**Article:**
Hernandez Alava, M., Pudney, S. and Wailoo, A. (2020) The EQ-5D-5L value set for England: Findings of a quality assurance program. Value in Health, 23 (5). pp. 642-648. ISSN 1098-3015

https://doi.org/10.1016/j.jval.2019.10.017

# The EQ-5D-5L value set for England: findings of a Quality Assurance Programme

## ABSTRACT

### Objectives

EQ-5D-5L values for several countries now exist. Decision makers require confidence in the underlying data and statistical analyses prior to advocating their use. Independent quality assurance of the published English value set is reported.

### Methods

Data from 996 participants, and code to run published statistical models, were provided for inspection. The main elements of the study were ten lead-time trade-off (TTO) experiments and seven discrete choice experiments (DCEs). Data quality were examined and tested with respect to subsequent assumptions made in the statistical analysis. We examined the statistical analysis including model specification and estimation methods.

### Results

The TTO experiments covered under 3% of the possible 3125 5L health states. There is strong evidence, both direct (self-reported) and indirect (poor data quality), that many participants found tasks difficult or did not engage effectively. 47% of respondents valued more than 20% of states inconsistently, double the 3L rate. DCEs covered 12.5% of possible states, 0.01% of possible 2-state comparisons. The design precludes examination of inconsistent responses. Several aspects of the statistical model conflict with the data and underlying experimental design. The model is unidentified. The Bayesian approach relies on unjustified, informative priors. There is clear failure to achieve convergence.

### Conclusion

Significant limitations are identified with the quality of the valuation data and the subsequent statistical analysis that underpin the English EQ-5D-5L value set. 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 for healthcare decision making.

**Introduction**

EQ-5D covers five dimensions of health: mobility, self-care, usual activities, pain / discomfort, and anxiety / depression. The EQ-5D-3L (3L) allows responses of "no problems", "some problems" or "extreme problems". 3L has an associated set of utility values based on estimates of the preferences of the UK general population[1], and many other countries.

3L is the most widely used preference based measure used in economic evaluation internationally. In England, the National Institute for Health and Care Excellence (NICE) recommends the 3L be used in economic evaluations submitted to its Technology Appraisal Programme[2].

A new version, EQ-5D-5L (5L), includes five levels of severity for each dimension (no problems, slight problems, moderate problems, severe problems, and extreme problems) and intends to improve the instrument's sensitivity and reduce ceiling effects[3]. Utility values for 5L are now available for England[4].

Work undertaken for NICE via its Decision Support Unit[5,6,7] demonstrated that economic evaluations undertaken using 5L rather than 3L are likely to generate very different results. Differences stem from both the responses to the descriptive systems and the valuation systems. 5L utilities are shifted up the distribution towards full health and compressed into a smaller space than 3L. These differences are often relatively large, having profound effects on estimates of cost-effectiveness. Decreases in incremental health gained of up to 87% were demonstrated in case studies reported by Hernandez et al.[6] 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.

Policy makers must therefore consider how they recommend health utilities are calculated, and how to deal with evidence from different versions of EQ-5D. Clearly, 3L and 5L cannot be used interchangeably. It is prudent to ensure that changes to current policies are based on robust analyses,

more so given the risk associated with any move to 5L that might underpin hundreds of health care decisions for decades. These concerns ought not be unique to the UK, but UK government practice mandates such quality assurance[8]. Decision makers may additionally wish to subject other measures, like 3L, to similar interrogation in the future when formulating policy.

In 2017, NICE announced its support for an independent quality assurance of the 5L valuation, commissioned by the UK Department of Health and Social Care. This paper reports its findings. It covers examination of the data on which the English Value Set depends, the extent to which the data are manipulated or excluded before use in the statistical modelling, whether the application of the statistical methods aligns to that which has been reported by the value set authors, and the appropriateness of those methods. The overall aim is to assess the extent to which the value set for England can be considered a firm basis for policy decisions.

