ARTICLE IN PRESS

ScienceDirect

Contents lists available at **sciencedirect.com**
Journal homepage: **www.elsevier.com/locate/jval**

The EQ-5D-5L Value Set for England: Response to the "Quality Assurance"

Ben van Hout, PhD,[1,2] Brendan Mulhern, MRes,[1,3] Yan Feng, PhD,[4] Koonal Shah, PhD,[1,5] Nancy Devlin, PhD[1,6,*]

[1]School of Health and Related Research, University of Sheffield, Sheffield, England, UK; [2]Pharmerit International, York, England, UK; [3]Center for Health Economics Research and Evaluation (CHERE), University of Technology Sydney, Sydney, Australia; [4]Centre for Primary Care and Public Health, Queen Mary University London, London, England, UK; [5]Office of Health Economics, London, England, UK; [6]Melbourne School of Population and Global Health, University of Melbourne, Melbourne, Australia.

ABSTRACT

Objectives: To respond to the 'quality assurance' of the EQ-5D-5L value set for England study.

Methods: We provide a point-by-point response to the issues raised by the authors of the quality assurance paper, drawing on theoretical arguments, empirical analyses and practical considerations.

Results: We provide evidence to show that many of the points made by the authors of the quality assurance are misleading, suggest misunderstandings, or are irrelevant.

Conclusions: The modelling approaches which were used appropriately address the characteristics of the data and provide a reasonable representation of the average stated preferences of general public in England. We provide reflections on the conduct of stated preference studies, and suggestions for the way forward.

Keywords: discrete choice experiment, EQ-5D, quality assurance, time trade-off, valuation, value set.

VALUE HEALTH. 2020; ■(■):■–■

Introduction

The EQ-5D-5L value set for England[1,2] has potentially important implications for healthcare decisions that are informed by EQ-5D-5L data. Among those are recommendations about the reimbursement of new technologies made by the National Institute for Health and Care Excellence (NICE). It is therefore entirely appropriate that it be subjected to external review before being recommended for use.

In 2017 the UK Department of Health and Social Care commissioned a "quality assurance" from the Economic Evaluation Policy Research Unit (EEPRU), which is summarized in an article[3] in this issue of *Value in Health*. The authors of the article (hereafter, "EEPRU authors") concluded that the EQ-5D-5L value set estimates "fall short of the required standards for decision making," citing deficiencies in both the quality and the subsequent modeling of the data. The points made by the authors principally focus on the English value set but have implications for all EQ-5D-5L value sets developed using the EuroQol Group's international protocol.[4]

NICE's current advice[5] to those who have collected EQ-5D-5L data is not to use the value set reported in our articles in *Health Economics*[1,2] but rather to map between the EQ-5D-5L and EQ-5D-3L (using the crosswalk developed by van Hout et al[6]) and to continue to use the "Measurement and Valuation of Health" (MVH) UK value set for the EQ-5D-3L.[7,8] In effect, this advice involves mapping to an inferior and less sensitive descriptive system[9,10] and applying a value set that is more than 20 years old and has characteristics that have not been replicated elsewhere (a point we will return to later).

In this article, we provide evidence to show that many of the points made by the EEPRU authors are misleading, suggest misunderstandings, or are irrelevant. We summarize the key points the authors made and respond briefly to each. We provide results from additional analyses to support our assertion that the modeling approaches were used appropriately to address the characteristics of the data and provide a reasonable representation of the average stated preferences of the general public in England. We conclude with some reflections on the conduct of stated preference studies and provide suggestions for the way forward.

Conflict of interest: All authors are members of the EuroQol Group. The views expressed do not necessarily reflect those of the EuroQol Group.

* Address correspondence to: Nancy Devlin, PhD, Centre for Health Policy, Melbourne School of Population and Global Health, Level 4, 207 Bouverie Street, the University of Melbourne, Victoria 3010 Australia. Email: nancy.devlin@unimelb.edu.au

https://doi.org/10.1016/j.jval.2019.10.013

Data

The Design and Coverage

The design is inadequate because it provides inadequate coverage of the states in the EQ-5D-5L descriptive system. The design followed the international protocol (EQ-VT) developed by the EuroQol Group, which has been used in many studies to date.[11–14] Value sets generated from this protocol have been accepted for use by other national healthcare decision makers, such as in The Netherlands.[15] As outlined by Oppe and van Hout,[16] a blocked design was used to achieve a mix of states with respect to severity level representation. The design comprised 10 blocks of 10 health states. Each block included the worst health state in the descriptive system (55555) and 1 of 5 of the least severe states. This left 8 health states per block (80 states in total) to be generated. These 80 health states were selected using Monte Carlo simulation, demanding orthogonality. The set with the best results in terms of level balance and predictive power was chosen to allow the estimation of all severity levels from across the five dimensions. The discrete choice experiment (DCE) design is similarly based on optimal design procedures.[16]

