Technical Appendix

Response to: Quality review of a proposed EEPRU review of the EQ-5D-5L value set for England

In the text below the EQ-5D-5L value set for England research team responds point-by-point to the issues raised by the EEPRU review. Each set of points is identified by the corresponding section and/or page number in the review document.

Section 1. Background

Section 1.1 Introduction

Page 12. While we agree with the rationale for undertaking an independent review of the EQ-5D-5L value set, the decision about whether to recommend one value set over another should surely rest on an assessment of both value sets using the same criteria.

Section 1.2 An outline of the valuation methodology

Pages 12-13. H-A suggest that the method for selecting health states is problematic because it "is not necessarily aligned with the configuration of health states found in real-life studies". We do not agree that this is a necessary or desirable property for valuation studies, as only including these states may bias estimates for the dimension levels that appear less regularly in practice. Ongoing work by Marten et al. has also found that two-way combinations of EQ-5D-5L dimensions and levels that may appear *prima facie* to be implausible were observed in a sample of the UK general population.[1] The selection of health states to be included in the TTO and the DCE tasks was made via methods and algorithms based on well-established mathematical theories on experimental design. The work by Yang et al.[2] and Bonsel et al.[3,4,5] validates the choice of the algorithms applied and the resulting selection of health states underlying the EQ-VT design. Furthermore, it has been shown that the suggestion by H-A to select the set of health states based on prevalence rather than on experimental design theory results in highly biased value sets. The evidence shows that the statistical properties of the design are more important than coverage in generating accurate estimates.

Page 13. H-A say: *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*

[1] Marten, O., Mulhern, B., Bansback, N. and Tsuchiya A., 2017. *Modelling implausible EQ-5D-5L states: prevalence in the general public and its effects on health state valuation – preliminary results*. Paper presented at the EuroQol Plenary Meeting. Barcelona.
[2] Yang, Z., Luo, N., Bonsel, G., Busschbach, J. and Stolk, E., 2018. Selecting Health States for EQ-5D-3L Valuation Studies: Statistical Considerations Matter. *Value in Health*, 21(4), 456-461.
[3] Bonsel, G.J., Oppe, M. and Janssen, M.F., 2014. *Unexpected large misspecification effects of health profile selection and interaction analysis to obtain a value function from unsaturated valuation datasets, using the standard EuroQol approach*. Paper presented at the EuroQol Plenary Meeting. Stockholm.
[4] Bonsel, G.J., Oppe, M. and Janssen, M.F., 2015. *Unlikely health states: Evidence from healthy and diseased populations*. Proceedings of the 32nd Scientific Plenary Meeting of the EuroQol Group. Krakow.
[5] Bonsel, G.J., Oppe, M. and Janssen, M.F., 2016. *Optimization of the design of multi-attribute vignette studies: A simulation study based on the multinational EQ-5D-5L pilot studies*. Paper presented at the EuroQol Plenary Meeting. Berlin.

led to 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.

We disagree that -1 is an obvious 'realistic' lower bound – while it may appear more defensible than the minimum value of -39 that could be obtained by the protocol used by Dolan (1997), it is entirely plausible that someone could have values < -1, as indeed we found in our experimental work testing various forms of lead and lag time TTOs.[6] The bounding of values at -1 in the EQ-VT design is a limitation that could be important for a subset of respondents with strong opinions about 'worse than dead' health states and is one of the things we sought to address in our modelling.

Page 13. H-A note that at T=0, *no trade-off takes place*, implying a valuation below -1. This suggests a misunderstanding about the TTO procedure used. At T=0, the respondent has traded *all* of the time available to them, including the 10 years of lead time. In the Devlin et al. (2018) paper we refer to such respondents as having "exhausted their lead time". The health state value at T=0 is *equal to or less than* -1.

Page 13. H-A state that "*at T=10, the TTO outcome is exactly at the seam between the primary TTO time frame and the secondary lead-time".* It is worth clarifying that the task for T=10 is repeated twice if a respondent considers the state to be worse than dead. In the better than dead framework, T=10 is presented as the second question, and in the worse than dead (lead time) framework, T=10 is presented as the first question.

Page 13. H-A note that at T=20, the respondent is "*unable to distinguish the specified state from full health"*. Again, this suggests a misunderstanding of how TTO works. Respondents who provide such values are typically perfectly capable of distinguishing the specified state from full health, but they do not consider the state to be undesirable enough to choose a shorter life (effectively giving up time) in order to avoid it. In other studies respondents have reported that they would be willing to give up very short amounts of time (say, a few weeks) but not as much as six months (the minimum time that can be traded in the EQ-VT protocol) in order to avoid relatively mild health states.

Page 13. As well as T=20, H-A also suggest that there should be few outcomes at T=10 and T=0*.* We dispute this. Respondents may consider a variety of severe states to be about the same (in terms of undesirability) as dead, or bad enough to trade all of the time available to them, respectively.

