NICE Expert advice on the EQ-5D-5L valuation set for England[1]

Charles F. Manski

Department of Economics and Institute for Policy Research, Northwestern University

August 1, 2019

Opening Remarks

I write as an econometrician who has had considerable experience developing methodology for empirical study of the distribution of preferences in human populations. I have performed many applied studies of preferences, interpreting data on actual choices or stated choices in hypothetical scenarios. I have not, however, investigated or used the EQ-5D-3L and EQ-5D-5L approaches to determination of health preferences. Nor have I been knowledgeable about their implementation in England. In this sense, I write as an outsider.

When asked by NICE to provide expert advice, I initially thought that updating the valuation set for England should be a straightforward task. One rationale for updating is to base the valuation set on recently collected data, to express the current preferences of the public more accurately than the data collected decades ago. Another is to enable the public to express a more nuanced perspective on health status than previously, this by eliciting preferences for health states on a five rather than three-level ordinal scale. Both rationales are prima facie sensible. Yet the outcome has been considerable controversy.

I have examined the published articles and other materials provided by NICE. I have posed clarification questions to the authors of the valuation set and the EEPRU

[1] This work was funded by an unrestricted grant from the EuroQol Research Foundation. The data for the published valuation set are owned by OHE.

report. After receiving and reading their responses, I have reached a two-edged conclusion. First, I conclude that some of the controversy regarding the OHE/University of Sheffield effort to create and implement EQ-5D-5L for England may be unavoidable given the nature of the task. Second, I conclude that much of the controversy stems from questionable decisions regarding data collection and modelling.

Some controversy is unavoidable because, as serious econometricians and applied economists have long recognized, empirical study of preferences is a subtle matter. Findings rest not only on availability of data but also fundamentally on the assumptions that researchers make when interpreting the data. For an extended textbook discussion, see Part III of Manski (2007). Available data are typically imperfect, and researchers often lack persuasive reasons for the assumptions they maintain. Hence, disagreements frequently arise and are thrashed out in the research literature. Disagreements are of more than academic concern when findings of preference studies play prominent roles in government resource allocation, as they do in the English public health system.

While some controversy may be unavoidable, questionable data-collection and modelling decisions have substantially exacerbated the severity of the situation that NICE and DHSC now face. The EEPRU report raised many serious concerns, concluded that the existing 5L valuation set should be abandoned, and recommended that a new study commence. My own concerns differ from those expressed in the EEPRU report in some specific respects. Nevertheless, I concur strongly with the broad Conclusions and Recommendations that the EEPRU report makes in its Section 4. The EEPRU recommendations (R1 through R4) to NICE and DHSC are exemplary.

I regretfully do not find persuasive the documented efforts to defend the methods used to generate the valuation set. These efforts appear first in published articles and then later in responses to the EEPRU report and to the clarification questions that I posed. It is disheartening that these documents provide little formal analysis to justify their positions. They sometimes cite published methodological research performed by others, but they do not argue cogently for the applicability or quality of that research. They sometime assert, ex cathedra, that their data collection and modelling decisions are well-grounded or at least reasonable. They sometimes seek to shift the burden of proof to their critics, arguing that their decisions and perspectives should be accepted unless they can be proved wrong.

Primary Concerns

I describe below my primary concerns regarding the EQ-5D-5L valuation study. I could discuss other concerns as well, but I think it best to focus on ones that I consider most important as NICE and DHSC look ahead. I comment on some other concerns later, when I respond to the specific “Questions for Experts” posed by NICE and DHSC.

Before proceeding, I call attention to the fact that in its document posing “Questions for Experts,” NICE and DHSC separate questions regarding Data Quality from ones regarding Modelling. The EEPRU report likewise separates these categories. While certain issues lie purely within the domain of Data Quality and others purely in the domain of Modelling, I find that some basic problems with the EQ-5D-5L study jointly concern data quality and modelling. The reason is that the nature of the preference data that one would like to collect depends on the way one chooses to model preferences.

