The EQ-5D-5L valuation set for England
A view from the side

Commissioned Author: Werner Brouwer
With support from: Arthur Attema, Matthijs Versteegh

Executive summary

The EQ-5D is a well-established generic health-related quality of life instrument. Recently, a five level version of the instrument was developed (EQ-5D-5L), replacing the three level version. The valuation set for England for the EQ-5D-5L became available in 2016, but is not yet recommended by NICE. This seems due to questions about whether the current valuation set adequately reflects health state preferences of the English general public.

Previous discussions have highlighted concerns about data quality and (subsequent) modelling choices to estimate the valuation set. NICE has asked advice on this matter, which is provided in this brief report.

We conclude that the EQ-5D-5L valuation set was based on data with important quality issues. Following the new EuroQoL protocol, much of the gathered data would have been discarded. Although they may have affected modelling choices, these data issues do not necessarily translate into issues with the valuation set. The authors of the valuation set have used innovative approaches to arrive at the current valuation set, mitigating some of the data problems by combining TTO and DCE data. This required a number of (normative) assumptions and choices, which have been debated, including the issues of convergence failure of the final model and censored data. Notwithstanding these issues, whether the resulting EQ-5D-5L valuation set does or does not adequately reflect 'true' health state preferences for England is inherently difficult to answer in absence of a golden standard.

This implies it is important to look forward. The combination of poor TTO data quality, innovative modelling, and the lack of a golden standard to evaluate the

resulting valuation set with, results in a situation that adopting the current valuation set will probably remain surrounded by questions regarding its validity. This may negatively affect the support for decisions based on economic evaluations using this valuation set.

We therefore advise NICE to perform a new valuation study and not to use the current EQ-5D-5L value set.

This recommendation does not imply the current valuation set misrepresents health state preferences. Such a claim in our opinion cannot be made with the current information. The recommendation reflects the observed data issues, and the existing doubts about the (modelling of the) resulting valuation set. A new valuation set need not even be significantly different from the current one. However, it would ideally be based on data that meets current, higher standards of quality, facilitating its tractable analysis, and increasing the trust in a resulting valuation set. A new valuation study can build on the previous experiences. Some recommendations for a potential new valuation study are provided, also concerning its governance.

Securing a valuation set with stronger support in our opinion is in the interest of all involved parties. Not in the least NICE and the English population, given its intended use in decisions on the allocation of scarce health care resources.

Introduction

The EQ-5D instrument is a well-established generic health-related quality of life instrument. It is often used in the context of economic evaluations, and in some jurisdictions its use is advocated or prescribed in guidelines for economic evaluations aiming to inform policy decisions. The EQ-5D instrument typically comprises a descriptive system, a VAS valuation of current own health, and a valuation set or 'tariff'. The former consists of five main domains of health (i.e., mobility, self care, usual activities, pain & discomfort, anxiety & depression). The latter typically reflects the preferences of the general public for the health states described with the EQ-5D instrument. These preferences are normally anchored so that 0 denotes the value related to the state 'dead' and 1 denotes the value related to the state 'perfect health'. Preferences for health states are generally obtained in the general public for a subset of all potential health states, and statistical techniques are subsequently used to estimate the preferences for health states not directly valued. The algorithm that describes the estimation of the preference value for all health states is generally referred to as a valuation set or 'tariff', and often is country specific. The valuations can be combined with the EQ-5D health state descriptions obtained in patients. Such valuations, or quality of life weights, in combination with information on duration of particular health states, allow calculating Quality-Adjusted Life-Years. QALYs constitute a crucial outcome measure in economic evaluations taking the form of a cost-utility analysis. The results of such evaluations can influence subsequent decision making, which underlines the importance of the validity and reliability of the EQ-5D instrument and the related valuation set(s).

Initially, the EQ-5D instrument had five items with three levels per domain (corresponding with no, moderate or severe problems in that domain). This is labelled as the 'three level version', which we will refer to as the EQ-5D-3L henceforth. Valuation sets to value the 243 states this instrument can describe exist in many countries, including in England. Of course, having three answering categories limits respondents in expressing their health state, which might for instance lead to insensitivity to smaller changes. More recently, therefore, a five level version was developed (EQ-5D-5L), which allows respondents more answering possibilities (as it is able to describe 3125 different states using the same five health domains), arguably leading to more precision and sensitivity. Having this new five level instrument also necessitates having a corresponding new valuation set.

Most valuation sets aim to reflect the preferences of the general public for health states (sometimes called ‘health related utility’) described with the corresponding instrument. Different methods exist to do so, but it is important to note that no golden standard exists, neither in terms of methods nor in terms of outcomes (i.e. values or preferences), against which valuation sets could be judged. The valuation set for the EQ-5D-3L version for England (like, later, for other countries) was based on health state preferences obtained using the so-called Time Trade-Off (TTO) method. In this method, respondents are asked to trade-off length and quality of life (normally by choosing between living longer in some impaired health state or living shorter in perfect health). Through an iteration procedure, the point of indifference between these two streams can be obtained, which can subsequently be used to derive the relative value of the impaired health state. (We label this as ‘relative value’ since its value is expressed relative to that of perfect health (1) and dead (0).) It is well-known that the TTO method, like other methods, is not without methodological problems. Nonetheless, the resulting valuation sets were and are used in several countries, including in England.

The EQ-5D-5L valuation set, based on a large-scale study, for England became available in 2016. However, it has been the source of some controversy, mainly related to data quality and modelling, which casted doubts on whether the derived valuation set reflected general public preferences for health states in England. This controversy resulted in an extensive interaction between the authors of the EQ-5D-5L value set and a team of experts at EEPRU who performed an independent quality assessment of the study. Meanwhile, NICE does not recommend the EQ-5D-5L valuation set in their methods guides and manuals (although it does recommend the EQ-5D-5L descriptive system). For now, it continues to work with the EQ-5D-3L valuation set and asked advice on the identified issues with the EQ-5D-5L valuation set for England.

This report is written in that context and provides an expert advice on the EQ-5D-5L valuation set for England. It is stressed that this report does not aim to address all potential issues or explore all possible ways forward. It focuses on a number of issues related to data quality and modelling, as formulated in the questions posed by NICE, aiming to seek ways forward. Moreover, we address these questions from the angle of our own expertise. Given the selection of experts asked to reflect on the current issues, here most emphasis will be placed on preference measurement, the use of the Time Trade Off method (in combination with the Discrete Choice Experiment), and general evaluation criteria. We do not consider some of the econometric issues at stake to be our core expertise, and address these more briefly, knowing that they will be addressed in more depth by other experts.

The report is structured as follows. First, we highlight the questions posed by NICE and our working methods. Next, we address some general issues as well as the questions posed by NICE. Finally, we conclude this report and our advice.

Questions and methods

NICE sought advice on six main questions, which were formulated in relation to the previous discussion between the valuation set authors and the EEPRU team. These are briefly listed below (and provided in full in appendix A).

1. **Do the valuation set data reflect the preferences of the public in England adequately?**
2. **Considering the model that informs the published 5L valuation set:**
    a. **Is there evidence of convergence failure?**
    b. **Is it possible to achieve convergence?**
3. **Does the modelling approach chosen by Devlin et al. (2018) and Feng et al. (2018) adequately account for the characteristics of the data?**
4. **Are there particular choices in the model that cause you concern? Please also explain the magnitude of your concern […]. In particular, please consider the 4 concerns raised in the EEPRU quality assurance report, listed in the table.**
5. **In your opinion, should resource allocation decisions in England (including NICE evaluations) use utility values derived using the 5L valuation set for England?**
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 resource allocation decisions in England […]?**
    b. **In the interim […] should resource allocation decisions in England be based on the existing 5L valuation set for England?**

The answers to these questions were formulated based on three steps:

1. Studying the written materials provided by NICE, including the exchange between the valuation set authors and EEPRU team
2. Considering / analyzing the data, analyses, and materials provided by NICE (and OHE)
3. Deliberation among the three involved researchers.

These steps were performed within the available budget and time constraints as set by NICE. In the next paragraph we will answer these questions. The controversy has resulted in a debate about the validity of the data and EQ-5D-5L valuation set, especially between the original authors and the EEPRU experts. We take their arguments into consideration, without suggesting these arguments are the only relevant ones. We also note up front that our answers are not intended to 'take sides'. In fact, we feel that the authors have gone to great lengths to innovatively create a valuation set based on combined data (with inherent issues), while the EEPRU has provided very thorough feedback on the data and modelling approaches. Hence, we merely provide our answers to the questions posed and highlight some fundamental challenges in the underlying search for ensuring a valuation set that adequately reflects general public preferences.