Page 15. H-A state that DCEs *are much less informative than TTO* because they do not give any quantitative information on the margin by which one state is preferred to another. This is only true when referring to a single response or respondent – it is not true when

[6] Devlin, N., Buckingham, K., Shah, K.K., Tsuchiya, A., Tilling, C., Wilkinson, G. and van Hout, B., 2013. A comparison of alternative variants of the lead and lag time TTO. *Health Economics*, 22(5), pp.517-532.

referring to a large sample study with multiple observations from each respondent and for each pair that are then modelled at the aggregate level.

Page 15. H-A refer to "extrapolating" from the small set of health states. The appropriate term here is "interpolating".

Section 2 – Data Quality

We note a general point here that we were obliged to follow the international research protocol provided by the EuroQol Group.

The data quality section more or less suggests that we give inadequate attention to data issues in reporting our work. We reject this. Our modelling choices were determined by, and carefully chosen to reflect, the nature of the data we obtained.

Section 2.1 The EuroQol research protocol

Page 17. H-A refer to the changes made to the EQ-VT protocol since version 1.0. The following references are relevant in this context: Ramos-Goñi et al. (2017)[7] and Stolk et al. (forthcoming)[8].

Page 18. H-A report weak and strong inconsistencies in the data. The figure of 92.2% noted in Table 2.5 is based on a definition of inconsistency which includes any situation where one state is given a value that is *equal to* or greater than the value given to another state that logically dominates it. This definition of an inconsistency involves a number of assumptions that make little sense when analysing and interpreting TTO data. For example, it deems as logically inconsistent a situation where a value of -1 is given to both 55555 and a logically better state such as 44444. But such responses can be entirely consistent with underlying preferences: 44444 may be valued at -1, and 55555 < -1, because as the task is bounded at -1 by design it is not possible to assign 55555 a value lower than -1. This was one of the methodological factors considered in the modelling process. More generally, to define inconsistency as including equal values between logically ordered states imposes a strong value judgement. Respondents giving such data may be entirely aware that the state is 'worse' but regard it not to be sufficiently worse to trade off as much as the minimum time possible, which was set as six months by design. Researchers in this field generally do not judge tied values to be inconsistent. For example, he Korean paper cited by H-A (Kim et al., 2016) does not include ties in the way it defines logical consistency in the Korean data Using the Kim et al. criteria the equivalent proportion in the English data is 56.7%. Whilst high, it is not unprecedentedly so. For example, in the Danish and Spanish TTO studies of EQ-5D-3L values referred to by Kim et al. (2016), the

[7] Ramos-Goñi et al, J.M., Oppe, M., Slaap, B., Busschbach, J.J. and Stolk, E., 2017. Quality control process for EQ-5D-5L valuation studies. *Value in Health*, 20(3), pp.466-473.
[8] Stolk, E., Ludwig, K., Rand, K., van Hout, B.A., and Ramos Goñi, J.M., forthcoming. Overview, update and lessons learned from the international EQ-5D-5L valuation work: version 2 of the EQ-5D-5L valuation protocol. Value in Health.

rates of logical inconsistency were 79% and 59% respectively. Bearing in mind the nature of the TTO task – health states are valued one at a time, without reference to other states or previous answers – and the often subtle differences between states described by the EQ-5D-5L, logical inconsistencies are not unexpected.

Section 2.2 Sample size and sample coverage.

Section 2.2.1 An illustrative example

Pages 19-20. The example aims to illustrate but manages to confuse. The graphs indicate "covariate coverage", which is something completely different than "sample coverage". Indeed, the health states are chosen to cover the covariates, not the samples. The design article by van Hout and Oppe explains in more detail which procedure was followed.[9]

Section 2.2.2 and 2.2.3 Coverage of logically possible and empirically relevant states

A lot of attention is given to coverage and H-A talk about *"the robustness issue linked to coverage*". We dispute whether this is an issue for the EQ-VT design. The design over-samples the extremes, but does not under-sample any part of the scale.

It could be argued that what is important in cost-effectiveness research is the change in values between levels of the descriptive system. We should therefore value a set of states where all levels are represented.

Page 21. H-A's argument that the relevant indicator of coverage (in DCE) is the number of comparisons as a proportion of all possible comparisons assumes that pairs involving dominant/dominated states should be included in the denominator, which we consider to be misleading. Further, H-A fail to recognise that the number of degrees of freedom in the DCE (and indeed the TTO) experimental design was far greater than was needed given the number of parameters that we were seeking to estimate. In Table 2.1 H-A focus on the inclusion of 392 separate health states. In a DCE, the overall number of health states is difficult to compare without information on the difference in severity within the pairs of health states across the 196 choice sets, and therefore the level of information gained from each choice set. It is also worth noting that the inclusion of 196 choice sets to estimate values for EQ-5D-5L is generally at the high end in comparison to similar studies, and the EQ-VT design has an efficiency (compared with the complete factorial) to estimate the main effects model of 79.4%.