I sequentially describe my five primary concerns. Concern 1 relates purely to data quality. Concerns 2 and 3 jointly concern data quality and modelling. Concerns 4 and 5 relate purely to modelling.

1. **The English data collection should be considered a pilot study rather than a basis for implementation of EQ-5D-5L.**

Describing the English EQ-5D-5L data collection, Devlin (2018, p.10) states:

“Our study was one of the first to use the protocol, and comparable studies in other countries have either recently concluded or are currently in progress.”

It has become apparent that severe difficulties occurred with the English data collection, which used version 1.0 of the EuroQol valuation protocol for EQ-5D-5L, referred to as EQ-VT Version 1.0. EuroQol (2019) presents a devastating critique of interviewer effects. The EEPRU report expresses strong concern that many respondents did not understand the TTO questions and that data quality suffers for other reasons as well.

Given the prominent role that the valuation set plays in the English public health system, I believe that it was premature to use EQ-VT Version 1.0 to value EQ-5D-5L. It would be scientifically appropriate to consider Version 1.0 as a pilot study. The intent of Version 1.0 should be to learn lessons that would inform later data collection meant for implementation.

The pilot could be used to learn how survey respondents interpret and answer TTO and DCE questions, the aim being to improve question design. Without this learning experience, the valuation-set and EEPRU authors can do no more than speculate (and disagree) about the extent and nature of possible "response errors" that may affect the estimated valuation set. Rather than speculate about response errors, it would be better to engage in a sequence of data collection efforts that would yield a progressively better understanding of response patterns.

2. The designs of the TTO and DCE questions should be informed by a well-defined coherent temporal model of health decisions.

To motivate the designs of the TTO and DCE questions, the valuation-set authors informally use conventional economic concepts of tradeoff between quality and quantity of life. Quantity of life is inherently a temporal idea, asking a person to contemplate living for different future periods of time. To sensibly interpret preferences that weigh quality against quantity of life requires specification of a coherent temporal model of decision making. For example, one could specify a standard economic model of life-cycle decisions, which assumes that a person seeks to maximize a discounted sum of present and future utilities.

The valuation-set authors do not pose a standard economic model of life-cycle decisions, nor any other coherent temporal model. In the absence of such a model, the authors cannot justify use of a common scale to interpret the responses to the various questions posed. I explain below, considering the TTO and DCE questions in turn.

A. The TTO questions are of two types. The original TTO questions ask respondents to compare living with some form of poor health for 10 years against living with better health for shorter periods of time. The new TTO questions specify a 20-year time frame, with 10 years of lead time followed by 10 years in the health state under evaluation.
   The authors describe how they measure responses to the original 10-year TTO questions and to the later 20-year questions on a common cardinal scale running from -1 to 1. Yet the 10-year questions and 20-year questions contemplate very different scenarios in terms of total length of life, with life span in the range [0, 10] years or [10, 20] years respectively. The authors do not justify their measurement of responses to the two sets of questions on a common cardinal scale.

B. Introducing the DCE questions, Devlin (2018, p. 10) states:
   "In each DCE task (Figure 3), participants were presented with a pair of health states (labelled A and B), with no reference to the duration of the states, and asked to indicate which they considered to be "better" by clicking the appropriate button."
   The absence of reference to the duration of the states is startling. It is not clear how a respondent should interpret the comparison being posed. Would health states A and B last for a day, a year, 10 years, or what? In the absence of specification of the duration of states, I see no way to justify the manner in which the authors combine the TTO and DCE findings in a hybrid model.
   By not specifying durations for the health states being compared, the DCE questions ask respondents to make hypothetical choices when facing *incomplete scenarios*. Manski (1999) observes that questions posing

incomplete scenarios do not elicit pure statements of preference. They elicit preferences mixed with expectations for the values of the decision-relevant variables not specified in the statement of an incomplete scenario.

3. Both the valuation-set authors and the EEPRU report lack clarity on how best to select health states for a good experimental design.

