CrossMark

ORIGINAL RESEARCH ARTICLE

German Value Set for the EQ-5D-5L

Kristina Ludwig[1,2] · J.-Matthias Graf von der Schulenburg[2] · Wolfgang Greiner[1]

Abstract

Objectives The objective of this study was to develop a value set for EQ-5D-5L based on the societal preferences of the German population. As the first country to do so, the study design used the improved EQ-5D-5L valuation protocol 2.0 developed by the EuroQol Group, including a feedback module as internal validation and a quality control process that was missing in the first wave of EQ-5D-5L valuation studies. *Methods* A representative sample of the general German population ($n = 1158$) was interviewed using a composite time trade-off and a discrete choice experiment under close quality control. Econometric modeling was used to estimate values for all 3125 possible health states described by EQ-5D-5L. The value set was based on a hybrid model including all available information from the composite time trade-off and discrete choice experiment valuations without any exclusions due to data issues. *Results* The final German value set was constructed from a combination of a conditional logit model for the discrete choice experiment data and a censored at -1 Tobit model for the composite time trade-off data, correcting for heteroskedasticity. The value set had logically consistent parameter estimates ($p < 0.001$ for all coefficients). The predicted EQ-5D-5L index values ranged from -0.661 to 1. *Conclusions* This study provided values for the health states of the German version of EQ-5D-5L representing the preferences of the German population. The study successfully employed for the first time worldwide the improved protocol 2.0. The value set enables the use of the EQ-5D-5L instrument in economic evaluations and in clinical studies.

✉ Kristina Ludwig
  kristina.ludwig@uni-bielefeld.de

  J.-Matthias Graf von der Schulenburg
  jms@ivbl.uni-hannover.de

  Wolfgang Greiner
  wolfgang.greiner@uni-bielefeld.de

[1] Health Economics and Health Care Management, Bielefeld University, P.O. Box 10 01 31, 33501 Bielefeld, Germany

[2] Center for Health Economics Research Hannover (CHERH), Leibniz Universität Hannover, Otto-Brenner-Straße 1, 30159 Hannover, Germany

Key Points for Decision Makers
The German EQ-5D-5L valuation study is the first country study to use the improved EQ-5D-5L valuation protocol 2.0 developed by the EuroQol Group, including a feedback module as internal validation, and a continuous quality control.
The study provides evidence that the refined valuation protocol with its quality control process appears to be a solid basis for estimating national EQ-5D-5L value sets.
The resulting German EQ-5D-5L value set, based on the complementary composite time trade-off and discrete choice experiment data in a hybrid model, is recommended as the preferred value set for Germany.
The German EQ-5D-5L value set allows a more refined preference-based health-related quality of life measurement to describe patients' health and enables the use of the EQ-5D-5L in a range of applications such as cost-utility analysis for health care policy and clinical assessment in Germany.

△ Adis

1 Introduction

The three-level version of EQ-5D (EQ-5D-3L) is one of the most commonly used generic health-related quality of life (HrQoL) instruments in Germany and other countries [1, 2]. Recently, an improved version of this instrument was introduced by the EuroQol Group to increase its sensitivity, extending the number of response levels per dimension from three to five, namely EQ-5D-5L (the five-level version of EQ-5D) [3]. This instrument consists of five HrQoL dimensions (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression), with each dimension specifying five levels of severity [no (level 1), slight (level 2), moderate (level 3), severe (level 4), and extreme problems/unable (level 5)], which allows the description of 3125 health states. Available evidence on the comparative performance of EQ-5D-3L and EQ-5D-5L suggests that EQ-5D-5L (1) is a useful improvement upon the measurement properties in terms of reduced ceiling effects, and (2) provides an improved discriminative capacity with a greater ability to detect differences between groups, compared with EQ-5D-3L [4, 5].

In contrast to EQ-5D-3L, currently, no specific set of social health status preference valuations for the German version of the instrument exists to convert each health state described by EQ-5D-5L to a preference-based summary score on a 1 (full health) to 0 (dead) scale, which allows quality-adjusted life year (QALY) calculations. To date, there is only a crosswalk value set available as an interim solution, based on a mapping algorithm between the two instruments [6]. As a result, the EQ-5D-5L currently can only be used with limitations as an alternative for EQ-5D-3L in Germany and there is a strong need for a national value set ('tariff') for EQ-5D-5L.

The introduction of the new instrument was considered as an opportunity to develop an internationally standardized valuation protocol with improved methods for health state valuation and accompanying computer-based valuation software (the EuroQol-Valuation Technology, EQ-VT). This standardized protocol enables comparison of the resulting EQ-5D-5L value sets between countries, in contrast to the EQ-5D-3L value sets where inconsistent study designs and methods were applied to elicit health state preferences [2, 7]. In the first wave of EQ-5D-5L valuation studies applying the first version of the protocol (EQ-VT Version 1.0), major data issues were observed leading to EQ-VT Version 1.1, a comprehensive EQ-VT research program [8], and finally to the improved EQ-VT Version 2.0 [9].

The objective of this study was to estimate a value set for EQ-5D-5L based on the societal preferences of the German population using, as the first country worldwide to do so, the improved valuation protocol 2.0, including a feedback module for the respondents as an internal validity check of their answers, and an accompanying quality control (QC) process.