Page 21. H-A criticise the TTO experimental design for failing to include health states with 'misery scores' of 23 and 24 (the sum score of all levels, for instance 11113 has a misery score of 1+1+1+1+3 = 7). This is a surprising stance to take, given their earlier points about the importance of coverage (very few states with these misery scores exist compared to states with lower misery scores – e.g. only five health states in the descriptive system have a misery score of 24 – and in practice very few people report being in those states and they are less likely than most to be encountered in either population health surveys or practical cost-effectiveness applications). Further, for many respondents states

[9] Oppe, M. and van Hout, B., 2017. *The "power" of eliciting EQ-5D-5L values: the experimental design of the EQ-VT*. EuroQol Working Paper 17003.

such as 45555 would be near-indistinguishable, in terms of undesirability, from state 55555. So a design that addresses this point would likely lead to a greater number of outcomes that H-A consider to be "anomalous" (see Table 2.5).

Section 2.2.4 Coverage of states important to cost-effectiveness studies

Pages 23-24. To add to their disputed line of argument with respect to coverage, H-A access two economic evaluations and address how many of the sampled states are observed in the cost-effectiveness studies. It is to be expected that trials with many patients with good health have good "coverage". This is simply because our valuation sample includes all five health states with just one dimension at "mild" problems which are included in the sample.

Another example – just as informative as the trials reported here – could have been obtained by taking the data from the GP Patient Survey, which would have indicated that the coverage rate when excluding 11111 is 88%, and when including 11111 is 94.7%. But again, that is because most people are in the better health states.

Page 24. H-A write: *Further research is needed to link the design of valuation methods to actual technology appraisals.* We speculate that the reason for not having seen this research before is that it is irrelevant to the general purpose and application of value sets, which aims to provide a common denominator for the measurement and comparison of quality of life between completely dissimilar patient groups and indeed the population. Taken to the extreme, H-A's points might suggest that each intervention has its own coverage rate and that a different design and different value set is needed for each trial.

Section 2.3 Sampling of participants

Page 25. H-A compare the sample size and response rate of the EQ-5D-5L value set for England study with the sample size and response rate of the Health Survey for England. This comparison is irrelevant, as the two represent entirely different kinds of studies. The Health Survey for England is a large-scale annual survey which aims to monitor and compare the health of different sub-groups of people, e.g. by region, age, deprivation, etc. The value set study has the primary aim of representing the *average* preferences of the general public – not (given its intended applications) to produce different value sets for different types of people. A valuation study seeks to obtain reliable estimates for a model with (say) 20 to 30 parameters and for relatively narrow confidence intervals; a population survey seeks to do something completely different. The choice of sample size was based on simulation studies which addressed the benefit from obtaining more respondents in terms of the width of the intervals versus the costs of including more respondents. The design of the TTO experiment allowed for 400 observations per model parameter.[10]

With respect to response rate, it is also worth pointing out that the EQ-5D-5L value set study involved asking respondents to complete a series of cognitively demanding stated preference tasks involving the consideration of serious ill health and death. Individuals

[10] Oppe, M. and van Hout, B., 2017. *The "power" of eliciting EQ-5D-5L values: the experimental design of the EQ-VT*. EuroQol Working Paper 17003.

approached to participate were provided with information about what the valuation interviews would involve, and it was emphasised at a number of points that their participation was voluntary and that they were free not to participate or to withdraw at any time (indeed, it was a requirement stipulated by the University of Sheffield School of Health and Related Research Ethics Review Procedure that this information was made prominent to potential respondents). It is inevitable that the response rate would not be as high as for a study that did not impose such requirements. We made this clear in our discussions with and reporting to the project's Steering Group.

A fairer comparison regarding the response rate in the EQ-5D-5L value set study is with other TTO studies conducted in the UK, for example by Rowen et al. (2011) which achieved a 40% response rate.[11]

Page 25. The response rate of 47.7% was calculated using response rate 1 defined by the American Association for Public Opinion Research Standard Definitions (2011). Specifically, it refers to the number of interviews achieved divided by the number of addresses issued (excluding addresses known to be ineligible or out of scope, i.e. where it is known that there were no eligible adults at the address, but including addresses that were definitely or potentially eligible or where eligibility was unknown). We acknowledge a typo in the Feng et al. (2018) paper. The sample included 2,220 addresses, not 2,020 addresses.

Page 25. H-A state that a *"decision was made by the designers of the experiment to discard all information about individuals who refused participation or gave partial responses".* In fact, information could not be collected about individuals who could not be contacted, refused to participate or did not provide all required information (in the latter case, by definition). It is not clear how we would have been able to collect such information without following up with dedicated research on non-responders (which we did not receive funding to undertake). We were required to discard information provided by respondents who did not complete the interview in full – this was a requirement stipulated by the University of Sheffield School of Health and Related Research Ethics Review Procedure, and our experience is that other research ethics committees impose similar requirements for comparable health preference studies. We believe that it is the role of research ethics committees, rather than of health economists, to determine whether or not the ethical issues in handling non- and partial response are an insuperable barrier.