**Methods**

The methods used to derive the English 5L value set are described elsewhere[4,9]. We provide essential details here for the reader to understand the quality assurance steps we conducted.

The EuroQol Group has developed a valuation protocol known as the EuroQol Valuation Technology (EQ-VT) for 5L. EQ-VT comprises prescribed numbers of experimental subjects, experimental tasks and health states to be compared[10]. It provides a digital environment for computer-assisted personal interviewing. EQ-VT has two main elements: ten lead-time time trade-off (TTO) experiments and seven discrete choice experiments (DCEs). Participants complete both the TTO and DC tasks. EQ-VT has changed over time in response to major data issues in early versions[11]. The English value set uses version 1.0, as did Canada[12], Netherlands[13], China[14] and Spain[15].

*TTO*

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. 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. Equivalence points are determined approximately, in steps of 0.5 years.

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 timeframe 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.

*DCE*

For each DC task, the participant is presented with two EQ-5D health states and asked to rank them. Respondents are required to make a definite choice. There is no indifference option.

DC experiments are much less informative than TTO experiments, and therefore play a lesser role in informing the final value set, 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.

*The experimental data*

We examined the coverage of health states in the TTO and DCE experiments and the sampling of participants.

The designers of the TTO and DC experiments included questions asking subjects to self-assess the degree of difficulty they encountered in completing the TTO tasks using 5-point Likert scales. 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 as a function of respondent characteristics.

We defined a set of ten indicators for identifying individuals generating poor-quality TTO data. These could be the result of individuals failing to understand or engage in the tasks, or due to inherent problems with the descriptive system such as the inability to distinguish differences in health states, inter alia. 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. We consider violations of this, sometimes referred to as "inconsistency", as an indicator of potentially problematic responses. Other indicators consider anomalous response patterns and valuations, for example giving the same response for all TTO trials. We use logistic regression models to examine the respondent characteristics associated with providing problematic responses.

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 could be as a result of respondent learning or fatigue, for example. We look for evidence of dependence between the current TTO task and earlier ones.

The design of the DC experiments gives much less scope for assessing data validity. The DC choice situations presented to participants all involve choices between states that cannot be ordered unambiguously *a priori*. If a participant were unable to make logically consistent comparisons, we would never be aware of it.

The only clear test that we could make on the DC data is a test of the assumption of statistical independence within the sequence of seven tasks undertaken by each participant.

*The statistical analysis*

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 cost-effectiveness studies.

The model reported by Feng et al.[9] was used to construct the proposed value set in Devlin et al.[4]. It combines TTO and DCE responses and is estimated in WinBUGS[16], widely-used software for Bayesian analysis of statistical models, using Markov Chain Monte Carlo (MCMC) methods. First, we investigate the consistency of the model specification. Second, we examine the consistency of the estimated parameters. We examined the supplied statistical code, line-by-line, from 82 WinBUGS files supplied by Devlin et al.

**Results**

*The experimental data*

Robust statistical analysis requires good coverage of the range of relevant factors in the population; otherwise, extrapolation to cases not adequately represented in the sample is dangerous. There are $5^5$ = 3,125 logically possible states defined by the EQ-5D-5L health description, and the TTO experimental design examines 86 health states – under 3% of the full set. The DCEs involve 392 health states, 12.5% of the full set. However, we only observe rankings within specified pairs of states in the DCEs. There are 3,125×3,124/2 = 4.88 million possible comparisons. The DC tasks cover 0.01% of the potential pairwise comparisons.

The EQ-VT protocol recommends a sample size of 1,000 individuals. We found no formal analysis of sampling error and specification robustness that led to this recommendation. Slightly over 2,000 potential subjects were approached using a two-stage random sampling process of English addresses. 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 (for example, the Health Survey for England reports response rates of around 60%[17]) but bias caused by non-response depends on the pattern of non-response rather than its level. All information about individuals who refused participation or gave partial responses was discarded so we have no direct evidence on the personal characteristics related to a high risk of non-response.