2 Methods

This study used the most up-to-date EuroQol Group's valuation protocol, the EQ-VT 2.0 [9]. Version 2.0 is a refinement of the protocol that was used in the first wave of EQ-5D-5L valuation studies [7]. Across the different versions of the protocol the valuation tasks have remained the same, but later versions pay more attention to the optimal implementation of these tasks with the introduction of a QC procedure (since EQ-VT Version 1.1) and a feedback module allowing participants to review their responses (in EQ-VT Version 2.0). At the launch of this study, these modifications were still experimental. Hence, as part of an international EQ-VT research program [8], it was decided to conduct an initial experimental pre-study to test the protocol updates in Germany. The pre-study provided evidence of the feasibility and acceptability of the protocol, confirmed the positive impact of the modifications from EQ-VT 1.0 to EQ-VT 1.1, contributed evidence for the development of EQ-VT 2.0, and built the empirical basis for this valuation study [10].

2.1 Study Design

The health state valuations were collected in computer-assisted personal interviews (CAPI). In accordance with the protocol's sample size calculation [11], a minimum sample of 1000 respondents from the German general population aged over 18 years was targeted. Quota-based sampling with respect to age, sex, educational level, and employment status was applied using German official statistics [12]. Interviews were conducted in six cities and surrounds located in different parts of Germany to ensure a balanced geographical spread: Berlin, Leipzig, Hamburg, Bielefeld, Munich, and Frankfurt. Individuals were recruited through a mixed recruitment strategy, i.e., through personal contact and from public locations. Interviews were conducted either in a public venue (e.g., research offices) or at the participants' or interviewers' homes.

2.2 Valuation Interview and Methods of Eliciting Preferences

A sub-set of the 3125 health states described by EQ-5D-5L was included in two preference elicitation tasks: composite time trade-off (cTTO) and discrete choice experiment (DCE) without duration. The EQ-VT design included a set of 86 EQ-5D-5L health states, divided into ten blocks of

ten health states for the cTTO tasks (in which some states were present in multiple blocks) and 196 pairs of EQ-5D-5L health states, divided into 28 blocks of seven pairs for the DCE tasks. Each respondent was randomly assigned to one of the cTTO blocks and to one of the DCE blocks [11].

Each interview consisted of the following:

1. welcome and purpose of the study;
2. self-reported health using EQ-5D-5L and background questions;
3. cTTO valuation tasks (wheelchair example, three practice states, ten real tasks, debriefing questions, and feedback module);
4. DCE valuation tasks (seven tasks, methodological add-on with six tasks, debriefing questions);
5. comment box;
6. further background questions.

The cTTO aims to elicit the number of remaining life years in full health at which the respondent is indifferent between a longer period of impaired health (10 years in an EQ-5D-5L health state) and a shorter life span in full health. The cTTO approach involves commencing with the 'conventional' time trade-off (TTO) for all health states, and shifting to a 'lead time' TTO when the participants' responses indicate that they consider the health state to be worse than being dead [13–15]. The resulting cTTO values range from −1 (trading whole lead time) to 1 (trading no years in full health) in 0.05 increments. The interviewer used the example health state "being in a wheelchair" to explain the cTTO tasks. After the cTTO tasks, each respondent was presented with the rank ordering implied by their cTTO valuations in the feedback module. The respondents were asked to review their responses and to flag any health state they felt should be reconsidered. However, those health states could only be flagged but not re-valued [9, 10].

In the DCE tasks, respondents were presented with a pair of EQ-5D-5L health states, designated A and B. They were asked to decide which of the two states was better (forced choice). Additionally, six further paired comparisons comprising a duration attribute and a death alternative were included as a methodological add-on. These data will be reported elsewhere, as this article focusses on the estimation of the German EQ-5D-5L value set using the EQ-VT protocol.

2.3 Data Collection and Quality Control Process

Figure 1 describes the data collection and QC procedures. In total, 32 professional interviewers from a market research company were intensively trained and each interviewer performed practice interviews. Data were reviewed using the EQ-VT QC software to check the interviewers' performance and compliance with the interview script [16] and each interviewer received detailed feedback.

The following QC criteria were defined:

1. time spent on the wheelchair example was too short (less than 3 min);
2. no explanation of the worse than dead task ('lead time' TTO) in the wheelchair example;
3. clear inconsistency in the cTTO ratings (i.e., 55555 is not the lowest and at least 0.5 higher than the health state with the lowest value);
4. time spent for the ten cTTO tasks was too short (less than 5 min).

If any of the criteria were met, the interview was flagged as being of "suspect" quality. Each interviewer had to perform at least two non-flagged interviews before entering the field phase. At the start of the data collection, experienced interviewers from the pre-study [10] participated in one interview supervising each interviewer. During the entire study, the interviewers received feedback based on a daily evaluation using the QC software. Continuously underperforming interviewers were excluded from the study (40% flagged interviews at maximum). Moreover, a written debriefing of every tenth participant provided additional information on the respondent's perception of the interview and the interviewer's performance, thus ensuring QC in terms of formative evaluation.

2.4 Data Analysis and Modeling

Descriptive analyses were used to examine the sample characteristics and the responses to the cTTO and DCE tasks (proportions for discrete variables, mean, and standard deviation for continuous variables). Statistical modeling was used to estimate values for all health states described by EQ-5D-5L, drawing on the cTTO data and DCE data, and combining both in a hybrid model. In the case of cTTO, the dependent variable was defined as 1 minus the observed cTTO value for a given health state indicating disutility and hence coefficients expressed utility decrements. The observed cTTO values for the non-flagged health states after the feedback module were used (i.e., the respondent's flagged cTTO observations were excluded). For the DCE data, the dependent variable was the binary stated choice (i.e., 0/1 indicated the choice for each health state pair). No DCE data were excluded.