The fact that the final time trade-off (TTO) design only included 2.75% of all possible 5L health states and the final DCE design included 0.01% of all potential pairwise comparisons has very limited relevance given the purpose of the study, which was to produce values for all 3125 states. This required experimental designs with appropriate statistical characteristics based on well-established mathematical theories rather than percentage coverage of all possible combinations.

Other valuation studies, for different descriptive systems, derived using various approaches to optimal designs, have similar characteristics. For example, the value set for the SF-6D was based on values generated using standard gamble for 249 states, accounting for 1.38% of the 18 000 states described by the SF-6D.[17] In the valuation of the cancer-specific EORTC-8D using TTO, 85 states were included, which was 0.1% of all possible states described.[18]

The Sample Size and Response Rate

There is no basis for the sample size and the response rate was low, leading to potential bias.

The target sample size for this study, n = 1000, which has been used in many EQ-5D-5L value set studies internationally, follows a study design reported in detail by Oppe and van Hout.[16]

Interviews were conducted face-to-face in respondents' homes. Potential respondents were fully informed about the nature of the questions and the length of the interview before consenting to participate. To achieve a response rate of nearly 50% of those identified is, we would argue, reasonable, given that each respondent was being asked to allow a stranger into their home to query them for around 45 minutes on questions about severe illness and death. Most social surveys—which the EEPRU authors use as a comparison—are very different in nature and do not include such questions. A more appropriate comparison would be with other health valuation studies. For example, Rowen et al report a response rate of 40.3%.[18] The current MVH value set is based on interviews with 3395 individuals out of 6080 (55.8%) addresses selected for sampling.[19] Information on the individuals who did not complete the interview in full was discarded as required by the research ethics committee that approved the study, a common procedure in interview-based studies and not a flaw of ours.

Data Anomalies and Flaws

The long, damning list of potentially problematic responses include a number of strong judgments from the EEPRU authors as to what they think constitutes acceptable data. We would emphasize that anyone who has been confronted with health state valuation questions will appreciate that answering them is difficult. Errors and rounding (to whole number values) are to be expected. Moreover, there are only 41 possible values in the TTO tasks in the EQ-VT protocol (meaning that the greatest precision that is possible is at 6 monthly intervals), and if a given respondent values the worst health state at, say, 0.8 (thereby expressing a strong preference for length of life compared to quality of life, which is a legitimate response), there are only 5 values greater than or equal to 0.8 left to choose from. Respondents who do so are deemed by the EEPRU authors to be "problematic" and described as producing data "flaws" and "anomalies."

The EEPRU authors judge data as "problematic" when respondents only give integer values. This kind of "lack of precision" is quite common to stated preference exercises; it is not exclusive to TTO valuation or to EQ-5D or to our study. It is important to note that unless that lack of precision biases upwards or downwards the values that are elicited, this type of rounding has virtually no impact on the *average* values calculated or modeled for the health states, which was the purpose of the study.