Answers

In this paragraph we will provide answers to the questions posed by NICE, in a succinct manner, as requested. We answer each question one by one. Before we do so, we make some general remarks to give some context to the questions and, especially, our answers.

General remarks

Preference measurement is not an easy task, certainly not in relation to health states. Any common method used to derive health state preferences, such as Standard Gamble, Time Trade-Off or Discrete Choice Experiment, has its own advantages and disadvantages. Golden standards in terms of methods or by which to judge whether derived preferences using such methods, or a valuation sets based on the observed preferences, are accurate do not exist, since these preferences cannot be directly observed. Indeed, the preferences revealed by these methods are, arguably, the result of actual preferences and the valuation technique applied. This means that any judgment regarding derived preferences is likely to be indirect, based on aspects such as elicitation methods and process chosen, data quality (which can be defined in several ways), accuracy of modelling (for instance judged by statistical performance of models), and comparison with other evidence (e.g. previous studies or in other countries). All such evaluations only indirectly answer the question about whether any given valuation set is 'accurate' or even 'good enough'. One may evaluate whether a valuation set adequately reflects underlying data (if 'adequately' is defined), but this is already more difficult when using more than one data source, or when the underlying data is considered biased or imperfect. Whether the data adequately reflects unobserved preferences is again more difficult, as is judging whether the valuation set that has been modelled 'adequately' reflects societal preferences (as this requires some assumption on the 'true' unobserved preferences).

Given this, two aspects need to be considered. First, it is important to have sufficient scrutiny applied to the process and outcomes of any valuation study. Second, an absolute answer as to whether a valuation set is 'good enough' is inherently difficult to provide.

Regarding the first issue, the process followed in England, including an independent quality assurance assessment, is commendable, also given the various stakeholders with own and divergent interests, involved in the project in different capacities (e.g. researchers, EuroQoL Research Foundation, NICE). At the same time, however, given that no method or dataset is perfect and golden standards as to when data and a valuation set are 'good enough' do not exist, it raises questions about which norms and standards to use in judging resulting methods, data, and valuation sets. This is relevant here in relation to the positions of both the EEPRU team as well as the valuation set authors. It is also relevant in looking forward, e.g. in the context of deciding whether a potentially collected new dataset would meet standards, or a new valuation set would be sufficiently 'trusted' to reflect general public preferences. Raising the bar high enough on each single possible criterion may lead to a situation in which any dataset / valuation set would be rejected as lacking sufficient quality. Moreover, some disputes cannot be readily solved with more analyses on the same data, or even with new evidence. We will also address this issue in our recommendations.

Some issues regarding the process leading up to the current debates also deserve attention. At several points, the authors of the valuation set and, subsequently, the EEPRU team, mention the role of the Steering group. Much of the current discussion is a scientific one, about data collection and quality, modelling choices, etc. However, the authors appear to reflect that some of the choices, also regarding reporting, were importantly influenced by the Steering group. As the authors write: "The project was overseen by a Steering Group chaired by the head of R&D at the Department of Health (DH) for England; its members comprising senior economists from the DH, senior members of the NICE technology appraisal team, a NICE technical appraisal committee chairperson and UK academics with experience in conducting value set studies and their use in economic evaluation. All aspects of the study design, the characteristics of the data generated, and a wide variety of alternative modelling approaches were presented in detail and discussed at Steering Group meetings. The work as reported in our papers in Health Economics reflects the guidance we received." (Devlin et al., 2018b)

This quote suggests involvement of the Steering Group in many choices, and highlights the distinct roles and interests represented in the Steering Group.[1] The latter may have led to choices (if only in presentation) that reflect a tension between the scientific desire for openness about the performed work (such as the impact of alternative specifications of models), and the intended policy use of a final model. The authors for instance write: "It had been our intention to report a number of these alternative models in our 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 potential users." (Devlin et al., 2018b) Some of the comments of the EEPRU team rightfully emphasized the

[1] See also the response of the authors to Q2, as posed by Manski.

scientific desirability of reporting more models (as well as reporting other elements influencing results or their interpretation). In this case, therefore, some of the choices, to some extent ‘guided by’ the Steering Group (like not reporting alternative models to reduce ‘gaming by potential users’), may have partially fueled the current debate.

Without wishing to criticize the composition or guidance of the Steering Group, these quotes make it relevant to at least recognize the role of this group in the context of the current debate. Moreover, they raise the issue of optimal governance of a project like this, acknowledging the different and potentially conflicting roles, responsibilities and interests of involved parties, as well as the power balance between them. This is clearly important for potential future studies.

One of the most fundamental issues in the current debate is that of data quality, also because part of the modelling choices (and issues) relate to it. It needs noting that the EuroQoL Research Foundation has improved their standards of data collection and quality control since the study conducted in England, upgrading to the EQ-VT 2.0 protocol (Stolk et al., 2019). They also highlight serious issues with the current data when evaluated using the new standards (EuroQol, 2019). One may view this as a clear signal of a desire to improve data quality.

We also note that next to issues of data and modelling, a number of normative choices were made in designing the research protocol. Next to general aspects (such as, like commonly the case and in line with current NICE guidelines, taking the general public as source), this relates to the choice of methods (TTO in combination with a DCE), the choice for and design of the lead time TTO for states considered to be worse than dead (in which respondents could not specify values lower than -1 – which was possible in the 3L valuation)[2], etcetera. Some of these choices impact subsequent choices (i.e. the need/desire/possibility to combine DCE and TTO results), the observed values below 0, etcetera. This implies that some of the

[2] See also authors response to Q3 as posed by Manski.

differences compared to the currently used 3L value set will be driven by normative, methodological choices, while others are due to (the process of collecting) the data.

The choice for TTO and DCE as sources of preferences in this context is certainly defendable, although, separately and jointly, not without problems. Both are stated preference methods and are described in detail elsewhere (e.g. Drummond et al., 2015; Jones, 2006). Through TTO exercises, points of indifference between living longer in poorer health and living shorter in perfect health can be derived. Known issues with the TTO and the common way in which answers to TTO questions are transformed into health state valuations (based on linear utility) are violations of procedural invariance (e.g. Bleichrodt et al., 2003; Attema and Brouwer, 2012), and not accounting for biases like loss aversion or discounting (e.g. Bleichrodt, 2002; Attema and Brouwer, 2014; Lipman et al., 2019). Moreover, the way in which a given operationalization of the TTO is applied is important and can importantly influence results (such as choice versus matching, mode of elicitation – web versus face to face, and instruction/interviewer). The advantage of TTO is that quite some experience exists with performing TTOs in this context, important biases are known, and the answers to TTO questions can be anchored to the states ‘dead’ and ‘perfect health’.

DCEs typically ask people to (repeatedly) choose between two alternative options, described using relevant aspects (attributes) of those options. The choices are subsequently used to derive preferences for the attributes. DCEs are based on Lancaster’s theory of value (Lancaster, 1966), and the Random Utility Model (McFadden, 1974; Manski, 1977). The DCE is an increasingly used preference elicitation method to obtain preferences with, also in the field of health care (De Bekker-Grob et al., 2012). DCEs can also suffer from issues with reliability and validity, including being susceptible to small methodological changes, questions about external validity, and violations of rational choice theory (e.g. Rakotonarivo et al., 2016). Arguably, less experience with performing DCEs in the context of deriving health state preferences exists, potential biases/disadvantages may be less known or directly observable at the level of the individual respondent, and, importantly, answers to DCE questions are normally not directly anchored on the same utility scale (with dead and perfect health as anchor points) as TTO values, which is required to allow QALY calculations.

In the EuroQoL valuation protocol the DCE did not include a ‘duration’ or ‘dead’ parameter, and, as a consequence, the DCE data needed to be combined with the simultaneously gathered TTO data to anchor it on the same utility scale. Combining both methods is innovative and interesting, but also raises new questions, like: is the combination of both sources of data ‘better’ than either source separately? Can these different methods of data be combined and, if so, how should they be combined to come to a valuation set? Whether combining data obtained with multiple methods is better (in terms of ultimately reflecting preferences of the general public for health states) remains unclear, and will also relate to the methods used. Just to illustrate, would we also consider a combination of Standard Gamble or Willingness to Pay data with TTO or DCE data better than only TTO or DCE data? It is likely that any evaluation of such questions also depends on the (relative) trust placed in both methods to produce sound preference data; a matter on which differences of opinion may exist.