As the EQ-VT design was optimized for main effects models [11], the dependent variable was explained by 20 independent variables: four variables for each EQ-5D-5L dimension, each representing the four levels beyond level 1 ("no problems") as the reference category. The coefficients presented the decrement from level 1 to the respective level
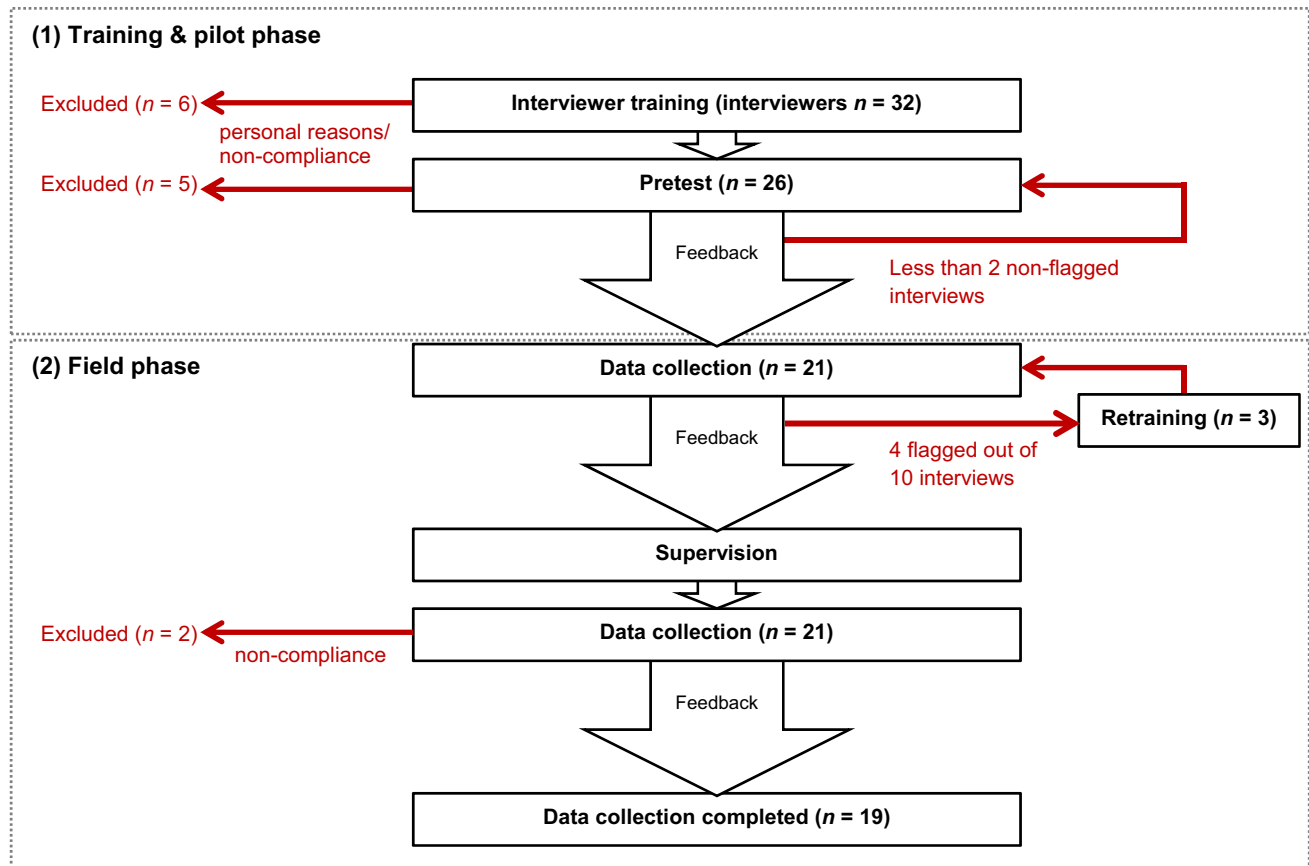
Fig. 1 Data collection and quality control of interviewers

(regular dummies). Thus, the regression equation was as follows:

$$Y = \beta0 + \beta1 * MO2 + \beta2 * MO3 + \beta3 * MO4$$
$$+ \beta4 * MO5 + \beta5 * SC2 + \beta6 * SC3 + \beta7 * SC4$$
$$+ \beta8 * SC5 + \beta9 * UA2 + \beta10 * UA3$$
$$+ \beta11 * UA4 + \beta12 * UA5 + \beta13 * PD2$$
$$+ \beta14 * PD3 + \beta15 * PD4 + \beta16 * PD5$$
$$+ \beta17 * AD2 + \beta18 * AD3 + \beta19 * AD4$$
$$+ \beta20 * AD5 + \varepsilon$$

Firstly, to illustrate how the hybrid model combined both types of data, the cTTO and DCE data were modeled separately and their agreement of the utility decrements and the predicted index values was compared. A censored regression approach (Tobit) assuming normally distributed errors was used for the cTTO data. According to the EQ-VT design, the maximum lead time to be traded was restricted to 10 years (minimum value of −1). However, it cannot be excluded that respondents would have traded off more years, thus values may have been in the range (−∞;1) [7, 10, 17]. Thus, the cTTO responses were considered to be left-censored at −1 [10, 18] and hence a

Tobit model was selected for the cTTO data (Model 1). With the constant term being marginal and non-significant, it was decided to suppress the constant in the models. In the case of the DCE data, a conditional logit model was used (Model 2). The values derived by the DCE valuations were estimated on a latent scale, and consequently cannot be used independently as a basis for value set generation. To allow direct comparisons, the coefficients were transformed to a 1 (full health) to 0 (dead) scale by dividing them by the rescaling parameter Theta of the hybrid model 3a, under the assumption that the cTTO model coefficients were proportional to the DCE coefficients [19].

