sequential dependency suggests that a significant number of participants in the TTO experiments were generating repetitive sequences of implausible or completely nonsensical responses.

[9] The initial condition variables which are used in the Wooldridge estimator are also highly statistically significant, so there is clear evidence of a lack of statistical independence within the sequences of TTO trials.

Table 2.9 Dynamic modelling of sequences of indicators of problematic TTO tasks

Coefficient	Type of problematic outcome				
	(1a)	(1b)	(2)	(3)	(4)
Lagged dependent variable	0.409***	0.498***	0.414***	1.556***	1.788***
	(0.065)	(0.121)	(0.065)	(0.378)	(0.229)
Mobility >3	0.513***	0.298***	0.294***	0.260	-0.070
	(0.079)	(0.103)	(0.083)	(0.464)	(0.186)
Self-care > 3	0.335***	0.304***	0.099	0.272	-0.525**
	(0.080)	(0.102)	(0.077)	(0.528)	(0.206)
Usual activities > 3	0.255***	0.232***	0.092	-0.220	-0.610***
	(0.064)	(0.083)	(0.062)	(0.421)	(0.172)
Pain/discomfort > 3	0.734***	0.528***	0.446***	0.475	-0.382*
	(0.074)	(0.102)	(0.075)	(0.366)	(0.220)
Anxiety/depression > 3	0.436***	0.370***	0.258***	0.418	-0.026
	(0.072)	(0.094)	(0.071)	(0.309)	(0.204)
Misery index	0.037**	0.023	0.036**	0.123	0.037
	(0.018)	(0.024)	(0.017)	(0.121)	(0.051)
Mild state (misery = 6)	-0.347***	-0.012	-0.568***	-4.473***	-
	(0.116)	(0.187)	(0.157)	(0.919)	-
Age	0.005	0.009**	0.002	0.018**	0.007
	(0.004)	(0.004)	(0.003)	(0.008)	(0.005)
Male	-0.002	-0.089	-0.165*	-0.004	0.017
	(0.096)	(0.111)	(0.085)	(0.200)	(0.139)
Never married	0.162	0.427***	-0.104	-0.236	-0.028
	(0.125)	(0.150)	(0.109)	(0.288)	(0.171)
Religion	0.089	0.196	-0.036	-0.302	-0.026
	(0.108)	(0.131)	(0.094)	(0.244)	(0.157)
Ethnic minority	0.212	0.088	0.366**	-0.946***	0.491**
	(0.184)	(0.195)	(0.163)	(0.307)	(0.233)
Limiting disability	-0.027	-0.007	-0.264**	-0.224	0.197
	(0.163)	(0.165)	(0.129)	(0.347)	(0.164)
Degree	-0.096	0.155	-0.227**	0.178	0.024
	(0.122)	(0.138)	(0.105)	(0.267)	(0.155)
Children	0.067	-0.027	-0.299***	0.158	0.136
	(0.114)	(0.135)	(0.103)	(0.227)	(0.145)
Experience of caring	-0.137	-0.237**	-0.060	0.536***	0.028
	(0.100)	(0.116)	(0.081)	(0.191)	(0.135)
English not native language	0.327	0.242	0.080	1.548***	-0.504
	(0.230)	(0.255)	(0.206)	(0.355)	(0.344)
TTO problems	-0.258***	-0.063	-0.013	-0.040	-0.091
	(0.095)	(0.108)	(0.082)	(0.181)	(0.122)
Elapsed time	1.365	-2.672	-1.860	1.172	-6.943**
	(1.414)	(1.893)	(1.341)	(4.372)	(3.431)
Variance of random effect	1.440***	1.259***	0.892***	0.025	0.254
	0.161)	(0.236)	(0.104)	(0.322)	(0.177)
No. individuals	986	986	986	860	950
No. of TTO tasks	7,100	7,100	8,874	1,705	2,507

Standard errors in parentheses. Statistical significance: * = 10%, ** = 5%, *** = 1%. Initial values and individual means of covariates are also included (see Wooldridge 2005)

2.7 The DC experiments

The design of the DC experiments give much less scope for assessing data validity than does the TTO design. There are two reasons for this:

- The DC choice situations presented to participants all involve choices between states that cannot be ordered unambiguously *a priori* : for example, no-one is asked to rank 11122 against a logically inferior state like 33234. Consequently, logically fallacious responses are ruled out as part of the experimental design. If a participant were unable to make logically consistent comparisons, we would never be aware of it.
- The individual is required to make a definite choice and there is no provision for responses like "*states A and B roughly equivalent*", or "*don't know*" or "*unable to judge*". When faced with this situation, some participants may follow the policy of making a random choice – which would be a valid response in cases of indifference between the two states. But others might resolve the difficulty by following an arbitrary policy, such as always picking alternative A. The latter type of problematic response behaviour would go undetected.

The only clear test that we can make on the DC data is a test of the assumption of statistical independence within the sequence of seven tasks undertaken by each participant. Table 2.10 shows results from a simple dynamic model:

$$\begin{aligned}\Pr(c_{it}=1) \; &= \alpha c_{it-1} + \beta_1(Mo_{Ait} - Mo_{Bit}) + \beta_2(Sc_{Ait} - Sc_{Bit}) + \beta_3(Ua_{Ait} - Ua_{Bit}) \\ &+ \beta_4(Pd_{Ait} - Pd_{Ait}) + \beta_5 Mis_{Ait} + \beta_6 Mis_{Bit} + \beta_7 t + \beta_8 + u_i \\ &+ \varepsilon_{it} \qquad \text{for } t = 2 \ldots 7 \end{aligned} \tag{3}$$

where c_{it} is a binary indicator of individual *i* choosing state B rather than A in task *t* and subscripts *A* and *B* indicate alternative states A and B. Table 2.10 shows the estimates of parameters α and $\beta_1 \ldots \beta_8$, obtained using the Wooldridge (2005) method of allowing for the endogeneity of the lagged dependent variable c_{it-1} and initial condition c_{i0} in short panel estimation.

Unlike the TTO experiments, there is no evidence of any statistical dependence between the outcomes, nor of any systematic association with the position of the task within the sequence. We also find no evidence of a persistent individual effect, since $var(u_i)$ is estimated as essentially zero. Consequently, the independence assumption made by Devlin *et al.* (2018) seems reasonable for the DC data.

Table 2.10 Dynamic modelling of sequences of responses in DC experiments

Covariate	**Coefficient** (standard error)
Alternative B chosen in previous task (c_{it-1})	-0.024
	(0.036)
Mobility difference ($Mo_{Ait} - Mo_{Bit}$)	-0.031**
	(0.013)
Self-care difference ($Sc_{Ait} - Sc_{Bit}$)	-0.085***
	(0.012)
Usual activities difference ($Ua_{Ait} - Ua_{Bit}$)	-0.102***
	(0.012)
Pain/discomfort difference ($Pd_{Ait} - Pd_{Ait}$)	0.004
	(0.012)
Misery index for state A	0.220***
	(0.010)
Misery index for state B	-0.237***
	(0.010)
Position in sequence (t)	0.005
	(0.010)
Intercept	-2.337***
	(0.982)

Statistical significance: * = 10%, ** = 5%, *** = 1%. Initial values and individual means of covariates are also included (see Wooldridge 2005)

SECTION 3: SPECIFICATION AND ESTIMATION OF THE VALUATION MODEL

In this section, we investigate the consistency of the model specification and the reliability of the estimated model parameters which underpin the Devlin *et al.* (2018) value set for England.