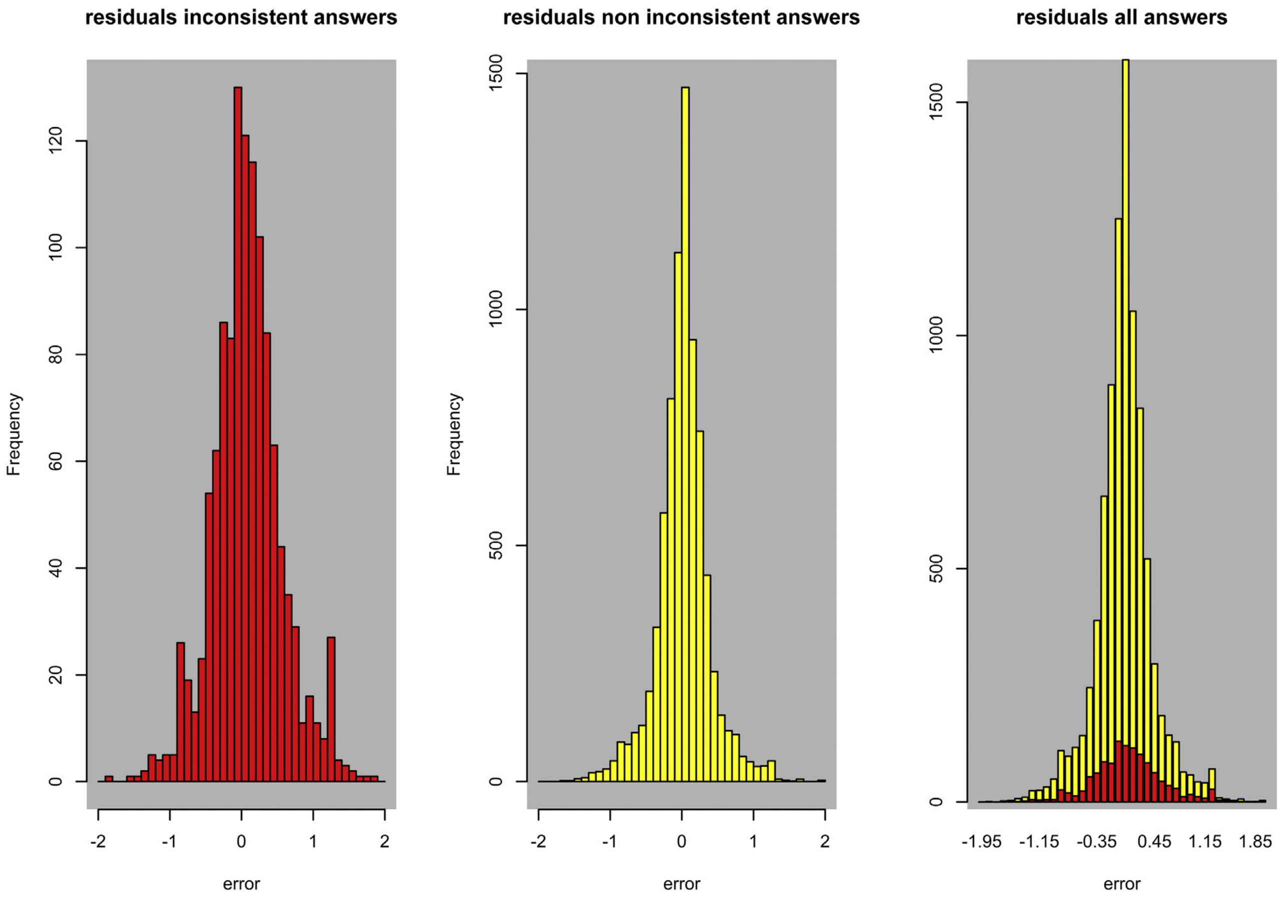
The EEPRU authors also judge data as being "anomalies" in cases where there is a logical ordering of health states, but the same TTO value is given to both; however, such responses may indicate entirely plausible and logical respondent preferences. For example, a state may be logically worse, such as how "no problems with anything other than mild problems with mobility" is logically worse than full health, but not worse *enough* for the respondent to be willing to sacrifice any length of life to avoid it. Consider another example: the EEPRU authors' definition of logical inconsistency deems as problematic a situation where a value of −1 is given to both 55555 and a logically better state, such as 44444. Yet −1 is a plausible value for 44444 and, because the task is bounded at −1, it is not possible to assign 55555 a value lower than −1. These methodological factors were considered in the modeling process.

The key point here is that many of the responses the EEPRU authors deem to be problematic—and upon which they base their "gee whiz" figure of 94% of the data being flawed—may actually represent people's preferences.

In the abstract the EEPRU authors note that "47% of respondents valued more than 20% of states inconsistently, double the 3L rate" and that there is "strong evidence, both direct (self-reported) and indirect (poor data quality), that many participants found tasks difficult or did not engage effectively."[3] In responding to this, it is worth noting that the percentage of responses in which respondents *can* be inconsistent differs between the studies slightly: 47% in the 5L study; 45% in the 3L study. It is also worth noting that respondents in the MVH study valued 12 health states using TTO, whereas respondents in the 5L study valued 10 health states. Importantly, the EEPRU authors have not compared their inconsistency findings to the *entire* MVH TTO data set—just to the MVH data that were used in modeling once exclusions had been made. We have done this, and the results for all respondents in both studies (N = 3395 for the 3L vs N = 996 for the 5L) are reported in the supplementary appendix. In the MVH study, 75.2% of respondents had at least one inconsistency in their responses, compared with 56.7% of respondents in the EQ-5D-5L value set for England study.