Page 26. H-A correctly point out that some of the background characteristics are missing for a small minority of the respondents. This information is missing due to a data upload issue whereby the data pertaining to the main questionnaire (which included a small number of background questions common to all EQ-VT studies) were uploaded successfully but the data pertaining to the supplementary questionnaire (which included further background questions specific to the value set for England study) did not upload successfully.

[11] Rowen, D., Brazier, J., Young, T., Gaugris, S., Craig, B.M., King, M.T. and Velikova, G., 2011. Deriving a preference-based measure for cancer using the EORTC QLQ-C30. *Value in Health*, 14(5), pp.721-731.

Page 26. H-A criticise the weighting strategy for failing to correct for over/underrepresentation of respondent subgroups based on certain background characteristics – characteristics which they later (on page 37) report had very little statistical impact on the valuation data. We also found – in analyses that could not be included in the Devlin et al. (2018) paper due to space constraints – that these background characteristics had very little statistical impact on the valuation data. Further, we wish to point out that the aim of the study was not to recruit a sample that was strictly representative of the general population with respect to all observable background characteristics, but to use a systematic recruitment strategy that offered all eligible individuals in England (regardless of geographic location) an opportunity to be invited to participate in the research. The sample procedures were discussed with and accepted by the project's Steering Group.

Section 2.4 Participants' experience of ill health

Page 28. H-A note that the argument for basing utility scores on views held by the general population is weakened if those views are not well-grounded or based on experience of illness. In various other research currently underway, we and others are exploring these issues, including the effect of experience on valuations. However, it was not an aim of the value set study to do so. NICE's methods guide states that the value set to be used should be based on the preferences of the general public.[12] This was also the clear requirement for the value set study we undertook, and at no point did our Steering Group indicate we should be recruiting a sample of people with experience of ill health, rather than a general public sample.

Section 2.5. Participants' self-assessment of difficulties

Page 29. H-A suggest that DCE tasks are simpler than TTO tasks since the former require only the ability to rank two states in terms of quality of life. This may well be true for some people, but our experience is that many people find TTO tasks easier to complete because they are only asked to evaluate one health state in each task, and have a series of questions about the same state, rather than considering two health states simultaneously.[13] Each method therefore requires different cognitive processes, and elicits preferences from an individual from different perspectives.

Page 31. H-A refer to problematic DCE outcomes, but it is not clear what problematic outcomes they are referring to.

Section 2.6 The TTO experiments

Table 2.5 (page 33). We question many of the "anomalous outcome" types proposed by H-A. In particular, H-A make little attempt to explain why they regard outcomes 2, 4 and 8 are potentially problematic, yet two of these (outcomes 4 and 8) are included in the

[12] NICE, 2013. *Guide to the methods of technology appraisal 2013*. London: NICE.
[13] Mulhern, B., Bansback, N., Brazier, J., Buckingham, K., Cairns, J., Devlin, N., Dolan, P., Hole, A.R., Kavetsos, G., Longworth, L., Rowen, D. and Tsuchiya, A., 2014. Preparatory study for the revaluation of the EQ-5D tariff: Methodology report. *Health Technology Assessment*, 18(12).

highlighted rows at the bottom of the table indicating outcomes that may regarded – in their words – as "serious".

Take, for example, outcome type 8: H-A are suggesting that a respondent's values are 'anomalous' if they only have integer values. Integer values, and digit preference arising from framing effects, are the norm in this field of research, so this comment is not specific to our study or indeed to the methods in the EQ-VT protocol.

Outcome 10a is based on an unusual definition of inconsistencies that judges as 'inconsistent' values which may be entirely valid representations of preferences. It also contrasts the way in which H-A define inconsistencies when referring to other valuation studies. Yet this outcome also appears to be included in the highlighted rows indicating serious outcomes. We also feel that it is misleading for H-A to fail to distinguish between small inconsistencies (which are more likely to reflect imprecision or measurement error) and larger inconsistencies (which are more likely to reflect misunderstanding or difficulty with the tasks).

Pages 34-35. The reviewers report data from 30 respondents. We conducted comparable analyses on the unmodified TTO responses for each and every respondent in the sample, examining visual representations of the valuation data at the individual respondent level (see page 13 of the Devlin et al. (2018) paper). This allowed us to assess patterns in the data and informed our data exclusion choices and subsequent modelling process. We presented these individual respondent analyses in a number of public presentations, and to our Steering Group, in order to seek feedback from others on the quality of our data and our proposed methods of analysis. Note that the 30 respondents selected are from relatively early in the data collection. In any interview-based stated preference study initial data are rarely, if ever, representative of the full sample data, for example, due to interviewer learning effects.