In The Netherlands, while DCE data was gathered next to TTO data as well, the final EQ-5D-5L valuation set was based on the TTO model only (Versteegh et al., 2016), which seemed to have sufficiently favorable characteristics (and allowed direct anchoring). The DCE data only informed the choice between different TTO based model specifications. Of course, when one of the sources appears less valid (e.g. the TTO data), the data from the other source (e.g. DCE) may be used in an attempt to overcome the observed problems. Such a combination may nonetheless yield new problems and potentially arbitrary choices regarding whether the combined data is a better description of underlying preferences. As illustration, in the English valuation study, the DCE data could be used to better capture valuations of health states on the lower end of the utility scale (i.e. considered worse than dead), for which the TTO data may be considered less reliable. However, then, especially how the length of scale (i.e. how far below 0 the scale runs, so the value of state 55555) can be adequately estimated, becomes an important question which the TTO data can no longer inform, given that the TTO estimate for the length of the scale was questioned in the first place.

The interesting general issue with using two methods is that, in absence of clear indications of better performance of one over the other (in terms of theory or data quality), any difference leads to a question of which source to trust. Combining data from both implies assuming that the 'truth' is somehow and somewhere in the middle and the combination is better than one of the two sources separately. One could link this to Segal's law stating that "A man with a watch knows what time it is. A man with two watches is never sure." Whether an average time is better than picking either watch is unsure, in absence of a golden standard.

Answering the questions

Below we provide succinct answers to the questions posed by NICE.

Q1: Do the valuation set data reflect the preferences of the public in England adequately?

We believe this question cannot be answered unequivocally because of the lack of a golden standard. Put simply: which test (on the existing data) would prove beyond any doubt that the current preference data or valuation set does or does not reflect the preferences for health states in the English general public adequately? What does 'adequately' even mean in this context and how would we assess adequacy, especially in a context of unobserved 'true preferences'? Answering this question therefore requires relying on indirect evidence of appropriateness of the estimates, which may take the form of assessing whether estimates are 'sufficiently trustworthy' rather than directly observing or evaluating their 'correctness' or 'accuracy'.

It is useful here to distinguish between the data on which the valuation set is modelled and the valuation set itself. The data obtained has some properties that clearly warrant caution, as we will elaborate on below. Some of the subsequent choices in selecting and combining data and the statistical modelling appear to have been aimed at overcoming these issues. Despite the data problems, this may have resulted in a valuation set that may reflect general public preferences fairly 'adequately'. If comparison to other valuation sets would be a standard to evaluate the English valuation set with, it for instance does not appear to be a pertinent outlier (Janssen et al., 2018). But, of course, not being an outlier cannot be seen as proof that the valuation set reflects English health state preferences. And indeed, as the

EEPRU team also rightfully notice, in itself small differences per parameter in a valuation set can still amount to a meaningful difference in a final calculation and economic evaluation.

In absence of a golden standard or a definite test to prove the data reflects preferences, the answer to questions such as the above is normally sought through confidence in the process (which can range from who performed the study to its design and the methods used), quality of the data (which can be based on several types of inspection and tests), modelling techniques (and statistical tests and properties), and the resulting outcomes (including some face validity of the outcomes and relation with the underlying data).

The methods used to obtain the preference data (i.e. TTO and DCE) are, in themselves, defendable and logical choices and both are often used methods of preference elicitation in the field of health care, as noted above. In the current context, the use of TTO is more common than that of DCE, especially given the need to anchor health state valuations to dead (with value 0) and perfect health (with value 1) to allow subsequent QALY calculations. The data obtained in the valuation set study is a reflection of the chosen methods and their operationalization. How to optimally combine the data from both methods, also given the differences in underlying scales, and whether any combination is better (in theory) than either of the two independently, is unclear.

The data obtained from the TTO appears to have noteworthy quality issues. Most prominent issues include the apparent interviewer effects in the data and low protocol compliance by interviewers (EuroQoL, 2019). The EuroQoL Research Foundation writes that: “The team would have kept only 138 of the 998 interviews if they had followed the current guidance. The remaining 860 interviews would not be part of the final dataset, because the low levels of protocol compliance would have prompted early intervention.” While the underlying issues and interviewer effects do not necessarily imply bad or unusable data, or, also given modelling techniques and choices, an inadequate valuation set, they clearly signal substantial issues with the data.3

In addition, the spikes in the observed TTO responses (especially for 1, 0.5 and 0, jointly making up for almost 40% of responses – see EuroQoL, 2019), also raise questions about the process of obtaining the data, accuracy of responses, and the use of the full valuation space by the respondents. The low use of the worse than dead task, potentially related to the explanation provided about this possibility by the interviewers, adds to this (EuroQoL, 2019). The latter also implies that the relationship between the observed data and the anchors on the utility scale used in QALY calculations (anchored on perfect health and dead) is disturbed. Running a simple (or even naïve) TTO model on the English data indicates that the implied range in the TTO data is almost from 0-1 (see appendix B). This is at odds with the EQ-5D-3L valuation set, the current EQ-5D-5L valuation set, and those from several other (comparable) countries, suggesting data issues, which were indeed confirmed (EuroQoL, 2019).

The EEPRU team writes: "The DC experiments do not explore variations in length of life alongside health states and they only give rankings of states rather than quantitative differences. DC data are therefore less informative than TTO data, and play a more limited role in driving the results of the overall valuation." (Hernández-Alava et al., 2018) Given that DCE responses are analyzed at the aggregate level to come to valuations (as the method does not iterate towards a point of indifference per respondent), the quality checks on the DCE data are a bit more limited. The EEPRU team write: "There is little scope to examine validity of the DC data because the experimental design ruled out any combinations of choices capable of displaying logical inconsistencies. … Unlike the TTO experiments, there is no evidence of any statistical dependence between the outcomes, nor of any systematic association with the position of the task within the sequence." (Hernández-Alava et al., 2018) The EuroQoL Research Foundation (EuroQoL, 2019) additionally tests for strange choice

[3] See also Q9, as posed by Manski, and the response from the authors.

patterns (always A, always B, or always alternating A and B). The results do not indicate large data problems, but the test is more limited.

Hence, we conclude that there are clear data quality issues in the English valuation study, especially those obtained with the TTO. This casts serious doubts on whether the obtained data in and of itself can be seen as an adequate representation of English health state preferences. The TTO data suffers from serious interviewer effects and implies relative high valuations for poor health states due to the low use of the worse than dead protocol. According to comparative standards (e.g. experience in other countries / studies) data issues are apparent, and according to the new EuroQoL protocol much of the data would have been discarded.

Briefly put, the Time Trade-Off data in our opinion in itself does not adequately reflect the preferences of the English general public for health states.

The fact that the final valuation set also differs from the observed TTO valuations, underlines that the above observation may even be shared by the valuation set authors. The more difficult issue is whether the use of this non-optimal (TTO) dataset necessarily implies that the resulting valuation set also does not adequately reflect the preferences for health states in the English general public. This is more difficult, if not impossible, to answer. The impact of the data quality on the final valuation set may be mitigated by some of the modelling and data selection choices and must be considered in light also of the DCE data. In appendix C we briefly explore the relationship between the TTO and DCE data to highlight some of the issues with the TTO data. Some, but not all, of these may be mitigated by also using the DCE data. One may consider the choice for combining the TTO data with DCE data as sensible in light of the limitations of the TTO data. At the same time, opting for that combination, rather than only TTO data, may also be seen as indirect evidence or acknowledgement of the limitations of the TTO data, for which solutions other than combined modelling of TTO and DCE data may also be considered (notably gathering new data).