3.1 Specification issues

In this review, we concentrate primarily on the final model estimated by Feng *et al.* (2018), which was used to construct the proposed value set in Devlin *et al.* (2018). Appendix A2.1 presents that model, based primarily on the WinBUGS code supplied, since published papers only provide an incomplete mainly verbal account of the model. The model assumes that there is a shared utility function underlying TTO and DC responses. This assumption allows estimation of a common set of parameters for the combined dataset. Twenty parameters (5 dimensions × 4 levels) measure the utility decrements from full health (all EQ-5D-5L dimensions at level 1). It is assumed that there are three distinct latent groups of individuals in the population with unobserved group membership. Each latent group shares the same underlying decrements in utility up to a proportionality constant, which Feng *et al.* (2018) termed a disutility scale. The degree of randomness in the TTO responses (the variance of the error terms) is allowed to differ across latent groups.

In arriving at this final model, Feng *et al.* (2018) estimated a sequence of models and specifications using TTO responses in isolation and combined with DC responses. Some of these preliminary models exhibit inconsistencies where the decrements in utility are higher for level 4 (severe) than for level 5 (extreme) in the usual activities and the anxiety/depression dimensions (these problems are consistent with the TTO data quality concerns detailed in Section 2). The final specification imposes restrictions on the parameters that force decrements in utility to conform to the expected level ordering.

The TTO valuation protocol can only produce 41 discrete values, from v = -1 to 1 in steps of 0.05. Feng *et al.* (2018) handle this discreteness by assuming that the true TTO response is only observed as the interval of radius 0.025, centred around the TTO response. Furthermore, TTO responses are assumed censored[10] at 3 possible values, -1 (respondents might have traded more

[10] Censoring occurs when a variable is only partially observed; the exact value is unobserved but we know that its value is at or beyond a limit.

time in full health if given the choice), 0 (respondents avoid using values below 0) and 1 (full health)[11].

DC responses are modelled using a binary logit model where it is assumed that the health state with the highest utility is chosen. DC responses only give rankings – they provide information about preferences but not their strength. For this reason, the parameters identified using DC and TTO data may differ in scale. To align the parameters, Feng *et al.* (2018) introduce a linear transformation for the parameters modelling the DC responses.

To complete the model, individual responses to tasks are assumed independent within each set of TTO and DC tasks as well as across them, conditional on latent class membership. An age-specific calibration weight is also used in the estimation (see Appendix 2.1 for details).

There are several significant concerns about the specification of this model, which are summarised in Table 3.1. The consequences of model misspecification are uncertain, but Table 3.1 indicates some of the possibilities.

[11] Due to the interval treatment of the data, the bottom and top censoring values used are -0.975 and 0.975 instead of -1 and 1.

Table 3.1 Potential concerns in the specification of the valuation model

Issue	Nature of problem	Potential consequences
Inadequate allowance for poor quality TTO responses	There is strong evidence (see section 2.6) of widespread lack of engagement with TTO experiments or inability to carry out TTO tasks coherently. Apart from a proportionately small number of sample adjustments, the model assumes that all TTO responses are accurate within the resolution of the measurement software.	Potentially serious biases in parameter estimates and valuation predictions
Inappropriate interpretation of limit at $v = 1$ as censoring	Valuations exceeding 1.0 are deemed possible but unobserved because of a censoring process. In fact valuations above 1 are ruled out theoretically, and the upper bound should be modelled as an inherent limit, not as censored observation.	No implications for estimates of model parameters, but systematic over-valuation, particularly of mild health states.
Heteroskedasticity assumption	TTO valuations are assumed heteroskedastic, with error variance proportional to a weight which is calculated as a calibration weight aligning the sample and population age composition. This confuses weighting for nonresponse and weighting for heteroskedasticity, which are two different statistical procedures, intended to address different statistical problems.	Overstatement of estimation precision and possible bias in parameter estimates
Independence of TTO tasks	Evidence in section 2.6 suggests that there is strong statistical dependence between the set of TTO responses made by any individual (conditional on latent class membership).	Overstatement of estimation precision, since the TTO sample contains less independent information than standard methods assume.
Inconsistency of distributional assumptions	Utility error terms assumed heteroskedastic and normally distributed in TTO experiments but homoskedastic and type I extreme value in DC experiments	Bias in parameter estimates
Intercept in DC model	Intercept in DC choice probability is interpretable as a difference between alternative-specific intercepts in the utility functions for the states being compared. It is mathematically impossible for all differences between a set of constant intercepts to have the same value.	Bias in parameter estimates

3.2 Bayesian estimation

Feng *et al.* (2018) estimate the model in WinBUGS (Lunn *et al.*, 2000), widely-used software for Bayesian analysis of statistical models, using Markov chain Monte Carlo (MCMC) methods. The Bayesian approach involves statistical inference based on a posterior distribution for the model parameters. The posterior distribution combines sample information (captured by the likelihood function) with other external information (captured by the prior distribution). The MCMC method does not calculate the posterior parameter distribution directly, instead it generates a sequence of values which eventually display the properties of random draws from the posterior distribution.

WinBUGS makes the estimation of many complex models relatively straightforward but reliable inferences require significant input from the user including careful consideration of prior distributions, the model and computation issues. Warnings on the WinBUGS website (https://www.mrc-bsu.cam.ac.uk/software/bugs/) and the first page of the WinBUGS User Manual alert users to the potential pitfalls (see Figure 3.1).

"Health warning The programs are reasonably easy to use and come with a wide range of examples. There is, however, a need for caution. A knowledge of Bayesian statistics is assumed, including recognition of the potential importance of prior distributions, and MCMC is inherently less robust than analytic statistical methods. There is no in-built protection against misuse." (https://www.mrc-bsu.cam.ac.uk/software/bugs/)
"Beware: MCMC sampling can be dangerous!" (WinBUGS User Manual Version 1.4)

Figure 3.1 Warnings on the WinBUGS website and User Manual.

Following specification of the full probability model (discussed in section 3.1 and Appendix 2.1), Bayesian data analysis involves (*i*) stating one's beliefs about the model parameters using prior probability distributions and; (*ii*) drawing inferences from the posterior distribution of the model parameters given the observed data. Issues (*i*) and (*ii*) are addressed in sections 3.2.1 and 3.2.2.

3.2.1 Specification of prior distributions

Priors reflect the information available before examining the data. When no information is available, noninformative priors are specified so that inferences are mainly based on the current

dataset. Given the potential impact of the priors on the results, it is important to state them explicitly and justify their choice. In the case of noninformative priors, it is prudent to assess the sensitivity of the results to the choice of priors to ensure that they play a minimal role. The priors used by Feng *et al.* (2018) are detailed in Appendix 2.2.

One important concern is the origin of the priors used in estimation. We were able to find no justification for the choice of priors in the published papers, nor any evidence of sensitivity analysis in the materials we received. Feng *et al.* (2018) state that "[the excluded data] were not used to define the prior probability distributions in the Bayesian regression analysis". We are not clear what exactly was meant by this and we could not find any indication of the source of the priors.

A second major concern is that some parts of the prior distribution appear to be both highly informative and in conflict with sample information. This is particularly true for the latent class aspect of the model, where the choice of priors is even more important than in standard models. Priors can help overcome some of the well-known difficulties in estimating these models by maximum likelihood, but they need to be selected carefully, as they may exercise considerable influence on the posterior distribution (Frühwirth-Schnatter, 2006). In particular, the prior distribution relating to the probabilities of latent class membership appears informative but is not justified. Priors for the TTO error variances appear to be in conflict with the data for some latent classes, since there are very large differences between prior and posterior means for those parameters.