Secondly, the cTTO and DCE data were combined in a hybrid model (Model 3a). The assumptions of the hybrid model were the same as for the Tobit model (cTTO data) and for the conditional logit model (DCE data). The combination was based on multiplying the likelihood functions of the cTTO model by the likelihood function of the DCE model (for a more detailed description of the hybrid model see [20]).

Thirdly, as the observed variance of the cTTO values increased with the severity of the health state [18], a test for homoskedasticity in a Tobit regression was applied [21].

Because homoskedasticity was rejected, another hybrid model censoring at −1 and accommodating heteroskedasticity was performed by relaxing the variance for each parameter (as suggested by Ramos-Goñi et al. [20]) (Model 3b), whereby the assumptions of the hybrid model 3a were kept.

The model selection was based on theoretical considerations, logical consistency of the parameter estimates (i.e., the higher the dimensional level, the higher the utility decrement), and significance of the parameters ($p < 0.05$). Prediction accuracy was not assessable in terms of mean absolute error because of the missing counterfactual for hybrid model predictions [18, 19]. Thus, the value range, the ranking based on the size of the coefficient for the worst level on each dimension, and the relative importance of the dimensions were compared. The predictions for the EQ-5D-5L health states between models were compared using scatterplots and Pearson product-moment correlation analysis.

The plotted kernel distributions of the 3125 possible EQ-5D-5L health states for the selected EQ-5D-5L value set, the crosswalk value set [6], and the 243 attainable values of the TTO-based EQ-5D-3L value set [22] were used to compare the different value sets for Germany.

The exclusion of single persons, as in previous EQ-5D-5L valuation studies, was tested. This resulted in: the same cTTO value for all health states ($n = 1$), non-traders ($n = 6$), suspect DCE response patterns (i.e., straight-liners who always clicked the same option in all DCE tasks, or variations thereof such as ABABABA) ($n = 26$), and flagged interviews by the QC software ($n = 83$). A sensitivity analysis of the resulting models with the excluded persons was performed. As there were only non-significant marginal differences, it was decided to use all available information and not to exclude any respondents. Statistical analyses were performed using STATA Version 14 (StataCorp. 2015. Stata Statistical Software: Release 14. College Station, TX: StataCorp LP).

3 Results

In total, 1158 interviews were conducted nationwide from December 2014 until March 2015 by 19 interviewers (Fig. 2). The mean interview time was approximately 48 min. 83 interviews (7.17%) were flagged according to the defined quality criteria. The final QC report can be found in the Electronic Supplementary Material (ESM) 1.
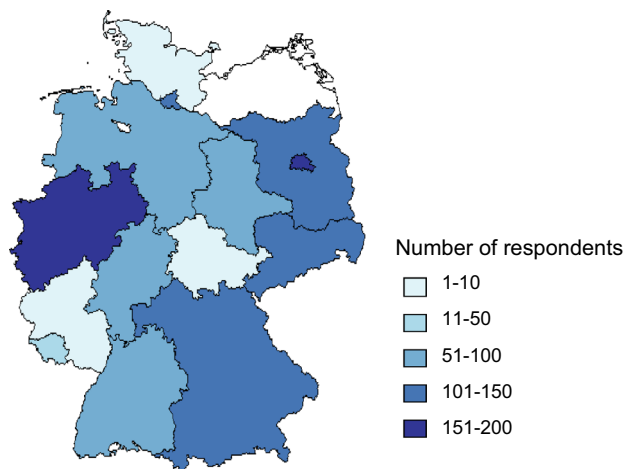

Fig. 2 Place of residence of respondents in the German EQ-5D-5L valuation study

3.1 Characteristics of the Sample

The study sample was representative for the German population in terms of age, sex, education, and employment status (Table 1). Self-reported health using EQ-5D-5L showed that the frequencies of reported problems varied from 6.39% in self-care to 55.61% in pain/discomfort, whereas 36.4% of respondents reported no problems in any dimension (11111).

3.2 Data Characteristics: Composite Time Trade-Off and Discrete Choice Experiment

On average, respondents had 7.39 iterative steps before they reached their point of indifference in the cTTO tasks. The observed mean cTTO values ranged from −0.461 for health state 55555 to 0.973 for health state 21111.

In the feedback module, a total of 6.16% of cTTO responses ($n = 713$) were removed by respondents from the rank ordering. 17.69% of respondents had at least one inconsistency (i.e., health state A defined as better than health state B but A having a lower cTTO value) in their responses (3.8% involving 55555). After the feedback module, 12.6% of respondents were still inconsistent (2.33% involving 55555) ($p < 0.001$). In total, 2.2% of all cTTO responses were inconsistent and this was significantly reduced by 0.72% after the feedback module ($p < 0.001$). The following results include all cTTO valuations after the feedback module (11,580–713 removed observations = 10,867 observations).

17.3% of mean cTTO values were negative, and most of these worse than dead responses were elicited at −1 (9.65%) (Fig. 3). The proportion of values clustered at 0 was 3%. The higher the severity level (i.e., sum of levels across dimensions), the lower the mean cTTO value,

