This is a repository copy of *A systematic review of the methodologies and modelling approaches used to generate international EQ-5D-5L value sets*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/186903/

Version: Accepted Version

Article:
Rowen, D. orcid.org/0000-0003-3018-5109, Mukuria, C. and McDool, E. (2022) A systematic review of the methodologies and modelling approaches used to generate international EQ-5D-5L value sets. PharmacoEconomics, 40 (9). pp. 863-882. ISSN 1170-7690

https://doi.org/10.1007/s40273-022-01159-1

This is a post-peer-review, pre-copyedit version of an article published in PharmacoEconomics. The final authenticated version is available online at: https://doi.org/10.1007/s40273-022-01159-1.

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

A systematic review of the methodologies and modelling approaches used to generate international EQ-5D-5L value sets

Donna Rowen*, Clara Mukuria, Emily McDool
School of Health and Related Research, University of Sheffield
*Donna Rowen, PhD, d.rowen@sheffield.ac.uk
School of Health and Related Research
University of Sheffield
Regent Court, 30 Regent Street
Sheffield, S1 4DA
UK
+44 (0)114 222 0728

Keywords: EQ-5D-5L; preference elicitation; TTO; DCE; hybrid

Running title: Review of methods for EQ-5D-5L value sets

Author contributions:
Concept and design: Rowen, Mukuria, McDool.
Acquisition of data: Not applicable.
Analysis and interpretation of data: Not applicable.
Drafting of the manuscript: Rowen, Mukuria, McDool.
Critical revision of the paper for important intellectual content: Rowen, Mukuria, McDool.
Statistical analysis: Not applicable.

Conflict of Interest Disclosures: Professor Rowen and Drs Mukuria and McDool reported receiving grants from the EuroQol Group during the conduct of the study. Professor Rowen and Dr Mukuria are members of the EuroQol Group.

Funding/Support: This article received funding from the EuroQol Group. The views expressed by the authors in the publication do not necessarily reflect the views of the EuroQol Group.

Acknowledgements: We would like to thank Anne Cunningham and Esther Chanakira for undertaking the systematic search of the literature, and initial screening of titles and abstracts. We would also like to thank Ruth Wong for advising on the literature search and Anju Keetharuth for providing advice on the early stages of the project.

Abstract

Background: The international protocol for valuing EQ-5D-5L focuses upon prescribed preference elicitation methods and design. However, there are no recommendations around sampling, recruitment, data analysis or modelling to generate the EQ-5D-5L value set. This review examines methods used to generate international EQ-5D-5L values sets, across sampling, recruitment, data analysis, modelling, assessing model performance and selection of the recommended value set.

Methods: All published EQ-5D-5L value sets were identified by a systematic search and confirmed by the EuroQol Group. Data were extracted to assess sampling, recruitment, preference elicitation techniques and design, data analysis, modelling, assessing model performance, and vale set selection. These are summarised in tables.

Results: The review included 29 studies with 27 value sets generated using time-trade-off (TTO) data (n=10) only or using a hybrid model that combines TTO and discrete choice experiment data (n=17). TTO data was most commonly estimated using a heteroscedastic Tobit model with censoring at -1, and the hybrid model using a specially created program. Model performance was generally assessed using goodness of fit, logical consistency and significance of coefficients, suitability of the model for the data characteristics and parsimony, though not all selected models account for the specific characteristics of the data.

Discussion: Different assessments of model performance and reasoning are provided for the selection of the recommended value set for a country. This raises the question of valid criteria for selecting a recommended value set and whether this should rely upon researchers' recommendations when value sets are often used to inform public policy.

Key points

- The international protocol for valuing EQ-5D-5L makes no recommendations around sampling, recruitment, data analysis or modelling to generate the EQ-5D-5L value set.
- This review paper of published EQ-5D-5L value sets demonstrates variability in the methods used to generate international EQ-5D-5L values sets, across sampling, recruitment, modelling, assessing model performance and selection of the recommended value set.

- This raises the question of valid criteria for selecting a recommended value set and whether this should rely upon the researchers' recommendation when these value sets are often used to inform public policy.

1 Introduction

The EQ-5D-5L is the five level version of the EQ-5D, a measure of health that is commonly used in countries across the world[1, 2]. The EQ-5D-5L has five dimensions: mobility; self-care; usual activities; pain/discomfort; and anxiety/depression, each with five levels (i.e. response options): no problems; slight problems; some problems; severe problems; extreme problems/unable to do[3]. For each completion of the EQ-5D-5L measure, an "off-the-shelf" utility value from a country value set can be generated that provides a score on a 1-0 full-health to dead scale that reflects the desirability of that EQ-5D-5L health state. These values can be used for a range of different purposes, including to generate Quality Adjusted Life Years (QALYs) for use in economic evaluation (see [4] for an overview), to ultimately inform resource allocation decisions in healthcare systems.

The "off-the-shelf" utility values, called value sets, are usually generated with a representative sample of the general public for a single country, since evidence has shown that preferences for different states of health varies across countries [5, 6]. A range of preference elicitation methods can be used to generate the values including time trade-off (TTO) and discrete choice experiments (DCE). TTO involves an iterative approach where respondents select the number of years in full health that are equivalent to living in a given health state for a fixed longer period, usually 10 years. DCE involves selection of a preferred state from a pair or triplet. There are variations in how each method is used (see [7] for an overview of DCE studies). Furthermore, only a subset of the possible health states described by the measure (3125 possible states for EQ-5D-5L) are included in the preference elicitation tasks. Considerations on how to select this subset differ by elicitation task and studies may vary in the selection of a subset. Evidence shows that the results of preference elicitation studies can differ according to the preference elicitation techniques[8, 9], and even different protocols for the same preference elicitation technique could generate dissimilar results [10-12]. For these reasons, there is an advantage to prescribing an international protocol for valuing measures that are available for use internationally (see [13, 14]). This also enables an assessment of how preferences differ across the countries without any differences being driven by disparities in elicitation method or protocol.

The international protocol for valuing EQ-5D-5L to generate a country value set focuses upon the prescription of preference elicitation methods and the subset of states (or profiles) that are valued, to ensure consistency[12, 14]. The current protocol uses bespoke software and an accompanying interviewer script to collect data. The software has digital presentation of two preference elicitation methods, TTO and DCE. Lead-time TTO is used for states that

are considered worse than being dead, and there is no time period mentioned in the DCE tasks.

Some issues were raised regarding data collected in the bespoke software, EuroQol Valuation Technology version 1 (EQ-VTv1) around data quality, including inconsistent values (where utility increases as health worsens) and clustering of values (such as at 1, 0.5, 0, -0.5, -1)[15-18]. Version 2 (EQ-VTv2) of the software implemented a series of improvements to EQ-VTv1 including enhancements to the TTO practice questions, a 'feedback module', and quality control monitoring and reporting. The number of TTO practice questions was increased and it was ensured that participants are shown the TTO task for states considered as both better than and worse than dead (i.e. the lead-time TTO task) during the practice questions [18]. In the 'feedback module' participants are shown the implied ranking of all health states from their TTO responses, and asked to highlight any where they do not agree with the ranking. These health states are flagged but participants do not provide a new value for them. This module is optional for inclusion in each study, though if selected for a study all participants answer the feedback module.

The EQ-VTv2 quality control reports provide details on each interviewers' performance, their protocol compliance and the elicited TTO values [19]. This process enables poorly performing interviewers to be identified in initial and subsequent rounds of data collection. Indicators of poor protocol compliance by interviewers include: 1) providing no explanation of lead-time TTO in practice questions (i.e. not explaining the task for states considered worse than dead); 2) spending less than the defined minimum time on the initial practice questions; 3) inconsistent TTO ratings where the most severe state described by the EQ-5D-5L (55555) is valued at least 0.5 higher than the lowest valued state; and 4) spending less than the defined minimum time on the TTO tasks in total. Interviewers with more than 40% of interviews flagged due to lack of protocol compliance and data quality are retrained and their prior data is dropped.

The combination of the protocol and quality control features ensure comparability and consistency in the data collected across countries, and ensure high quality data that is compliant with the protocol. It should be noted that there is flexibility within the application of the protocol for applications at the local context. However, within this protocol there are no recommendations around sampling and recruitment of participants (except for recommended data exclusions where there is no protocol compliance), preference elicitation data analysis, modelling to generate the EQ-5D-5L value set, or selection of the recommended country value set.

This review examines the methods used to generate international EQ-5D-5L values sets, across sampling, recruitment, data analysis, modelling, assessing model performance and selection of the recommended value set. This can be informative for the generation and selection of future EQ-5D-5L value sets and the value sets of other measures.

2 Methods

2.1 Search strategy, data identification and extraction

The objective was to identify all published EQ-5D-5L value sets. A wider systematic search was conducted in PubMed and Scopus databases in March 2021 to identify all time-trade-off studies published 2016 onwards for a larger project and the subset relating to EQ-5D-5L were retained in this project (no EQ-5D-5L value sets were published prior to 2016) (see Supplementary Materials Figure A1 for the search strategy). The criteria for inclusion were: reporting a country value set for EQ-5D-5L, or supplementary paper for another paper that reports a country value set for EQ-5D-5L (for example, reporting additional details on the modelling approaches) (see Supplementary Materials Table A1 for detailed inclusion criteria). Initial sifting was undertaken to establish included studies, and these were checked for accuracy with the EuroQol Group who own the copyright for EQ-5D-5L. Two additional value sets that were published after the search was conducted were later added.

Three papers were independently extracted by all reviewers, and the extractions were compared and amended by consensus (differences related to the level of detail extracted). The remaining papers were extracted by one of three reviewers (EM, CM, DR).

2.2 Analytic strategy

Data was extracted on the following:

- The sample: Sample size, sample population, country, recruitment, sampling method, sampling characteristics, whether there was an assessment of sample representativeness, response rate, exclusion criteria in addition to the EuroQol quality control exclusion criteria on the grounds of interviewer quality, sample, mode of administration, and location of data collection.
- Design: TTO health state selection method and number of states, DCE profile selection method and number of choice sets.
- TTO modelling: Model specification, models estimated, interaction effects, other variables included in the model, heteroscedasticity adjustments, heterogeneity adjustments, and robustness analyses.