3.2.2 Implementation of the simulation estimator

In Bayesian analysis, it is important to pay careful attention to the parameterisation of the model and check convergence diagnostics. The MCMC algorithm can be run for a set number of iterations and will produce results. However, for reliable inferences, it is critical to ensure that convergence to a stationary distribution has taken place and then decide on the number of additional Monte Carlo samples required to obtain the necessary precision.

It is helpful to choose a parameterization of the model that ensures reasonably quick convergence of the algorithm. If convergence problems cannot be solved by increasing the number of samples, then it is possible that a reparameterization or even different specification of the model is needed. There are two issues related to the parameterization of the model used in Feng *et al.* (2018) which deserve attention.

A model with three proportionality constants for the three latent classes is unidentified. Setting a strong prior (see Appendix A2.2) does not solve the problem and a normalising restriction is

needed. Although this type of identification does not bias predictions, it may lead to convergence problems.

Consistency of the valuation process requires that the statistical model should give lower values as the severity indicator in any domain rises. This was found to be a problem for the increase from level 4 to level 5 for two of the EQ-5D domains. Consistency was imposed on the model by specifying utility decrements as the squares of basic parameters (since a square can never be negative). It is unclear why this approach was used since, in a Bayesian framework, it would arguably be more natural to use the prior distribution to impose the restriction.[12] The drawback of using a squared term is that the sign of the underlying parameter is indeterminate; this can induce convergence problems for MCMC as the samples in the chain switch from positive and negative values, producing a bimodal distribution for the parameter.

In addition, the difficulties of identifying mixture distributions (such as the one used here to model latent classes) are well known (McLachlan and Peel, 2000). The Bayesian approach has some practical difficulties, particularly *label switching*, which may arise because the mixture distribution is invariant to interchanging the order of the components. Label switching has to be addressed explicitly because, in the course of sampling from the mixture posterior distribution, the ordering (labelling) of the unobserved categories may change (Frühwirth-Schnatter, 2006).

Feng *et al.* (2018) report using a single chain with a burn-in of 2,000 iterations and 5,000 additional iterations to compute the posterior distributions. Convergence is reported to be assessed using (*i*) visual inspection of the autocorrelation graphs and (*ii*) assessing whether the last 2,000 iterations gave the same estimates (to within +/- 0.01) as the previous 2,000 iterations. It is unclear whether this was carried out for all parameters or a subset, as we found no record of these checks in the documentation provided. A uniform +/- 0.01 tolerance for all parameters is inadvisable as some estimated means are large (e.g. 18.9) but others are small – several much smaller than 0.02. We re-ran estimation of the model as reported and saved the output to assess convergence using the CODA package.[13] A warning message appears on loading the CODA files produced by WinBUGS (see Figure 3.2), indicating a lack of convergence.

---

[12] In the conclusion, Feng *et al.* (2018) report the following "…we formulated the prior distributions of the level parameters to guarantee that in each dimension, the coefficients between two adjacent levels are logically consistent". However, in the material we received, there appears to be no prior distributions incorporating this information; the restriction is imposed through the model parameterisation.

[13] CODA stands for Convergence Diagnosis and Output Analysis. It is an R package routinely used for the purpose of assessing convergence.

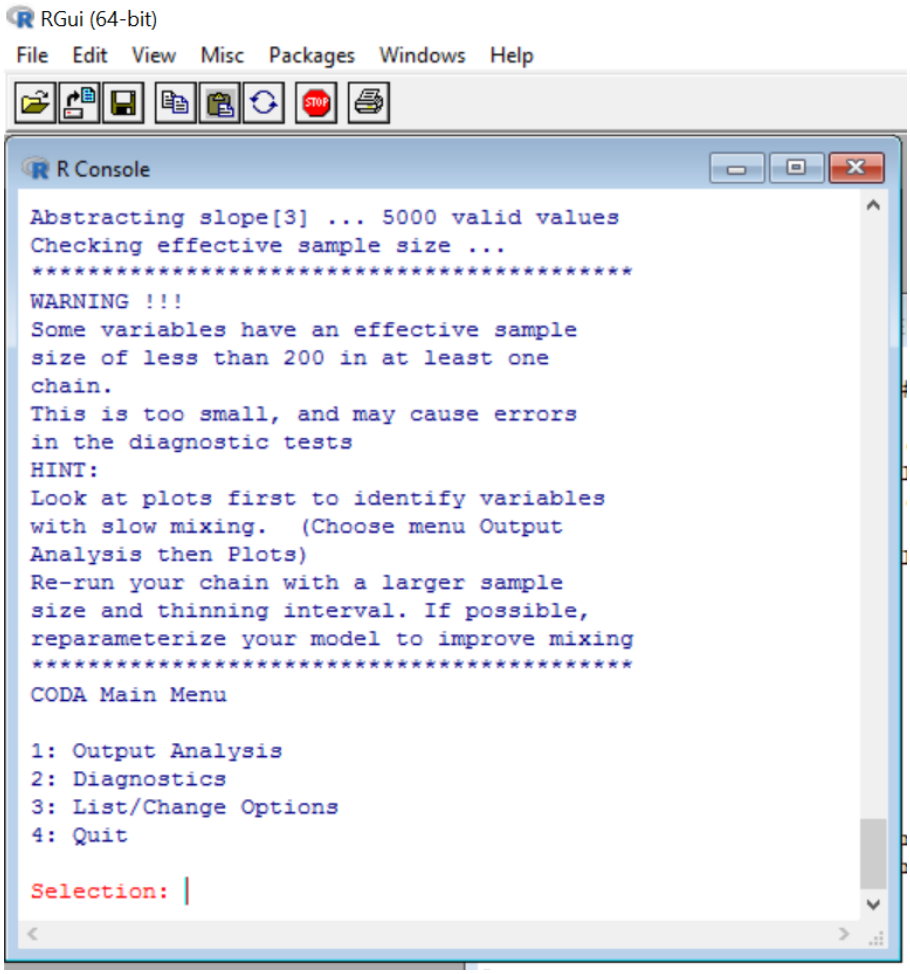


Figure 3.2: Warning message after loading the CODA files produced by WinBUGS.

We examined different convergence diagnostics and plots provided by CODA for other signs of non-convergence but only present here a few examples of evidence which strongly suggests convergence failure. Geweke (1992) Z-scores indicate problems of convergence for the two parameters (α_1 and α_2) of the linear transformation for the DC responses. Raftery and Lewis (1992b)'s convergence diagnostic measures the increase in the number of iterations needed to reach convergence due to the dependence between the samples of the chain. Raftery and Lewis (1992a) suggest values higher than 5 indicate convergence problems. More than half of the parameters have values above 5, including parameters which are important for construction of the value set. For example, the level 2 decrements give values which are all above 8 and the minimum number of iterations reported for them ranges from 30,000 to just above 47,000. The

minimum number of iterations to estimate the default 2.5% quartile[14] is above 5,000 for all parameters except the 3 latent group probabilities which have informative priors.

Bimodality of the posterior distribution is evident in the parameters governing the level 2 mobility and anxiety/depression dimensions, pointing towards label switching problems (see Figure 3.3).