Table 2.8 (page 39). As noted previously, we disagree with H-A's focus on tied values as a problematic outcome. We believe that outcome type 1b is a relevant outcome to investigate in this context, but not outcome type 1a. Further, we are surprised by H-A's implication that a minimum value given to "a state with misery index <15" is a problematic outcome. It is perfectly plausible that some people would consider a state such as 15151 (unable to wash and dress and extreme pain/discomfort; level sum score = 13) to be extremely undesirable, and would prefer not to live at all than to have to experience that health state for 10 years (even if preceded by a period of good health). Here, as elsewhere in their review, we believe that H-A's analyses are being guided by a number of their own (rather strong) value judgements.

Overall, with respect to the TTO data, while H-A make much of what they consider to be problems, and cast doubt on the validity of the results based on these data, we would point out the high degree of concordance between the TTO model results and the DCE results. We will return to this point later, when we respond to the reviewers' section on modelling.

Section 2.7 The DC experiments

Page 43. H-A refer to arbitrary strategies that respondents might adopt, such as always picking alternative A, and assert that the type of problematic response behaviour would go undetected. In fact, we routinely monitored the number of respondents always picking one alternative (A or B) as part of our quality control procedures, and found that this type of response behaviour very rarely occurred (and could represent genuine preferences anyway).

Section 3. Specification and estimation of the valuation model.

Section 3.1 Specification issues.

Table 3.1 (page 47). H-A list a set of "Potential concerns in the specification of the valuation model". We respond to each in turn.

H-A say under the issue labelled *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."* They note *"Potentially serious biases in parameter estimates and valuation predictions"* as the potential consequence of this issue.

The model does NOT assume that responses are accurate. Rounding is expected and errors are expected. The critique would hold if there had been any indication that errors and rounding would have always pointed in an upward or downward direction, and this would not have been taken into account in the error distributions. Neither the researchers nor H-A have found any indication that this has been the case. As such there does seem to be any ground for a suspicion of potential serious biases in parameter estimates and valuation predictions.

H-A say: *"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."*

Nowhere, in any of the articles we have published, has it been suggested that values above 1 are possible. Indeed, the fact that we use the word 'censoring' may imply this and may need some further explanation. Our reasoning is as follows. When imagining the data generating process behind the errors we imagined someone standing at a billiard table and trying to throw a billiard ball from one end to a specific value (close to the right hand corner) at the opposite end. We played with the idea that the ball might bounce after hitting the end and that errors would result after bouncing. And we played with the idea that after the ball would hit the right wall that it would keep on rolling sticking to that wall of the table. Additionally, we modelled that there would be a distribution of opinions which has a natural limit of 1 (or the right hand corner of the billiard table). The problem is the identification of what the distribution is of errors and what the distribution is of opinions.

After many simulations we recognised that in almost all cases the estimates after censoring were only to be improved by having strong priors on the correct parameters (informed by

the correct parameter distributions as we simulated them). We may seek to publish these analyses separately but considered this degree of detail not to be relevant to the two papers already published.

The suggestion that this process has resulted in a systematic over-valuation of the mild health states can easily be checked by comparing the estimates with some summary statistics for health states 11112, 11121, 11211 12111 and 21111.

	Min	25%	Median	Mean	75%	Max	SD	Max
21111	0	0.9	0.95	0.8896	1	1	0.17	1
12111	0	0.8	0.95	0.8666	1	1	0.21	1
11211	0	0.9	0.95	0.8928	1	1	0.18	1
11121	-0.20	0.9	0.95	0.8854	1	1	0.19	1
11112	-0.65	0.8	0.95	0.8533	1	1	0.24	1

Naturally, the model estimates are based not only on the data for these health states but also on those for other health states holding a "2" within the specification of the model. But given these statistics, it remains difficult to conclude that the model estimates systematically over-value the scores.

H-A say: *"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 reviewers appear to have misunderstood what is being modelled. We explored models that were explicitly designed to capture the heteroscedasticity (these were included in the code sent to H-A, but the model results were not included in our final published papers).

It is known that the variance surrounding the lower values is greater than that surrounding the higher values. This was captured – in the heteroskedasticity models – by linking the variance (in a variety of functional forms) to the expectations. This captures the fact that the variance surrounding 'good' health states is smaller than that surrounding 'bad' health states. Alternatively, assuming that the slope varies per respondent, a slope which starts at 1 for 11111 and ends at different points for 55555 obtains a natural explanation for the fact that there is a higher variance in the worse health states. This is what is done in the value set as reported in our *Health Economics* papers. All coefficients, which H-A assume to reflect heteroskedasticity, reflect population weights.