- DCE modelling: Model specification, models estimated, interaction effects, other variables included in the model, method of anchoring on 1-0 full health-dead scale, heterogeneity adjustments, and robustness analyses.
- Hybrid TTO and DCE modelling: Checks to assess whether TTO and DCE data can be combined, model specification, models estimated, interaction effects, other variables included in the model, heterogeneity adjustments, and robustness analyses.
- Assessments of model performance: Assessments undertaken, model selection criteria, plots or summary of data, assessment of interviewer effects, out of sample validation.
- Selected model to generate the country value set: Model specification, model estimated, reason for selection.

Many studies explained the quality control process and how they used it to identify interviewers and drop interviews from the completed sample data. This information was not extracted since this is a common feature across studies using the EQ-VTv2 protocol and quality control process (as described above). The review did not extract results, since the aim of the review was to examine the methods used and not the findings generated in the different value sets, which is the subject of a recent book[20].

3 Results

3.1 Search results

A total of 1,550 records were identified by the search; 1,000 records after duplicates were removed. Following a title and abstract sift, twenty-seven studies were identified as meeting the inclusion criteria, and these were verified by the EuroQol Group. Two studies were published online after the search was conducted and were added to the included studies, making twenty-nine studies in total (see Supplementary Materials Figure A2 for the PRISMA diagram outlining the selection of studies).

3.1.2 Included studies

Sampling and recruitment of study participants for included studies are summarised in Table 1. These twenty-nine studies cover twenty-seven different country value sets for EQ-5D-5L with countries from Africa, Asia, Europe, North America and South America. The countries with two studies included in the review are England and Spain.

Data was collected using the EQ-VT (versions 1.0[15-17, 21-24], 1.1[25-29], 2.0[30-37] or 2.2[38-45] – where this information was extracted from Devlin et al [20] rather than the studies themselves as this was often not reported in the studies directly) with the exception of one study[46].) Data was collected using computer-assisted personal interview system via face-to-face interviews, with the exception of one study that was conducted via video conferencing[39], and one study via a postal survey[46] (for one study the details were not reported[25]). The face-to-face interviews were often conducted in the participants home (n=7 papers, 6 studies), participants home or public venue (n=3), participants or interviewers home (n=1), participants or interviewers home or public venue (n=1), public venue (n=1), survey centre (n=1), community centre (n=1), convenient location (n=2 papers, 1 study), and for some studies details were not provided (n=9).

The majority of studies followed the EQ-5D-5L valuation protocol for the selection of health states and DCE profiles, which involves the valuation of 86 health states using TTO and use of a DCE with 196 choice sets. Two studies used a different study design [38, 46]. The Swedish valuation study asked respondents to value their own health, not hypothetical health states, to generate experienced utility values[46]. The Peruvian valuation valued 31 health states using TTO (25 from an orthogonal array, plus 5 mild states plus the worst state) and 80 choice sets in the DCE for one part of the sample (n=300), and for the remainder of the sample used only a DCE with 180 choice sets (n = 700)[38].

3.1.3 The samples

All studies involved a general population sample. The EuroQol Group recommend a sample size of 1,000 for a standard EQ-5D-5L value set. The studies were broadly in line with this recommendation ranging from 805[25] to 1,451[31]. One study was an outlier for sample size at 25,967 participants[46], though this Swedish valuation asks respondents to value their own health[46].

Across the 27 value sets, a range of sampling methods were used: quota sampling (n=10); multi-stage stratified quota sampling (n=4); two-stage stratified sampling (n=1); multi-stage sampling (n=1); random sampling (n=2); sampling frame (n=2); stratified quota sampling (n=1); stratified sampling (n=3); no sampling initially followed by purposive sampling (n=1); and for two studies the methods were not reported. The majority of studies selected the sample to ensure representativeness for age and sex as the sample characteristics (n=26), with additional criteria selected in some studies by socioeconomic group (n=5), geographic region (n=15), education level (n=10), urbanicity (n=4) ethnicity (n=3) and religion (n=2). Five studies reported a formal test of sample representativeness, using Chi-square test[33, 37], t-

test[16] and Z-test[36] (and for one study the method was not reported[34]). Participants were recruited in many different ways, with some studies using multiple methods to recruit study participants. For many studies the response rate of participants invited to participate in comparison to interviewed participants was not reported, as depending upon the method of recruitment it is not always possible to calculate this, and where it was reported it was not always clear whether this was a response rate or a rate of completed interviews relative to all interviews conducted.

3.2 TTO analysis

3.2.1 TTO exclusion criteria

Data from interviewers that did not meet the quality control criteria was excluded from TTO data analysis. Additional exclusion criteria was used by some studies (see Supplementary Materials Table A1): TTO values flagged during the feedback module (n=11); respondents with a positively sloped relationship between TTO values and misery index (misery index is the summed score of all levels of the 5 dimensions, e.g. the misery index for state 23142 = 2+3+1+4+2=12) e.g. using regression (n=8); respondents with pits TTO value ≥ TTO value for mildest state (n=4); respondents who valued all states at the same value, except non-traders (i.e. subjects who value all states as 1) (n=2); and respondents who valued all states at the same value (regardless of whether all states were valued at 1) (n=7). Other exclusion criteria were reported including around respondent understanding, and four studies did not apply any additional exclusion criteria. The application of exclusion criteria meant that for some studies the sample size for the modelling differed for TTO and DCE analyses.

3.2.2 TTO model specification

Table 2 reports TTO modelling specifications across the 22 studies where TTO data is analysed and reported separately. Eight studies estimate a model specification of 20 parameter incremental dummies, consisting of 20 parameters with 4 dummy variables for each health dimension where dummies for the increments between consecutive levels are used to capture the disutility associated when moving from one level of the health dimension to another (i.e. from level 1 to level 2, from level 2 to level 3, from level 3 to level 4, from level 4 to level 5). Fourteen studies estimate a model specification of 20 parameter level dummies, where this consists of 20 parameters with dummies for levels 2 to 5 of each of the 5 dimensions, leaving level 1 as the reference category (note that two studies merge adjacent inconsistent levels within a dimension). Both of these methods generate estimates that are equivalent for generating the value set utility (i.e. the level 5 parameter in the level dummies model equals the sum of the four incremental dummies for the equivalent dimension) but the standard errors and p values differ by the model specification. Three

studies also estimate further specifications: an 8 parameter multiplicative model (5 parameters for level 5 for each dimension and 3 parameters for levels 2, 3 and 4 multiplied by the respective dimension parameters); a 9 parameter multiplicative model (5 parameters for level 5 for each dimension and 3 parameters for levels 2, 3 and 4 multiplied by the respective dimension parameters and an additional parameter for level 5 for anxiety/depression and pain/discomfort); and a 5 parameter dimension model (where each variable has a value of 1 to 5 that is equal to the dimension severity level). None of the 22 studies included interaction effects between dimensions in the model specification. Eleven studies included a constant term.

Five studies included other variables in the model specification, where these typically reflect severity captured by having one or more dimensions at the most severe level(s) or the number of severe dimensions.

3.2.3 TTO models

Across the 22 studies analysing and reporting TTO data separately, the Tobit model with censoring at -1 is the most commonly estimated, with studies estimating a Heteroscedastic Tobit model with censoring at –1 (n=5), a random effects Tobit model with censoring at -1 (n=6), or a Tobit model where it is unclear if (or how) the structure of the data with repeated observations for each individual is taken into account (n=4). One study estimates both a pooled homoscedastic Tobit model with censoring at -1, and a pooled heteroscedastic Tobit model with censoring at –1. A range of other models are also reported as being estimated by the authors (though it is not always possible to identify which models have been estimated): OLS (n=7, with robust standard errors in one of these studies); random effects GLS (n=2); linear mixed models (n=2); generalised linear model (n=3); robust estimation; GLS; GLS random intercept model; Tobit-GLS regression; linear heteroscedastic model; random effects nonlinear mixed effects; multilevel regression models; pooled homoscedastic linear model; pooled heteroscedastic linear model; mixed-effects models with random intercepts; random coefficient model; nonlinear mixed model; heteroscedastic model with Bayesian estimation; heteroscedastic censored (at -1) model with Bayesian estimation.

3.2.4 TTO heterogeneity and heteroscedasticity

None of the studies accounted for heterogeneity. Fifteen studies explicitly referred to heteroscedasticity and adjustments made to account for this. The most common approach was to estimate a heteroscedastic Tobit model (n=4) though heteroscedastic models were also estimated using robust estimation and robust standard errors, a linear heteroscedastic model, Bayesian estimation and using a log link with polynomials.

3.2.5 Assessing robustness of TTO models

There was no common approach for assessing robustness. Thirteen studies assessed robustness, most commonly comparing the reported models or preferred model to models estimated on a subset of the dataset or without the excluded participants or responses. This included comparing models estimated using: the dataset with and without data excluded by the feedback module (n=3) [41, 43, 44]; with and without inclusion of all preference data/participants (n=4) [24, 35, 40, 43]; with and without censoring the data (n=2) [24, 38]; excluding data with different levels of inconsistencies (n=1)[26] and with specific inconsistencies (n=1)[16]; with and without non-traders and interviews flagged for quality control (n=1) [35]; with full data and excluding data for some states (n=1)[38]; examination of interviewer effects (n=1) [43]; and estimating models with and without adjustments for heteroscedasticity (n=1)[38]. One study used an ANOVA test and the inclusion of dummy variables to examine the impact of education level [25]. Other studies assessed split sample validation and bootstrap analysis [46], or undertook cross-validation where all observations excluding observations for one health state were modelled, then predicted for that state using models and by excluding observations from a single block, then models fitted and missing block predicted [22]. Another study assessed out-of-sample predictive accuracy over mean TTO health state values using a cross-validation approach by sequentially splitting the dataset into two subsets, fitting the models to one set, using the fitted models to predict the other set, and comparing the predicted and observed values [37].