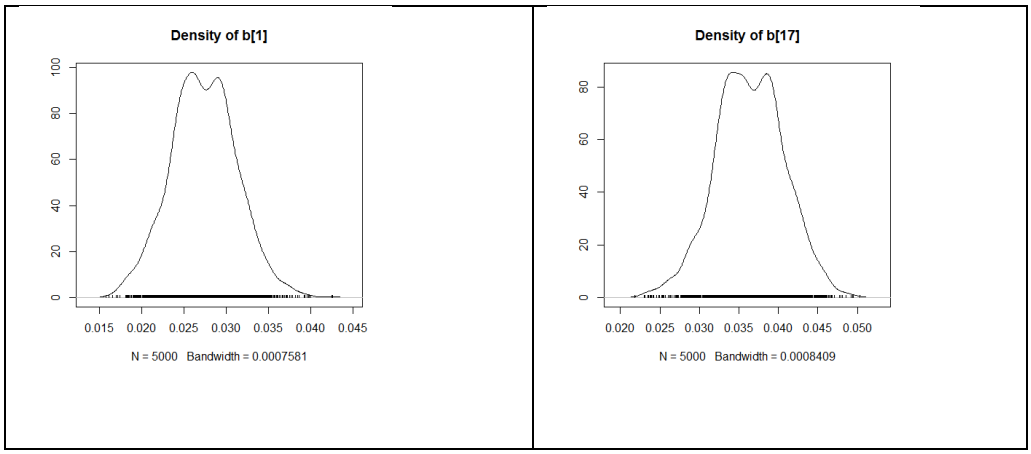


Figure 3.3: Density of the parameters governing the level 2 mobility parameter (b[1]) and anxiety/depression parameter(b[17])

To ensure that problems are not due to the lack of identifiability of the model or the small number of iterations, we re-ran the computations for 60,000 iterations with the missing normalisation constraint imposed. We still find significant evidence of lack of convergence. Particularly problematic is the parameter α_2 (see Appendix 2.1) linking the TTO and DCE data which shows extremely high MCMC autocorrelations even after many lags (see Figure 3.4).

Table 3.2 summarises our conclusions on the Bayesian analysis: not only do we find clear evidence of convergence failure but also more fundamental specification/parametrisation problems that cannot be solved by increasing the length of the MCMC chain.

[14] with default probability (0.95) of attaining the default accuracy (±0.005).

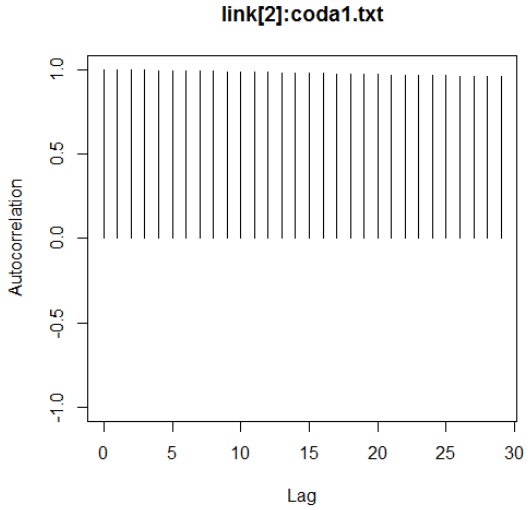


Figure 3.4: Autocorrelation plot of the parameter linking the TTO and DCE responses, α_2.

Table 3.2 Potential issues in the Bayesian analysis

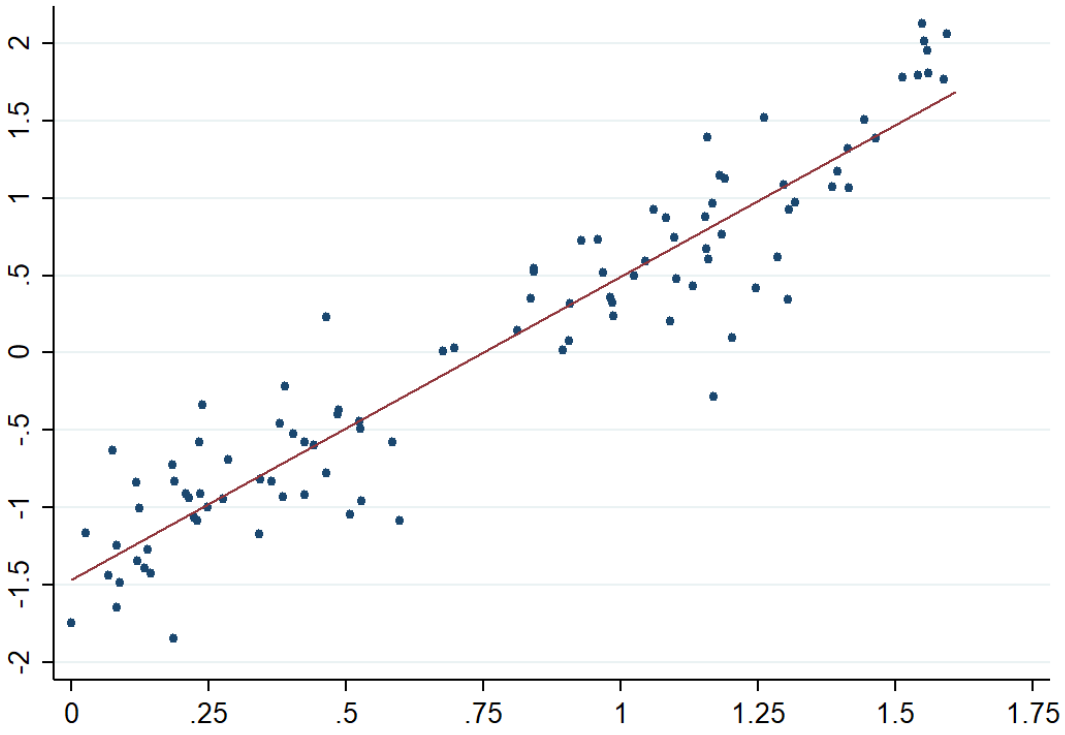
Issue	Nature of problem	Potential consequences
Choice of priors	Priors on key parameters are informative. There is no justification for the priors used or sensitivity analysis presented.	Results dependent on priors which may be unreliable.
Unidentified model	A model with proportionality constants for all latent groups is theoretically unidentified.	No direct implications for prediction of utility values, but inflated parameter uncertainty and problems of convergence may distort results.
Parameterization of the model	Specification of some parameters may cause problems for the algorithm.	Slow mixing and convergence failure, leading to unreliable estimates.
Label switching	The labelling of the unobserved categories changes when sampling from the mixture posterior distribution.	The posterior marginal densities estimated from the samples may be poorly estimated
Single vector of initial values	The MCMC sampler could get trapped in a spurious mode.	Inference regarding parameters of interest may not be reliable
Convergence failure	Insufficient number of iterations to ensure convergence to the stationary distribution, possibly as a result of inappropriate model specification or parametrisation.	Inference regarding parameters of interest may not be reliable