To consider the impact of the problems that are present in the data and represent "errors"—such as logical inconsistencies (as conventionally defined, *not* as per the EEPRU authors' definition)—it is helpful to examine the distribution of residuals. These residuals are the differences between predicted values and observed values. In Figure 1, we distinguish between the distribution of

Figure 1. Residual distributions for inconsistent responses, consistent responses, and all responses. Distributions of residuals are near-symmetric surrounding the value of zero.



errors of the logically inconsistent responses, the non-inconsistent responses, and the sum of those; and find a near-symmetric distribution of residuals surrounding zero.

Sequential Dependency

The EEPRU authors question the sequential dependency between responses from the same individual and therefore the estimations based on TTO data. For this, they analyze whether "anomalies" are sequentially dependent. This is not the same as analyzing whether *responses* are dependent. To address the latter, we plot the 9 pairs of sequential residuals from the 10 responses for each respondent. Figure 2 includes results from all 912 respondents with each circle representing a pair of residuals from TTO tasks i and i+1 for a respondent. The 8208 (= 912 × 9) circles are scattered through the figure with a rather flat fitted line. These results do not support the EEPRU authors' argument regarding strong sequential dependency.

Modeling

Inconsistency in the Distributional Assumptions Applied to TTO and DCE Data

The EEPRU authors note that utility error terms are assumed heteroscedastic and normally distributed in the TTO experiments but homoscedastic and type 1 extreme value in the DCEs, and state that this leads to mis-specification and inconsistent parameter estimates.[3] First, on a theoretical level, TTO and DCE are different tasks and therefore assuming a type I distribution (with wider tails) for the one task and a normal error for the other is not inconsistent. Why this would lead to inconsistent parameter

Figure 2. Sequential residuals from 912 respondents. Each of the 8,208 circle represents a pair of residuals from TTO task i and i+1 for a respondent. Fitted line is rather flat, which does not support the EEPRU authors' argument regarding strong sequential dependency.

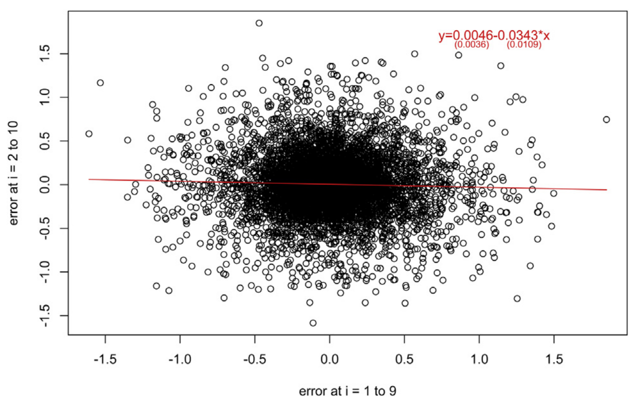
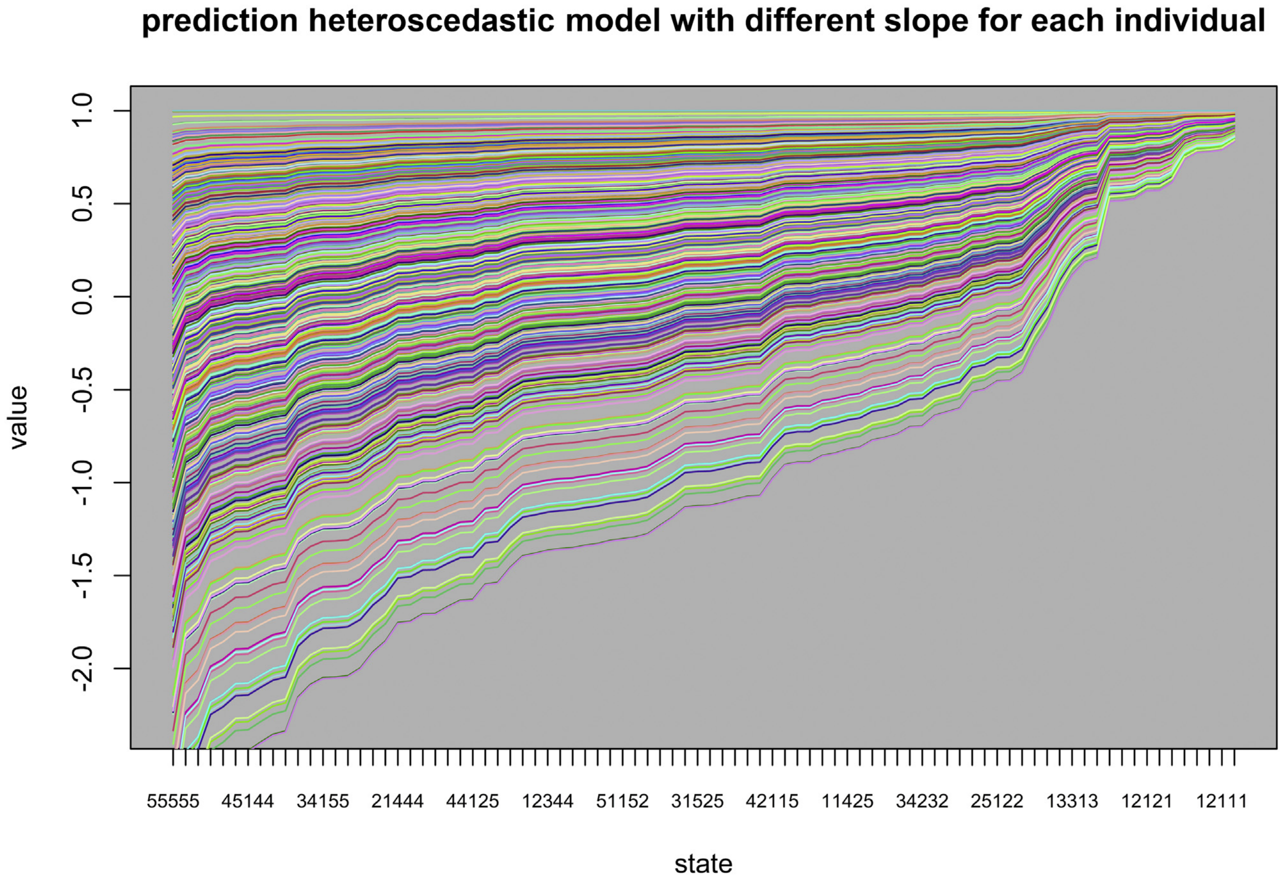