Table 1 Demographics of the respondents in the German valuation study

		Study sample (n = 1158)	German general population (%) [12]	Proportional difference (%)
Sampling characteristics, n (%)				
Age, years	18–24	94 (8.1)	9.3	− 1.2
	25–29	73 (6.3)	7.3	− 1.0
	30–39	155 (13.4)	14.2	− 0.8
	40–49	226 (19.5)	19.2	+ 0.3
	50–64	320 (27.6)	24.9	+ 2.7
	65–74	164 (14.2)	13.4	+ 0.8
	≥ 75	126 (10.9)	11.7	− 0.8
Sex	Female	618 (53.4)	51.6	+ 1.8
	Male	540 (46.6)	48.4	− 1.8
Education	Still in education	5 (0.4)	1.2	− 0.8
	Lower education[a]	410 (35.4)	40.7	− 5.3
	Middle education[b]	396 (34.2)	29.8	+ 4.4
	Higher education[c]	347 (30)	28.3	+ 1.7
Employment status	Employed	608 (52.5)	49.9	+ 2.6
	Non-employed	550 (47.5)	50.1	− 2.6
Self-reported health using EQ-5D-5L, n (%)				
11111		421 (36.4)	N/A	
Any other health state		737 (63.6)	N/A	
Self-rated health using EQ VAS, n (%)				
< 80		381 (32.9)	N/A	
80–89		294 (25.39)	N/A	
90–99		411 (35.49)	N/A	
100		72 (6.22)	N/A	
EQ VAS, mean (SD)		79.45 (17.05)	N/A	

N/A not available, *SD* standard deviation, *VAS* visual analogue scale

[a]Lower education: with or without secondary general school certificate

[b]Middle education: intermediate school certificate

[c]Higher education: entrance qualification for universities of applied sciences, university entrance qualification

whereby the standard deviation increases with the severity level ($p < 0.001$) (Fig. 4).

In the DCE tasks, the proportion of the choice of A or B was strongly correlated to the difference in the severity level between the health states. The greater the difference in severity between the two states, the more likely respondents were to choose the state with the lower severity level. In total, 26 respondents (2.2%) answered with a suspect response pattern across all DCE tasks.

3.3 Modeling

Both single-method models, the Tobit model for the cTTO data (Model 1) and the conditional logit model for the DCE data (Model 2), had one inconsistent order in levels 2 and 3 (i.e., a smaller parameter estimate than the adjacent one; Table 2). The absolute difference between the estimated parameters of the cTTO- and DCE-only models was on average 0.02 (largest difference for mobility level 3 with 0.048). The estimated predictions of both single-method models were strongly correlated ($r = 0.998$, $p < 0.001$). Both hybrid models (3a and 3b) had logically consistent parameter estimates ($p < 0.001$ for all coefficients).

Comparing all the models, there were only small differences in the parameter estimates. The preference ranking of the dimensions was consistent across the four models (ordered from most to least important): (1) pain/discomfort, (2) anxiety/depression, (3) self-care, (4) mobility, (5) usual activities.

The scatterplots of the predictions of the different models suggest the compatibility of the cTTO and DCE data (Fig. 5a) and show the effect of adding the DCE data to the cTTO valuations in the hybrid model (Fig. 5b). There was still a large agreement between the predictions

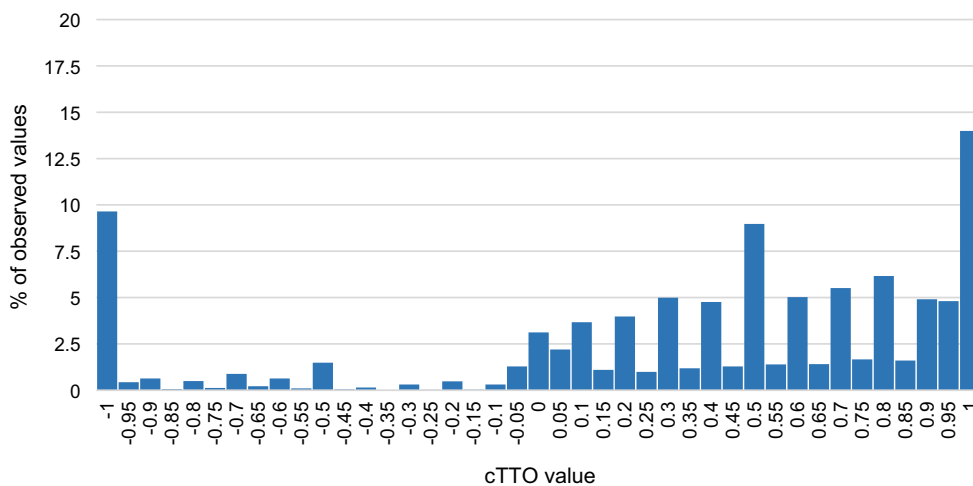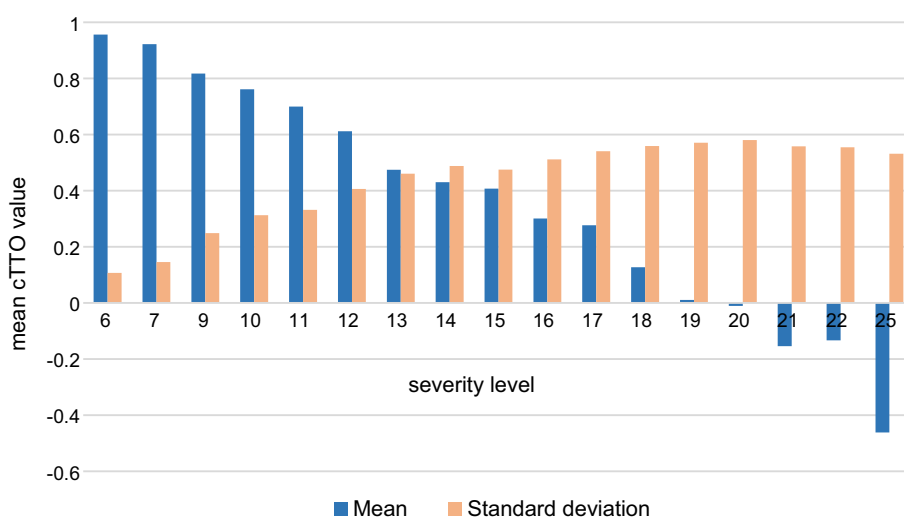**Fig. 3** Observed composite time trade-off (cTTO) value distribution

Fig. 4 Mean composite time trade-off (cTTO) value by severity level

(mean absolute difference of parameter estimates: 0.01). Moreover, accounting for heteroskedasticity spread out the range of the predicted utilities as the decrements for extreme problems were higher (Fig. 5c).