3.3 DCE analysis

Across the 28 studies, 16 report DCE data analyses (see Table 3). Regarding model specification, studies estimate 20 parameter incremental dummies (n=6), 20 parameter level dummies (n=8), or both (n=2), with no constant term. Thirteen studies estimate a conditional logit model, one study also estimates a heteroscedastic conditional logit model[41], one study estimates a mixed logit model[43], and one study estimates a Zermelo-Bradley-Terry model with a power function[38], and for one study it is unclear[42]. Only one study includes additional terms in the model specification, exploring D1 as the number of dimensions at levels 2, 3, 4, or 5 beyond the first; IJ as the number of dimensions at level J beyond the first; K45 as the number of dimensions at level 4 or 5, and squared of all terms were also introduced to assess nonlinear effects on the dependent variable[23]. One study included lifespan in years [38].

Ten of the studies anchored the DCE estimates. The most common approach was to rescale estimates using theta from the hybrid model (see below). Other approaches were: rescaling

parameter of modelled TTO estimates (n=3); exponential of the hybrid Tobit heteroscedastic model theta parameter (n=1); rescaled using (worst health state DCE – 1)/(worst health state TTO – 1) (n=1); anchoring to worst TTO state by multiplying the DCE dis-score by a constant γ = (1 − mean worst score of TTO for worst state)/worst DCE dis-score for worst state) (n=1); via mapping DCE utility decrements to mean TTO utilities (n=1); using an estimate from the hybrid model (unclear if this is theta) (n=1); and rescaling parameter obtained from a line fit of scatterplot of TTO censored means and DCE mixed logit latent health values assuming linear relationship (n=1). Only one study accounted for heterogeneity[41]. Three studies assessed model robustness: by the impact of moving responses following a pattern (e.g. AAAAAAA from always picking the left-hand option in the DCE pair which is designated as A) (n=1)[16]; including all participants and assessing interviewer effects (n=1)[43]; a number of models to assess sample size, subsets of DCE design, and functional form (n=1)[38].

3.4 Hybrid model analysis

Most of the studies (n= 20 papers, 18 studies) estimated a hybrid model where TTO and DCE data was modelled jointly (see Table 4). Hybrid models assume that the TTO and DCE data is measuring the same utility function but on different scales and that the data can therefore be combined so long as there is a relationship between the data (assuming constant proportionality). Ten studies assessed whether this was the case although this was not always explicitly to assess whether the data could be combined. One study concluded that it was not appropriate to combine the data and no hybrid models were estimated[38]. As with TTO and DCE models, the 20 parameter level models (n=13) and 20 parameter incremental models (n=7) were the most common specifications. Two studies [15, 21] included 5, 9 and 10 parameter models alongside 20 parameter models. The majority of the studies did not include other variables (n=17), with one study including age and gender[30] and one including additional terms (D1, IJ, K45 and squared terms)[17]. One study[21] included other variables but these were not reported. Two studies included a constant term, and three studies estimated models with and without constant terms.

Most of the studies (n= 15) that estimated a hybrid model used a specially created program in Stata, hyreg[47], which combined the TTO and DCE data based on a common likelihood function. This approach was developed in earlier studies[15, 17, 21, 23] in part to address data quality issues identified in the TTO data. The hyreg Stata command offers alternatives including running the models with or without censoring of TTO data including at additional points to the common censoring at -1, accounting for heteroscedasticity in both the TTO and DCE data and treating the TTO data as continuous or interval data. Studies did not always

set out which aspects they tested or used from these options. Two studies implemented the hybrid approach independently using Bayesian approaches [28, 32].

Most of the studies did not account for heterogeneity though two studies used latent groups [15, 21] and one study used a scaling parameter to represent religiosity [32]. Robustness checks were undertaken which replicated robustness checks undertaken when modelling TTO and DCE data separately i.e. based on fitting the models to different groups of data (n=10). One study also assessed robustness based on analysing the data while sequentially removing censoring at -1,0 and 1 as well as not censoring any points of the TTO data [21] and another study assessed the impact of weighting the results using results from a boosted sample to increase under-represented age and gender groups [33].

3.5 Assessments of model performance

Goodness of fit was assessed using AIC (n=6), BIC (n=5) and in some studies was referred to but specific details not reported (see Table 5). A large number of studies assessed the logical consistency of coefficients, where utility does not increase as health worsens (n=18), and some studies examined the significance/insignificance of coefficients (and/or the constant term) (n=11). Four studies assessed DIC (Deviance information criterion).

Studies also looked at the underlying theory of the model and the appropriateness of the model for the characteristics of the data being analysed, where more appropriate models were preferred. Studies often made comparative assessments across the models, and some studies examined the ranking of dimensions across different models (for example by the size of the utility decrement for the worst level of each dimension). Some studies assessed MAE (n=9) and MSE or RMSE (n=4) across different distributions of the scale, and others reported the number of studies with MAE higher than a specified size, for example >0.05 or >0.10. Other studies referred to assessments of predicted values (n=9), for example examining correlation of observed and predicted values or predictions were compared using scatterplots, sometimes across different models or preference elicitation methods. Many studies used parsimony as a criteria to select between different model specifications, for example where models that were more parsimonious were preferred (for example [24, 43]. Five studies had no assessments of model performance or did not report any.

3.6 Selection of value set

Reasons stated for the model selected as the value set (see Supplementary Materials Table A2) should be taken alongside the assessments of model performance since for many studies these were considered jointly. Some studies made an a priori decision around the

model on the basis of using a hybrid model [27, 31], and one study discussed the selection with the EuroQol Executive committee[38]. However, the majority of studies used model performance in terms of goodness of fit, the significance and logical consistency of coefficients, the suitability of the model for the data characteristics, and parsimony.

3.6.1 Recommended value set model and model specification

Ten studies select a recommended value set using only TTO data, and seventeen studies generate a recommended value set using a hybrid model that combines TTO and DCE data (see Table 6). Across the studies selecting TTO for the value set, there is no common approach in the model that is used though three studies use a heteroscedastic Tobit model with censoring at -1. The majority of studies which recommend a hybrid model used the same user-written command to implement the hybrid models and took into account censoring at -1 (Tobit model) while taking into account heteroscedasticity for the TTO data with a logit model fitted for the DCE data. Thirteen studies have a model specification of 20 parameter level dummies, and eight studies use 20 parameter incremental dummies. Three studies include additional terms reflecting severe levels in the model specification [24, 26, 46], two studies merge adjacent inconsistent coefficients [43, 46] and two studies estimate an 8-parameter multiplicative model [22, 37].

4 Discussion

A common international protocol for EQ-5D-5L[12, 14] means that there are commonalities across the studies. This included standardisation across target sample size, preference elicitation methods, data collection methods and the states or profiles presented in the TTO and DCE tasks respectively. Earlier studies indicated that it was not sufficient to rely on this standardisation to ensure valid, high quality data was collected[15, 18, 21], therefore a standardised quality control process was built into the protocol[19]. However, there are no formal requirements around the selection of participants in the preference elicitation study, exclusion of participants or observations (with the exception of where there is not protocol compliance), data analysis and modelling and assessments of model performance, or selection of the recommended country value set from the estimated models.

This means that despite the commonalities of protocol, there is heterogeneity in the published studies. In addition, there are no requirements around the sampling and recruitment of participants, and in particular around the characteristics that the sample is recruited to be representative of, and the methods used vary considerably across the value sets. This heterogeneity is often advantageous, since it means that the protocol can be

adapted to take into account the local context (for example [38, 46]) which impacts directly on the data that is collected. Some of the protocol considerations in the conduct of a study to elicit a value set for EQ-5D-5L, and indeed any preference-based measure, will be taken to meet the local requirements of agencies for that country that will receive economic evaluations with benefits valued using the value set. Furthermore, important characteristics to reflect in the sample are expected to vary by country, and the feasibility and acceptability of different recruitment methods may also differ in the local context.

However, some of the protocol considerations are determined by the researchers conducting the study. For example, the exclusion criteria applied to TTO data (as an addition to quality control criteria implemented in the EQ-VT v2 quality control process) varied widely across studies. This can incorporate the researchers' views around the TTO elicitation tasks and what reflects valid preferences, which could be considered contentious, meaning that guidance on this could be beneficial.

The review highlights the heterogeneity in the models applied to the TTO data in particular, despite the use of a common protocol for 25 of the 27 value sets, and not all studies separately report models estimated using the TTO data alone[15, 21, 23, 29, 31-33]. Whilst the most appropriate models vary for different data, there are a range of models employed across similar datasets, and not all models account for the specific characteristics of the data. The majority of the hybrid models used the same modelling code that was originally developed to allow TTO and DCE data to be combined. Although this was based on the assumption that there is a relationship between TTO and DCE data, the reasons for combining the data were initially to deal with potential problems with earlier TTO data studies [15, 17, 21, 23]. Although combining the data has been shown to improve models even where the data is not problematic[23], studies did not always check whether the data could or should be combined.

Whilst there were differences in the performance assessments used to assess model performance, there was a general approach using goodness of fit (including comparisons of predictions and observed TTO, AIC and BIC, errors including the number of larger errors), logical consistency and significance of coefficients, suitability of the model for the data characteristics and parsimony. However, not all papers report assessments of model performance[27, 31, 33, 36, 42], which may be of concern since it is not confirmed how well the models perform in these studies, though it is unclear whether these were conducted but not reported. Furthermore, some papers (for example [31]) noted that some measures were not appropriate for comparison across all the models including the hybrid models. For

example, AIC and BIC rely on log likelihood and by design, the hybrid models will have larger likelihood than TTO models for instance[17].