Nonetheless, we have no standard by which to judge whether or not the current EQ-5D-5L valuation set for England reflects ‘the preferences of the public in England adequately’. Still, it is important to note two things. First, the subsequent use and modelling of the data by the valuation set authors appears to have relieved some of the data issues (even if the modelling itself is not without problems), resulting in a valuation set which has some face and comparative validity (even if both criteria are not sufficient conditions to accept the valuation set).4 Second, it is likely that new data collection, following current standards of data collection, would at least result in better (TTO) data quality due to less protocol violations. This may lower the need for sophisticated modelling techniques and may better address important issues, such as the length of the utility scale (i.e. its lowest value). It may perhaps even reduce the need to combine TTO and DCE data in one model to derive the valuation set. (This was for instance the case in the Netherlands, where the quality of the data derived with the TTO was more favourable and the new valuation set was based on that data alone. Note that we do not mean to imply that preferences derived with a TTO in general are to be preferred over those obtained with a DCE or, under the right conditions, a combination of both. Merely, using one method of data collection in modelling can facilitate the analysis - and the TTO has the advantage of directly using the common anchor points of 0 and 1.)

However, even though new data collection is likely to result in data of higher quality (certainly for the TTO), it is unsure whether this would result in a significantly 'better' or even different valuation set, also given the efforts of the authors to take maximum advantage of the available data. Indeed, the authors have been resourceful in using the obtained data, in an attempt to correct for some of the indicated anomalies, into a valuation set. The resulting valuation set appears fairly aligned with those reported in other countries (and the sources of observed differences can reflect many things, ranging from differences in true preferences to problems with the data or modelling). Moreover, as illustrated in appendix C, simpler, transparent, and easily reproducible methods result in a similar length of scale lending some general credibility to the approach taken in combining DCE and TTO data (for the purpose of overcoming problems in the TTO data). (Whether combining DCE and TTO data would be

[4] See also Q2, as posed by Manski, and the response by the authors.

necessary or preferable when (new) TTO data would not be considered to be problematic can be debated). Nonetheless, the EEPRU team point to some noteworthy issues in the modelling of the data, including convergence failure of the final hybrid model and the way the value of +1 is treated. Moreover, as noted, relatively small differences in a valuation set can still meaningfully impact results of economic evaluations.

Notwithstanding the above, in the absence of a gold standard, it is important to emphasize that gathering new data (if performed correctly) is likely to increase the confidence in a resulting valuation set, which does not necessarily imply that the resulting valuation set more adequately reflects (unobserved) 'true' health state preferences.

Some of the (correctly noted) concerns with the dataset and the modelling exercises, in the context of deriving a valuation set, ultimately need to be judged in terms of their (net) impact on that valuation set. Any change in data or modelling techniques will need to be evaluated in terms of whether or not its impact is deemed an improvement or not. This requires statistical testing, but ultimately likely also normative judgments. The same holds for the evaluation of the current EQ-5D-5L valuation set. Hence, we conclude that while it is impossible to judge with current information whether or not the EQ-5D-5L valuation set reflects the preference of the public in England adequately, there are important concerns about the underlying data quality, which the modelling needed and sought to overcome. Whether or not it fully succeeded in doing so, and whether all modelling choices can be defended, subsequently became matter of debate. In our opinion, it is unlikely that a further discussion on these matters will lead to consensus or result in the ultimate goal: confidence in the valuation set to reflect health state preferences adequately. Some disputes are simply difficult to settle using the same dataset.

Q2: Convergence failure of the model that informs the published 5L valuation set

There indeed appears to be convergence failure in the final model. The EEPRU team provides an in depth discussion and analysis of this issue (Hernández-Alava et al., 2018). The authors appear to acknowledge this in their response to the EEPRU appraisal, by indicating that "… 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 … with respect to MCMC and guided us in judging whether there were genuine problems with the estimations." (Devlin et al., 2018b) In other words, the authors do not deny the convergence failure, but indicate this does not pose serious problems by comparing the results of the model to other estimates, providing figures that offer some insight into the differences in terms of the parameter estimates. The EEPRU team make a clear case for the fact that the plausibility of results is not a full "… reassurance that the correct model has been estimated and that the differences in utilities reflect the preferences of the sample or the general public." (Hernández-Alava et al., 2019) We agree, yet also highlight that convergence failure is also not a proof of an inadequate valuation set in terms of reflecting preferences of the sample or general public. In other words, the magnitude of the potential problem for the resulting valuation set is difficult to quantify. Nonetheless, clearly reporting convergence failure of a final model, as well as its potential causes and implications, should be part of reporting/publishing the model.

Could convergence failure have been avoided? Convergence failure might be avoided by a different specification of the final hybrid model, but also by opting for other models, like the maximum likelihood model. This model had, according to the response from the valuation set authors, only minor differences compared to the final model. (Although a full assessment of differences should be based on more than the information provided, i.e. not only on parameters, but also on the constant and slope parameters; so predicted values.) A clear indication on why the Bayesian model was preferred as the final model over other models (without convergence failure issues) was not found, which may be related to the fact that not all models were originally presented (as requested by the Steering group, according to the authors). However, we again note that while it may be fairly straightforward to opt for another model, which may avoid issues like convergence failure, this does not imply that the associated valuation set is, somehow assessed, better.

Q3. Does the modelling approach chosen by Devlin et al. (2018) and Feng et al. (2018) adequately account for the characteristics of the data?

In our opinion, the valuation set authors appear to have been quite thorough and resourceful in analyzing and modelling the data to generate the valuation set, given the inherent limitations of the data. One may for instance see the combination of the TTO data with the DCE data as a way of dealing with some of the issues with the TTO data, by ‘smoothing’ the data further over the ‘full’ utility space, which as illustrated in Appendix C. Three issues deserve mentioning in this context.

First, the first reports on the valuation set may not have provided sufficient details about the different models tested, and some characteristics of the model performance and choices. While many aspects may play a role in this context (ranging from word limits of journals to preferences of a Steering group), this may have fueled some of the subsequent debate.

Second, in evaluating whether a modelling approach ‘adequately accounts for the characteristics of the data’ also depends on the goal one sets out to reach and (related) the evaluation criteria used. Two extremes are highlighted. One could use only statistical tests to see how a model fits the data and performs (almost independently from the practical implications or purpose). One could also consider whether or not the outcomes of the model appear appropriate in light of its intended use, broader knowledge on / assumptions about health state preferences, and the availability of other (present and past) valuation sets (almost independently from statistical properties). Ideally, of course, the two positions coincide; in practice they may not, necessitating normative choices (which are again open for debate). This especially requires transparency about and justification of (the reasons for) such choices. For example, some of the choices in the modelling (e.g. the treatment of +1 values as being censored) may partly have been driven by consideration of the outcomes of alternative specifications and an underlying notion of what adequate outcomes ‘should’ look like. To illustrate, the authors of the valuation set for instance write, in the context of the +1 censoring: “It does not suggest that we believe the true TTO values are higher than one, but rather a combination of the true value and an error term which follows a normal distribution. The observed average TTO value for the mild health states is clearly too low. These values might be unable to reflect the true average value for mild states in the English population. Fitting the TTO data into a normal distribution with assumption of right censored at 1 (which is not too different from taking the median) could better represent the "real" average values." (Feng et al., 2016; underlining added) As highlighted above, such evaluations of outcomes against prior knowledge and other sources, are a logical part of an exercise as deriving a valuation set in absence of golden standards, but can create tensions with other evaluation criteria, requiring normative trade-offs.

Third, whether the data as such sufficiently captures health state preferences, or whether it is a reflection of the (imperfect) way the elicitation methods were applied, is an important issue. In the latter case, the question is whether modelling can completely correct for the quality issues with the data or whether the effort could better be put into collecting better quality data. An answer to this question should consider not only the technical aspects of the issue, but also the intended use of the resulting valuation set and therefore political and social aspects. Given the current dispute and the lack of a conclusive external criterion to test the appropriateness of the current data and value set, replication and confirmation, by gathering new data, could address both these aspects. This is a preferred way forward.

Q4. Are there particular choices in the model that cause you concern?

Here we address the four concerns raised in the EEPRU report, with an emphasis on the first point.

(i) Treatment of +1 values.

The TTO method in principle defines +1 as being 'perfect health'. By definition within the underlying theoretical model, no other health state can have a higher value. Commonly, we assume that perfect health in the EQ-5D nomenclature is represented by state 11111 (i.e. no problems on any domain). This issue of how to treat observations of 1 in the data, raises a number of interesting questions. The issue is important, also because of the spike in the TTO data at 1 (i.e., many observations of 1 were observed).