3.3 Derivation of the value set: prediction of limited and censored variables

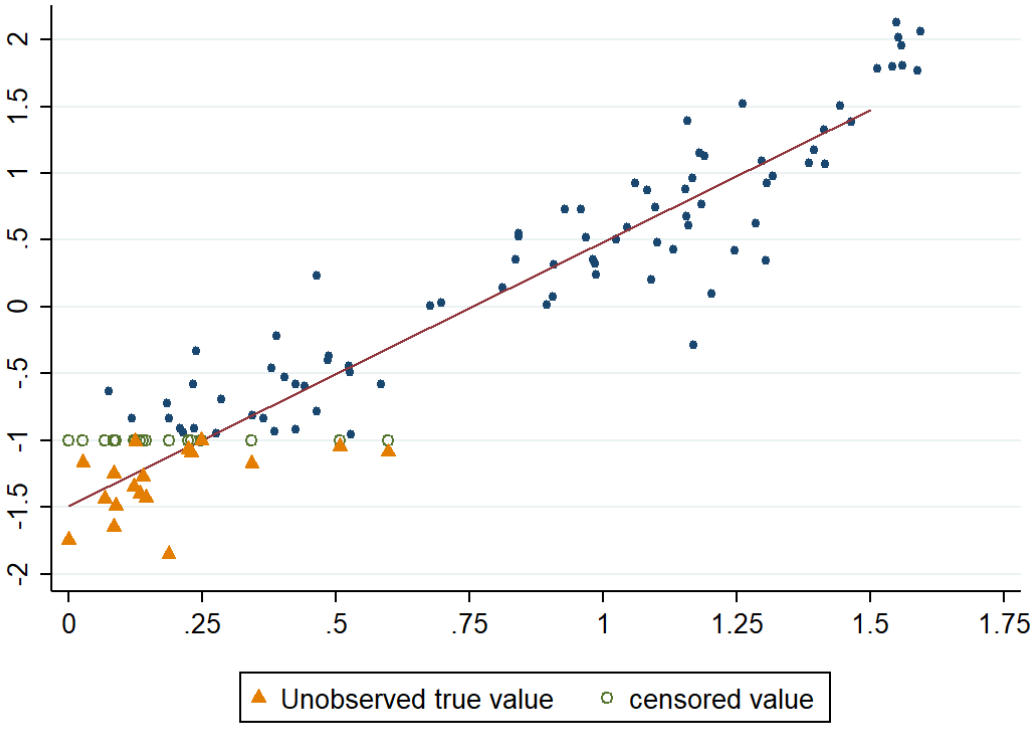
There is an important distinction between two quite different processes that can lead to the same distribution of observable data. They are *censoring* and *conceptual limitation*. Although these lead to exactly the same data distribution (*i.e.* likelihood function), the appropriate way to predict from the fitted model depends on which of the two processes is at work. We use a simple illustrative example of a linear regression model. Figure 3.5(a) shows the linear regression function and scatter of sample points for a standard regression model.

Suppose (as is the case for the TTO experiments) that the dependent variable y is exactly observed if it takes a value above -1, but the limitations of the TTO measurement process means that the outcome is coded as -1 if the true value of y lies below -1. This is known as censoring from below at -1. Figure 3.5 illustrates this, using a simulated sample. Panel (a) shows the true data on y generated from a linear regression model; in panel (b), the outcomes below -1 (shown as yellow triangles) are not observed, instead the outcomes are recorded as the hollow circles at $y = -1$.

To avoid bias, the estimation process must take account of the pile-up of outcomes at the discrete value -1, since otherwise the slope of the relationship will tend to be underestimated. Once the censored regression model has been estimated, it is appropriate to predict values of y from the straight line relationship, since we are interested in the true values of y (which may lie below -1), rather than the censored observations.



(a) Data before censoring



(b) Data after censoring

Figure 3.5: The linear regression model with censoring at -1: straight-line predictions may correctly lie below -1

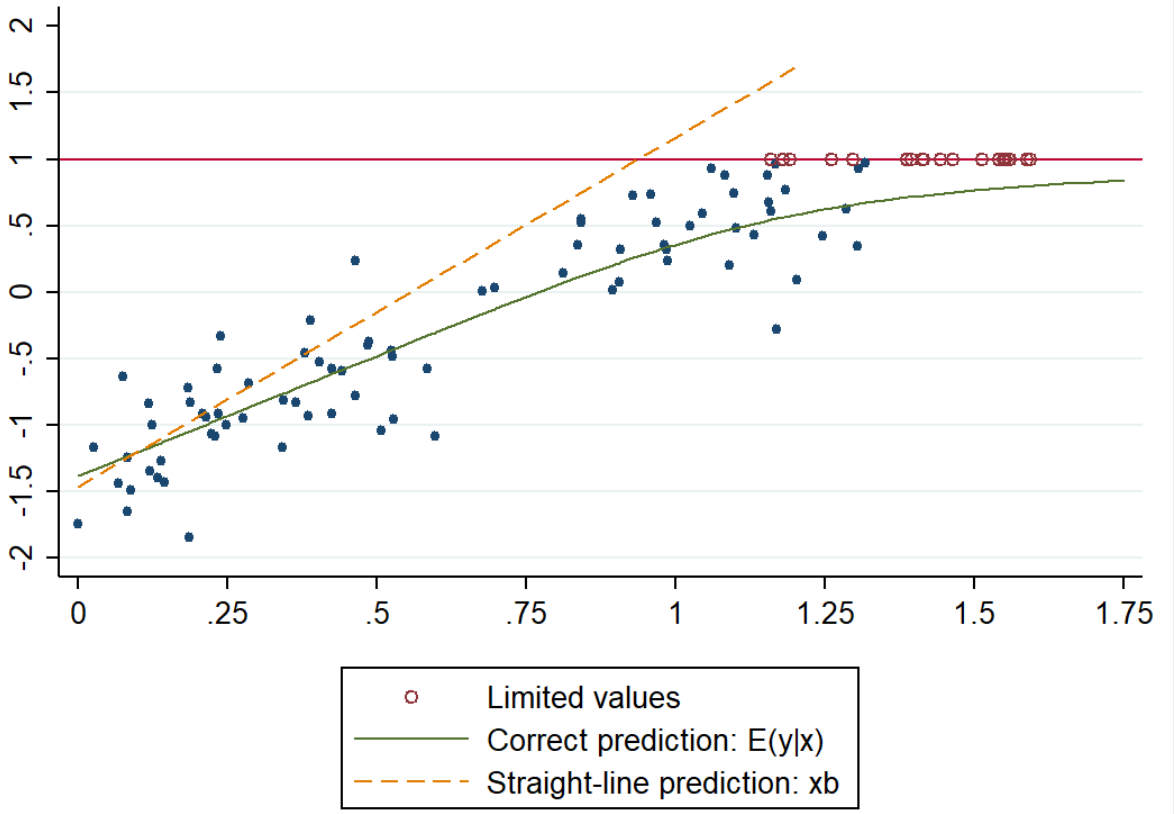


Figure 3.6: **The limited (Tobit) regression model with conceptual constraint at +1: the straight-line prediction formula is incorrect and may generate values above 1; the correct conditional expectation predictor always lies below 1 (Pudney 1989, p. 141).**

The prediction method used by Devlin *et al.* (2018) uses the straight-line predictor ($x\beta$ in technical language) which would be quite accurate at the lower limit where values below -1 cannot be observed because of censoring but also, inaccurately, at the full health level where values above +1 are logically impossible. The correct conditional expectation prediction formula, $E(v|x)$, is more complex and corresponds to the solid curved line in Figure 3.6.

SECTION 4: CONCLUSIONS AND RECOMMENDATIONS

We undertook a review of the methods used to generate the EQ-5D-5L value set for England. With access to the raw data and the code to implement the statistical models, we have interrogated the analysis in detail.

In section 2 we examined the raw data in order to understand the challenges faced for the subsequent modelling. This examination revealed a number of serious deficiencies in relation to the TTO data. These deficiencies lead us to have doubts about the extent to which the respondents to these experimental tasks were able to understand what was being asked of them, were able to distinguish the health states they were asked to value or engaged in the tasks. We found that even casual inspection of the individual-level TTO data revealed immediately that much of the data treated as perfectly accurate in the valuation analysis is logically inconsistent or otherwise potentially misleading. The same may be true of the DC experiments but the experimental design precludes any assessment of data quality.

Our analyses do not allow us to identify the reasons why the data suffer these limitations. It may be a consequence of the EuroQoL valuation protocol but there are other potential sources. The valuation analysis performed by Devlin et al. (2018) is based on experimental data generated under version 1.0 of the EuroQol valuation protocol (EQ-VT), which has been found to yield poor quality data in a number of international applications and in EuroQol's own in-house evaluation. That version of EQ-VT has been superseded in two subsequent revisions of the protocol that have led to improvements in some aspects of data quality. However, it should not be assumed that the EQ-VT is the sole source of subsequent data limitations. It may be the case that there are problems with the descriptive system of 5L itself, the TTO procedure in general as a means of valuing health states, or other aspects of the study inter alia, each of which require further investigation to inform any potential future new data collection.