Different assessments of model performance and reasoning are provided for the selection of the recommended value set across the different studies. Whilst model performance was often cited, as well as the suitability of the model for the data and parsimony, the validity of the model was rarely discussed in terms of its acceptability for informing public policy despite EQ-5D-5L being widely used to inform healthcare resource allocation decisions. The EQ-VTv2 protocol is implemented with support from the EuroQol Group who provide external assessment and support which often includes discussions regarding the performance of models and selection of an appropriate model. However, none of the studies mentioned that study steering committees, decision makers (for example, for the international agency that the value set will be used to inform health technology assessment submissions), patients or members of the public, for example via public involvement, were involved in the selection of the recommended value set. Though of course not mentioning this does not mean this did not happen, nor does it mean that independent quality assurance is not undertaken on the recommended country value set (for example [48]). The importance and influence of a country value set for informing public policy is likely to mean that international agencies require an acceptable, high quality value set, and therefore their role in the approval or selection of a value set is an issue for consideration. Overall, this raises the question of valid criteria for selecting a recommended value set and whether this should rely upon the researchers' recommendations when these value sets may be frequently used to inform public policy.

Limitations of the review include that the review is limited to EQ-5D-5L, though there are other preference based measures with international protocols including EQ-5D-Y-3L[49] and EORTC-QLU-C10D[13]. We did not use a recommended checklist such as CREATE[50] in this review as our aim was not to assess all the aspects of EQ-5D-5L studies, especially as many of the aspects would be standardised since they follow the same protocol.

This review summarises the methods used in the generation of EQ-5D-5L value sets to date, and is intended to be informative for researchers developing statistical analysis protocols for the generation of value sets of preference-based measures, and policy makers assessing the quality and comparability of value sets to the existing literature. Recommendations arising from this review are to ensure clear reporting and transparency of methods and decisions, since these were not always able to be determined. It is also recommended that there is greater thought around valid criteria for selecting (and potentially validating) a

recommended value set, and whether this should rely (solely) upon the researchers' recommendations where value sets are expected to be commonly used to inform public policy.

Table 1: Characteristics of included studies

Study	Country	Sample size	Sampling method	Sample characteristics selected for representativeness						Recruitment of participants	Data collection location
				Age	Sex	Socioec. status	Geog. region	Education level	Other		
Andrade, 2020[30]	France	1143	Quota sampling	✓	✓	✓				Market research agency, methods not reported	Participant's home
Augustovski, 2016[25]	Uruguay	805	Stratified quota sampling	✓	✓	✓	✓			NR	Participant's home
Augustovski, 2020[38]	Peru	1000	Stratified sampling	✓	✓	✓				NR	NR
Burström, 2020[46]	Sweden	25,867	Sampling frame	✓	✓		✓			Postal survey included in larger survey "Life & Health 2017"	N/A
Devlin, 2018[15]	England	996	Random sampling				✓			Letter (unclear whether door-to-door used)	Participant's home
Feng, 2018[21]	England	1004	Random sampling				✓			Letter (unclear whether door-to-door used)	Participant's home
Ferreira, 2019[31]	Portugal	1451	Random sampling	✓	✓					NR	Participant's home
Finch, 2021[39]	Italy	1182	Quota sampling	✓	✓		✓			Existing panel and local recruiters	N/A

Study	Country	Sample size	Sampling method	Sample characteristics selected for representativeness						Recruitment of participants	Data collection location
				Age	Sex	Socioec. status	Geog. region	Education level	Other		
Golicki, 2019[32]	Poland	1252	Quota sampling	✓	✓		✓	✓		Public locations, personal contact	Participant's home or public venue
Gutierrez-Delgado, 2021 [40]	Mexico	1000	Sampling frame	✓	✓	✓				NR	NR
Hobbins, 2018[33]	Ireland	1160	None initially, followed by purposive sampling	✓	✓					Door-to-door, personal contacts, snowballing	Participant's home
Jensen, 2021[41]	Denmark	1014	NR	✓	✓		✓	✓		Letter, phone, email to panel members	Participant's home or public venue
Kim, 2016[26]	Korea	1085	Quota sampling	✓	✓		✓	✓		NR	Participant's home
Lin, 2018[34]	Taiwan	1000	Multi-stage stratified quota sampling	✓	✓			✓		Referrals from local leaders, persons-in-charge in community centres, owners of street shops, school teachers, pharmacists	Participants home or public venue

Study	Country	Sample size	Sampling method	Sample characteristics selected for representativeness						Recruitment of participants	Data collection location
				Age	Sex	Socioec. status	Geog. region	Education level	Other		
										in community pharmacies, or participants themselves	
Ludwig, 2018[35]	Germany	1158	Quota sampling	✓	✓			✓		Public locations, personal contacts	Participant's or interviewer's home or public venue
Luo, 2017[22]	China	1296	Quota sampling	✓	✓		✓	✓		Public locations and places with restricted access such as schools and factories	NR
Mai, 2020[42]	Vietnam	1200	Multi-stage stratified cluster sampling	✓	✓	✓	✓		Urbanicity by geographic region	Door-to-door	NR
Pattanaphesaj, 2018[27]	Thailand	1207	Multi-stage quota sampling	✓	✓		✓			Identified and invited by an area coordinator	NR
Pickard, 2019[43]	US	1134	Quota sampling	✓	✓				Ethnicity and race	Web-based recruitment, promotion by ISPOR,	NR

Study	Country	Sample size	Sampling method	Sample characteristics selected for representativeness						Recruitment of participants	Data collection location
				Age	Sex	Socioec. status	Geog. region	Education level	Other		
										community platforms including flyers, online, local community centres	
Purba, 2017[36]	Indonesia	1056	Multi-stage stratified quota sampling	✓	✓			✓	Urbanicity; religion and ethnicity used separately	Personal contacts, local leader assistance, public locations, snowballing	Participant's or interviewer's home
Ramos-Goñi, 2017[17]	Spain	1000	Two-stage stratitified sampling	✓	✓		✓			Panel from a market research company	Convenient location
Ramos-Goñi, 2018[23]	Spain	1000	Two-stage stratitified sampling	✓	✓		✓			Panel from a market research company	Convenient location
Rencz, 2020[44]	Hungary	1000	Quota sampling	✓	✓					Personal contacts, organisations (e.g. civil organisations, companies, libraries, senior clubs	NR

Study	Country	Sample size	Sampling method	Sample characteristics selected for representativeness						Recruitment of participants	Data collection location
				Age	Sex	Socioec. status	Geog. region	Education level	Other		
										and sports clubs)	
Shafie, 2019[37]	Malaysia	1125	Quota sampling	✓	✓		✓		Ethnicity, urbanicity	Public locations	Public venue
Shiroiwa, 2016[28]	Japan	1098	Stratified sampling	✓	✓		✓			Market research agency, methods not reported	Survey centre
Versteegh, 2016[16]	Netherlands	979	Stratified sampling	✓	✓		✓	✓		Panel from a market research company	NR
Welie, 2020[45]	Ethiopia	1050	Multi-stage stratified quota sampling	✓	✓		✓		Religion, urbanicity	Demographic Health Surveillance Site (DSS) and personal contacts	NR
Wong, 2018[29]	Hong Kong	1014	NR	✓	✓			✓		Leaflets and posters in community centres	Community Centres
Xie, 2016[24]	Canada	1073	Quota sampling	✓	✓			✓		Phone, publically posted flyers	NR

Notes: NR=not reported, N/A=Not Applicable.

Table 2: TTO regression analyses

Study	TTO model specification	Other variables included in model	TTO models estimated	TTO heteroscedasticity adjustments	Constant term
Andrade, 2020[30]	20 parameter incremental dummies		Heteroscedastic Tobit model with censoring at –1	Modelling the variance, Heteroscedastic Tobit model	No
Augustovski, 2016[25]	20 parameter level dummies	✓[1]	OLS; Robust estimation	In the robust estimation model "the impact of the outliers is reduced and the heteroscedasticity problem is addressed"	Yes
Augustovski, 2020[38]	20 parameter incremental dummies		Heteroscedastic Tobit model with censoring at –1	Heteroscedastic Tobit model	No
Burström, 2020[46]	20 parameter incremental dummies	✓[2]	OLS with robust standard errors; Generalised linear models (GLMs) through binomial distribution with logit link function	Robust standard errors	Yes
Finch, 2021[39]	20 parameter level dummies		Random effects Tobit model with censoring at -1; GLS random intercept model; linear heteroscedastic model;	Linear heteroscedastic model and argues possibly via random effect Tobit	Yes
Gutierrez-Delgado, 2021 [40]	20 parameter level dummies		GLS; Tobit model with censoring at -1 (presumed with random effects); heteroscedastic model with Bayesian estimation (accounting for multiple observations per respondent); Heteroscedastic censored (at -1) model with Bayesian estimation (accounting for multiple observations per respondent).	In Bayesian models, allowing for an exponential relationship between the TTO variance and health state severity	No

Study	TTO model specification	Other variables included in model	TTO models estimated	TTO heteroscedasticity adjustments	Constant term
Jensen, 2021[41]	20 parameter level dummies		Random effects Tobit model with censoring at -1; GLS random intercept model; Interval regression	Yes, details not reported	Yes
Kim, 2016[26]	20 parameter level dummies	✓[3]	Linear mixed models	No	Yes
Lin, 2018[34]	20 parameter incremental dummies		OLS; Tobit model with censoring at –1; GLS; Tobit-GLS regression	Yes, in model though not reported	No
Ludwig, 2018[35]	20 parameter level dummies		Tobit model with censoring at –1	Model tested for homoscedasticity	No
Luo, 2017[22]	20 parameter level dummies; 8 parameter multiplicative model; 9 parameter multiplicative model	✓[4]	OLS; Additive model: Random effects GLS; Multiplicative model: Random effects nonlinear mixed effects	No	No
Mai, 2020[42]	20 parameter incremental dummies		Heteroscedastic Tobit model with censoring at –1; Random effects Tobit model with censoring at -1; Tobit model with censoring at –1	Heteroscedastic Tobit model	No
Pattanaphesaj, 2018[27]	20 parameter level dummies		Multilevel regression models	Yes (in modelling, no details provided)	No
Pickard, 2019[43]	20 parameter level dummies (merged for UA Levels 4 and 5)		Random effects Tobit model with censoring at -1	Investigated modelling the heteroscedasticity of the error term using a log link with polynomials	No
Purba, 2017[36]	20 parameter incremental dummies		Heteroscedastic Tobit model with censoring at –1	Heteroscedastic Tobit model	No
Ramos-Goñi, 2017[17]	20 parameter incremental dummies		OLS[6]	No	Yes