Figure 3. Predicted values for each individual. Each line in the figure shows the predicted values from each respondent based on the heteroscedastic model with different slope. The results show that the variance increases with the increased level of severity for TTO health states.



estimates is not explained. Second, on a more practical level, the parameter estimates using either a probit or logit model are identical up to 3 digits, indicating that the assumption used does not make any meaningful difference. We chose the assumption of type I extreme value distribution because of the slightly lower deviance information criterion statistics reported for the model.

It is observed that the variance surrounding the lower TTO values is greater than that surrounding the higher TTO values. This feature of the data is captured, in the heteroscedasticity models for TTO data, by linking the variance to the expectations. For the DCE data, the task is to rank order and choose between 2 health states. There is no clear motivation, based on the observed binary data, that one should also apply this relationship. Nevertheless, we experimented with this using a probit specification in the DCE models, assuming an identical relationship between the expected value and the variance in the DCE data as in the TTO data and found that this had a negligible effect on the parameter estimates.

Furthermore, it seems the EEPRU authors have misunderstood our final hybrid model where the increasing variance with decreasing health is captured by heterogeneity rather than heteroscedasticity. We discuss this point in the following section.

Confusing Weighting for Non-Response with Weighting for Heteroscedasticity

We tried to capture the fact that there is increasing variance with decreasing values in 2 distinct ways. First, we estimated models where the variance is a function of the expected value of the health state ("heteroscedasticity models"). Second, we assumed heterogeneity in the use of the scale as modeled by a shape parameter ("heterogeneity models"). Three different models for the heterogeneity of the shape parameter were used: a normal, log normal, and multinomial distribution. Subsequently, we also used a gamma distribution. Figure 3 illustrates this estimating a different shape parameter for each individual. Here, respondents are estimated to have the same relative weights for the different dimensions and the levels on those dimensions but differ with respect to the gradient towards dead and negative health states. Assuming heterogeneity offers a very natural explanation for why the variance around worse health states is greater.

We would emphasize that the final model is a model with heterogeneity, not with heteroscedasticity. As such, the EEPRU authors' argument that we have mixed weighting with heteroscedasticity is inaccurate. Comparing the heterogeneity models with and without weighting allows for an understanding of how the weighting for age is introduced, and these are reported in Appendix 1 in Supplemental Materials (found at https://doi.org/10.1016/j.jval.2019.10.013).