3.4 Preferred Model (Value Set)

Following the agreement of the cTTO and DCE data, both data types can be combined in a hybrid model. In contrast to the cTTO- and DCE-only models, both hybrid models show a logical order for all parameter estimates ($p < 0.001$ for all coefficients). However, the value set has been based on Model 3b as this model takes into account the heteroskedasticity of the error terms in the cTTO data, and has the highest precision of the model coefficients (i.e., smaller standard errors and best model fit).

Applying the value set, EQ-5D-5L health state utilities can be estimated by subtracting the relevant decrement for each problem on each dimension from 1. A full set of the predicted EQ-5D-5L values is available in ESM 3. For example, the predicted EQ-5D-5L index value for health state 12345 was calculated as follows:

$$U(12345) = 1.000 - 0.000 - 0.050 - 0.049 - 0.404 \\ - 0.356 \\ = 0.141$$

The magnitude of utility decrements and the resulting relative importance of the functional dimensions (mobility, self-care, and usual activities) were similar to each other. The decrements of pain/discomfort and anxiety/depression were higher, especially levels 4 and 5. However, symptoms of pain/discomfort had by far the greatest impact on HrQoL. All moves between the five consecutive levels within each of the five dimensions were statistically significant except for the move from level 2 to 3 in the functional dimensions.

Table 2 Parameter estimates for main effects models

Independent variables of the model[a,b]	Model 1: Tobit (cTTO model)		Model 2: Conditional Logit (DCE model)			Model 3a: Hybrid censoring at -1 (cTTO + DCE model)[c]		Model 3b (Value Set): Hybrid censoring at -1 and correcting for heteroskedasticity (cTTO + DCE model)[d,e]	
	β (SE)	p	β (SE)	p	rescaled β	β (SE)	p	β (SE)	p
MO2: slight problems	0.028 (0.015)	0.062	0.135 (0.057)	0.019	0.023	0.028 (0.008)	0.000	0.026 (0.006)	0.000
MO3: moderate problems	**0.015** (0.017)	0.379	0.370 (0.069)	0.000	0.063	0.051 (0.009)	0.000	0.042 (0.009)	0.000
MO4: severe problems	0.130 (0.018)	0.000	0.834 (0.069)	0.000	0.141	0.139 (0.009)	0.000	0.139 (0.009)	0.000
MO5: unable	0.207 (0.017)	0.000	1.349 (0.077)	0.000	0.228	0.216 (0.009)	0.000	0.224 (0.009)	0.000
SC2: slight problems	0.035 (0.014)	0.013	0.408 (0.063)	0.000	0.069	0.058 (0.008)	0.000	0.050 (0.006)	0.000
SC3: moderate problems	0.050 (0.018)	0.006	**0.393** (0.070)	0.000	**0.067**	0.062 (0.009)	0.000	0.056 (0.008)	0.000
SC4: severe problems	0.174 (0.017)	0.000	1.034 (0.072)	0.000	0.175	0.174 (0.009)	0.000	0.169 (0.009)	0.000
SC5: unable	0.244 (0.016)	0.000	1.520 (0.071)	0.000	0.257	0.248 (0.008)	0.000	0.260 (0.008)	0.000
UA2: slight problems	0.034 (0.015)	0.024	0.119 (0.059)	0.044	0.020	0.025 (0.008)	0.001	0.036 (0.006)	0.000
UA3: moderate problems	0.069 (0.016)	0.000	0.232 (0.066)	0.000	0.039	0.049 (0.009)	0.000	0.049 (0.008)	0.000
UA4: severe problems	0.121 (0.017)	0.000	0.669 (0.070)	0.000	0.113	0.117 (0.009)	0.000	0.129 (0.008)	0.000
UA5: unable	0.203 (0.016)	0.000	1.130 (0.073)	0.000	0.191	0.191 (0.009)	0.000	0.209 (0.008)	0.000
PD2: slight problems	0.061 (0.013)	0.000	0.421 (0.063)	0.000	0.071	0.066 (0.008)	0.000	0.057 (0.006)	0.000
PD3: moderate problems	0.098 (0.018)	0.000	0.739 (0.070)	0.000	0.125	0.119 (0.009)	0.000	0.109 (0.009)	0.000
PD4: severe problems	0.423 (0.016)	0.000	2.264 (0.079)	0.000	0.383	0.397 (0.010)	0.000	0.404 (0.010)	0.000
PD5: extreme problems	0.558 (0.017)	0.000	3.516 (0.098)	0.000	0.595	0.577 (0.010)	0.000	0.612 (0.011)	0.000
AD2: slight problems	0.036 (0.014)	0.012	0.183 (0.067)	0.007	0.031	0.033 (0.008)	0.000	0.030 (0.005)	0.000
AD3: moderate problems	0.106 (0.017)	0.000	0.439 (0.070)	0.000	0.074	0.085 (0.009)	0.000	0.082 (0.008)	0.000
AD4: severe problems	0.250 (0.016)	0.000	1.378 (0.078)	0.000	0.233	0.236 (0.009)	0.000	0.244 (0.008)	0.000
AD5: extreme problems	0.345 (0.016)	0.000	1.985 (0.080)	0.000	0.336	0.334 (0.009)	0.000	0.356 (0.009)	0.000
Observations included in model	10,867		8,106			18,973		18,973	
Continuous uncensored (cTTO)	9,818		-			9,818		9,818	
Continuous left-censored (cTTO)	1,049		-			1,049		1,049	
Dichotomous observations (DCE)	-		8,106			8,106		8,106	
Estimated values by health state									
U(11111)	1		1			1		1	
U(21111)	0.972		0.977			0.972		0.974	
U(12111)	0.965		0.931			0.942		0.950	

