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

EEPRU response to comments from the 5L valuation team (Devlin et al).

Monica Hernandez Alava, Steve Pudney, Allan Wailoo.

Overview:

We received a report and accompanying technical appendix that challenges the analyses of the EQ-5D-5L value set for England that we presented in our report. There are many, many comments made there that we take issue with. We reiterate our recommendation that the data, code and statistical analyses be made available without restriction to the research community so others can form their own judgments on these issues. We restrict our response to the main issues:

1. **Coverage of health states** – Both the TTO and the DC experiments are based on a very small proportion of the feasible EQ5D-5L states. The value set for 3,125 health states is based on TTO experiments covering 2.8% of those and DCE's covering 0.01% of the feasible pairwise choices. Therefore, Devlin et al extrapolate from data covering relatively few health states, to all the 5L health states. These are statements of fact that are part of our report and are not in dispute.
2. **The quality of the TTO data** – We reported what we consider to be serious concerns about the quality of the TTO data, which indicate that many respondents either did not engage with the experiments or did not understand them. The figures that appear in our report are statements of fact.

   As stated correctly in our report, studies using later versions of the EQ-VT that reported the number of inconsistencies in the published papers (namely Korea and Germany) have substantially lower rates than the English study which used the earlier version of the valuation protocol. Indeed, improving data quality was one of the reasons for developing subsequent versions of the EQ-VT.

   Devlin et al suggest that they "developed strategies to deal with these", but there is nothing in their statistical model that takes account of the inability of many respondents to understand or engage with the experiments. Data quality is the foundation of good, robust analysis. We recommend a new data collection exercise, not in the pursuit of perfection, but because there is no modelling approach that can overcome such serious data deficiencies. Our analyses, based on access to the individual response data, show just how serious these deficiencies are.

3. **Data quality over time** - Devlin et al's responses to our report raise another issue that we were not previously aware of. We presented data for the first 30 participants in the dataset supplied to us by Devlin et al to illustrate the poor quality of the TTO data (Table 2.6 of the report) and to provide the reader with a sense of the types of data issues that are presented in summary form in Table 2.5. Devlin et al state that this is misleading because these were the first interviews conducted and that interviewer learning effects over time improve the quality of the data. This is deeply concerning. If there were such a difference

in the quality of the data collected at different times as Devlin et al claim, this is a further weakness in both the data quality and the subsequent modelling approaches (which ignore this), not a strength.

We looked for evidence of improvement in data quality over time. Data for the last 30 respondents are provided in Table 1 at the end of this report, with the types of problematic outcomes for each respondent listed below it. As with the original report, this refers to respondents 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 data in Table 1 were treated as fully accurate data in the Devlin et al analysis. Only 2 out of the 30 respondents exhibit no problematic responses. Comparing the first 30 with the last 30 respondents we find no evidence for an improvement in data quality (see Table 2). Even using the strict definition of inconsistency that Devlin et al prefer we find that the data quality deteriorates, not improves (83% vs 70% for the last and first 30 respondents respectively). We also compare the first and second half of the data in Table 2. There are no differences.

4. **The statistical modelling and estimation methods** – We found that there are fundamental flaws both in the set-up of the statistical model and the approach to estimation, which are related issues. We should reiterate here that – unlike journal referees or advisory board members – we had access both to the published accounts of what was done (which Devlin et al's responses indicate do not in fact provide adequate descriptions of the modelling that they undertook) and the statistical code, which we examined line-by-line. We therefore know the precise form of the model and, for the avoidance of any ambiguity, we wrote it down, algebraically, in Appendix A2. There is no misinterpretation of the model on our part.

   Amongst the most worrying aspects of Devlin et al's counter claims are:

   a. ***"Censoring" versus "limited variable".*** In the report, we identified a major error in the way the valuation model deals with observations of values at 1. Devlin et al's response to our criticism counters this with the claim: "Nowhere, in any of the articles we have published, has it been suggested that values above 1 are possible." This is clearly wrong. In Feng et al, section 2.5.3 is titled "Censoring at 1". That section includes the following text "values at 1 could be considered as being either 1 or greater than 1 (i.e., "right censored") and this is how we treat all values at 1." Our review of the computer code used to implement the model shows that it does indeed treat those values as censored. We reiterate our view that this is an error: the observed valuations should be viewed as limited at 1, not censored. This error will lead to a systematic over valuation.
   b. Heterogeneity. It is a fact that TTO valuations are assumed heteroskedastic with variance proportional to a calibration weight in the Devlin et al model. Equation (4) in Appendix A2 presents the TTO part of the model in algebraic form where it is clearly seen that the specified variance of each class is a function of the calibration weight, exactly as we state in the report. The authors appear to confuse this with discrete heterogeneity in the form of latent classes, both in their original model and again in

their responses to our report – these are separate issues. For weights to act as population weights they would need to be weighting each observation in the summation defining the loglikelihood, not appearing as part of the structure assumed for the error variance.[1] It makes no sense to suggest that the degree of inherent variability in some particular individual's experimental outcome depends on the degree to which his or her age group happens to be over- or under-represented in the overall sample.