Study	TTO model specification	Other variables included in model	TTO models estimated	TTO heteroscedasticity adjustments	Constant term
Rencz, 2020[44]	20 parameter level dummies		Pooled homoscedastic linear model; Pooled heteroscedastic linear model; Pooled homoscedastic Tobit model with censoring at -1; Pooled heteroscedastic Tobit model with censoring at –1	Yes in models estimated	Yes
Shafie, 2019[37]	20 parameter level dummies; 8 parameter multiplicative model		Mixed-effects models with random intercepts	NR	Yes
Shiroiwa, 2016[28]	20 parameter level dummies		Linear mixed model	NR	Yes
Versteegh, 2016[16]	20 parameter level dummies		OLS; Random effects Tobit model with censoring at -1	NR	Yes
Welie, 2020[45]	20 parameter incremental dummies		OLS: Random effects Tobit model with censoring at -1; Generalised linear model; Random coefficient model;	Yes, details not reported	No
Xie, 2016[24]	20 parameter level dummies; 5 parameters for dimensions (each taking value 1 to 5)	✓[5]	Nonlinear mixed model	NR	Yes

Notes: 20 parameter incremental dummies = 20 parameters with 4 dummy variables for each health dimension where dummies for the increments between consecutive levels are used to capture the disutility associated when moving from one level of the health dimension to another. 20 parameter level dummies = 20 parameters with dummies for levels 2 to 5 of each of the 5 dimensions, leaving level 1 as the reference category. NR = not reported.

[1]The study included variables “D1”, that captured the number of movements away from full health beyond the first; “I2”, the number of dimensions at level 2 or 3 beyond the first; “C3”, the number of dimensions at level 3, 4 or 5 beyond the first; “K45”, the number of dimensions at level 4 or 5; “I45”, the number of dimensions at level 4 or 5 beyond the first; “O2”, which takes the value of 1 if all dimensions are at level 1 or

2 and 0 otherwise; “Z2”, which takes the value of 1 if at least one dimension is at level 2 or 3 and one is at level 4 or 5 and 0 otherwise; and “Z3”, the number of dimensions at level 2 or 3 given that at least one dimension was at level 4 or 5[25].

[2] The study included N2, N3, N4, and N5 terms that reflect when one or more dimensions is at the given level or worse (i.e. N5 equals 1 if one or more dimensions are at level 5) [46].

[3] The study explored: N2, N3, N4 and N5 terms; Mk, Sk, Uk, Pk, and Ak for dimension mobility (M) self-care (S) usual activities (U) pain/discomfort (P) and anxiety/depression (A) which equals 1 if the health state contains level k for that dimension and 0 otherwise; Lk which equals 1 if the health state contains level k at any dimension and 0 otherwise; Ck which reflects the number of dimensions equal to or above level k minus 1 (for k = 3, 4, and 5); and Ik which reflects the number of dimensions equal to level k minus 1 (for k = 2, 3, 4, and 5)[26].

[4] The study also explored N2, N3, N4, N5, as well as i2, i3, i4, i5 which represent the number of dimensions beyond the first at levels 2, 3, 4 and 5 respectively, and each of their square terms, and the square root of the number of movements away from full health [22].

[5] The study explored, “Num45”, which equals the additional number of level 4 or 5 beyond the first level 4 or 5 in any dimension, and dummies where there is any level 4 or 5 within a dimension (ie, MO45, SC45, UA45, PD45, and AD45), and Num45sq (additional number of level 4 or 5 beyond the first level 4 or 5 in any dimension, squared) [24].

[6]This has been assumed to be OLS but is referred to as “linear regression model assuming normal distribution in its errors”[17].

Table 3: DCE regression analyses

Study	DCE model specification	DCE models estimated	DCE Interaction terms e.g. N5	DCE Other variables included in model	DCE anchoring method
Andrade, 2020[30]	20 parameter incremental dummies	Conditional logit	No	No	No
Augustovski, 2020[38]	20 parameter incremental dummies	Zermelo-Bradley-Terry model with a power function	No	Lifespan in years	No
Ferreira, 2019[31]	20 parameter incremental dummies; 20 parameter level dummies	Conditional logit	No	No	No
Finch, 2021[39]	20 parameter level dummies	Conditional logit	No	No	1) Rescaling parameter of TTO model estimates; 2) exponential of the hybrid Tobit heteroscedastic model theta parameter
Jensen, 2021[41]	20 parameter level dummies	Conditional logit and heteroscedastic conditional logit	No	No	No
Kim, 2016[26]	20 parameter level dummies	Conditional logit	No	No	No
Lin, 2018[34]	20 parameter incremental dummies	Conditional logit	No	No	Rescaled using theta from the Hybrid model
Ludwig, 2018[35]	20 parameter level dummies	Conditional logit	No	No	Rescaled using theta from the Hybrid model
Mai, 2020[42]	20 parameter incremental dummies; 20	Logit model	No	No	Rescaling parameter of TTO model estimates

Study	DCE model specification	DCE models estimated	DCE Interaction terms e.g. N5	DCE Other variables included in model	DCE anchoring method
	parameter level dummies				
Pattanaphesaj, 2018[27]	20 parameter level dummies	Conditional logit	No	No	All coefficients divided by a scalar: (worst health state DCE – 1)/(worst health state TTO – 1)
Pickard, 2019[43]	20 parameter level dummies	Mixed logit, random parameter logit that can account for repeated observations	No	No	Rescaled using parameter obtained from a line fit of scatterplot of cTTO censored means and DCE mixed logit latent health values assuming linear relationship
Purba, 2017[36]	20 parameter incremental dummies	Conditional logit	No	No	Rescaled using theta from the Hybrid model
Ramos-Goñi, 2017[17]	20 parameter incremental dummies	Conditional logit	Yes	No	Rescaled using theta from the Hybrid model
Shiroiwa, 2016[28]	20 parameter level dummies	Conditional logit	No	No	1) anchoring to the worst TTO state multiplies the DCE dis-score by a constant γ = (1 − mean worst score of cTTO)/worst DCE dis-score 2) via mapping DCE utility decrements to mean TTO utilities; 3) using an estimate from the Hybrid model
Versteegh, 2016[16]	20 parameter level dummies	Conditional logit	No	No	NR
Welie, 2020[45]	20 parameter incremental dummies	Conditional logit	No	No	Rescaling parameter of TTO model estimates

Table 4: Hybrid model regression analyses

Study	Test appropriateness of combining or comparison of DCE and TTO	Model specification	Models estimated	Other variables included in models	Heterogeneity model or assessments	Robustness assessments
Andrade, 2020[30]	Yes - scatterplots for DCE, cTTO and hybrid model reported	20 parameter incremental dummies	Hybrid model using hyreg. No details on implementation.	Age and sex	No	No
Devlin, 2018[15]	No	5, 9, 10 and 20 parameter models.	Bayesian hybrid regression model with latent classes	No	Latent classes and allowing variance to vary by age groups	No
Feng, 2018[21]	No	5, 9, 10 and 20 parameter models, see TTO model specification.	Bayesian hybrid regression model with latent classes and different assumptions regarding the slope distribution. Censoring and heteroscedasticity	Yes but not reported – interaction terms and N3 term	Latent groups and allowing variance to vary by age groups	Yes – (1) inclusion of excluded data; (2) impact of censoring TTO data by sequentially (one at a time) removing the TTO data censoring at −1, 0, and 1; (3) impact of no censoring
Ferreira, 2019[31]	Scatter plot of DCE and cTTO data	20 parameter incremental dummies; 20 parameter level dummies	Hybrid model using hyreg. Censoring -1 and heteroscedasticity	No	No	Yes – inclusion of excluded data (flagged feedback data, value all health states same, positive slope of regression values and misery index)
Finch, 2021[39]	No	20 parameter level dummies	Hybrid model using hyreg With and without censoring -1 and heteroscedasticity	No	No	Yes – (1) excluding pilot data; (2) excluding flagged states in feedback

Study	Test appropriateness of combining or comparison of DCE and TTO	Model specification	Models estimated	Other variables included in models	Heterogeneity model or assessments	Robustness assessments
Golicki, 2019[32]	No	20 parameter level dummies	Bayesian regression with random parameters, error scaling with fat tails, censoring at − 1, unwillingness to trade in time trade-off by the religious people (for TTO only) and Cauchy distribution in DCE.	No	Scaling parameter for those who are religious	No
Hobbins, 2018[33]	No	20 parameter level dummies	Hybrid model using hyreg Censoring -1 and heteroscedasticity	No	No	Yes – (1) Weighted and unweighted results based on cTTO and DCE results; (2) 'boosted' sample to increase under-represented age and sex groups; (3) removing over-represented age and sex groups
Jensen, 2021[41]	Yes - scatter plots from DCE and cTTO data	20 parameter level dummies	Hybrid model using hyreg Censoring -1 and heteroscedasticity	No	No	Yes – inclusion of excluded data (flagged feedback data)
Lin, 2018[34]	Not formally but hybrid models results are compared to DCE and CTTO results	20 parameter incremental dummies	Hybrid model using hyreg Censoring -1	No	No	No
Ludwig, 2018[35]	Yes agreement of the utility decrements for cTTO and DCE and the	20 parameter level dummies	Hybrid model using hyreg Censoring -1 and heteroscedasticity	No	No	Yes – excluding some participants (equal values TTO, non-