In Section 3, we examined the specification and estimation of the valuation model. This process identified numerous serious concerns. There are flaws and inconsistencies in the specification of the statistical model; Bayesian prior distributions are specified without justification or robustness checks; and the simulation procedure used to generate draws from the posterior distribution of the model parameters does not meet recommended standards for Bayesian analysis.

Many of the identified limitations of the English 5L valuation, both relating to the data generated and the subsequent statistical modelling approach, are common with those for other countries.

4.1 Recommendations to NICE and DHSC on the proposed English value set

R1 A 5L value set for use in policy applications must be based on good quality data. A new programme of further development, including a new data collection initiative, should be considered to put EQ-5D-5L on a sufficiently firm evidential basis.

The objections to the proposed value set cannot be overcome by re-analysis of the existing experimental data, either taken as a whole or in part. Any new data collection exercise should use the opportunity to consider the range of issues set out in this report. It is unlikely that implementation of the updated EQ-VT alone gives sufficient consideration to all these issues. These issues include sample size and coverage of health states included in the experimental tasks, the value of inbuilt checks on data quality in the design including for the DC experiments, consideration of the nature of the tasks that respondents are being asked to perform, the number and format of these tasks. Section 2 of our report gives the evidence for this conclusion.

R2 A value set that is to be used in decision making must be based on a statistical modelling process that is robust and fit for purpose. If new data is collected, the statistical analysis should not simply replicate the analysis that has been reviewed here.

There are numerous, serious concerns about the quality of the statistical modelling process that underpins the proposed value set. These concerns cover several important aspects: the specification of the statistical model, the use of Bayesian prior distributions, and the simulation procedure used to estimate model parameters. These and other concerns are set out in detail in section 3 of the report.

4.2 General recommendations

Two further recommendations relate to general research issues arising from our review. They are intended to inform wider decisions by NICE and DHSC relating to research on the valuation of EQ-5D-5L health states.

R3 This review has demonstrated the value of in-depth assessment for research findings which are critical to policy. The academic peer review system cannot provide the depth of review that is required to comply with the recommendations of MacPherson, since journal referees do not have access to underlying data or computer codes, nor do they have the time or professional incentive to review in full detail.

Additionally, for policy research such as this, which involves complex statistical modelling, it would be reasonable to expect the specification and conduct of statistical work to additionally be validated by publication in a statistical or econometric journal which uses specialist statistical reviewers.

R4 It is good practice in scientific research for all relevant evidence to be made freely available to facilitate replication and secondary studies. This is even more important when the evidence underpins critical policy decisions. Full datasets, coding scripts and statistical analyses should be open to scrutiny to the fullest extent possible, within the bounds set by the need to protect personal information and respect research ethics.

The data used here (which was partly publicly funded by the Department of Health and Social Care) should be properly documented and made available without restriction to the research community under the auspices of a body like the UK Data Archive. Any future dataset considered for decision making in the UK health service should follow these same principles, making data and analyses available immediately upon completion of the research and prior to policy decisions being made.

REFERENCES

Alshreef A., Wailoo A.J., Brown S.R., Tiernan J.P., Watson A.J.M., Biggs K., Bradburn M., Hind D. (2017) Cost-Effectiveness of Haemorrhoidal Artery Ligation versus Rubber Band Ligation for the Treatment of Grade II–III Haemorrhoids: Analysis Using Evidence from the HubBLe Trial, *PharmacoEconomics Open*, doi:10.1007/s41669-017-0023-6

Augustovski F, Rey-Ares L, Irazola V, Garay OU, Gianneo O, Fernandez G, Morales M, Gibbons L, Ramos-Goni JM. (2016) An EQ-5D-5L value set based on Uruguayan population preferences. *Quality of Life Research*, 25(2):323-33.

Devlin N.J., Shah K.K., Feng Y., Mulhern B., van Hout B. (2018). Valuing health-related quality of life: An EQ-5D-5L value set for England. *Health Economics*, 27:7–22. https://doi.org/10.1002/hec.3564

Dolan, P. (1997). Modeling valuations for EuroQol health states. *Medical Care*, 35:1095-1108.

Feng Y, Devlin NJ, Shah KK, Mulhern B, van Hout B. (2018). New methods for modelling EQ-5D-5L value sets: An application to English data. *Health Economics*, 27:23–38. https://doi.org/10.1002/hec.3560.

Frühwirth-Schnatter, S. (2006). Finite Mixture and Markov Switching Models. Springer Series in Statistics. New York/Berlin/Heidelberg: Springer.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In *Bayesian Statistics,* 4:169-193 (J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, eds.) Oxford, U.K.: Oxford University Press

Hernández Alava, M., Wailoo A., Grimm S., Pudney S. E., Gomes M., Sadique Z., Meads D., O'Dwyer J., Barton G. and Irvine L. (2018). EQ-5D-5L versus 3L: the impact on cost-effectiveness in the UK, *Value in Health*, 21:49-56.

Kim SH, Ahn J, Ock M, Shin S, Park J, Luo N, Jo MW. (2016) The EQ-5D-5L valuation study in Korea. *Quality of Life Research,* 25(7):1845-52.

Ludwig, K., Graf von der Schulenburg, J.-M. and Greiner, W. (2018). German value set for the EQ-5D-5L. *PharmacoEconomics*, forthcoming. https://doi.org/10.1007/s40273-018-0615-8.

Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000) WinBUGS — a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10:325–337.

Macpherson, N. (2013). Review of quality assurance of Government analytical models: final report. London: HM Treasury.

McLachlan G.J., Peel D. (2000) Finite mixture models. New York: Wiley.

NICE (2013). Guide to the methods of technology appraisal 2013 (PMG9). National Institute for Health and Care Excellence: https://www.nice.org.uk/process/pmg9/resources/guide-to-the-methods-of-technology-appraisal-2013-pdf-2007975843781

Oppe, M., Devlin, N.J., van Hout, B., Krabbe, P.F.M. and de Charro, F. (2014). A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. *Value in Health*, 17: 445-453.

Oppe, M., van Hout, B. (2017). The "power" of eliciting EQ-5D-5L values: the experimental design of the EQ-VT. EuroQol Working Paper Series Number 17003

Pennington B, Hernandez Alava M, Pudney S, Wailoo A. (2018). Comparing the EQ-5D-3L and 5L versions. What are the implications for model-based cost effectiveness estimates? NICE DSU report, forthcoming.

Pudney, S. E. (1989). *Modelling Individual Choice. The Econometrics of Corners, Kinks and Holes*. Oxford: Blackwell.

Raftery, A. E., & Lewis, S. M. (1992a). Comment: One Long Run with Diagnostics: Implementation Strategies for Markov Chain Monte Carlo. *Statistical Science*, 7(4), 493-497.

Raftery, A.E. and Lewis, S.M. (1992b). How Many Iterations in the Gibbs Sampler? In *Bayesian Statistics*, 4: 763-773 (J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, eds.) Oxford, U.K.: Oxford University Press

Ramos-Goñi JM, Craig B, Oppe M, Ramallo-Fariña Y, Pinto-Prades JL, Luo N, Rivero-Arias O (in press) Handling data quality issues to estimate the Spanish EQ-5D-5L Value Set using a hybrid interval regression approach. Value in Health

Shah, K., Rand-Hendriksen, K., Ramos, J.M., Prause, A.J. and Stolk, E. (2014). Improving the quality of data collected in EQ-5D-5L valuation studies: a summary of the EQ-VT research methodology programme. *31st Scientific Plenary Meeting of the EuroQol Group, Stockholm September 2014; Proceedings: 1-18*. http://eq-5dpublications.euroqol.org/download?id=0_53918&fileId=54332

van Hout, B., Janssen, M., Feng, Y.-S., Kohlmann, T., Busschbach, J., Golicki, D., Lloyd, A., Scalone, L., and Pickard, A. (2012). Interim scoring for the EQ-5D-5L: Mapping the EQ-5D-5L to EQ-5D-3L value sets. *Value in Health*, 15:708-715.