First of all, there is a difference between the error term in eliciting preferences and the values themselves. In the TTO exercise, people cannot express a preference for a health state beyond the health state of 11111, since one cannot express the equivalence of living more than 10 years in perfect health and 10 years in the imperfect health state under valuation. In that sense, higher values could not be observed by design of the TTO. It seems that the issue of censoring was especially dealt with in relation to the error term, although some of the reports are not completely clear about this. Indeed, the EEPRU team (Hernández-Alava et al., 2018) reference text from the authors suggesting the potential of higher values than 1. To give another example, Feng et al. (2016) write: "*However, to assume the errors follow a normal distribution is incorrect as the assumption denies the fact that the theoretical TTO values could exceed 1.*" The first part is understandable: for very mild states, people can err to lower values more than to higher values, given that one cannot go beyond the value of 1 (i.e. equal to perfect health defined as 11111), resulting in a right truncated distribution. However, that the theoretical TTO values could exceed 1, is less obvious and allowing it has broader consequences (as explained below).

Second, correcting for the distribution of the error term through treating +1 as being censored has an upward effect on the valuation of mild health states as highlighted by the EEPRU team. It needs noting that the valuation set authors already highlighted this by performing a sensitivity analysis on the censoring assumptions as shown in Feng et al. (2016; Table 5). In response to the EEPRU criticism, the authors of the valuation set highlight that the resulting values may not lead to 'overvaluation' of the mild health states by comparing it to some summary statistics for all health states with all 1's and one 2. Such comparison is informative and relevant, but at the same time not a complete proof that the correction is justified, also since a similar criterion may not be used across the full range of values, such as around 0, and also depends on the notion that the underlying data adequately reflect health preferences. Whether the upward effect of the correction (compared to no censoring at 1) is deemed 'appropriate' depends on the approach taken to define this, as also highlighted above. In our opinion, the effects of the assumptions of censoring (e.g. shown by Feng et al. 2016) **are non-negligible**. Moreover, treating +1 as a censored observation does implicitly allow states higher than 1.

Finally, the issue of theoretical values above 1 is interesting in and of itself. In principle, the theory behind TTOs assumes that perfect health (commonly defined as 11111) has the highest value. The assumption of strong monotonicity requires people to prefer something that is better in at least one aspect and worse in none. If one accepts this, health state 11111 should always be preferred over any other health state (and indeed any health state dominating another health state by being better in at least one domain and worse in none, should be preferred). Any observation of +1 for any other state than 11111 in that context would imply a measurement error, as it should be lower than 1 (if only marginally). In practice, given a TTO procedure that may have fairly large steps in the amount of time sacrificed, observing a value of 1 for another health state would simply indicate that the disutility of some health problem relative to 11111 was so small that the reductions in lifetime shown in the TTO exercise all were too high to accept and 1 was a better approximation of the value than any other implied value. It may also reflect imprecise preferences. Under the assumption of weak monotonicity, people would value a health state A that is better in one domain and worse in none than health state B, as at least as good as health B. Then, a state like 11211 could be seen as equivalent in terms of utility value to 11111. Still, values above 1 would not be possible.

Relaxing the monotonicity assumption, so that a state with strictly more health problems could be preferred over a health state with strictly fewer health problems (e.g. preferring 11211 over 11111) is uncommon. In theory, this could be possible (even if such preferences may be unlikely) and would need to be observed already in a ranking exercise. Relaxing the assumption of monotonicity has several disturbing implications, including that the TTO exercise (which sets perfect health as the upper limit by design) becomes invalid and that some of the common data quality checks (i.e. assessing whether TTO responses are logically ordered given the health state profiles) can no longer be performed. We do not assume the valuation set authors imply this theoretical possibility, as their response indeed seems to confirm.

(ii) The approach to heteroscedasticity and heterogeneity

We agree with the EEPRU team (Hernández-Alava et al., 2018) that accounting for non-response is different from accounting for heteroscedasticity. The exact influence of the current specification on parameter-estimations versus potential other specifications remains unclear from the current exchange between the valuation set authors and the EEPRU team, however. Moreover, as a more general point, it is unclear which influence would be viewed as ‘problematic’. Without further sensitivity analyses, using different approaches to these issues, it is not possible to provide any clear indication of the nature and magnitude of the problem.

(iii) Possible conflict between distributional assumptions for TTO and DCE

First of all, the exchange between the two groups does not add much clarity regarding the problem at hand. The described issue is viewed by the valuation set authors as a misinterpretation, writing that the EEPRU team “…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.” (Devlin et al., 2018b)

Besides some of the statistical issues and distributional assumptions, the meaning of combinations of TTO and DCE responses is important to consider. TTO responses are placed on a defined utility scale, but the scale underlying a DCE model is not directly defined. As the EEPRU team note: “In particular, there is no reason for them to be comparable with the scale of the TTO model.” (Hernández-Alava et al., 2018) Although the TTO and DCE valuations are correlated, and arguably attempt to measure the same general construct of health state preferences, as illustrated to some extent in Appendix C, the independent TTO and DCE models lead to significant differences in parameter estimates, as shown in Feng et al. (2016) and referenced by the EEPRU team (Hernández-Alava et al., 2018). The question becomes how to combine and whether any combination is better in reflecting preferences than a separate model. This is an important issue that goes beyond ‘only’ distributional assumptions. An important, illustrative question in this context would be: would one opt for the use of a hybrid model if the TTO data did not give any cause for concern? If so, why?

(iv) Prior distributions in the model: justification, information and sensitivity.

The choice and impact of priors indeed needs to be properly transparent and justified. The exchange between the EEPRU team and the valuation set authors emphasizes this. While the information and justification could have been more extensive in original reports, we feel that the choice for priors has been sufficiently justified and the impact on outcomes highlighted by the authors in the response, also suggesting that the Wishart prior used was not very informative or influential. This issue especially seems to relate to reporting. The convergence issue was previously addressed, and we share the opinion expressed by EEPRU that any convergence failure should be reported clearly, including potential causes and consequences.

Q5. Should resource allocation decisions use the 5L valuation set for England?

Before providing our answer to this question, let us briefly recap our view. The current EQ-5D-5L valuation set was based on a large study, which yielded data with quality issues, including interviewer effects, few observations of values below dead, and protocol violations. These quality issues may have strengthened the wish to combine DCE and TTO data, and seem to have triggered the development of a new EuroQoL valuation protocol. According to this new EuroQoL protocol, much of the gathered data would have been discarded. Notwithstanding these issues, the valuation set authors have been resourceful in creating a valuation set from the available data, combining TTO and DCE data in complex modelling exercises, necessarily requiring a number of assumptions and choices. Both the data and some of the methods were subsequently criticised, based on a thorough quality assessment study performed by EEPRU.

Whether the gathered data reflect health state preferences in the general public in England is, in our opinion, doubtful given the mentioned issues with the data. These quality issues have been confirmed convincingly, by the valuation set authors, the EEPRU team, and in the EuroQoL report (2019), and these issues do not seem to be contested.

The authors have been innovative in dealing with this data in order to come to a valuation set. It is more difficult to conclude something about whether the resulting valuation set reflects health state preferences of the general public in England. This cannot be readily confirmed or rejected using the existing data. Also comparing to other valuation sets implies taking the latter as a norm and, even when not strikingly different, cannot exclude the possibility of misspecification. Not all of the highlighted and debated choices made by the authors may have been equally influential. Even when this is the case, one may question whether or not any influence (also given the gathered data) improves or deteriorates the adequacy of the resulting valuation set in describing health state preferences. Again, this is difficult to judge and cannot be confirmed or rejected with the current data set. Some of these choices may have based on or evaluated with pre-existing notions of a good valuation set (e.g. having higher values for mild health states or more values below 0), which should then be (more) clearly justified.

Whether the current EQ-5D-5L valuation set adequately reflects ‘true’ health state preferences in our opinion is difficult to answer in any case, certainly with the current data. It is likely that any valuation set can be criticised, including potential future ones. However, the combination of poor TTO data quality, which importantly anchors the utility scale, innovative modelling and the lack of golden standards to evaluate the resulting valuation set with, results in a situation that adopting the current valuation set will probably always be surrounded by questions regarding its validity, which may negatively affect the support for decisions based on economic evaluations using this valuation set.