Study	Test appropriateness of combining or comparison of DCE and TTO	Model specification	Models estimated	Other variables included in models	Heterogeneity model or assessments	Robustness assessments
	predicted index values was compared					traders, suspect DCE responses, flagged interviews)
Mai, 2020[42]	Yes – scatter plot of predictions from cTTO and DCE	20 parameter incremental dummies; 20 parameter level dummies	Hybrid model using hyreg Censoring -1 and heteroscedasticity	No	No	No
Pattanaphesaj, 2018[27]	Not formally but comparison of results across different data and correlation of predicted values	20 parameter level dummies	Hybrid model using hyreg Censoring -1 and heteroscedasticity	No	No	No
Pickard, 2019[43]	Yes - bland altman plots, correlation and concordance measures of TTO and DCE values	20 parameter level dummies	Hybrid model using hyreg Censoring -1 and heteroscedasticity	No	No	Yes – (1) inclusion of excluded data (flagged feedback and other exclusions); (2) examination of interviewer effects
Purba, 2017[36]	Yes Comparison of results across different data and correlation of predicted values	20 parameter level dummies	Hybrid model using hyreg Censoring -1 and heteroscedasticity	No	No	No
Ramos-Goñi, 2017[17]	No	20 parameter incremental dummies	Hybrid model - likelihood function obtained multiplying the likelihood functions of a normal distribution for the C-TTO data by the likelihood function of a conditional logit	D1, IJ, K45[1]. Squared of all terms	No	No

Study	Test appropriateness of combining or comparison of DCE and TTO	Model specification	Models estimated	Other variables included in models	Heterogeneity model or assessments	Robustness assessments
			distribution for DCE data. Assumed normality, homoscedasticity			
Ramos-Goñi, 2018[23]	No	20 parameter incremental dummies	Hybrid model using hyreg Censoring and heteroscedasticity. Also incorporated modelling 1) interval responses 2) interviewer-specific violations	No	No	Yes – analysis in separate data (follow-up study)
Shafie, 2019[37]	No	20 parameter level dummies; 8 parameter model with and without constant	Hybrid model using hyreg Homoscedasticity	No	No	Yes - cross validation
Shiroiwa, 2016[28]	Not explicitly stated by probability density functions of cTTO and DCE are compared for overlap	20 parameter level dummies	Bayesian hybrid regression model	No	No	No
Welie, 2020[45]	No	20 parameter incremental dummies	Hybrid model using hyreg Censoring -1 and heteroscedasticity	No	No	No
Wong, 2018[29]	No	20 parameter level dummies	Hybrid model using hyreg Censoring -1. Unclear whether heteroscedasticity accounted for.	No	No	Yes – including excluded data (flagged feedback)

Notes: The hyreg Stata programs allows the TTO and DCE data to be combined with additional modifications including selecting the distribution of the data, censoring data, modelling TTO data as interval data and accounting for heteroscedasticity.

[1] Interaction terms were: D1 which represents the number of dimensions at levels 2, 3, 4, or 5 beyond the first; IJ which represents the number of dimensions at level J beyond the first; and K45 which represents the number of dimensions at level 4 or 5, and others.

Table 5: Assessments of model performance

Study	AIC	BIC	Logical consistency	Insignificant coeffs	MAE	MSE or RMSE	DIC	Predictions	Notes
Andrade, 2020[30]			✓	✓				✓	Scatterplots of predictions by different elicitation methods and observed TTO
Augustovski, 2016[25]	✓	✓			✓	✓			
Augustovski, 2020[38]			✓	✓				✓	Power parameter less than 1
Burström, 2020[46]			✓	✓	✓	✓			Plot of observed and modelled TTO; adjusted R-squared
Devlin, 2018[15]							✓		
Feng, 2018[21]							✓		
Ferreira, 2019[31]									None
Finch, 2021[39]			✓	✓	✓				
Golicki, 2019[32]							✓ (and penalised deviance)		Potential scale reduction factors were monitored to diagnose convergence for individual parameters
Gutierrez-Delgado, 2021 [40]	✓	✓	✓	✓			✓		AIC and BIC not reported
Hobbins, 2018[33]									NR
Jensen, 2021[41]			✓						Goodness of fit
Kim, 2016[26]			✓		✓				Generalised R-squared
Lin, 2018[34]	✓	✓	✓					✓	Goodness of fit; log likelihood
Ludwig, 2018[35]			✓	✓				✓	Value range; scatterplots of predictions
Luo, 2017[22]			✓		✓			✓	
Mai, 2020[42]									NR

Study	AIC	BIC	Logical consistency	Insignificant coeffs	MAE	MSE or RMSE	DIC	Predictions	Notes
Pattanaphesaj, 2018[27]									NR
Pickard, 2019[43]			✓	✓					
Purba, 2017[36]									NR
Ramos-Goñi, 2017[17]	✓	✓	✓						
Ramos-Goñi, 2018[23]									External validation
Rencz, 2020[44]			✓	✓	✓			✓	
Shafie, 2019[37]			✓		✓	✓		✓ (Out of sample)	
Shiroiwa, 2016[28]					✓	✓		✓	Summary statistics of all 3,125 states; Kernel density functions compared including with EQ-5D-3L
Versteegh, 2016[16]			✓	✓	✓			✓	Agreement across TTO and DCE predictions
Welie, 2020[45]	✓	✓	✓	✓					
Wong, 2018[29]			✓	✓ (constant)					Goodness of fit using AIC/n
Xie, 2016[24]	✓		✓					✓	Face validity

Notes: AIC = Akaike information criterion; BIC = Schwarz information criterion; DIC = deviance information criterion, NR = not reported; MAE = mean absolute error ; MSE = mean squared error ; RMSE = root mean squared error.

Table 6: Model and model specification for value set

Study	Model specification selected for value set	Model selected for value set e.g. RE GLS	TTO, DCE or hybrid
Andrade, 2020[30]	20 parameter incremental dummies	Hybrid of Tobit and logit model adjusted for sex and age	Hybrid
Augustovski, 2016[25]	20 parameter level dummies	Robust estimation regression model	TTO
Augustovski, 2020[38]	20 parameter incremental dummies	Heteroscedastic Tobit model with censoring at -1	TTO
Burström, 2020[46]	17 parameter incremental dummies with levels 4 and 5 combined in the mobility, self-care and usual activities dimensions, and N5 term (equals 1 if there is a level 5 in at least one dimension; 0 otherwise)	OLS with robust standard errors	TTO
Devlin, 2018[15]	20 parameter level dummies	Hybrid Bayesian linear regression model	Hybrid
Feng, 2018[21]	NA	NA	NA
Ferreira, 2019[31]	20 parameter incremental dummies	Censored heteroskedastic hybrid model	Hybrid
Finch, 2021[39]	20 parameter level dummies	Hybrid Tobit heteroscedastic without constant model with censoring at -1	Hybrid
Golicki, 2019[32]	20 parameter level dummies	Hybrid model, Bayesian, with error scaling and religion scaling	Hybrid
Gutierrez-Delgado, 2021 [40]	20 parameter level dummies	Bayesian heteroscedastic model with censoring at -1	TTO
Hobbins, 2018[33]	20 parameter level dummies	Hybrid regression model accounting for heteroskedasticity	Hybrid
Jensen, 2021[41]	20 parameter level dummies	Hybrid Tobit heteroscedastic model with censoring at -1	Hybrid
Kim, 2016[26]	20 parameter level dummies with N4 term (equals 1 if the health state contains level 4 or 5 at any dimension; 0 otherwise)	Random-effects (unclear)	TTO
Lin, 2018[34]	20 parameter incremental dummies	Hybrid model assuming TTO censoring at -1	Hybrid

Study	Model specification selected for value set	Model selected for value set e.g. RE GLS	TTO, DCE or hybrid
Ludwig, 2018[35]	20 parameter level dummies	Hybrid model with censoring at -1 and allowing for heteroskedasticity.	Hybrid
Luo, 2017[22]	8 parameter multiplicative model	Random effects nonlinear mixed effects	TTO
Mai, 2020[42]	20 parameter incremental dummies	Hybrid model censored at -1	Hybrid
Pattanaphesaj, 2018[27]	20 parameter level dummies	Hybrid regression model	Hybrid
Pickard, 2019[43]	19 parameter level dummies where usual activities level 4 and 5 are combined	Heteroscedastic Tobit model with censoring at -1 and random effects	TTO
Purba, 2017[36]	20 parameter incremental dummies	Hybrid model	Hybrid
Ramos-Goñi, 2018[23]	20 parameter incremental dummies	Hybrid interval regression model	Hybrid
Rencz, 2020[44]	20 parameter level dummies	Pooled Tobit model with censoring at -1, heteroscedastic, constrained	TTO
Shafie, 2019[37]	8 parameter multiplicative model	Hybrid model	Hybrid
Shiroiwa, 2016[28]	20 parameter level dummies model	Hybrid model with Bayesian approach	Hybrid
Versteegh, 2016[16]	20 parameter level dummies	Constrained random effects Tobit model with censoring at -1	TTO
Welie, 2020[45]	20 parameter incremental dummies	Hybrid model	Hybrid
Wong, 2018[29]	20 parameter level dummies	Hybrid model	Hybrid
Xie, 2016[24]	20 parameter level dummies with MO45 (level 1 where there is level 4 or 5 in any dimension; 0 otherwise), SC45, UA45, PD45, AD45 and Num45sq (additional number of level 4 or 5 beyond the first level 4 or 5 in any dimension, squared)	Nonlinear mixed model	TTO

Notes: Feng, 2018[21] and Ramos-Goñi, 2017[17] do not generate value sets and are hence excluded from this table.

Supplementary Materials

Figure A1: Search strategy

The following search was undertaken using the PubMed and SCOPUS databases:

(((((("time trade off"[Title/Abstract]) OR ("time trade-off"[Title/Abstract])) OR ("time-tradeoff"[Title/Abstract])) OR ("TTO"[Title/Abstract])) OR ("cTTO"[Title/Abstract]))

AND (("2016/01/01"[Date - Publication] : "3000"[Date - Publication]))

The PubMed search returned 588 results. An equivalent search on SCOPUS produced 962 results.