Table 2 continued

	Model 1	Model 2	Model 3a	Model 3b (<u>Value Set</u>)
Estimated values by health state				
U(11211)	0.966	0.980	0.975	0.964
U(11121)	0.939	0.929	0.934	0.943
U(11112)	0.964	0.969	0.967	0.970
U(55555)	-0.557	-0.607	-0.566	-0.661
Model performance				
#illogically ordered	1	1	0	0
#non-significant	2	0	0	0
Ranking of dimensions	PD – AD – SC – MO – UA	PD – AD – SC – MO – UA	PD – AD – SC – MO – UA	PD – AD – SC – MO – UA

AD anxiety/depression, *cTTO* composite time trade-off, *DCE* discrete choice experiment, *MO* mobility, *PD* pain/discomfort, *SC* self-care, *SE* standard error, *UA* usual activities

[a]The constant $\beta0$ was suppressed in the models, i.e., $\beta0 = 0$

[b]Bold figures indicate logical inconsistencies

[c]Theta of Model 3a = 5.911

[d]Theta of Model 3b = 5.526

[e]The full heteroskedastic Model 3b including the model for lnsigma can be found in ESM 2

(a) Model 1 vs. Model 2 (r=0.998; p<0.001)

(b) Model 1 vs. Model 3a (r=0.999; p<0.001)

(c) Model 3a vs. Model 3b (value set) (r=0.999; p<0.001)
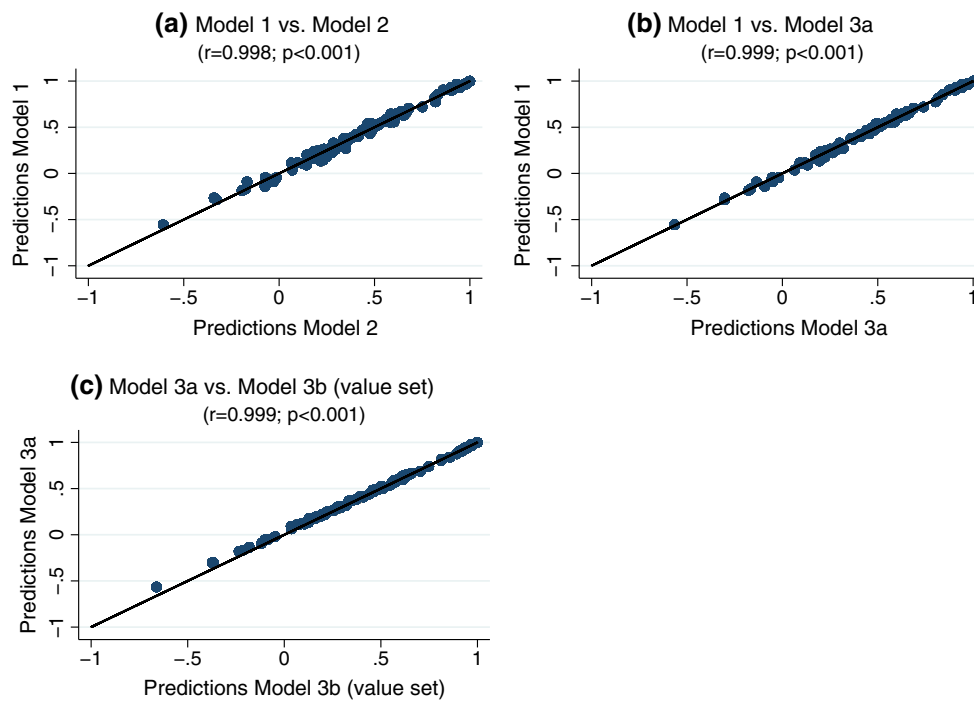
Fig. 5 Comparison of model predictions (86 states included in the composite time trade-off design)
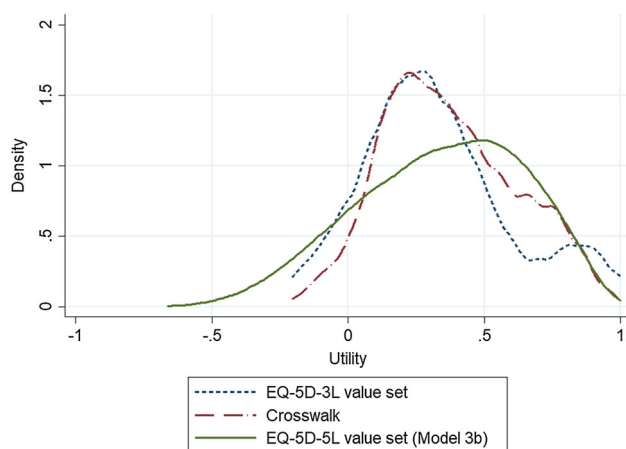
Fig. 6 Kernel density plot of all possible EQ-5D-3L and EQ-5D-5L values

3.5 Comparison of EQ-5D-3L and EQ-5D-5L Values

In contrast to the German EQ-5D-3L value set [22], each level of this EQ-5D-5L value set can be ascribed a certain utility weight that is consistent throughout. The kernel density plot of the index values for the 3125 attainable states using the EQ-5D-5L value set shows a left-skewed distribution, whereas the EQ-5D-3L value set is characterized by two peaks (bimodal distribution). The EQ-5D-5L value set covers a larger evaluation space without a constant as a deviation from full health (−0.661 to 1) than the EQ-5D-3L value set and the crosswalk (−0.205 to 1) [Fig. 6].