Forcing Utility Decrements in the Ordering

The EEPRU authors note the potential drawbacks of using a squared term to guarantee choice inconsistency. We tested many priors and found that the choice of the prior distribution can make quite a difference. For example, when using a gamma distribution, decrements that were expected to be close to zero (on the basis of

Table 1. Observed and predicted values for the 5 mildest health states.

Health States	Minimum	25% percentile	Median	Mean	75% percentile	Maximum	Predicted value
21111	0	0.9	0.95	0.8896	1	1	0.942
12111	0	0.8	0.95	0.8666	1	1	0.950
11211	0	0.9	0.95	0.8928	1	1	0.950
11121	–0.2	0.9	0.95	0.8854	1	1	0.937
11112	–0.65	0.8	0.95	0.8533	1	1	0.922

the unrestricted models) differed substantially from zero; a uniform distribution and using quadratic terms with normal distributions showed similar results. We chose the latter and this indeed results in jumps from positive to negative values in the MCMC chains. Nevertheless, this does not give any problems in the MCMC results for the relevant parameter values.

Censoring for TTO Data

The EEPRU authors state that "the interpretation of the limit at 1 as censoring is inappropriate. Censoring means that values exceeding 1 are possible but unobserved."[3] We do not say or imply that values greater than 1 are possible. Rather, the point is that whereas for values along the continuum between 1 and –1 there is an error distribution around the values, at the top end of the scale, the error distribution is necessarily asymmetric, biasing central estimates of the utility of very mild states downwards. Censoring methods are used to explicitly address this asymmetry in the errors. The rationale for doing so neither implies nor requires the possibility of values > 1. Note also that we did not censor at 1 and –1 but at 0.975 and -0.975. The EEPRU authors also suggest that this censoring approach would bias estimates for the better health states upwards. Table 1 presents the observed and predicted values for the five mildest health states. There is no support for the EEPRU authors' claim that the estimates are biased.

Bayesian Modelling and Convergence Failure

The EEPRU authors provide a list of potential issues regarding the Bayesian analysis process, ranging from the choice of priors to the selection of initial values to the implementation of the simulation estimator. The extent to which we could report *all* of our analyses and results in peer-reviewed journal articles was limited. We conducted sensitivity analyses to check the effect of substantively different priors, initial values, multiple chains in the MCMC, and frequentist approaches.

The EEPRU authors state that they "found no justification for the choice of priors, nor any evidence of sensitivity analysis."[3] In particular, their concern is with the prior distribution relating to the probabilities of latent class membership. Our priors for the probabilities of being members of each of the 3 latent groups are 0.3, 0.3, and 0.4, following a Dirichlet distribution. There is both a theoretical and a practical note of relevance here. The theoretical note is that the Dirichlet distribution is the natural conjugate of a Dirichlet prior and a multinomial distribution. So suppose that individuals can be in 3 groups and the prior follows a Dirichlet (α_1, α_2, α_3), and we find N_1, N_2, and N_3 people observed in the 3 groups, then the posterior distribution is again Dirichlet with $\alpha_1 + N_1$, $\alpha_2 + N_2$, $\alpha_3 + N_3$. Now, this is not completely applicable here, where the group membership has to be estimated per individual, but it does suggest, where N = 912, that priors below 1 are unlikely to be informative. This is confirmed when exploring different priors in our extensive sensitivity analyses. As might be expected, the estimated membership parameters are very robust and the posterior distribution for the parameters is highly driven by the data and not by the prior distribution. We therefore question the claim that these priors are highly informative.

The EEPRU authors assert that they "found significant evidence of lack of convergence, which is indicative of fundamental shortcomings of the Devlin et al (2018) model specification."[3] In selecting the best modeling method, we explored both Bayesian and maximum likelihood approaches. Within the maximum likelihood approach, we also included an iterative approach in which we first estimated the maximum likelihood estimates without heterogeneity (the standard parameters), then estimated a shape parameter for each individual given the maximum likelihood parameters. We then, given the shape parameters, re-estimated the standard parameters and then again re-estimated the shape parameters. This was repeated until no improvement was achieved. The estimates obtained in this way are strikingly similar to the ones obtained using the multinomial model, the log-normal model, and the (unpublished) gamma model for heterogeneity. We are more than happy to share the Convergence Diagnosis and Output Analysis (CODA) results from our Bayesian analyses—accompanied with a R-program—in the public domain for judgment about whether convergence was achieved.