c. ***Convergence and priors.*** Devlin et al appear to suggest that public policy can rely on results from their model, despite the clear evidence of convergence failure, because the results seem plausible. Plausibility of results provides scant reassurance that the correct model has been estimated and that the differences in utilities reflect the preferences of the sample or the general public. "Plausible results" could have been obtained without the inconvenience of collecting or analysing any data by this rationale.

   Bayesian analysis requires justification of the priors, sensitivity analyses and careful checking of convergence diagnostics to ensure reliable inferences. This is standard practice. Our review uncovered clear evidence of convergence failure which could not be solved by increasing the number of iterations (see section 3.2.2 of our report). Devlin et al claim that such checks were undertaken but were neither reported in the published papers nor included in any of the 82 WinBUGS files they sent us. In section 3.5, Feng et al states "All models are fitted in WinBUGS with one chain, each with a burn-in of 2,000 iterations followed by 7,000 iterations." Furthermore, if these sensitivity analyses were undertaken yet the MCMC iteration still did not converge (as our report demonstrated) then this points towards more fundamental problems of model specification (which our report listed). It is certainly not evidence that the results are reliable.

d. ***Consistency of results for different models***. Devlin et al include a diagram in the "Technical Appendix" to their response which, they claim, shows that a) DCE and TTO results are broadly aligned, and b) Maximum likelihood estimation broadly aligns with the Bayesian final model and consequently that their results are robust.

   - Despite the fact that this appears in the "Technical Appendix" there is no description, technical or otherwise, that explains how the calculations presented were made.
   - Maximum likelihood estimation of an unidentified model is not possible. The three latent class model specified in the original studies does not have an identifying restriction. So some identifying constraint presumably has been imposed in order to get any output.
   - DC models do not determine the scale of utility - scaling in DC models depends entirely on an arbitrary normalisation of the error variance. In particular, there is no reason for them to be comparable with the scale of the TTO model.

[1] Note that the sum of logarithms of the TTO likelihood terms defined by our equation (4) does not have the form of a sum weighted by w_i , since the log of the sum over latent classes c is not the sum of log for each class c. Consequently, the specification of the error variance as a function of w_i is not equivalent to weighting the likelihood for differential response.

- This can be seen by comparing the output presented in the response by Devlin et al with the results reported in Table 3 of Feng et al., which shows large differences between DCE and TTO model estimates.
- The Bayesian analysis suffers from non-convergence problems identified in our original report. Therefore, the "Bayesian estimates" presented in the figure are not proper estimates.
- Even if we were to accept the calculations in this new figure, some of the differences are in fact large on the utility scale. The median health gain from all NICE Technology Appraisals is less than 0.3 QALYs, calculated from models which run over long time horizons. Seemingly small differences in the values of health states described by EQ-5D matter.
- To really understand what this figure tells us, we would need to see the computer code that generated it because, for the reasons above, we have no confidence that it supports the conclusions Devlin et al wish to draw.

Summary:

Some of our criticisms may result from choices that were not made by the study team. The responses by Devlin et al give information about the process leading to some of those choices. Our review does not cover such issues. Our task was to assess the scientific quality of the proposed value set for England and there should be no ambiguity about that assessment: both the data quality and the statistical modelling fall short of the standards required for robust policy decisions. Further reviews of our report commissioned by DHSC from (non-conflicted) experts have strongly supported our analysis and its conclusions. Our view is that the work of Devlin et al should be considered part of the development process towards a value set that is fit for use, not the end point of that process. There is nothing in the responses from Devlin et al that changes our recommendations.

