Responses to referee reports on "Quality review of a proposed EQ-5D-5L value set for England"

Full copies of the reviews received are available on the EEPRU website.

Reviewer 1:

"Hernández-Alava et al. indicate that a difficulty of the EQ-VT protocol lies in the fact that the protocol has changed over time and they define those changed as "[…] superseded successively […]" (p.17). This interpretation must be corrected as the valuation tasks have remained the same across the different version of the protocol. However, later protocol versions pay more attention to the implementation of the valuation tasks with the introduction of a quality control procedure (since EQ-VT 1.1) and a feedback module allowing participants to review their time trade-off (TTO) responses (in EQ-VT 2.0) (Ludwig et al. 2018)."

Response: p. 18 of the report states "Version 1.1 introduced a quality control procedure based on four criteria, three of which relate to the process of the interview" and "Version 2.0 built on v1.1 by the addition of a feedback module to respondents as an internal check on their responses."

No changes made.

"Experimental design: The coverage of possible health states is criticized by Hernández-Alava et al. However, the reasons behind the selected approach by Oppe & van Hout (2017) should be clarified: a certain number of observations per health state is needed to estimate a robust model. Therefore, there is a trade-off between coverage and statistical considerations. The proposed coverage by Hernández-Alava et al. based on severity level is not an appropriate alternative as it has other important limitations: severity is not a good proxy for health state selection as quite different health states can have the same severity level (e.g. 22222 versus 11152)."

Response: We do not "propose coverage". We describe the coverage (lack of) in the 5L design. No changes made.

TTO data: Hernández-Alava et al. analyzed the data quality of the TTO data in detail. Table 2.5 summarizes the proportions of individual participants displaying potentially problematic response behavior. Five of the mentioned criteria are well-chosen and comprehensible (i.e. 1, 3, 5, 10a, 10b). However, the other criteria might also be the (partly) respondent's preferences: According to the preference of respondents an individual threshold can be achieved if a certain dimensional level is reached and an additional increase in severity will result in the same value (ties) (2). There might be people who cannot distinct between more than five values but this can also be low task engagement by the respondent and the interviewer (4, 6). Criteria 7, 8 and 9 can be simply the preference of respondent and therefore conclusions must be made with carefulness. Even though the degree of inconsistencies on a TTO response level (table 2.8) relativizes the extent of inconsistencies in the TTO data, the level of inconsistencies and the clustering of values in the data are concerning. Post hoc analysis of the TTO data of two other EQ-5D-5L valuation studies in Spain and the

Netherlands using EQ-VT 1.0 have shown that a high proportions of inconsistencies and clustering per respondent can be related to the interviewer performance (issues of protocol non-compliance and low engagement with the task) (Ramos-Goni et al. 2016).

Response: Many response types could reflect the preferences of individuals. We refer to potentially problematic responses that we would expect to see few of. Many of the types of responses that the referee seeks to dismiss as acceptable would lead to the data displayed in Table 2.6 to be deemed perfectly legitimate, as Devlin et al implicitly do. The most cursory view of that data shows that position is highly questionable. No changes made.

Discrete Choice (DC) data: Hernández-Alava et al. correctly indicate that no assessment of the inconsistency level of the DC data are possible due to the experimental design (i.e. no dominant pairs included). Possible data quality checks could include response pattern across all DC tasks (e.g. AAAAAAA, BBBBBBB, ABABABA, BABABAB) and analyzing if the proportion of the choice of A or B was correlated to the difference in the severity level between the health states.

Response: This is what is reported in section 2.7

Hernández-Alava et al. judge the modeling choices to be non-transparent. However, the authors of the English EQ-5D-5L value set published an extra paper on the modeling where the main choices are described (Feng et al. 2016). The modelling approaches to the EQ-5D-5L TTO data are innovative in that they account for truncating, censoring, heteroskedasticity, and preference heterogeneity. The modelling advances were driven by considerations obtained from carefully investigating the TTO task and the resulting data, and by matching these to the assumptions underlying the regression models.