Three questions were asked about the difficulty of the TTO tasks. Only around 10% of participants reported serious difficulty (points 4 and 5 on the scale) in "understanding the questions" and "distinguishing between hypothetical lives". Over half the participants agreed that they had "difficulty in deciding on the equivalence point" (see Figure 1a). Some of the differences 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. Acquiescence bias may have contributed to a more positive response for the first two questions.

DCEs are potentially simpler than TTO, since they require only the ability to rank two states in terms of quality of life. Nevertheless, Figure 1b shows that there were only slightly fewer participants reporting the two highest levels of difficulty (49% rather than 54%).

We estimated logistic regression models of the probabilities of reporting either of the highest two levels of difficulty in making the TTO and DCE choices. Table 1 shows the marginal effects, defined as the sample mean predicted change in probability as each personal characteristic changes in turn by one unit. Results differ between the TTO and DCEs. 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 – which may reflect lower average willingness to admit difficulty, rather than an actual difference in difficulty.

Ten indicators of potential TTO data flaws are listed in Table 2, 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.[4].

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 sample deletions and other special treatment used by Devlin et al.[4] to deal with problematic TTO responses make relatively little difference to this picture.

Logistic regression models for the occurrence of each of the ten problematic outcomes reported in Table 2 were conducted (see Appendix Table 1). The most consistent effect, statistically significant in seven of the ten problem indicators, is for self-reported difficulty with TTO. This might have been 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. 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.

None of these tests can be performed on the DCE data because the design did not require respondents to rank health states that can be ordered unambiguously a priori.

*Model specification*

The model assumes a shared utility function underlying TTO and DCE responses. This 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 5L dimensions at level 1). However, there is inconsistency in the distributional assumptions between the TTO and DCE responses, leading to potentially biased parameter estimates. Utility error terms are assumed heteroskedastic and normally distributed in the TTO experiments but homoskedastic and type I extreme value in DC experiments.

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. 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.

The model imposes restrictions on the parameters that force decrements in utility to conform to the expected level ordering of the descriptive system. TTO responses are assumed censored at 3 possible values, -1 (respondents might have traded more time in full health if given the choice), 0 (respondents avoid using values below 0) and 1 (full health). The interpretation of the limit at 1 as censoring is inappropriate. Censoring means that values exceeding 1 are possible but unobserved. In fact, valuations above 1 are ruled out theoretically, and the upper bound should be modelled as an inherent limit, not as censored observation. Although these two different processes 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. Treating the limit of 1 as censoring leads to systematic over-valuation, particularly of mild health states. For a fuller description see section 3.3 of Hernandez Alava et al[18].

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

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. We found evidence of strong sequential dependency for the TTO experiments when tested in a dynamic panel data model, contrary to this assumption. This suggests that a significant number of participants were generating repetitive sequences of implausible or even nonsensical responses. No serial correlation was found for the DCE data.

*Bayesian estimation*

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.

Priors reflect the information available before examining the data. When no information is available, non-informative priors are specified so that inferences are mainly based on the current data. 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. We found no justification for the choice of prior, nor any evidence of sensitivity analysis. Some parts of the prior distribution appear to be both highly informative and in conflict with sample information. This is true of the latent class aspect of the model, where the choice of priors is particularly important. The prior distribution relating to the probabilities of latent class membership is highly informative but is not justified. Priors for the TTO error variances are in extreme conflict with the data for some latent classes, since there are very large differences between prior and posterior means for those parameters.

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 that convergence to stationary distribution has taken place, with an adequate number of additional Monte Carlo samples to obtain the necessary precision.

Feng et al.[9] report using a single chain with a burn-in of 2,000 iterations and 5,000 additional iterations to compute the posterior distributions. We examined a range of convergence diagnostics and plots[19,20,21] provided by the widely-used Convergence Diagnosis and Output Analysis (CODA)

package[22] in R. We found clear evidence of convergence failure. For example, the autocorrelation plot for the parameter linking the TTO and DCE responses shows extremely high MCMC autocorrelation even after many lags. There are several aspects of the model specification / parameterisation which mean that convergence problems are not overcome by increasing the length of the MCMC chain.