Versteegh, M.M., Vermeulen, K.M., Evers, S.M.A.A., de Wit, G. A., Prenger, R. and Stolk, E. A. (2016). Dutch tariff for the five-level version of EQ-5D. *Value in Health*, 19: 343-352.

Wailoo, A., Hernández Alava, M., Grimm, S., Pudney, S., Gomes, M., Sadique, Z. et al. (2017) Comparing the EQ-5D-3L and 5L versions. What are the implications for cost effectiveness

estimates? Report by the DSU Available from http://nicedsu.org.uk/wp-content/uploads/2017/05/DSU_3L-to-5L-FINAL.pdf

Wooldridge, J.M. (2005). Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *Journal of Applied Econometrics* 20: 39-54.

Xie F, Pullenayegum E, Gaebel K, Bansback N, Bryan S, Ohinmaa A, Poissant L, Johnson JA. (2016) A Time Trade-off-derived Value Set of the EQ-5D-5L for Canada. *Medical Care* 54(1):98-105.

APPENDICES

A1 Materials accessed in the review

We were supplied with a range of materials by the authors of the proposed value set, who also responded to a request for further clarification. These materials included data files with accompanying variable lists, and computer code. We also derived some information from pre-publication versions of the published sources.

A1.1 Data files

Five data files were supplied to us in text format:

- `respUK1000.txt`: containing original data on the personal characteristics and experimental assignment of 999 individuals (the file has 1,000 data rows, but all entries for row 344 are coded "NA")
- `respUKweight.txt`: containing information on a range of personal characteristics of 999 individuals, together with a weight variable. This appears to contain a subset of the variables in the original data file with row 344 deleted. The origin of the weight variable is unclear; it is not documented and appears not to have been used in the published analysis.
- `ttoUK1000.txt`: containing the results of ten TTO tasks for each of 1,000 individuals. It appears that the individual occupying row 344 of file `respUK1000.txt` carried out the full set of TTO tasks but did not supply personal details.
- `dceUK1000.txt`: containing the results of seven DC tasks for each of 996 individuals
- englandwales.txt: containing the age/gender breakdown of the England and Wales population in mid-2015, derived from population data from the Office for National Statistics.

Accompanying these data files were:

- a description of the files and a spreadsheet describing the variables covered by the datasets
- interviewer instructions and the slides that had been used in interviewer briefings

We also requested and received data files from the pilot study (n = 49) conducted prior to the main experiments.

A1.2 Programme code

We received the following files:

- 82 different BUG model specification files
- 6 R files which process the BUG files in batches
- one excel file containing summary results of the estimated models.

A1.3 Unpublished sources

Devlin, N., Shah, K. K., Feng, Y., Mulhern, B. and Van Hout, B. (2016). Valuing health-related quality of life: An EQ-5D-5L value set for England *Health Economics & Decision Science (HEDS) Discussion Paper Series No. 16.02*.

Feng, Y., Devlin, N., Shah, K. K., Mulhern, B. and Van Hout, B. (2016). New methods for modelling EQ-5D-5L value sets: An application to English data. *Health Economics & Decision Science (HEDS) Discussion Paper Series No. 16.03*.

A2 Technical aspects of the specification of the valuation model

A2.1 The algebraic form of the valuation model

For simplicity, we postpone consideration of three complications: the censoring of TTO observations at v = -1; the upper limit of values at v = +1; and the observation of TTO timings only to the nearest 0.5 years. Initially, we treat the value v as an unbounded, continuously variable, accurately observed quantity.

Note that the algebraic notation used by Devlin *et al.* (2018) and Feng *et al.* (2018) is not fully consistent, so we use our own notation here.

First consider the TTO data. The econometric model of the value v is a latent class regression model with the following features.

$$v_{it} = 1 - \gamma_i\left[\sum_{d=1}^{5}\sum_{k=2}^{5}\beta_{dk}x_{itdk}\right] + \varepsilon_{it} \quad (1)$$

$$\gamma_i = \gamma^c \quad \text{with probability } p^c \text{ for } c = 1 \ldots 3 \quad (2)$$

where $p^1 \ldots p^3$ are non-negative constants summing to 1 and:

$i = 1 \ldots N$	indexes individual participants
$t = 1 \ldots 10$	indexes TTO tasks
$d = 1 \ldots 5$	indexes the five health domains covered by EQ-5D
$k = 1 \ldots 5$	indexes individual the five severity levels of EQ-5D in each domain
$c = 1 \ldots 3$	indexes the three latent classes of response style

The covariate x_{itdk} is a binary variable taking the value 1 if the health state presented to participant i in task t has domain d set at level k.

The conditional distribution of the error terms is specified to have the following normal distribution within latent class c:

$$\varepsilon_{it} \mid \boldsymbol{x}_i, \gamma_i \sim N\left(0, \frac{\sigma_c^2}{w_i}\right) , \quad (3)$$

where $\boldsymbol{x}_i$ **is** shorthand for $\{x_{itdk}, t = 1 \ldots 10, d = 1 \ldots 5, k = 1 \ldots 5\}$ and w_i is the age-specific weight constructed as the ratio of the population to sample proportion of respondent i's age group.

The dependence of the variance of ε_{itc} on w_i is described by Feng *et al.* (2018) as a model of heteroscedasticity, but this is a definite flaw in the specification, since the construction of w_i is dependent on a sample statistic. The interpretation would be that the degree of randomness in

participant *i* 's response behaviour is related to the number and type of individuals that were recruited for the experiments – which is an indefensible assumption.

The TTO component of the model is completed by an assumption that responses are statistically independent, across individuals and across the ten TTO tasks for any given individual. This allows the likelihood for all tasks to be constructed as a product of the marginal likelihood for each task individually. Independence across individuals is a reasonable assumption which is automatically satisfied by random sampling procedures. However, independence between tasks within individuals is a strong assumption which appears untenable on the basis of results in section 2.6 (see Table 2.9).

Under these assumptions, the likelihood function for the individual *i* based purely on TTO data is:

$$L_i^{TTO} = \sum_{c=1}^{3} p^c \prod_{t=1}^{10} \frac{1}{\sigma_c/\sqrt{w_i}} \varphi\left(\frac{v_{it} - 1 - \gamma^c\left[\sum_{d=1}^{5}\sum_{k=2}^{5}\beta_{dk}x_{itdk}\right]}{\sigma_c/\sqrt{w_i}}\right) \quad (4)$$

where $\varphi(.)$ is the density function of the N(0, 1) distribution.