Figure A2: PRISMA diagram outlining the selection of studies

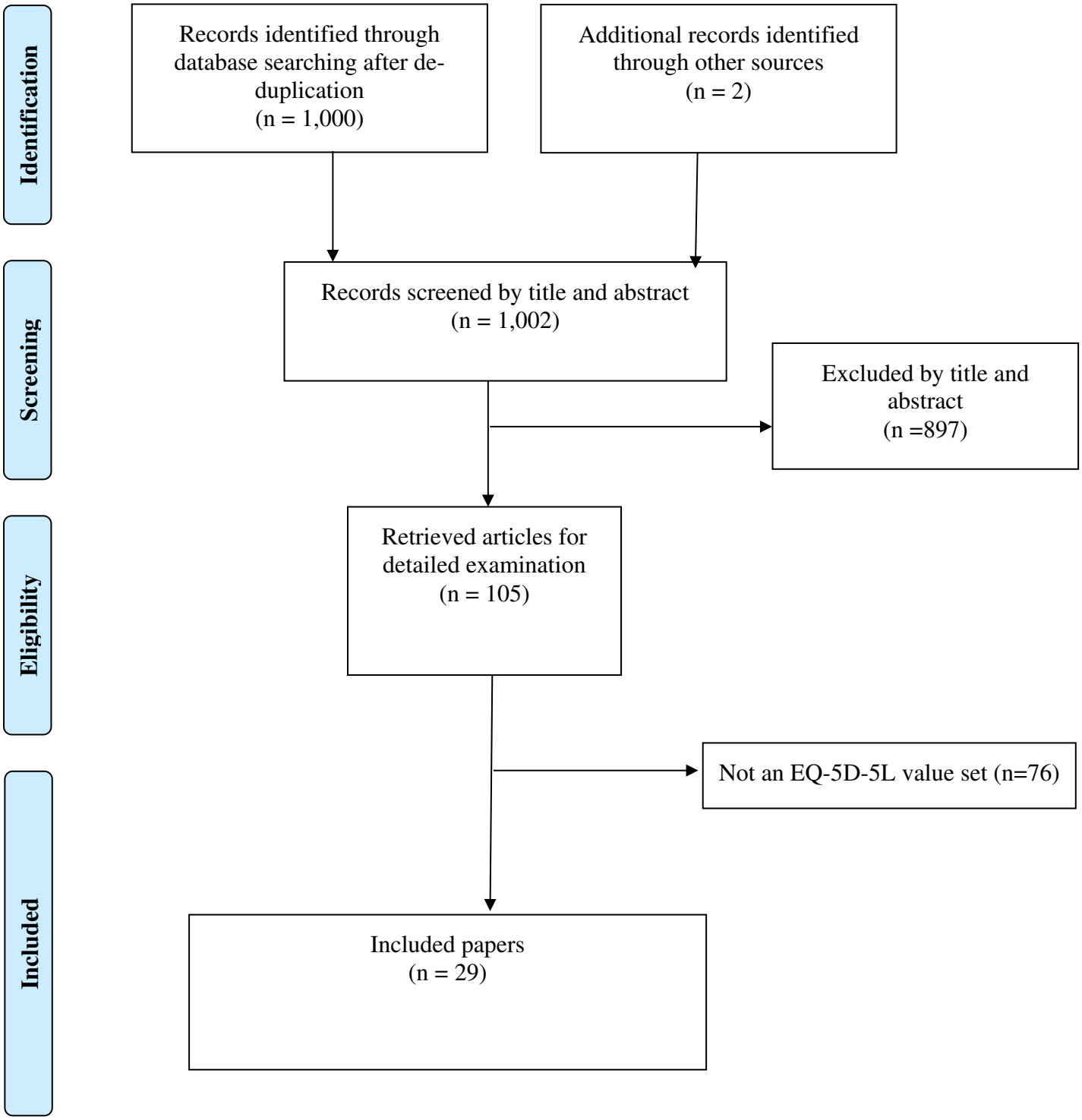

Table A1: Inclusion criteria

	Inclusion criteria	Exclusion criteria	Additional Notes relating to study eligibility
Measure	EQ-5D-5L	Any other measure, including EQ-5D-3L	Paper can also report a value set for other measures
Study design	Valuation study	Any other study design, for example collecting self-report data for EQ-5D-5L with no intention of generating a value set	Studies allowed that report on wider aspects of data analyses or data quality to accompany a value set publication
Language	English	Non-English	

Table A2: Exclusion criteria of TTO data, related to exclusion of participants or a subset of their TTO responses

Study	Respondents with pits TTO value ≥ TTO value for mildest state	TTO values flagged during the feedback module	Respondents with a positively sloped relationship between TTO and misery index (e.g. using regression)	Respondents who valued all states at the same value, except non-traders (i.e. subjects who value all states as 1)	Respondents who valued all states at the same value (regardless of whether all states were valued at 1)	Other
Andrade, 2020[30]		✓				Respondent with pits value > mildest health state
Augustovski, 2016[25]			✓	✓		
Augustovski, 2020[38]						None
Burström, 2020[46]						NR
Devlin, 2018[15]	✓				✓	
Feng, 2018[21]	✓				✓	
Ferreira, 2019[31]		✓	✓		✓	Pits TTO value inconsistencies where is not valued lowest
Finch, 2021[39]						None
Golicki, 2019[32]		✓				
Gutierrez-Delgado, 2021 [40]		✓			✓	
Hobbins, 2018[33]		✓				
Jensen, 2021[41]		✓				
Kim, 2016[26]					✓	

Study	Respondents with pits TTO value ≥ TTO value for mildest state	TTO values flagged during the feedback module	Respondents with a positively sloped relationship between TTO and misery index (e.g. using regression)	Respondents who valued all states at the same value, except non-traders (i.e. subjects who value all states as 1)	Respondents who valued all states at the same value (regardless of whether all states were valued at 1)	Other
Lin, 2018[34]						Respondents who rushed to complete the interview
Ludwig, 2018[35]		✓				
Luo, 2017[22]						Interviewed respondents aged under 18 years
Mai, 2020[42]		✓				
Pattanaphesaj, 2018[27]			✓		✓	Very inconsistent responses e.g. value of 1 for pits and better state had lower value.
Pickard, 2019[43]		✓ (some)				Respondents who didn't understand the task(s), as judged by the interviewer
Purba, 2017[36]		✓	✓			Irrational TTO responses e.g. stating health state as better than full health
Ramos-Goñi, 2017[17]			✓			Respondents who valued all states equal to dead
Ramos-Goñi, 2018[23]			✓			Respondents who valued all states equal to dead
Rencz, 2020[44]						None
Shafie, 2019[37]			✓			
Shiroiwa, 2016[28]						None

Study	Respondents with pits TTO value ≥ TTO value for mildest state	TTO values flagged during the feedback module	Respondents with a positively sloped relationship between TTO and misery index (e.g. using regression)	Respondents who valued all states at the same value, except non-traders (i.e. subjects who value all states as 1)	Respondents who valued all states at the same value (regardless of whether all states were valued at 1)	Other
Versteegh, 2016[16]						Respondents reported by the interviewer as unable to understand task
Welie, 2020[45]	✓				✓	
Wong, 2018[29]		✓	✓	✓		
Xie, 2016[24]	✓					Respondents giving the same or lower value for the very mild health state compared with the majority of the health states that are dominated by the very mild health state within the same block

Table A3: Model selection criteria, rationale or justification

Study	Model selection criteria, rationale or justification
Andrade, 2020[30]	Strong agreement between TTO and DCE data, goodness of fit and logical consistency, and to comply with the representativeness of the general population.
Augustovski, 2016[25]	Four criteria to evaluate the performance of the model: (1) logical consistency of parameters, (2) goodness of fit, (3) prediction accuracy and (4) parsimony. Also considered adjustments for heteroscedasticity and presence of outliers.
Augustovski, 2020[38]	Discussion with the EuroQol Executive Committee.
Burström, 2020[46]	Criteria of consistency (monotonicity), predictive performance (goodness of fit), simplicity of the model (parsimony) and ease of understanding by non-experts in statistics (transparency). Also referred to significance.
Devlin, 2018[15]	Not explicit, selected based on model statistics, logical consistency and allowing values to vary by different levels.
Ferreira, 2019[31]	Prior decision to use both cTTO and DCE data. Decision was reinforced by the ability of this model to estimate consistent and statistically significant coefficients.
Finch, 2021[39]	Performance assessments, MAE and MAE for mild states, accounting for censoring at -1, heteroscedasticity.
Golicki, 2019[32]	Statistical criteria (model fit), pragmatic reasons (what the estimation results are used for) or beliefs concerning how the elicitation tasks work.
Gutierrez-Delgado, 2021 [40]	A priori decision to use TTO data only. Theoretical considerations relating to the characteristics of the TTO data, logical consistency of parameters, significance of parameters, and relevant information criteria (i.e. BIC, AIC, DIC).
Hobbins, 2018[33]	To optimise use of available utility data, though note only one model reported.
Jensen, 2021[41]	Logical consistency and appropriateness of the model.
Kim, 2016[26]	Three criteria: 1) logical consistency; 2) goodness of fit; 3) parsimony (for models with similar performance otherwise) similar MAE.
Lin, 2018[34]	Not explicitly stated beyond reference to performance assessment.
Ludwig, 2018[35]	Agreement between cTTO and DCE data, logical consistency, taking into account heteroskedasticity and highest precision (smaller standard errors)
Luo, 2017[22]	Predictive accuracy and parsimony.
Mai, 2020[42]	Logical consistency and the degree to which models used all the available data.
Pattanaphesaj, 2018[27]	Model selected a priori to address limitations of TTO and DCE data, which are seen as complementary.
Pickard, 2019[43]	Statistical significance, ability to handle left censored data, account for panel data and heteroscedasticity, additional complexities of a hybrid model deemed unnecessary.
Purba, 2017[36]	Implicit that model selected a priori to maximise information from TTO and DCE data, which are seen as complementary.