A model such as this, with three proportionality constants for the three latent classes, is inherently unidentified. Setting a strong prior does not solve the problem and some normalising restriction is needed. Choice consistency is built into the model by specifying utility decrements as the squares of basic parameters (since a square can never be negative). The drawback of using a squared term is that the sign of the underlying parameter is indeterminate; this induces non-convergence in the MCMC algorithm, as the samples in the chain switch from positive and negative values, producing an unstable bimodal distribution for the parameter. The difficulties of identifying mixture distributions (such as the one used here to model latent classes) are well known[23]. 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[24]. We re-ran the computations for 60,000 iterations with the missing normalisation constraint imposed. We still found significant evidence of lack of convergence, which is indicative of fundamental shortcomings of the Devlin et al (2018) model specification.

**Conclusions**

With access to the raw data and statistical code, we conducted thorough, independent quality assurance of the 5L value set for England. In our view, the findings suggest that the current estimates fall short of the required standards for decision making. This is because of the highlighted deficiencies in both the quality of the underlying data and the modelling of that data. Both sets of issues must be addressed, jointly, informed by the findings of the 5L international programme of work to date, if a 5L variant of EQ-5D is to be developed for use in England. Major data problems were also identified by the EuroQoL group, including in the English study, leading to a one year moratorium on value set

development and substantial revisions to the EQ-VT to improve data quality and address some, but by no means all, of the issues raised here.

Compared to 3L that the 5L is intended to supersede, the 5L value set is based on data from fewer respondents (912 vs 2,997), giving values for proportionally fewer health states (2.75% by TTO and 12.5% by DCE vs 17.7%), each of which was valued less often (mean 116 by TTO vs 856). Similarly high proportions of respondents make at least one inconsistent response but twice the proportion rank 20% or more of the health states inconsistently (0.47 vs 0.24). The proportion of inconsistent responses is higher (mean 0.26 vs 0.15). There were more problematic responses of all other types in the 5L sample than for the 3L.

These issues are not all restricted to the English value set and may not be unique to EQ-5D. Similar data collection methods have been used in all countries with 5L value sets. Robust, independent and transparent interrogation of data quality should be encouraged. Making data and analyses available immediately upon completion of research and prior to policy decisions being made facilitates such studies. Modelling methods vary more substantially. Several of the modelling concerns are not unique to the hybrid modelling approach. There are potential implications for all 5L value sets, even in those countries where estimates are based only on TTO data.

Most HTA agencies ensure rigorous quality assurance of decision models used to inform decision making for individual health technologies. In contrast, many of the parameters and methods which recurrently underpin those analyses are not subject to the same interrogation but instead rely on academic peer review, without access to underlying data and models, and without resources to conduct full review. Good practice now required across UK government means this situation should change.

The importance of extensive piloting is obvious. Future studies should consider, as part of the research design, approaches that facilitate the identification of problematic responses both to TTO and

DCE tasks, and the implications for the research if large proportions of respondents exhibit such responses.

Our analysis focusses solely on the data quality and statistical analysis but some of the problems cannot be entirely divorced from the design of the descriptive system for 5L and the valuation tasks. For example, there are concerns about the potential ambiguity between levels 4 and 5, at least in the English language version of the instrument. It is a limitation of our analysis that we are unable to provide insights into the source of the data problems. Those who seek to develop 5L to the point where it can be used to inform decision making will want to consider processes of obtaining more reliable valuation data including interviewer training and monitoring, piloting to establish if problems relate to the 5L descriptive system itself, and expanded coverage of health states to avoid excessive extrapolation.