Table 1: TTO outcomes for the last 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	12111	0.95	11122	0.95	13224	0.6	42321	0.6	35311	0.8	34232	0.5	52335	0.5	24445	-0.5	43555	-0.7	55555	-0.3
2	11121	1	11414	0.1	25222	0.6	31514	0.5	25331	0.6	21444	-0.8	35143	0.2	53243	-0.7	53244	0.2	55555	-0.95
3	11112	1	21334	0.95	12334	1	23242	0.95	32314	1	53412	1	24342	1	33253	0.95	55225	1	55555	0.95
4	21111	1	11421	1	13313	1	25122	1	12244	0.35	31525	0.8	45233	0.35	55233	0.8	52455	0.8	55555	-0.1
5	21111	1	11421	0.95	13313	1	25122	1	12244	0.25	31525	1	45233	0.5	55233	0.7	52455	0.4	55555	-0.65
6	11112	0.95	23242	0.5	32314	0.3	21334	-0.7	12334	-0.05	53412	1	24342	0.5	33253	0.3	55225	0	55555	-0.9
7	11211	0.9	13122	0.95	42115	0.5	11425	0	51152	-0.6	22434	0	35332	0	45413	0.95	24553	0	55555	0
8	11121	1	21112	0.9	12513	0.8	53221	0.9	12344	0.55	44125	0.9	54342	0.5	14554	0.6	44345	0.6	55555	-0.5
9	21111	0.95	11212	0.95	12112	0.9	23152	0.8	21345	0.8	34244	0.5	43514	0.7	55424	0.8	44553	0.8	55555	0.6
10	12111	1	11221	1	11235	0.5	12514	0	54231	0.5	51451	0.5	34515	0.5	45144	-0.2	35245	0	55555	-0.5
11	21111	1	11212	1	12112	1	23152	0.7	21345	0.2	34244	0.25	43514	0.75	55424	0.6	44553	0.5	55555	0.15
12	21111	0.9	11421	0.9	13313	0.9	25122	0	12244	0.9	31525	0.4	45233	0.5	55233	0	52455	-0.6	55555	-0.2
13	11112	0.7	32314	0.65	23242	0.7	12334	0.7	21334	0.7	24342	0.8	53412	0.7	33253	0.65	55225	0.7	55555	0.65
14	21111	0.95	11421	0.9	13313	0.95	25122	0.6	12244	0.5	31525	0.95	45233	0.5	55233	0.9	52455	0	55555	0.5
15	11121	1	11414	0.8	25222	0.8	25331	1	31514	0.75	21444	0.8	35143	0.4	53243	0.9	53244	0.7	55555	0.5
16	21111	0.95	11212	1	12112	1	23152	1	21345	1	43514	1	34244	1	55424	1	44553	0.1	55555	0
17	11121	0.6	21112	0.6	12513	0.6	53221	0.6	12344	-0.5	44125	0.6	54342	0.5	14554	0	44345	0.6	55555	0.5
18	11112	0.95	32314	0.95	12334	1	23242	1	21334	1	24342	0.95	53412	1	33253	1	55225	1	55555	0.75
19	21111	1	11421	0.8	13313	0.9	25122	0.9	12244	0.5	31525	0.9	45233	1	55233	0.5	52455	0	55555	0.8
20	21111	1	11212	0.5	12112	0.9	23152	0.5	21345	0.6	43514	1	34244	0.6	55424	0	44553	0.5	55555	0.95
21	21111	0.5	11421	0.95	13313	0.8	25122	0.9	12244	0.8	31525	0.95	45233	0.6	55233	0.95	52455	0.8	55555	0.4
22	12111	0.8	11122	0.6	42321	1	13224	0.45	35311	0.8	34232	0.7	52335	0.9	24445	0.8	43555	0.7	55555	-0.55
23	11121	0.95	11414	0.9	25222	0.95	25331	0.9	31514	0.95	21444	0.5	35143	0.9	53243	0.95	53244	0.8	55555	0
24	11112	0.8	14113	0.7	21315	0.6	15151	0.5	52431	0.7	31524	0.7	43315	0.7	24443	0.7	54153	0.7	55555	0
25	21111	0.9	11421	0.95	13313	0.6	25122	0.7	12244	0.4	31525	0.5	45233	0.5	55233	0.5	52455	0.9	55555	-0.5
26	11211	0.8	13122	0.65	42115	0.2	11425	0.4	51152	0.3	22434	0.1	35332	0.4	45413	0.2	24553	0.4	55555	0.1
27	11121	1	21112	0.9	12513	0.9	53221	0.8	12344	0.6	44125	0.6	54342	0.5	14554	0.5	44345	0.4	55555	0.05
28	11211	0.9	13122	0.7	42115	0.4	11425	0.5	51152	1	22434	0.6	35332	0.8	45413	0.6	24553	0.5	55555	0.4
29	12111	1	11122	1	42321	0.5	13224	0.5	35311	0.5	34232	0.7	52335	0.3	24445	0.3	43555	0.3	55555	0.3
30	12111	0.8	11122	0.9	13224	0.05	42321	0.05	35311	0.5	34232	0.6	52335	0.05	24445	0.05	43555	0.05	55555	0

The following list of problematic TTO outcomes illustrates the characteristics of the last 30 respondents in the dataset:

- Individual 1. 2 states ranked worse than 55555. Inconsistent (weak and strong)
- Individual 2. Inconsistencies not related to 55555, weak and strong
- Individual 3. Undiscriminating with 6 states ranked at 1, 4 at 0.95. Many inconsistent values, weak and strong
- Individual 4. Undiscriminating. 4 states at 1, 3 at 0.8. Inconsistent, weak and strong
- Individual 5. 4 states at 1, inconsistencies weak and strong
- Individual 6. Inconsistent (weak and strong)
- Individual 7. 5 states valued at 0, including 55555. One other rated far lower. Inconsistencies, weak and strong.
- Individual 8. Inconsistent (weak and strong). 12344 valued worse than 2 inferior states
- Individual 9. Inconsistent, weak and strong. 4 states at 0.8. 34244 rated worse than 55555
- Individual 10. 4 states at 0.5. Inconsistencies, weak and strong.
- Individual 12. 4 values at 0.9. 52455 valued worse than 55555 and other inconsistencies, weak and strong. All integer responses
- Individual 13. 6 values at 0.7. Inconsistencies weak and strong, not limited to 55555
- Individual 14. Many inconsistencies, weak and strong, not limited to 55555.
- Individual 15. 35143 valued lower than 55555.
- Individual 16. Undiscriminating. 7 values at 1. Inconsistencies, weak and strong not involving 55555
- Individual 17. Undiscriminating. 6 values at 0.6. 2 states far lower than 55555. Inconsistencies weak and strong.
- Individual 18. Undisscriminating. 6 values at 1.
- Individual 19. All integer responses. Many inconsistencies, strong and weak.
- Individual 20. Many inconsistencies, weak and strong, including 55555 valued at 0.95 with other states as low as 0.
- Individual 21. Many inconsistencies, strong and weak, not related to 55555.
- Individual 22. Many inconsistencies, strong and weak, not related to 55555.
- Individual 23. 31514 valued higher than 11414. Other inconsistencies not including 55555.
- Individual 24. Non discriminating. 6 values at 0.7. All integer responses. Inconsistencies, strong and weak, not including 55555.
- Individual 25. Inconsistencies weak and strong, not including 55555.
- Individual 26. 22434 equal to 55555.
- Individual 28. 42115 valued the same as 55555. Other inconsistencies, weak and strong.
- Individual 29. All integer responses. 4 states valued at 0.3, including 55555. Inconsistencies.
- Individual 30. Inconsistent and 5 states valued at 0.05.

Table 2: Percentages (from the dataset after deletions[1] and special treatment[2]) of individual participants displaying potentially problematic response behaviour

Anomalous outcome type	First 30 respondents	Last 30 respondents	1st half (n=302)	2nd half (n=302)
(1) All individual's TTO trials result in same value of T[3]	0.00%	0.00%	0.00%	0.00%
(2) Individual reports at least 1 non-55555 trial with same or lower value of T than trial of 55555	53.33%	46.77%	50.99%	48.68%
(3) Individual reports at least 1 non-55555 trial with strictly lower value of T than trial of 55555	26.67%	30.00%	28.15%	29.14%
(4) Individual reports fewer than 5 distinct values for T	16.67%	23.33%	20.53%	21.19%
(5) Individual reports mild trial (1-point difference from 11111) with same or lower T result as trial of 55555	0.00%	0.00%	0.00%	0.00%
(6) Individual reports values T =0, 10 or 20 in every trial	0.00%	0.00%	0.00%	0.00%
(7) Individual reports all ten trial values T as multiple of 5 years	0.00%	0.00%	0.66%	0.33%
(8) Individual gives only integer values for T	30.00%	23.33%	29.47%	32.78%
(9) 'Seam' outcome of T = 10 in at least two trials with no outcome below 10	0.00%	0.00%	0.00%	0.00%
(10a) Individual with any inconsistencies between the logical ordering of health states and the TTO valuation	93.33%	93.33%	88.74%	88.08%
(10b) Individual with inconsistencies in more than 20% of tasks between the logical ordering of health states and TTO valuation	43.33%	40.00%	42.05%	36.09%
Individual displays any of anomalies (1), (3), (4) or (5)	40.00%	50.00%	43.38%	45.03%
Individual displays any of anomalies (1), (3), (4), (5), (7), (8) or (9)	56.67%	60.00%	60.26%	61.59%
Individual displays any of anomalies (1), (3), (4), (5), (7), (8), (9) or (10b)	63.33%	70.00%	66.56%	68.21%
Individual displays any of anomalies (1), (3), (4), (5), (7), (8), (9) or (10a)	93.33%	93.33%	89.74%	92.38%

[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)
[3] Zero by definition since Devlin et al deleted these cases