The DC responses are modelled using a binary discrete choice framework. Given the *t*th DC task of ranking two health states *A* and *B*, characterised by EQ-5D health descriptors x_{itdk}^A and x_{itdk}^B, individual *i* ranks state *A* above state *B* if:

$$v_{it}^A - v_{it}^B > 0 \quad (5)$$

where v_{it}^A and v_{it}^B are the subjective values assigned to states *A* and *B*. To allow both TTO and DC data to contribute to a common valuation, v_{it}^A and v_{it}^B are assumed to be related to the characteristics x_{itdk}^A and x_{itdk}^B in the same way as in the TTO experiments, via equation (1). If TTO and DC states are valued in exactly the same way, this would imply:

$$v_{it}^A - v_{it}^B = \gamma_i\left[\sum_{d=1}^{5}\sum_{k=2}^{5}\beta_{dk}\left(x_{itdk}^B - x_{itdk}^A\right)\right] + \left(\varepsilon_{it}^A - \varepsilon_{it}^B\right) \quad (6)$$

Making assumption (3) about the heteroskedastic normality of the random errors ε_{it}^A and ε_{it}^B, this would give a heteroskedastic probit model for ranking of *A versus B*, since $\varepsilon_{it}^A - \varepsilon_{it}^B \sim N\left(0, \frac{2\sigma_c^2}{w_i}\right)$. However, Devlin *et al.* (2018) differ from this in two ways:

(*i*) The error distribution in (6) is assumed to be homoskedastic and logistic, which would imply that ε_{it}^A and ε_{it}^B have type I extreme value distributions (Pudney 1989, appendix 1). This conflicts with the heteroskedastic normal distribution (3) assumed for utilities in the TTO experiments. It

would be entirely feasible to maintain the normality assumption for the DC experiments, so it is unclear why this switch of distributional assumption was made. It introduces a complication over scaling, since the standard logistic and normal distributions have different variances.

(*ii*) A linear transformation is introduced into the utility difference:

$$v_{it}^{A} - v_{it}^{B} = \alpha_1 + \alpha_2\gamma_i\left[\sum_{d=1}^{5}\sum_{k=2}^{5}\beta_{dk}\left(x_{itdk}^{B} - x_{itdk}^{A}\right)\right] + \left(\varepsilon_{it}^{A} - \varepsilon_{it}^{B}\right) \quad (7)$$

There are two difficulties with this:

- There is no mathematical basis for the intercept α_1. If expression (6) is the difference between two utilities, then the intercept α_1 must be the difference between two state-specific intercepts: $\alpha_1 = \alpha_1^A - \alpha_1^B$. But α_1 is specified as a constant across all pairwise comparisons, which means that all state specific intercepts α_1^A, α_1^B *etc.* must differ from one another by a universal constant. This is mathematically impossible, unless $\alpha_1 = 0$.
- The slope coefficient α_2 is interpreted as a scale factor adjusting for the difference in scaling of the normal TTO errors (3) and the logistic error differences in (7). However, that interpretation would require α_2 to be heteroskedastic across individuals *i* and to differ across latent classes *c*. Instead, it is specified as a constant.

Define Y_{it} as a binary variable equal to 1 if individual *i* rates state *A* as better than state *B* in DC task *t*, and 0 otherwise. Then, under the assumptions of Devlin *et al.* (2018), the (unweighted) likelihood for the seven DC tasks undertaken by individual *i* would be:

$$L_i^{DC} = \sum_{c=1}^{3} p^c \prod_{t=1}^{7} P_{it}^{c\,\mathrm{Y}_{it}}(1 - P_{it}^{c})^{1-\mathrm{Y}_{it}} \quad (8)$$

where P_{it} is the logit probability:

$$P_{it}^{c} = \frac{\alpha_1 + \alpha_2\gamma^c\left[\sum_{d=1}^{5}\sum_{k=2}^{5}\beta_{dk}\left(x_{itdk}^{B} - x_{itdk}^{A}\right)\right]}{1 + \alpha_1 + \alpha_2\gamma^c\left[\sum_{d=1}^{5}\sum_{k=2}^{5}\beta_{dk}\left(x_{itdk}^{B} - x_{itdk}^{A}\right)\right]} \quad (9)$$

Devlin *et al.* (2018) make the further assumption that, conditional on latent class membership, individuals' behaviour in the TTO and DC tasks are independent. They also introduce the calibration weights w_i into the DC likelihood, to give an overall likelihood:

$$L_i = \sum_{c=1}^{3} p^c \left\{\prod_{t=1}^{10} \frac{1}{\sigma_c/\sqrt{w_i}}\varphi\left(\frac{v_{it} - 1 - \gamma^c\left[\sum_{d=1}^{5}\sum_{k=2}^{5}\beta_{dk}x_{itdk}\right]}{\sigma_c/\sqrt{w_i}}\right)\right.$$

$$\times \prod_{t=1}^{7} P_{it}^{c\,w_i Y_{it}} (1 - P_{it}^{c})^{w_i (1 - Y_{it})} \Bigg\} \qquad (10)$$

Note that this involves the weights w_i being used in two conflicting ways – as adjustments for heteroskedasticity in the TTO component and for sample post-calibration in the DC component.

A.2.2 Bayesian priors for the valuation model[15]

Feng *et al.* (2018) report the following priors:

- a normal prior N(0.1, 1) is assumed for each of the five level 2 coefficients ($\beta_{12} \ldots \beta_{52}$)
- The coefficients for the remaining levels 3 to 5 ($\beta_{13} \ldots \beta_{55}$) are constrained to impose consistency. Each is built as the coefficient for the previous level plus a squared parameter: $\beta_{dk} = \beta_{dk-1} + \theta_{dk}{}^2$, which ensures that $\beta_{dk} \geq \beta_{dk-1}$ for each domain *d* and levels *k* = 3 ... 5. A normal prior N(0.01, 1) is assumed for each of the 15 parameters θ_{dk}.
- Normal priors N(0, 10) and N(1, 100) are assumed for the intercept α_1 and slope α_2 of the linear transformation introduced to combine the DC and TTO responses.
- Gamma priors are used for the proportionality constants γ^c for the three latent groups. The prior for γ^1 is Γ(1,000, 1,000); for the remaining two groups, priors Γ(0,1, 0,1) are used.
- The probabilities of latent class membership p^c are given a Dirichlet prior Dir(0.3,0.3,0.4)

Not reported in Feng *et al.* (2018), Gamma priors Γ(0.1,0.1) for the three precision parameters ($1/\sigma_c^2$) of the three latent groups are also assumed.

For several parameters, it is difficult to justify the assertion that these priors are non-informative. It is good practice to carry out sensitivity analysis to investigate the role of each prior in determining the final estimates. The following issues seem particularly important:

---

[15] In this appendix, we refer to the normal distribution characterised using its mean and variance. Note that in WinBUGS the normal distribution is parameterised using its mean and precision (reciprocal of the variance). Feng *et al.* (2018) report the parameterization inconsistently sometimes using the WinBUGS convention and at other times reporting the variance.

- Priors relating to parameters determining latent groups need to be chosen very carefully. Selecting a prior on precision parameters imposes structure on the variances. Priors can help avoid spurious local modes which plague latent class models but may exercise considerable influence on the posterior distribution (Frühwirth-Schnatter, 2006).The priors chosen for the precision parameters of the three latent groups ($1/\sigma_c^2$) seem to be at odds with the data as some of the posterior mean values would occur with near zero probabilities according to the priors. The mean and the variance of the prior distributions for $1/\sigma_c^2$ are 1 and 10 respectively but the posterior means of two of the precision parameters are 9.5 and 18.9. This apparent conflict is not discussed in the published papers and we have no evidence that any sensitivity analysis has been carried out.
- There is no clear justification for using the vector (0.3, 0.3, 0.4) as the parameters of the Dirichlet prior for the probabilities of group membership. The Dirichlet parameters can be any positive numbers and do not have to sum to one (indeed, their sum can be used as a measure of their informativeness of the prior distribution). If there is no prior knowledge about the components one could use a symmetric Dirichlet with the same value for all three parameters, such as (1,1,1). Sensitivity analysis with different priors is recommended.