The issue of data quality, especially for the TTO exercise, in our opinion lies at the root of the current debates. This issue of data quality casts serious doubts on the process of deriving a valuation set, in spite of all the innovative work in modelling the data (which may partly be initiated due to data issues). Given this situation, the fact that better data can be obtained fairly straightforwardly, and the fact that new data may also be used to address some of the questions that cannot be answered with just one dataset (including whether better data leads to a significantly different valuation set), and the desire (if not need) to have a valuation set that is free from discussions regarding its validity, **we would not recommend NICE to use the current valuation set.**

We stress that this answer does not imply that we feel the current valuation set necessarily misrepresents health state preferences or even that gathering new data will necessarily result in a significantly different valuation set. However, it would ideally increase the trust in such a valuation set, and start with data that would meet current, higher standards of quality, which something as important as a valuation set for allocation decision in health care deserves. (As a side product, it may allow addressing important methodological issues highlighted in the debate.)

Our recommendation also does not imply that we feel that the EQ-5D-3L valuation set is necessarily better than the current 5L version. However, if deciding to collect new data, we would consider it to be ill-advised, given a clear benefit of consistency, to now change to the current 5L valuation set, only to (expectedly) move to a new one after a short period of time.

Q6a. What action do you recommend to create a 5L valuation set

We briefly highlight our recommendations.

1. We recommend to perform a new valuation study for England.

2. We recommend deciding on a clear governance structure, specifying the distinct roles, competences, responsibilities, and freedoms of the involved parties. If one opts for a 'steering group', its role, composition, and scope, and even name (steering group vs. advisory group) needs to be defined, also in relation to the researchers responsible for performing the study. It is advisable to anticipate potential conflicts, power imbalances, and tensions between parties and attempt to make arrangements in dealing with these (e.g. through a publication plan and agreeing on publication freedom).

3. The valuation study could be arranged in two phases: (i) data collection and (ii) modelling a valuation set, with phase (ii) starting, once it is established that the gathered data is of sufficient quality to proceed.

4. A crucial part of the new valuation study relates to new data collection (phase (i)). It is recommended to follow the most recent quality control protocol of the EuroQoL group in collecting the data.

5. We recommend to set explicit, deliberated, realistic criteria by which to judge the resulting data before starting phase (i), i.e. in the overall research protocol. Given that new data will never be perfect, such criteria are ideally based on existing, accepted, similar data-sources, e.g. from other countries having used the new EuroQoL protocol. This prevents setting the bar too low or raising it too high.

6. For phase (ii), a clear analysis plan can be formulated and evaluated through quality control. Analyses and modelling require choices that to a certain degree can be seen as normative or 'arbitrary'. It is good research practice to define an analysis protocol prior to data collection and analysis.

Hence, we recommend a new valuation study and consider it important to decide on processes and criteria to guide analytical and design choices and potential subsequent evaluation or assessment, to avoid (as much as possible) potential future debates about a final valuation set.

Q6b. In the interim […] should resource allocation decisions in England be based on the existing 5L valuation set for England?

As indicated above, we do not recommend this, mainly for reason of consistency.

Conclusion and recommendation

In this report we have addressed the questions raised by NICE concerning the EQ-5D-5L valuation set developed for England. We have emphasised the difficulty of addressing the most pressing question: does the current EQ-5D-5L valuation set adequately capture health state preferences of the English population? This question cannot be directly answered, since no golden standard exists against which a valuation set can be evaluated.

In absence of a golden standard to prove the valuation set reflects preferences, one normally considers aspects like the process, quality of the data, modelling techniques, and the resulting outcomes. These issues have been addressed in this report.

The issue of data quality is highly important in this context. There are clear indications of problems in the data obtained in the EQ-5D-5L valuation study, especially in relation to the TTO data. Indeed, using the current quality standards of the EuroQoL group, a large part of this data would have been discarded.

Although they may have affected modelling choices, these data issues do not necessarily translate into issues with the valuation set. The authors of the valuation set have used innovative approaches to arrive at the current valuation set, mitigating some of the data problems by combining TTO and DCE data. Whether this fully counters the data problems, and whether the final modelling approaches can be fully justified, remains a matter of (ongoing) debate and also depends on the frame of reference.

Given this situation, the fact that better data can be obtained fairly straightforwardly, which may also be used to answer some of the open questions, **we therefore recommend NICE to perform a new valuation study**.

We have provided some general recommendations regarding how to proceed with such a new study, should NICE decide to follow our recommendation.

We end this report by emphasising that our recommendation does not imply that the current valuation set misrepresents health state preferences or even that gathering new data will necessarily result in a significantly different valuation set. However, it would ideally increase the trust in such a valuation set, and start with data that would meet current, higher standards of quality, which something as important as a valuation set for allocation decision in health care clearly deserves and researchers should strive for.

Securing a valuation set with stronger support in our opinion is in the interest of all involved parties. Not in the least NICE and the English population, given its intended use in decisions on the allocation of scarce health care resources.

Acknowledgment and disclosure

This work was commissioned by NICE. (NICE received an unrestricted grant from the EuroQol Research Foundation to cover the costs of the project.) The data for the published valuation set are owned by OHE.

Werner Brouwer was the commissioned expert and author for this report. The content of the report fully reflects his opinions and recommendations. He received valuable (technical) advice and support from dr. Arthur Attema and dr. Matthijs Versteegh, who both agree with the opinions and recommendations made in this report. He also gratefully acknowledges helpful comments (linguistic and clarification) from NICE.

As disclosure, it is noted that Matthijs Versteegh is a member of the EuroQoL research foundation and was an author of the Dutch EQ-5D-5L valuation study.

References

Attema, A.E., Brouwer, W.B.F., 2009. The correction of TTO-scores for utility curvature using a risk-free utility elicitation method. Journal of Health Econonomics 28, 234–243

Attema, A.E., Brouwer, W.B.F., 2012. The way that you do it? An elaborate test of procedural invariance of TTO, using a choice-based design. European Journal of Health Economics 13: 491–500

Attema, A.E., Brouwer, W.B.F., 2014. Deriving time discounting correction factors for tto tariffs. Health Economics 23(4): 410-425

Bleichrodt, H., 2002. A new explanation for the difference between time trade-off utilities and standard gamble utilities. Health Economics 11: 447-456.

Bleichrodt, H., et al., 2003. A consistency test of the time trade-off. Journal of Health Econonomics 22, 1037–1052

De Bekker-Grob E.W., et al., 2012. Discrete choice experiments in health economics: a review of the literature. Health Econ 21:145–72

Devlin N et al. 2018a. Valuing health-related quality of life: An EQ-5D-5L value set for England. Health Economics 27(1): 7-22

Devlin N et al. 2018b. Response to: Quality review of a proposed EQ-5D-5L value set for England [online].

Drummond, M.F., et al., 2015. Methods for the economic evaluation of health care programmes. Oxford university press

EuroQoL Office. QC report for England. EuroQol, Rotterdam, The Netherlands. March 2019, unpublished report

Feng Y et al. 2018. New methods for modelling EQ-5D-5L value sets: An application to English data. Health Economics 27(1):23-38.

Hernandez Alava M et al. 2018. Quality review of a proposed EQ-5D-5L value set for England. EEPRU report [online]

Hernandez Alava M et al. 2019. EEPRU response to comments from the 5L valuation team (Devlin et al). EEPRU report

Janssen et al., 2018. Is EQ-5D-5L Better Than EQ-5D-3L? A Head-to-Head Comparison of Descriptive Systems and Value Sets from Seven Countries. PharmacoEconomics 36(6): 675-697

Jones, A.M. (ed). The Elgar Companion to Health Economics. Edward Elgar, UK, 2006

Lancaster KJ, 1966. A new approach to consumer theory. J Polit Econ 74:132–57

Lipman, S.A. et al., 2019. QALYs without bias? Non-parametric correction of time trade-off and standard gamble utilities based on prospect theory. Health Economics, in press

Manski, F. (1977). The Structure of Random Utility Models. Theory and Decision, 8, 228-254

McFadden, D (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), Frontiers in Econometrics (pp. 105-142). New York: Academic Press

NICE 2018. Position statement on use of the EQ-5D-5L valuation set for England [online]

Rakotonarivo O.S., et al., 2016. A systematic review of the reliability and validity of discrete choice experiments in valuing non-market environmental goods. J Environ Manage. 183: 98-109

Ramos-Goñi JM et al. 2017. Quality Control Process for EQ-5D-5L Valuation Studies. Value in Health 20(3):466-473.