[1] Dolan, P. (1997). Modeling valuations for EuroQol health states. Medical Care, 35:1095-1108.
[2] 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
[3] Herdman M, Gudex C, Lloyd A, et al Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res. 2011; 20(10): 1727–1736.
[4] 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
[5] 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
[6] 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.
[7] Pennington B, Hernandez Alava M, Pudney S, Wailoo A. (in press) "The Impact of Moving from EQ-5D-3L to -5L in NICE Technology Appraisals", Pharmacoeconomics,
[8] Macpherson, N. (2013). Review of quality assurance of Government analytical models: final report. London: HM Treasury.
[9] 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.
[10] 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.

[11] 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

[12] 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.

[13] 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.

[14] Luo N, Liu G, Li M, Guan H, Jin X, Rand-Hendriksen K. Estimating an EQ-5D-5L Value Set for China. Value Health. 2017 Apr;20(4):662-669

[15] 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

[16] 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.

[17] NHS Digital (2018) Health Survey for England 2017 Quick Guide, available at http://healthsurvey.hscic.gov.uk/media/78596/HSE17-Quick-Guide-rep.pdf (last accessed 9th July 2019)

[18] Hernández-Alava, M., Pudney, S., Wailoo, A. (2018) "Quality review of a proposed EQ-5D-5L value set for England" Policy Research Unit in Economic Evaluation of Health and Care Interventions. Universities of Sheffield and York. EEPRU Research Report 060. Available at http://www.eepru.org.uk/wp-content/uploads/2017/11/eepru-report-eq-5d-5l-27-11-18-final.pdf (last accessed 9th July 2019)

[19] 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

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

[21] Raftery, A.E. and Lewis, S.M. (1992). 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

[22] Plummer, M., Best, N., Cowles, K., et al. (2018) Output Analysis and Diagnostics for MCMC. Package 'CODA'. Available at https://cran.r-project.org/web/packages/coda/coda.pdf

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

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

## Tables

**Table 1: 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.

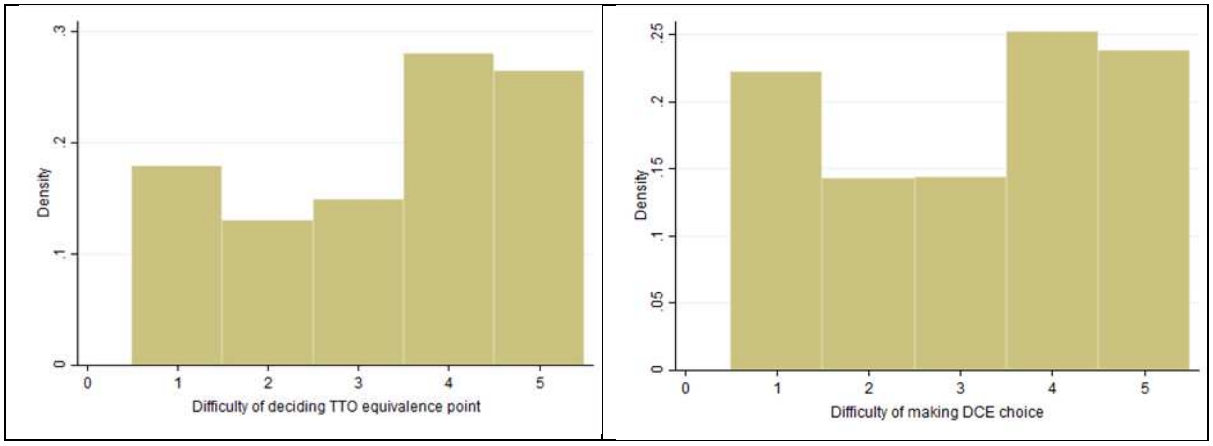
**Table 2: Proportions of individual participants displaying potentially problematic response behaviour**

| 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)

## Figures

**Figure 1: Distribution of participant responses to the degree of difficulty in a) deciding the TTO equivalence point and b) making DCE choices**



Appendix tables

**Appendix Table 1: 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) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | (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%