We shared 82 different BUG specification models and six R files with the EEPRU authors as part of the quality assurance review process, excluding models that were discarded and the many sensitivity analyses. Run time of the most complex models with 2000 burn-in iterations and 5000 subsequent iterations is between 30 and 90 minutes, and when running so many models, one obtains an understanding of when a model converges. The EEPRU authors only address the multinomial model and seem to have missed that alternative specifications such as the log-normal model (or the gamma-distribution for the shape parameter) lead to very similar results without the admittedly temperamental—but quite logical—results when identifying which group people belong to. This might have tempered their claim that the model is unidentified. Indeed, the multinomial model allows for 3 groups and the ultimate (TTO) model is p1*(1-a1*x'b)+p2*(1-a2*x'b)+p3*(1-a3*x'b).

Although a1 is forced to be always close to 1, different combinations b, a1, a2, and a3 can be obtained, all leading to the same predictions, so we question whether the lack of model identification is as much of a problem as is suggested. Also we do not find any convergence problems in the parameters of the link function where the constant term, following Bansback et al,[20] is reflecting left-right bias.

What This Means for the England EQ-5D-5L Value Set

The pertinent question is whether our model is a fair reflection of the average values for the English general public with respect to the EQ-5D-5L. We believe that it is for the following reasons.

First, answering questions about life and death and different dimensions of health is difficult. People make errors, may be inconsistent, and may show weak cognitive or empathic ability to carry out TTO tasks. The question is whether the answers of such individuals should be included and whether doing so would lead to biased results. The results in Figure 1 show relatively symmetric error distributions in both consistent and inconsistent answers, giving *no* indication of any structural biases. Although the TTO data may be "messier" than we would like, that has not affected the central estimates of the values, the production of which is the purpose of the exercise.

Second, recall that the data quality problems the EEPRU authors allege relate only to the TTO data. Nevertheless, there is a striking similarity in the findings from the TTO models and the DCE models, as shown in the supplementary appendix. Both methods point toward similar weights for the dimensions and similar values for the levels within the dimensions. Our final model, the one criticized by the EEPRU authors, is just one of many that we tested. We ran hybrid models with maximum likelihood methods (with heteroscedasticity and with heterogeneity) and used Bayesian methods with the heterogeneity modeled using a variety of distributions. The results of the various models were in line with each other.

Third, the data show distributions that are broadly similar to those from comparable studies undertaken in other countries. For example, Abel Olsen et al report "striking similarities" between the England value set and those of Canada, The Netherlands, and Spain with respect to the relative importance of the 5 dimensions, the relative utility decrements across the 5 levels, and the scale length.[21] We are confident that our statistical approach captures the error distributions in such a way that the mean estimates are a reliable representation of the average values of the general public.

Future-Proofing NICE for Transitional Issues in Switching Value Sets

The catalyst to the EEPRU review of the value set appears to have been the observation that implementing the EQ-5D-5L value set would lead to different estimates of quality-adjusted life-year gains and incremental cost-effectiveness ratios compared with the widely used MVH value set. Therefore there were concerns about consistency in health technology assessment (HTA) decisions.

The EEPRU authors provide 3 references to their own work showing that economic evaluations undertaken using 5L rather than 3L are likely to generate very different results. Their research confirms that new technologies that improve quality of life appear less cost-effective if health gains are valued using 5L rather than 3L, whereas technologies that extend life can appear more cost-effective.

The fact is that such a difference was predictable before our study: although the MVH value set has been recommended for use by NICE for more than 20 years, the unusual nature of this value set, especially the high proportion of negative values and the wide utility range, are widely known and reported. Studies undertaken in the United Kingdom using the same protocol in the years immediately after its publication were unable to replicate its properties.[22] A new value set, whether for the EQ-5D-3L or EQ-5D-5L, is unlikely to recreate the characteristics of the MVH value set. Ironically, the EEPRU authors' concerns about the England value set implies that the MVH value set will continue to be used for the foreseeable future, yet it was never subjected to formal quality assurance.

In the light of the EEPRU review, NICE has recommended against using the current EQ-5D-5L value set.[23] EQ-5D-5L data are instead recommended to be mapped to the EQ-5D-3L value set.[24] A new UK-wide EQ-5D-5L value set is now being commissioned. We would urge NICE and other HTA bodies to consider the switch away from old value sets as both inevitable and something that should not be regarded as a "one off." No value set should remain in use for the lengthy period that the MVH value set has. Preferences change, the composition of the general public changes, and methods develop and improve. In the future, online data collection will facilitate more frequent updating of value sets, and we encourage researchers, both in the United Kingdom and elsewhere, to continue to test and develop methods in this area to facilitate comparisons across value sets developed in different eras and using different protocols. We hope as much attention is paid to addressing the transitional challenge for HTA as is being paid to the value sets themselves.