Response: There is no Feng (2016) in the references supplied by the reviewer. There is Feng (2017) in that list, but this refers to the paper published in "Health Economics" in 2018. This is one of the two key publications we base our review on and is referenced throughout the report. Feng et al (2016) is a University of Sheffield Discussion paper that is an earlier version of the 2018 publication, and this also informed our review as stated on P63. We also examined the actual code used to estimate the models and report assumptions of the model and indeed the full algebraic form of the model (appendix 2), none of which is reported by either of the Feng et al papers.

The referee claims the modelling approach is "innovative", "assumptions match the data" and results from careful investigation of the tasks. No evidence is provided for any of these claims, nor any rationale that questions the detailed critique we provided. No changes made.

Hernández-Alava et al. criticize that too many TTO data with quality issues are included unmodified in the modeling. However, the level of modification of preference values should be carefully traded-off as the resulting value set should still represent the opinion of the general population.

Response: The referee implies that we think the analysis would have been better had more manipulation of the data been undertaken. That is not our view. Our first recommendation states "A 5L value set for use in policy applications must be based on good quality data."

Even though modeling choices always have a "normative nature", in the following there are comments of three modeling choices made by the English team: (a) Even though the modeling choice for truncating at 1 is well described by the authors of the English value set, the criticism by Hernández-Alava et al. is justified as the maximum TTO value is theoretically bounded at 1 and a correction of the error term at 1 is by definition not inevitably needed. Response: It is unclear if the referee agrees with our approach or not. No justification is given for this view, so we make no changes.

(b) In contrast to the opinion of Hernández-Alava et al., the correction for existing hetereoskedasticity of the TTO data is strongly needed to be prevent biased model parameters.

Response: Table 3.1 row 3 addresses this point. We agree that correcting for heteroskedasticity may be required. Devlin et al claim their model does this. Examination of the code reveals this is not the case and has instead been confused with weighting the sample for nonresponse. These are totally different issues. No changes.

(c) The forced consistency criticized Hernández-Alava et al. is a comprehensible point of criticism. Other national EQ-5D5L value sets, like the Netherlands, included collapsed level in the final value set as this was the

preference of the respondents (Versteegh et al. 2016). Response: We appreciate that the referee accepts this point. No changes.

Hernández-Alava et al. criticize combining TTO and DC data in hybrid model. In accordance to the other mentioned modeling choices, there are advantages and disadvantages for combining the (complementary) information in one value set. However, the authors of the English value set well-justified the inclusion of both data types in a joint model.

Response: We do not criticise the concept of using a hybrid model, we show the problems with this hybrid model. No changes.

Reviewer 2:

No comments require a response.

Reviewer 3:

No comments require a response.

Reviewer 4:

3a) Lack of consistency between the EQ-5D-3L and -5L

As the authors point out, each is a five item instrument and covers the same domains and recall period (today). The primary difference is the number and labeling of the levels, which led to three inconsistencies:

1. Slight problems were not captured by the EQ-5D-3L, and this new level diminishes the frequency of "No problems" and "Some problems" responses.

2. Respondents are often confused by the 5L label orderings, for example, "Severely anxious or depressed" and "Extremely anxious or depressed," which decreased the number of extreme responses

3. The worst possible EQ-5D-5L state includes "confined to bed" and the worst possible EQ-5D-5L state includes "unable to walk about." Therefore, the ranges differ.

The authors are correct that the 3L and 5L are not consistent. The selection between the two versions depends on the relative merits of capturing slight or serious problems, and this choice may lead to different conclusions. I agree with the authors conclusions that "the EQ-5D-3L and -5L cannot be used interchangeably if consistent decision making is required."

Response: To be clear, this was not a conclusion from the current work but arises from previous research in the area.

Deficiencies in the data and statistical methods

The composite time trade-off task was described well in their report. However, the authors did not properly compare it to the TTO task used in the EQ-5D-3L valuation study, namely:

1. Compared to the TTO, the CTTO (CTTO) is inherently different in process and theoretical assumptions. Therefore, they should give different results, which is not addressed by the authors. Response: this was not the aim of our review.

(2) The discrete choice experiment was conducted using a separate descriptive system (no time attribute), and its merger with the CTTO may cause further difference.