H-A say: *"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."*

As indicated above, the evidence in section 2.6 concerns the question of whether there is a dependence between the occurrence of "problematic" values, not of the values themselves. Given H-A's definition of problematic values such dependence is easily found, for example due to respondents who only give relatively high values. An analysis using the

values themselves indicates a minute albeit significant dependency which is unlikely to have affected the estimates.

H-A are correct in stating that such dependence would overestimate the precision of the parameters. They also mention the quality improvements implemented by the EuroQol Group. Some of those quality improvements, leading to fewer inconsistencies in the TTO responses, have increased the potential that values are correlated and that the precision is overestimated. Without an ordering exercise included in the protocol, one may get more inconsistencies, more measurement error but also fewer worries about dependencies.

H-A say: *"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."*

As indicated above, H-A appear to have confused some of the variables which capture population weights with parameters which capture heteroskedasticity. In the only model that they have studied (the one reported in our papers in *Health Economics)* heteroskedasticity is modelled through the existence of different slopes.

H-A say under the issue labelled *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."* They note *"Bias in parameter estimates"* as the potential consequence of this issue.

H-A may not be aware of the fact that an intercept in a discrete choice model is often included to capture left-right bias. There is nothing mathematically impossible about that. If such tendency, in favour of left to right, exists, then *not* taking it into account leads to a bias in the parameter estimates.

Section 3.2 Bayesian estimation.

Page 48. Including the "health warning" and "Beware: MCMC sampling can be dangerous!" is both pedantic and condescending. We will not comment further on this or the general tone of the review, except to say that we are sure others will find it as surprising as we do.

3.2.1 Specification of prior distributions.

Pages 48-49. H-A express major concerns about the lack of documentation of the choice of prior distributions and the fact that they did not find any documentation of the sensitivity analyses which were performed. We note that they were provided with 82 different BUG model specification files and six R files which process the BUG files in batches.

We further note that such analyses are quite easy to perform: H-A could themselves have changed the priors – a simple change of code in R – to see whether this has any effect. The assumption that such analyses were not performed by us is both speculative and unjustified.

H-A go on to say (page 49):

"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."

We do not understand this critique. In the model that was given to H-A a Wishart distribution was used with priors of 0.3, 0.3 and 0.4 and the probabilities which are estimated are 0.332, 0.388 and 0.281. Choosing a prior with probabilities (0.1, 0.4, 0.5) obtains estimates of 0.332, 0.388 and 0.280. Such results do not suggest that the Wishart prior was especially informative.

Section 3.2.2 Implementation of the simulation estimator.

Pages 49-50. H-A note that: *"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."*

The reviewers seem to have missed that in the programmes, the priors for the slopes include two gamma-distributions with parameters (0.1, 0.1) and one with parameters (1000, 1000). The latter forces this slope to be very close to 1, effectively being a normalisation constant.

Page 50. H-A note that: *"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."*

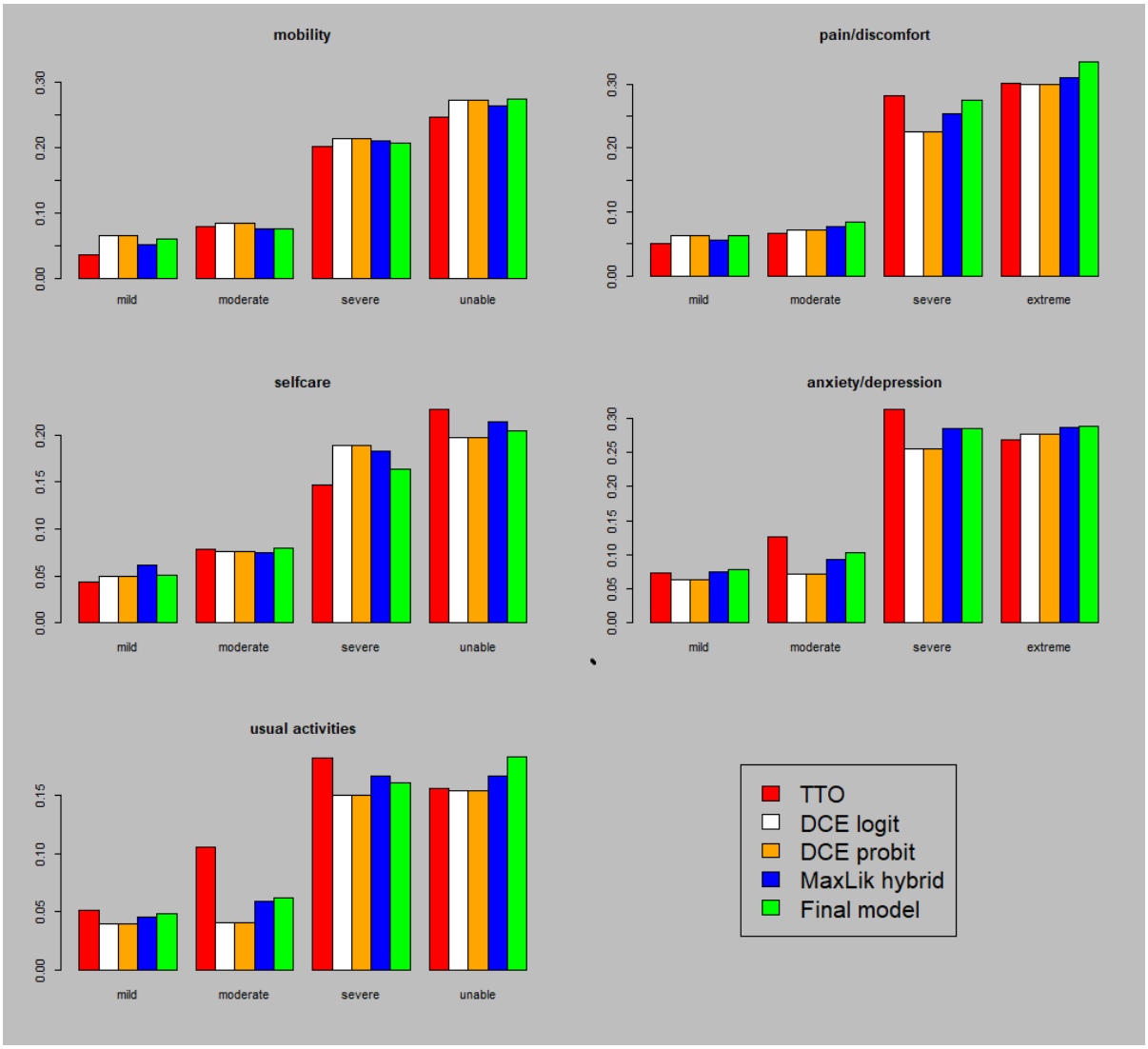
This suggests that H-A have not tried to apply the approach. If they had, they would have found that such an approach appears to be quite informative in forcing the differences between the steps on each dimension to be larger than expected, most notably those expected to be close to zero.