Study	Model selection criteria, rationale or justification
Ramos-Goñi, 2018[23]	Addresses data quality issues, as well as significance of coefficients and correlation with an external validation model.
Rencz, 2020[44]	Accounting for the censored nature of TTO, accommodating heteroscedasticity, reducing number of illogical and insignificant parameters, predictive accuracy.
Shafie, 2019[37]	On basis of model performance assessments on logical consistency and out-of-sample predictive accuracy.
Shiroiwa, 2016[28]	Not explicit, based on RMSE and proportion of states with smaller errors.
Versteegh, 2016[16]	Logical consistency, significance of parameters and predictive performance.
Welie, 2020[45]	Logical consistency of parameters, goodness of fit and significance levels, argument that hybrid model has highest validity due to combination of elicitation methods
Wong, 2018[29]	Logical consistency of parameters, goodness of fit, maximising data from both methods.
Xie, 2016[24]	Logical consistency and model fit.

Notes: Feng, 2018[21] and Ramos-Goñi, 2017[17] do not generate value sets and are hence excluded from this table.

References

1. Devlin, N. and R. Brooks, *EQ-5D and the EuroQol group: Past, present, future.* Applied Health Economics and Health Policy, 2017. **15**(2): p. 127-137.
2. Richardson, J., J. McKie, and E. Bariola, *Multiattribute utility instruments and their use.*, in *Encyclopedia of health economics*, A.J. Culyer, Editor. 2014, Elsevier: San Diego, USA. p. 341-357.
3. Herdman, M., et al., *Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L).* Qualith of Life Research, 2011. **20**: p. 1727–36.
4. Whitehead, S.J. and S. Ali, *Health outcomes in economic evaluation: the QALY and utilities.* British Medical Bulletin, 2010. **96**(1): p. 5-21.
5. Roudijk, B., A.R.T. Donders, and P.F.M. Stalmeier, *Cultural Values: Can They Explain Differences in Health Utilities between Countries?* . Medical Decision Making, 2019. **39**(5): p. 605-616.
6. Olsen, J.A., A.N. Lamu, and C. J., *In search of a common currency: a comparison of seven EQ-5D-5L value sets.* Health Economics, 2018. **27**(1): p. 39-49.
7. Mulhern, B., et al., *One Method, Many Methodological Choices: A Structured Review of Discrete-Choice Experiments for Health State Valuation.* Pharmacoeconomics, 2019. **37**(1): p. 29-43.
8. Ferreira, L.N., et al., *Do Portuguese and UK health state values differ across valuation methods?* Quality of life research, 2011. **20**(4): p. 609-619.
9. Rowen, D.L., et al., *Estimating a Dutch value set for the paediatric preference-based CHU-9D using a discrete choice experiment with duration.* Value in Health., 2018. **21**(10): p. 1234-1242.
10. Arnesen, T. and M. Trommald, *Are QALYs based on time trade-off comparable?--A systematic review of TTO methodologies.* Health Econ, 2005. **14**(1): p. 39-53.
11. Attema, A.E., et al., *Time trade-off: one methodology, different methods.* European Journal of Health Economics, 2013. **14**: p. 53-64.
12. Oppe, M., et al., *EuroQol Protocols for Time Trade-Off Valuation of Health Outcomes.* Pharmacoeconomics, 2016. **34**(10): p. 993-1004.
13. King, M.T., et al., *Australian Utility Weights for the EORTC QLU-C10D, a Multi-Attribute Utility Instrument Derived from the Cancer-Specific Quality of Life Questionnaire, EORTC QLQ-C30.* Pharmacoeconomics, 2018. **36**(2): p. 225-238.
14. Oppe, M., et al., *A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol.* Value in Health, 2014. **17**(4): p. 445-453.
15. Devlin, N.J., et al., *Valuing health-related quality of life: An EQ-5D-5L value set for England.* Health Econ, 2018. **27**(1): p. 7-22.
16. M, M.V., et al., *Dutch Tariff for the Five-Level Version of EQ-5D.* Value Health, 2016. **19**(4): p. 343-52.
17. Ramos-Goñi, J.M., et al., *Valuation and Modeling of EQ-5D-5L Health States Using a Hybrid Approach.* Med Care, 2017. **55**(7): p. e51-e58.
18. Stolk, E., et al., *Overview, Update, and Lessons Learned From the International EQ-5D-5L Valuation Work: Version 2 of the EQ-5D-5L Valuation Protocol.* Value Health, 2019. **22**(1): p. 23-30.
19. Ramos-Goñi, J.M., et al., *Quality Control Process for EQ-5D-5L Valuation Studies.* Value in Health, 2017. **20**: p. 466-473.
20. Devlin, N., B. Roudijk, and K. Ludwig, *Value sets for EQ-5D-5L: A compendium, comparative review & user guide*. 2022, Springer.

21. Feng, Y., et al., *New methods for modelling EQ-5D-5L value sets: An application to English data.* Health Econ, 2018. **27**(1): p. 23-38.
22. Luo, N., et al., *Estimating an EQ-5D-5L Value Set for China.* Value Health, 2017. **20**(4): p. 662-669.
23. Ramos-Goñi, J.M., et al., *Handling Data Quality Issues to Estimate the Spanish EQ-5D-5L Value Set Using a Hybrid Interval Regression Approach.* Value Health, 2018. **21**(5): p. 596-604.
24. Xie, F., et al., *A Time Trade-off-derived Value Set of the EQ-5D-5L for Canada.* Med Care, 2016. **54**(1): p. 98-105.
25. Augustovski, F., et al., *An EQ-5D-5L value set based on Uruguayan population preferences.* Qual Life Res, 2016. **25**(2): p. 323-333.
26. Kim, S.H., et al., *The EQ-5D-5L valuation study in Korea.* Qual Life Res, 2016. **25**(7): p. 1845-52.
27. Pattanaphesaj, J., et al., *The EQ-5D-5L Valuation study in Thailand.* Expert Rev Pharmacoecon Outcomes Res, 2018. **18**(5): p. 551-558.
28. Shiroiwa, T., et al., *Comparison of Value Set Based on DCE and/or TTO Data: Scoring for EQ-5D-5L Health States in Japan.* Value Health, 2016. **19**(5): p. 648-54.
29. Wong, E.L.Y., et al., *Assessing the Use of a Feedback Module to Model EQ-5D-5L Health States Values in Hong Kong.* Patient, 2018. **11**(2): p. 235-247.
30. Andrade, L.F., et al., *A French Value Set for the EQ-5D-5L.* PharmacoEconomics, 2020. **38**(4): p. 413-425.
31. Ferreira, P.L., et al., *A hybrid modelling approach for eliciting health state preferences: the Portuguese EQ-5D-5L value set.* Qual Life Res, 2019. **28**(12): p. 3163-3175.
32. Golicki, D., et al., *Valuation of EQ-5D-5L Health States in Poland: the First EQ-VT-Based Study in Central and Eastern Europe.* Pharmacoeconomics, 2019. **37**(9): p. 1165-1176.
33. Hobbins, A., et al., *Utility Values for Health States in Ireland: A Value Set for the EQ-5D-5L.* Pharmacoeconomics, 2018. **36**(11): p. 1345-1353.
34. Lin, H.-W., et al., *Valuation of the EQ-5D-5L in Taiwan.* PLoS ONE, 2018. **13**(12).
35. Ludwig, K., J.M. Graf von der Schulenburg, and W. Greiner, *German Value Set for the EQ-5D-5L.* Pharmacoeconomics, 2018. **36**(6): p. 663-674.
36. Purba, F.D., et al., *The Indonesian EQ-5D-5L Value Set.* Pharmacoeconomics, 2017. **35**(11): p. 1153-1165.
37. Shafie, A.A., et al., *EQ-5D-5L Valuation for the Malaysian Population.* Pharmacoeconomics, 2019. **37**(5): p. 715-725.
38. Augustovski, F., et al., *Peruvian Valuation of the EQ-5D-5L: A Direct Comparison of Time Trade-Off and Discrete Choice Experiments.* Value in Health, 2020. **23**(7): p. 880-888.
39. Finch, A.P., et al., *An EQ-5D-5L value set for Italy using videoconferencing interviews and feasibility of a new mode of administration.* Soc Sci Med, 2021: p. 114519.
40. Gutierrez-Delgado, C., et al., *EQ-5D-5L Health-State Values for the Mexican Population.* Applied Health Economics and Health Policy, 2021. **19**(6): p. 905-914.
41. Jensen, C.E., et al., *The Danish EQ-5D-5L Value Set: A Hybrid Model Using cTTO and DCE Data.* Applied Health Economics and Health Policy, 2021. **19**(4): p. 579-591.
42. Mai, V.Q., et al., *An EQ-5D-5L Value Set for Vietnam.* Qual Life Res, 2020. **29**(7): p. 1923-1933.

43. Pickard, A.S., et al., *United States Valuation of EQ-5D-5L Health States Using an International Protocol.* Value Health, 2019. **22**(8): p. 931-941.
44. Rencz, F., et al., *Parallel Valuation of the EQ-5D-3L and EQ-5D-5L by Time Trade-Off in Hungary.* Value Health, 2020. **23**(9): p. 1235-1245.
45. Welie, A.G., et al., *Valuing Health State: An EQ-5D-5L Value Set for Ethiopians.* Value Health Reg Issues, 2020. **22**: p. 7-14.
46. Burström, K., et al., *Experience-Based Swedish TTO and VAS Value Sets for EQ-5D-5L Health States.* Pharmacoeconomics, 2020. **38**(8): p. 839-856.
47. Ramos-Goñi, J.M., et al., *Combining continuous and dichotomous responses in a hybrid model*, in *EuroQol Working Paper Series*. 2016.
48. Hernandez Alava, M., S. Pudney, and A. Wailoo, *The EQ-5D-5L Value Set for England: Findings of a Quality Assurance Program.* Value in HEalth, 2020. **23**(5): p. 642–648.
49. Ramos-Goñi, J.M., et al., *International Valuation Protocol for the EQ-5D-Y-3L.* PharmacoEconomics, 2020. **38**(7): p. 653-663.
50. Xie, F., et al., *A Checklist for Reporting Valuation Studies of Multi-Attribute Utility-Based Instruments (CREATE).* Pharmacoeconomics, 2015. **33**(8): p. 867-77.