The valuation-set authors and the EEPRU disagree sharply on the merits of the experimental design used to gather data for the EQ-5D-5L English valuation set. However, neither group provides formal analysis to justify their positions. Both groups do little more than make competing assertions regarding the presence or absence of bias and imprecision in valuation-set estimates obtained using the data collected.

One might initially think of experimental design as a matter that relates purely to data quality, but it relates to modelling as well. It is known that the characteristics of a good experimental design for discrete choice analysis depend on the form of the model, the available knowledge of the model parameters, the estimation method used, and the specific objectives of the analysis. See, for example, Manski and McFadden (1981), Sec. 1.9.

Lack of understanding of the features of a good experimental design may be problematic looking ahead, should NICE and DHSC decide to recommend new data collection. As far as I am aware, the existing literature in econometrics and statistics does not provide a basis for choice of an “optimal” design for estimation of a valuation set.

4. The estimated model makes unjustified strong assumptions of homogeneity in preferences across the English population.

In the 1970s and 1980s, early econometric analysis of discrete choice made strong assumptions of homogeneity in preferences, exemplified by McFadden’s multinomial logit model. The rationale was facilitation of computation rather than belief that preferences are homogeneous. Technological advances in computation have gradually made it feasible to estimate much more flexible models permitting expression of considerable heterogeneity in preferences, as exemplified by the mixed logit model of McFadden and Train (2000).

The valuation-set authors verbally express considerable concern about heterogeneity in the health preferences of the English public. Feng (2018, p. 36) states:

“The final model is not one model for all; rather, it is a compromise of different opinions, statistics, and trying to capture the opinions of a nation with different—sometimes very different—opinions.”

Despite this expression of concern, the estimated model assumes that all persons share the same utility parameters β. It permits heterogeneity only in a very limited way, by permitting some variation across persons in the scaling parameter γ.

It would have been deemed acceptable in the 1970s and 1980s to use a model permitting such limited scope for heterogeneity. I find it difficult to justify this modelling decision now, when it has become routine to use mixed logit and other random-coefficient models in applied econometric research.

The above discussion concerns expression of idiosyncratic variation in health preferences across the English population. A distinct problem, which goes beyond the scope of the EQ-5D-5L study, is that using a single valuation set for the entire public does not permit health policy to recognize systematic variation in preferences with observable personal attributes. Consider the TTO questions. It is reasonable to expect that a person’s valuation of living 10 or 20 years and then dying varies systematically with a person’s age and present health condition. Yet the valuation-set authors assume that all persons have the same preference parameters β.

5. The valuation-set authors use Bayesian methods to estimate their model. They unreasonably minimize the role that the assumed prior distribution plays in determining their findings.

Applied economists using data on actual or stated choices to estimate population distributions of preferences have mainly used frequentist statistical methods, primarily versions of maximum likelihood and the method of moments. Depending on the structure of the model and the nature of the available data, computation of estimates may be easy to accomplish or a quite complex task. Econometricians and computational economists have performed much research aiming to characterize the properties of different methods and attempting to make them transparent.

The EQ-5D-5L study use Bayesian methods, which specify a prior distribution on all unknown parameters, compute the likelihood of the observed data as a function of these unknowns, and use Bayes Theorem to compute the resulting posterior distribution. The fundamental difference between the Bayesian approach and frequentist methods is the specification and application of the prior distribution. The chosen prior logically must affect the posterior findings. Hence, an essential requirement for credible use of Bayesian methods is that the researcher be able to substantiate the reasonableness of the chosen prior in the application being performed.

A serious deficiency of the EQ-5D-5L study is that the authors provide essentially no justification for their choice of prior. Feng *et al.* (2018) merely states the prior used without comment, writing (p. 31):

“For the three hybrid models with different assumptions about the distribution in slope, we assume five Normal priors N(0.1, 1) for the Level 2 parameters and wide Normal priors N(0.01, 1) for quadric parameters. Normal priors N(0, 0.1) and N(1, 0.01) are assumed for the constant and slope terms, respectively, that link the TTO and DCE data together.