Ramos-Goñi JM et al. 2018. Handling Data Quality Issues to Estimate the Spanish EQ-5D-5L Value Set Using a Hybrid Interval Regression Approach. Value in Health 21(5):596-604.

Shah K. et al. 2014. Improving the quality of data collected in EQ-5D-5L valuation studies: a summary of the EQ-VT research methodology programme. 31th Scientific Plenary Meeting of the EuroQol Group, Stockholm, Sweden [online]

Stolk E. et al. 2019. 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 (22), 23-30.

Versteegh, M., et al., 2016. Dutch Tariff for the Five-Level Version of EQ-5D. Value in Health 19, 343–352

Appendix A: Questions for experts

Please ensure that your answers are succinct and use language that is comprehensible by a non-specialist. Your report will be published on the NICE website. If members of your research team provide technical support to you when answering these questions, please acknowledge their input in your report and ensure they submit a declaration of interests form.

Some questions are deliberately open-ended. Please answer as comprehensively as possible in the time available. If you would have liked to do further analysis but were unable to do so, please describe the further analysis briefly and explain why it would be informative.

Data quality

1. Do the valuation set data reflect the preferences of the public in England adequately?
   Explanatory note: concerns have been raised about interviewer effects, whether the respondents understood and engaged with the task, and the number, nature and distribution of possibly inconsistent responses in the time trade-off (TTO) data. There is a lack of scientific consensus about how to define an 'inconsistent response' in TTO tasks, a question that we do not seek to resolve. Instead, we request advice on whether the data quality issues raise concerns about the validity of the data set.

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)?

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). Does the modelling approach chosen by Devlin et al. (2018) and Feng et al. (2018) adequately account for the characteristics of the data?

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 resource allocation decisions in England? In particular, please consider the 4 concerns raised in the EEPRU quality assurance report, listed in the table [see table 1 of the document showing the questions set by NICE].

Conclusions and recommendations

5. In your opinion, should resource allocation decisions in England (including NICE evaluations) use utility values derived using the 5L valuation set for England?

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 resource allocation 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 resource allocation decisions in England be based on the existing 5L valuation set for England?

Appendix B - Simple regression on TTO data

Table 1 shows the parameters of a simple (naïve) linear regression using the TTO data, intended only to illustrate the location of the worst possible health state on the utility scale, if directly estimated on the TTO data (through linear regression). The regression is run on the subset of respondents that meet the inclusion criteria as provided in the R-code by the authors as supplied to us.

The predicted value for 55555, following the assumptions of ordinary least squares regression, is -0.08.

Table B1. Linear regression using TTO data

Predictors	Estimates of yTTO	CI	P
(Intercept)	0.88	0.85 – 0.92	**<0.001**
factor(TTO_MO)2	-0.03	-0.06 – 0.01	0.099
factor(TTO_MO)3	-0.06	-0.10 – -0.03	**<0.001**
factor(TTO_MO)4	-0.16	-0.19 – -0.12	**<0.001**
factor(TTO_MO)5	-0.20	-0.23 – -0.16	**<0.001**
factor(TTO_SC)2	-0.04	-0.07 – -0.00	**0.024**
factor(TTO_SC)3	-0.05	-0.09 – -0.01	**0.008**
factor(TTO_SC)4	-0.10	-0.13 – -0.06	**<0.001**
factor(TTO_SC)5	-0.16	-0.20 – -0.13	**<0.001**
factor(TTO_UA)2	-0.05	-0.08 – -0.01	**0.005**
factor(TTO_UA)3	-0.07	-0.10 – -0.03	**<0.001**
factor(TTO_UA)4	-0.13	-0.16 – -0.09	**<0.001**
factor(TTO_UA)5	-0.14	-0.17 – -0.11	**<0.001**
factor(TTO_PD)2	-0.03	-0.06 – 0.00	0.057
factor(TTO_PD)3	-0.06	-0.09 – -0.02	**0.003**
factor(TTO_PD)4	-0.24	-0.27 – -0.21	**<0.001**
factor(TTO_PD)5	-0.25	-0.29 – -0.22	**<0.001**
factor(TTO_AD)2	-0.06	-0.09 – -0.02	**0.001**
factor(TTO_AD)3	-0.10	-0.14 – -0.07	**<0.001**
factor(TTO_AD)4	-0.23	-0.27 – -0.20	**<0.001**
factor(TTO_AD)5	-0.21	-0.25 – -0.18	**<0.001**

Observations	9140
R^2/R^2 adjusted	0.252/0.251

Appendix C Exploration of lower end of the scale

This Appendix aims to illustratively explore the valuations of health states on the lower end of the utility scale, making use of both DCE and TTO data. As indicated in the EuroQol (2019) report, protocol violations resulted in underutilization of the 'worse than dead' protocol in the TTO exercise, especially affecting the range of the scale implied by the TTO values (as highlighted in Appendix B) and the valuations of poor health states.

The modelling exercise of the valuation set authors may have overcome this problem (combining TTO valuations with DCE data), but we wished to explore this a bit further. The results of this illustrative analysis are reported here.

In short, we used the DCE data and TTO data (for the UK and, as a comparator, for the Netherlands) to first estimate the linear relationship between the DCE predicted values and the observed TTO values for a subset of the data, i.e. health states with a misery index <13 and therefore fairly good health states.[5] We selected this subset, as the valuations of these health states arguably were not affected by above mentioned protocol valuations (as they would not often be seen as worse than dead). Subsequently, the estimated relationship was used to predict TTO valuations of worse health states (misery index >13) based on the DCE data. Analyses were run in RStudio.

In order to do so, we used the data set that resulted from following exclusions R code supplied by OHE, similar to appendix B. Moreover, Feng et al., 2018 (table 2) provide a 5 parameter DCE model. We used the coefficients of that model to estimate the DCE predicted values for the 86 health states used in the TTO exercise, allowing to estimate the linear relationship between the DCE predicted values and the observed TTO values for the subset of health states with a misery index <13. This linear relation is provided in table C1 for England.

[5] The misery index is the sum of the levels of the health state (i.e. 55555 = 5+5+5+5+5 = misery 25, while 11111 = 1+1+1+1+1 = misery 5).

Table C1. Linear prediction of English TTO values (ytto) with misery index <13 from DCE values

Predictors	Estimates	CI	p
(Intercept)	1.23	1.18 – 1.27	**<0.001**
dcepredicted	-0.19	-0.20 – -0.17	**<0.001**
Observations	2460		
R^2/R^2 adjusted	0.161/0.161		

A similar procedure was followed for the Netherlands, giving the results shown in Table C2.

Table C2: linear prediction of Dutch TTO values (Dutch.TTO_util) with misery index <13 from DCE values

Predictors	Estimates	CI	p
(Intercept)	0.75	0.72 – 0.79	**<0.001**
dcepredictedNL	0.17	0.16 – 0.18	**<0.001**
Observations	2634		
R^2/R^2 adjusted	0.336/0.336		

These results were used to generate predicted TTO values for all health states, also those with a misery index above 13 for both England and the Netherlands, and are shown in Figures C1 and C2, respectively. In these Figures, TTO values are plotted in relation to the 'misery index' of the health state. The boxplot follow the standard Tukey representation. From misery value 17, the boxplot show that the large majority of valuations are above 0 for England.

These Figures provide some interesting insights.

First, for England, mean TTO values does not follow the pattern of the extrapolated DCE values. Especially for worse health states, the extrapolated DCE values are lower than observed TTO values. Moreover, the value for state 55555 based on the extrapolated DCE data is -0.22 (compared to -0.08 in the regression shown in Appendix B), which is similar to the value of a hybrid model without left censoring as described by Feng et al (2018). We do note that the predicted value for 55555 based on the extrapolated DCE values strongly depends on the cut off point for the misery index (in this case 13) when scaling the DCE values. The R-squared of the linear relation between DCE predicted TTO values and the observed TTO values is 0.24 for the English data.

Second, Figure C2 shows that for the Netherlands, while the Dutch valuation study used the same valuation protocol as the English study, Dutch data showed a stronger declining trend in TTO values with a relatively large part of the valuations using the 'worse than dead' protocol. The DCE predicted TTO values were better aligned with the TTO data than was the case in the English data. The value for state 55555 of the extrapolated DCE data on the TTO utility scale is -0.444. The TTO based Tobit model used for the Dutch value set predicted a value for state 55555 of -.446. The R-squared of the linear relation between DCE predicted TTO values and the observed TTO values is 0.32.