(3) The EQ-5D-5L survey excluded parts of the UK and used computers with a strictive iterative procedure and well-known interviewer effects, which may cause differences. The differences between EQ-VT protocols across countries is particularly troublesome.

(4) The analyst of the original EQ-5D-3L study (XXXXX) intentionally changed its worst-than-death TTO responses (>25% of the data) prior to the analysis to improve the face validity of the results. This precedent has be criticised by many. The analyst of the EQ-5D-5L (XXXXX) did not change the data, but manipulated the analysis without justification (changing sample selection, priors, formulae and computational procedures) until it produced the desired results. To this day, no one (even the EQ-

5D-5L analyst) can reproduce the same EQ-5D-5L value set using the same approach and data. These actions by the analysts XXXXXXXXXXXX rather than scientific or methodologic debate. Response: our aim was to investigate the analysis of the 5L valuation data.

Apart from the gaps, the report is a logical and well thought out, and policy makers can have confidence in the quality of the findings presented. Nevertheless, "our analyses do not allow us to identify the reasons why the data suffer these limitations. " It is difficult to point to a list of lessons learned from their report, so here are a few:

1. The EQ-5D-5L is inherently different than the EQ-5D-3L, and each has limitations that impede their use in health measurement (i.e., seemingly obsolete).

2. The composite TTO is inherently different than the original TTO, and each has limitations that impede their use in health valuation (i.e., definitively obsolete).

3. The protocol did not specify the quality control, data management and analysis plan properly in advance of data collection, which led to "numerous serious concerns."

Each of these is a fatal flaw. Ideally, a future health valuation study will use a large representative sample, a coherent descriptive system, a simple choice-based task, and a clear protocol. The EuroQol investigators may argue in favor their instrument and method based on precedent, comparability and convention, but this work is not acceptable due to its limitations.

Response: None required.

Yes, if the recommendations are followed, this will be a severe loss for the EQ-5D-5L study team and the EuroQol Group more generally. Any clinical trial that used the EQ-5D-5L will not be able to submit its evidence to inform UK health policy. This report has the potential to make a large impact.

Within the field, the report adds to the knowledge base as an outside perspective (i.e., non-EuroQol members). Its dissemination within health policy is particularly relevant. The task given to the authors was challenging, and the conclusions are largely well founded. As an analogy in clinical trials, these two interventions (3L and 5L) have different mechanisms of action and both trials were conducted poorly, so the reviewers recommend a new trial.

The authors' recommendations assume that NICE and DHSC has an appetite for a new value set after this experience. Alternatively, they may choose to simply continue to use the value set for the EQ-5D-3L and wait for the next version of the EQ-5D. The authors did not provide any motivation to switch to the EQ-5D-5L (new value set or not). Why bother?

Response: We agree that it is an option to not produce a new 5L version. Our recommendation states only that a new data collection exercise be considered.

Assuming that NICE and DHSC want a new value set, its seems prudent that the new study include a valuation of the EQ-5D-3L and EQ-5D-5L (both descriptive system; head-to-head trial) and not be conducted or reviewed by members of the EuroQol Group or original study team, except in an advisory capacity. Its protocol should be stated in advance and its data and code should be distributed widely. There are many excellent research teams within England who are not members of the EuroQol Group and would be willing to take on a large discrete-choice study of this type.

Response: None required.

1. Involve psychometric and clinical expertise. This flaw is not critical, but would have added to the depth of the comparison, particularly beyond economics.

2. Involve a researcher who conducts health valuation studies regularly (apart from the EQ-5D). None of the study team members have ever produced a value set, so they are more like hecklers. This flaw is not critical, but it impedes the dissemination of their findings. They seem disgruntled.

3. Comparing English and UK values seems inappropriate, and was not addressed. This is not a critical flaw and would not have changed the recommendations.

4. Differences between ICERs may be due to error in the models that has nothing to do with the different value sets. This was not addressed by the authors, but is not a serious flaw. The report would have benefited from a limitation section

Response: This is extremely useful feedback for the development of additional research in this area. We do not see these as limitations relevant to the scope of the quality assurance programme we were asked to undertake.