For the multinomial slope model, we assume Gamma priors Γ(0.1, 0.1) for two slope parameters and the other slope parameter constrained to be one with Gamma prior Γ(1,000, 1,000). The three parameters for the probabilities of being in the three latent groups are assume with Dirichlet priors Dir(0.3, 0.3, 0.4). For the lognormal slope model, we assume the slope $\sim \ln N(\mu, \sigma2)$ with priors $\sigma2 \sim \Gamma(1,1)$ and $\mu = -0.5/\sigma2$. For the normal slope model, we assume the slope $\sim N(1, \sigma2)$ with prior $\sigma2 \sim \Gamma(1,1)$.”

The only attempt at justification for these choices is a brief and cryptic comment in Feng et al. (2018, p. 36).

An obvious question is how the choice of prior affects the posterior distribution they obtain for the parameters of interest. In their response to the EEPRU report, the valuation-set authors wrote (p. 4): “For example, with respect to priors, we tested the sensitivity of our models to alternative priors and found them to be robust.” In response to my request for clarification, they provided a more elaborate statement declaring that they wanted the prior to be “uninformative” and that, in any case, their findings are robust to choice of prior.

These responses are unsatisfactory for two reasons. First, the general notion of an “uninformative” prior is inherently ill-defined, including the specific idea that a uniform distribution is “uninformative.” Bayes Theorem transparently states that the prior and the likelihood jointly determine the posterior, making no distinction between “informative” and “uninformative” priors. Second, as a consequence of the structure of Bayesian Theorem, it logically cannot hold in general that posterior findings are robust to choice of the prior. Such a finding can at most hold locally, as one makes small perturbations in the prior.

Responses to Questions for Experts

I have discussed about my own primary concerns regarding the EQ-5D-5L study. NICE and DHSC have sought responses to a set of specific questions. In what follows, I quote these questions in *italics* and give succinct responses in normal font.

Data quality

1. **Are the data used to develop the valuation set likely to reflect the preferences of the public in England adequately?**

I am not confident that the data reflect the preferences of the English public adequately. My lack of confidence stems primarily from my Concerns 1, 2, and 3 discussed above.

Modelling

2. **Considering the model that informs the published 5L valuation set:**

a. **Is there evidence of convergence failure? If so, please comment on the strength of this evidence and the implications for the validity of the model.**
b. **Is it possible to achieve convergence (e.g. by changing the model parameters or specifications, or by estimating a model based on only TTO or discrete-choice experiment data instead of a hybrid model)?**

I am unable to answer these questions. I have had considerable experience using frequentist statistical methods to study the distribution of preferences in populations, but I have not applied Bayesian methods. Regarding the latter, I feel that I have a firm conceptual understanding of Bayesian concepts, but I am not expert in the use of MCMC algorithms to computer posterior distributions. I have no familiarity with the WinBUGS program.

Unfortunately, I found it impossible to understand the discussion of convergence problems provided by the valuation-set authors. It may be that WinBUGS is a “black-box” that defies transparent description. Or it may be that the authors did not implement the program competently. Or it may be that I am deficient in not being able to understand the issues. I am unable to discriminate among these possibilities.

While lacking knowledge of WinBUGS, I endorse the suggestion that the hybrid model be abandoned. For the reasons expressed in my Concern 2, I see no foundation for combining the TTO and DCE data in the manner of the hybrid model.

3. **The valuation set authors state that "modelling does not assume that all TTO responses are 'accurate'. The modelling approaches were selected to reflect the characteristics of the data, following careful assessment of individual respondent level data". They state that the modelling methods also account for interviewer effects (see page 4 of Devlin et al. response to the EEPRU report and the unpublished analysis of interviewer effects). Does the modelling approach chosen by Devlin et al. (2018) and Feng et al. (2018) adequately account for the characteristics of the data?**

I view the quoted statement to be unfounded. Also unfounded is the statement that the modelling methods account for interviewer effects.