Acknowledgments

The authors received no funding to write this response.

Supplemental Material

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.jval.2019.10.013.

REFERENCES

1. Devlin N, Shah K, Feng Y, Mulhern BJ, van Hout B. Valuing health-related quality of life: an EQ-5D-5L value set for England. *Health Econ.* 2018;27(1):7–22.
2. Feng Y, Devlin N, Shah K, Mulhern BJ, van Hout B. New methods for modelling EQ-5D-5L value sets: an application to English data. *Health Econ.* 2017;27(1):23–38.
3. Hernandez-Alava M, Pudney S, Wailoo A. The EQ-5D-5L value set for England: findings of a quality assurance program. *Value Health.* 2020;23(5):xxx–xxx.
4. Oppe M, Devlin NJ, van Hout B, Krabbe P, de Charro F. A programme of methodological research to arrive at the new international EQ-5D-5L valuation protocol. *Value Health.* 2014;17(4):445–453.
5. National Institute for Health and Care Excellence. Position statement on the use of the EQ-5D-5L valuation set for England. www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/technology-appraisal-guidance/eq-5d-5l. Accessed July 9, 2019.
6. van Hout B, Janssen MF, Feng YS, et al. Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. *Value Health.* 2012;15(5):708–715.
7. MVH Group. *Final report on the modelling of valuation tariffs.* York: Centre for Health Economics; 1995.
8. Dolan P. Modeling valuations for EuroQol health states. *Medl Care.* 1997;35(11):1095–1108.
9. Janssen MF, Bonsel GJ, Luo N. Is EQ-5D-5L better than EQ-5D-3L? A head-to-head comparison of descriptive systems and value sets from seven countries. *Pharmacoecon.* 2018;36(6):675–697.
10. Buchholz I, Janssen MF, Kohlmann T, Feng YS. A systematic review of studies comparing the measurement properties of the three-level and five-level versions of the EQ-5D. *Pharmacoecon.* 2018;36(6):645–661.
11. Augustovski F, Rey-Ares L, Irazola V, et al. An EQ-5D-5L value set based on Uruguayan population preferences. *Qual Life Res.* 2016;25(2):323–333.
12. Luo N, Liu G, Li M, Guan H, Jin X, Rand K. Estimating an EQ-5D-5L value set for China. *Value Health.* 2017;20(4):662–669.
13. Ramos-Goñi JM, Craig BM, Oppe M, et al. Handling data quality issues to estimate the Spanish EQ-5D-5L value set using a hybrid interval regression approach. *Value Health.* 2018;21(5):596–604.
14. Versteegh M, Vermeulen K, Evers S, de Wit GA, Prenger R, Stolk E. Dutch tariff for the five-level version of EQ-5D. *Value Health.* 2016;19(4):343–352.
15. Nederland Zorginstituut. Guideline for economic evaluation in health care. 2016. https://tools.ispor.org/PEguidelines/source/Netherlands_Guideline_for_economic_evaluations_in_healthcare.pdf. Accessed July 9, 2019.
16. Oppe M, van Hout B. *The "power" of eliciting EQ-5D-5L values: the experimental design of the EQ-VT. EuroQol Working Paper Series.* Rotterdam: EuroQol Group; 2017.
17. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. *Health Econ.* 2002;21:271–292.
18. Rowen D, Brazier J, Young T, et al. Deriving a preference-based measure for cancer using the EORTC QLQ-C30. *Value Health.* 2011;14(5):721–731.

19. MVH Group. *The measurement and valuation of health. First report on the main survey*. York: Centre for Health Economics; 1994.
20. Bansback N, Brazier J, Tsuchiya A, Anis A. Using a discrete choice experiment to estimate health state utility values. *Health Econ*. 2012;31(1):306–318.
21. Abel-Olsen J, Lamu A, Cairns J. In search of a common currency: a comparison of seven EQ-5D-5L value sets. *Health Econ*. 2018;27(1):39–49.
22. Tsuchiya A, Brazier J, Roberts J. Comparison of valuation methods used to generate the EQ-5D and the SF-6D value sets. *Health Econ*. 2006;25(2):334–346.
23. https://www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/technology-appraisal-guidance/eq-5d-5l. Accessed February 13, 2020.
24. van Hout B, Janssen B, Feng YS, et al. Interim scoring for the EQ 5D 5L: Mapping the EQ 5D 5L to EQ 5D 3L value sets. *Value Health*. 2012;15:708–715.