Pages 50-51. H-A continue their critique with an analysis of the convergence in WinBUGS and illustrate the problems with a copy of their output, rather than offering suggestions about how to solve the problem. It all sounds very alarming, but is it?

It must be noted that the results reported in our papers in *Health Economics* are among a variety of other models which have not been reported. For example, models were estimated using maximum likelihood, without taking account of the heterogeneity in the slope but including heteroskedasticity.

To address H-A's criticism it may be useful to look at some of the results in perspective. The following graph illustrates the parameter estimates from five different models. First,

the model based on TTO data only. Second and third are two DCE models, one using a logit specification, one using a probit specification. Fourth, the model with heteroskedasticity and estimated using maximum likelihood (for which there are no warning messages that this would be dangerous). Finally, the parameter estimates of the proposed model as reported in our papers in *Health Economics.*



First, (comparing white with orange) we note that the parameter estimates of the DCE models with logit and probit specifications are identical up to five decimal places, so we question H-A's conclusion that using a logit model instead of probit model may lead to biased estimates.

Second, (comparing red with white) we note that the results from the DCE model are generally in line with the results from the TTO model. The biggest difference is found in the weight for moderate problems in the usual activities dimension.

Third, (comparing blue with green), it is found that the estimates using a maximum likelihood approach are very much in line with those from the 'final' model we report in the *Health Economics* papers. We note that the final model is a mixture of three groups with different attitudes toward the relative values of length of life and quality of life. It should be noted that the research proposal for the EQ-5D-5L value set for England study

(as approved by the NIHR Policy Research Programme) specifically stated that the research would address the observed heterogeneity of the population in this respect. The aim was to show that the average value (which was expected to include values below zero) is indeed a weighted average of opinions, including people who always prefer length of life to quality of life and do not think that there is any state worse than dead.

So, convergence issues should be judged in view of the plausibility of the results. The credible intervals resulting from the Bayesian model are again very much in line with those from the maximum likelihood approach. The maximum likelihood estimates, in combination with common sense, functioned as safeguards against the dangers H-A are concerned about with respect to MCMC and guided us in judging whether there were genuine problems with the estimations.

On a more general point, it is worth emphasising that our study team spent considerable time investigating a wide range of alternative models and their properties. It was our intention to report a number of these alternative models in our principal manuscript from the project – but this suggestion was very firmly rejected by our Steering Group, who recommended we publish one 'final' model only, in order to avoid uncertainty and gaming by users.

Table 3.2 (page 53). H-A go on to list a number of potential issues in the Bayesian analysis.

H-A say under the issue labelled *Choice of priors*: *"Priors on key parameters are informative. There is no justification for the priors used or sensitivity analysis presented."* They note "*Results dependent on priors which may be unreliable"* as the potential consequence of this issue.

Earlier text from H-A seems to suggest that this most notably concerns the prior about the number of groups. A simple analysis with substantially different priors would have shown H-A that this is incorrect. All other priors within the range within which the estimates are expected to be found were highly uninformative.