4. **Are there particular choices in the model that cause you concern? Please provide your rationale, specific recommendations for alternative approaches and, where possible, supporting evidence (for example, outcome of sensitivity analyses performed by the valuation set authors or EEPRU). Please also explain the magnitude of your concern – are any issues grave enough to mean that the model should not be used to inform health and social care policy decisions in England? In particular, please consider the 4 concerns raised in the EEPRU quality assurance report, listed in the table.**

As described above, I have a severe Concern 2 about the lack of foundation for the hybrid model. Additionally, Concern 4 expresses my severe discomfort with the strong assumptions of preference homogeneity made in the valuation-set model.

Considering the 4 concerns listed in the table, I have these comments:

A. I agree with the EEPRU report that the discussion and treatment of purported “censoring” in the valuation-set study is not well motivated.

B. I have already commented on heterogeneity in preferences. Regarding heteroscedasticity, I would go beyond this concern and question more broadly the many distributional assumptions made in the valuation-set study. These include normality of distributions, linearity of distributional means in covariates, and homoscedasticity. The modern econometrics literature on semiparametric and nonparametric analysis has to a considerable degree moved away from such assumptions, which historically have been motivated by convenience rather than by belief in their realism. See, for example, Manski (2007).

C. As discussed in Concern 2, I have a severe worry about the hybrid model combining the TTO and DCE data. My concern is more fundamental than just the conflict in the distributional assumptions made in the model.

D. As discussed in Concern 5, I find the motivation for and interpretation of prior distributions in the valuation study to be severely deficient.

Conclusions and recommendations

5. **In your opinion, should health and social care policy decisions in England (including NICE evaluations) use utility values derived using the 5L valuation set for England?**

I am unable to answer this question in the absence of specification of feasible alternatives to consider. I have described my strong concerns with the existing 5L valuation set. However, the question asks me to compare the use of 5L in policymaking with unspecified alternatives. This leaves me unable to respond.

6. **If the answer to question 5 is NO:**
    a. **What action do you recommend to create a 5L valuation set that would be suitable for informing health and social care policy decisions in England (including NICE evaluations)? Please be explicit about whether you believe new data collection is required or if you recommend different modelling approaches of the current data set.**
    b. **In the interim, whilst the actions specified above are being done, should health and social care policy decisions in England be based on the existing 5L valuation set for England?**

A. I endorse Recommendations R1 through R4 of the EEPRU report. I recommend new data collection, to be undertaken in conjunction with rethinking the modelling. I do not recommend application of different modelling approaches with the current data set.

B. For the reason expressed in my response to Question 5, I feel unable to answer Question 6b. A constructive discussion of this important matter might be possible if NICE and DHSC are able to pose feasible alternatives to use of the existing 5L valuation set.

Declarations of interest

In line with NICE's Policy on Conflicts of Interest, I have no conflicts to declare.

References

Devlin, N., K. Shah, Y. Feng, B. Mulhern, and B. van Hout (2018), "Valuing Health-Related Quality of Life: An EQ-5D-5L Value Set for England," *Health Economics*, 27, 7-22.

EuroQol (2019), "QC Report for England," EuroQol Office, Rotterdam.

Feng, Y., N. Devlin, K. Shah, B. Mulhern, and B. van Hout (2018), "New Methods for Modelling EQ-5D-5L Value Sets: An Application to English Data," *Health Economics*, 27, 23-38.

Manski, C. (1999), "Analysis of Choice Expectations in Incomplete Scenarios," *Journal of Risk and Uncertainty* 19, 49-66.

Manski, C. (2007), *Identification for Prediction and Decision*, Cambridge, MA: Harvard University Press.

Manski, C. and D. McFadden (1981), "Alternative Estimators and Sample Designs for Discrete Choice Analysis," in C. Manski and D. McFadden (editors), *Structural Analysis of Discrete Data with Econometric Applications*, Cambridge, MA: MIT Press, 2-50.

McFadden, D. and Train, K. (2000), "Mixed MNL Models for Discrete Response," *Journal of Applied Econometrics* 15, 447-470.

© NICE 2019. All rights reserved. Subject to Notice of rights