These comparisons, while clearly not intended to be conclusive, appear to confirm the data issues, especially in the lower parts of the utility scale, with the English data. Moreover, they indicate that DCE data can be used to mitigate these problems, but how to do so optimally (e.g. in our analysis: which cut-off point to use) remains a difficult to answer question. In data less hampered by quality issues, like the Dutch data, such questions are less relevant and different analyses may sooner be used to confirm the findings of a main analysis.

Figure C1:

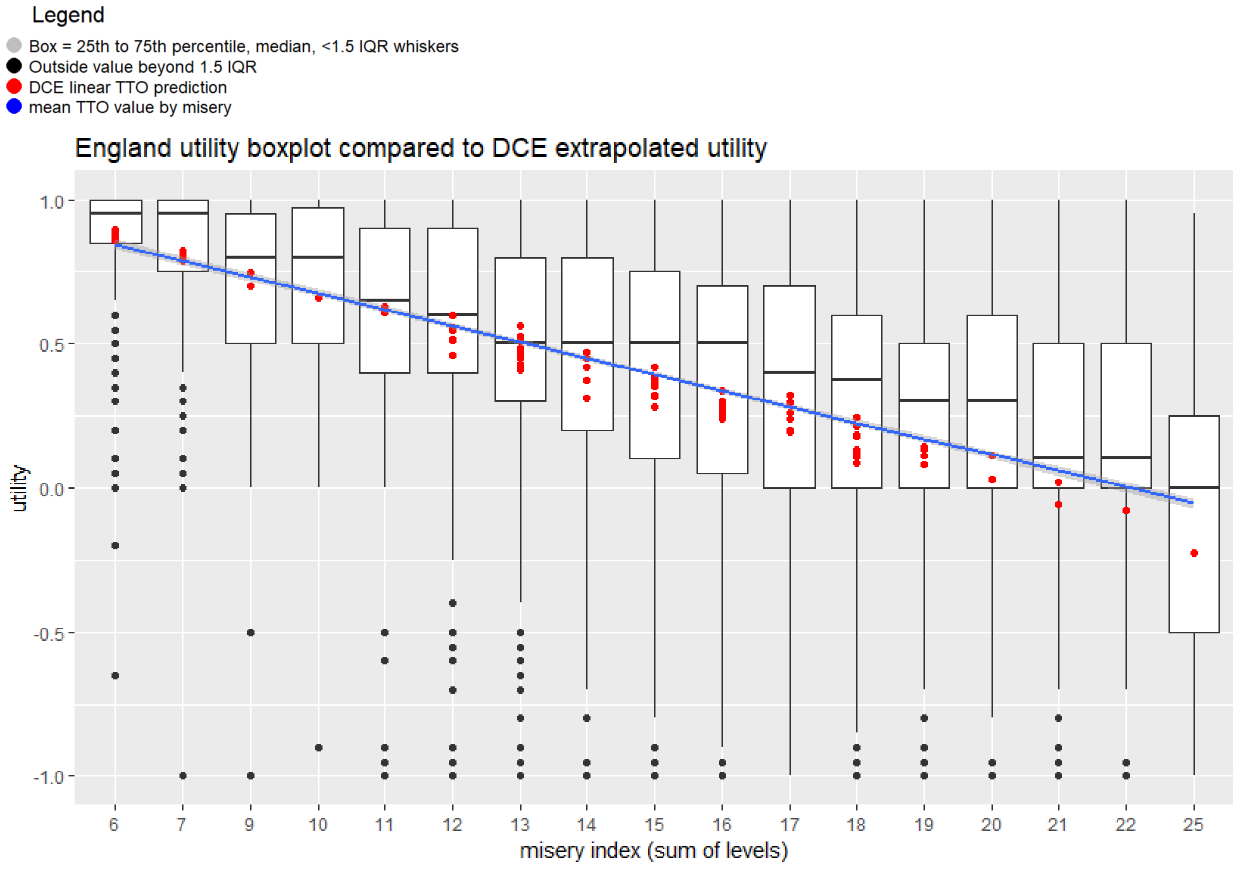
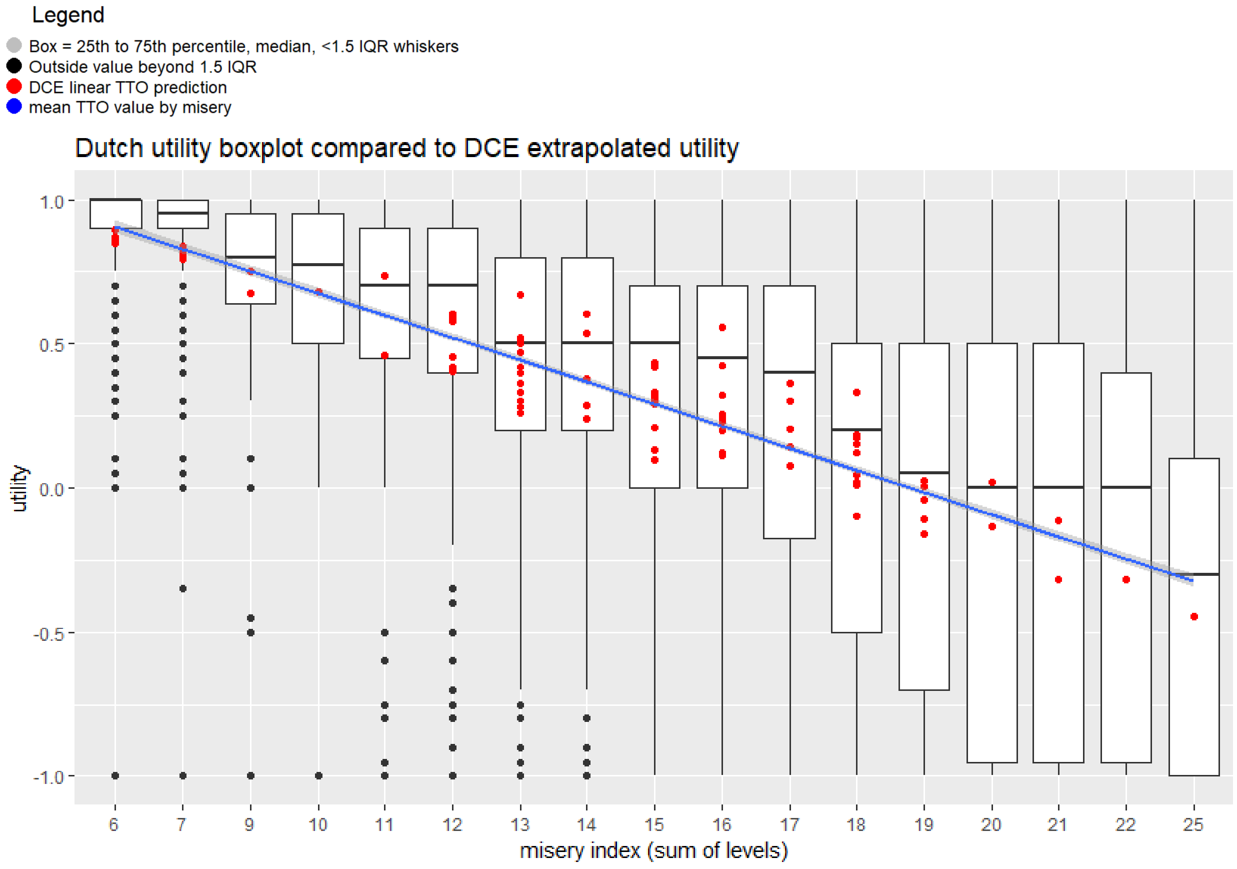

Legend
● Box = 25th to 75th percentile, median, <1.5 IQR whiskers
● Outside value beyond 1.5 IQR
● DCE linear TTO prediction
● mean TTO value by misery
England utility boxplot compared to DCE extrapolated utility
1.0
0.5
0.0
-0.5
-1.0
utility
6
7
9
10
11
12
13
14
15
16
17
18
19
20
21
22
25
misery index (sum of levels)

Figure C2:



*IQR = Inter Quartile Range

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