H-A say under the issue labelled *Unidentified model*": *"A model with proportionality constants for all latent groups is theoretically unidentified."* In the potential consequences column they note: *"No direct implications for prediction of utility values, but inflated parameter uncertainty and problems of convergence may distort results."*

As indicated above, the reviewers seem to have missed that one of the slope parameters was forced to be very close to one and as such we disagree that the model is not identified.

H-A say under the issue labelled *Parameterization of the model*: *"Specification of some parameters may cause problems for the algorithm."* They note "*Slow mixing and convergence failure, leading to unreliable estimates"* as the potential consequence of this issue.

While slow convergence may indeed cause problems and unreliable estimates, we would assess this problem in the view of alternative results such as those produced by using simple maximum likelihood (see figure above), where such problems are almost non-

existent. No significant differences are found, showing robustness of the results and no indication that the estimates are unreliable.

H-A say under the issue labelled *Label switching*: "*The labelling of the unobserved categories changes when sampling from the mixture posterior distribution.*" They note "*The posterior marginal densities estimated from the samples may be poorly estimated"* as the potential consequence of this issue.

The results indicate that there are three groups of respondents with different opinions about the trade-off between length of life and quality of life, the value of 55555 being an indication of this. The value of 55555 is estimated at 0.41 by 33.2% of the respondents, at -0.24 by 39% of the respondents, and at -1.15 by 28% of the respondents. Visual inspection of the values given by all respondents shows this interpretation to have face validity. Moreover, given that consistent estimates are obtained with different priors concerning the mixing distribution, there is no indication that the posterior marginal densities have been poorly estimated.

H-A say under the issue labelled *Single vector of initial values*: "*The MCMC sampler could get trapped in a spurious mode.*" They note "*Inference regarding parameters of interest may not be reliable"* as the potential consequence of this issue.

Multiple vectors were also tested. This did not lead to any changes in the results.

H-A say under the issue labelled *Convergence failure*: "*Insufficient number of iterations to ensure convergence to the stationary distribution, possibly as a result of inappropriate model specification or parametrisation.*" They note "*Inference regarding parameters of interest may not be reliable"* as the potential consequence of this issue.

Many of the models have been estimated with more iterations, more initial values and more chains, and they have not resulted in changes to the parameter estimates. Stationary distributions were obtained in the parameters for which the maximum likelihood approach did not offer a backbone. Other estimates were similar to those obtained using maximum likelihood also with respect to confidence/credible intervals.

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

H-A provide a lecture in econometrics with respect to limited and censored observations. Within this, they suggest the we think that people may have given values above 1.

As indicated earlier, the background behind this was that a correction is needed for the asymmetry in the potential to make errors. It was assumed that people have a vague notion of the value of a health state and that they need to give an answer. Again, our analogy is that they are standing at one end of a billiard table and aiming to throw the ball to the other end to indicate the value of the health state. They aim, but they are imprecise. The likelihood of the observations, when analysing what they aimed at, is that of censored observations. This data generating process provides the background of the approach that was chosen.

Appendix 2. Technical aspects of the specification of the valuation model.

Appendix 2.1 The algebraic form the valuation model.

H-A write:

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.

H-A seem to have misinterpreted the text and equations in the Feng et al. (2018) paper. The inclusion of the heterogeneity (in the slope of the curve) captures the heteroskedasticity. The variable *w* is to correct for the age distribution.

What follows in the next two paragraphs builds further on H-A's misinterpretation of how the heteroskedasticity is captured: that is, by assuming three groups with different slopes, which gives a logical explanation of the increased variance with decreasing health. In doing so, the derivation of the TTO and DCE models are perfectly in line with each other.

H-A say: *There is no mathematical basis for the intercept α_1.*

Indeed, this may apply in a perfect world with a perfect model and perfect respondents. The rationale for including this constant term is that a DCE value function leads to exactly the same relative values as TTO if the coefficients are connected by a linear transformation. That is a linear transformation including a constant term which has been included here.

What follows in the review confirms the H-A's misinterpretation of what constitutes heteroskedasticity and population weights, and should be disregarded.

Appendix 2.2 Bayesian priors for the valuation model.

H-A have not found any documentation of sensitivity analyses and raise the suggestion that it has not been carried out. We can confirm that such analyses were indeed carried out. We invite H-A to carry out these analyses themselves to confirm that the results are robust to changing any of the priors within the margins of logical uncertainty. It would have been particularly useful if they had done this with respect to the prior of the group distribution, since this is the subject of much attention in their review.

Here, as elsewhere, we find it surprising that so much of the review focuses on the listing of *potential* problems that 'might' exist, without having used the data we shared to investigate whether these problems actually exist in practice. We assumed that this was the purpose of the external validation of the model that H-A had been commissioned to undertake.