4 Discussion

This article presents the results of the German EQ-5D-5L valuation study. It has provided evidence of the successful completion of an EQ-5D-5L valuation study using the improved valuation protocol EQ-VT 2.0, combined with intensive interviewer training and data monitoring. The different QC steps ensured high data quality in terms of few inconsistencies, high values for mild states, little clustering of values, and high interviewer compliance compared with previous valuation studies using earlier protocol versions without QC [9, 16, 23, 24].

This EQ-5D-5L value set is based on a hybrid model judged to be theoretically preferred over a TTO-based model under the assumption that individuals each have a unique utility function that determines their health state preferences. Both techniques, cTTO and DCE, try to measure the same utility function for health and can be summed up to a linear transformation ($\text{Beta}_{\text{TTO}} = \text{Theta}*\text{Beta}_{\text{DCE}}$) [9, 13, 19]. However, values derived from both methods need not be identical and may not capture values perfectly: health states are valued against time in cTTO (matching task) whilst in DCE, health states differing on dimensions' severity levels are directly compared to another (choice task). cTTO data might therefore be influenced by scale compatibility and loss aversion [7, 25], whereas lexicographic preferences and attribute non-attendance have been reported for DCE [26, 27]. The idea that the true value cannot be derived from a single valuation technique provides an argument for the use of hybrid models. In addition, the parameter estimates may be derived with greater precision when informed by two types of (complementary) data, as evidenced in this article. The utility decrements, ranking, and relative importance of the dimensions were robust across the cTTO, DCE and hybrid models. However, using all valuation data in the selected hybrid model provided the highest precision on the parameter estimation (i.e., smaller standard error) [9, 19]. Hence the German EQ-5D-5L value set considered the complementary cTTO and DCE data in a hybrid model without excluding single respondents due to data issues. Innovative modeling approaches representing best practice [9, 18, 20] were used, taking into account that cTTO data are left-censored and the heteroskedasticity of the error terms to prevent biased parameter estimates.

This German study was not troubled with data issues reported in some studies using an earlier protocol version [23, 24]. The health state descriptions in the EQ-5D-5L were matched with a valuation method that was able to discriminate between health severity levels even when the differences were subtle. This stands in sharp contrast with the German TTO-based EQ-5D-3L value set [22], where some levels were collapsed. This inspires faith in this German EQ-5D-5L value set, and it may be noted that the kind of insights that guided this valuation study did not exist in the past when the EQ-5D-3L was valued. Therefore, the EQ-5D-5L value set with its larger evaluation space and the more differentiated distribution of attainable index values than its predecessor reinforces the (theoretical) potential of the EQ-5D-5L to better discriminate between patients, especially those with mild health conditions and for small changes in health. This might improve the applicability of the EQ-5D in general and in certain disease areas using the national value set in Germany. However, further research is required to examine the practically relevant evaluation space and the discriminative properties of the German EQ-5D-5L in patient populations.

An implication of the results is that a gradual shift in the basis for utility assessment in Germany might be anticipated, moving away from the EQ-5D-3L to the EQ-5D-5L. This would allow users to benefit from a better descriptive system [3] and from a valuation point of view, as the high quality EQ-5D-5L valuation data were derived from a

much larger and geographically more diverse sample. However, the sample was clustered within six regions and there was a small middle-class bias. The study sample was, nevertheless, predominantly representative in terms of age, sex, education, and employment status for the German general population [12]. In summary, the resulting value set has provided a robust and up-to-date basis for estimating EQ-5D-5L index values and should be the preferred value set for Germany.

5 Conclusions

This is the first country study to utilize the improved EQ-5D-5L valuation protocol 2.0 developed by the EuroQol Group, including a feedback module as internal validation, and a continuous QC. It provides evidence that the refined protocol with its QC process appears to be a solid basis for estimating national EQ-5D-5L value sets.

The German EQ-5D-5L value set, based on cTTO and DCE, is recommended as the preferred value set for Germany. It allows a more refined preference-based HrQoL measurement to describe patients' health and enables the use of the EQ-5D-5L in a range of applications such as cost-utility analysis for health care policy and clinical assessment in Germany.

This set of EQ-5D-5L preference values demonstrates the relative importance placed on different HrQoL dimensions according to the German general population, where the health problems of anxiety/depression and especially pain/discomfort should receive the greatest attention. These societal preferences have implications for the assessment of treatments that affect individuals' HrQoL as a patient-relevant outcome and should be reflected in health care decision making in Germany.

Compliance with Ethical Standards

References

1. Greiner W. Der EQ-5D der EuroQol-Gruppe. In: Schöffski O, Graf von der Schulenburg JM, editors. Gesundheitsökonomische Evaluationen. 4th ed. Heidelberg, Berlin: Springer; 2012. p. 411–22.
2. Devlin NJ, Krabbe PFM. The development of new research methods for the valuation of EQ-5D-5L. Eur J Health Econ. 2013;14(Suppl. 1):1–3.
3. Herdman M, Gudex C, Lloyd A, Janssen MF, Kind P, Parkin D, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res. 2011;20:1727–36.
4. Janssen MF, Pickard AS, Golicki D, Gudex C, Niewada M, Scalone L, et al. Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: a multi-country study. Qual Life Res. 2013;22:1717–27.
5. Pickard AS, de Leon MC, Kohlmann T, Cella D, Rosenbloom S. Psychometric comparison of the standard EQ-5D to a 5 level version in cancer patients. Med Care. 2007;45:259–63.
6. van Hout B, Janssen MF, Feng YS, Kohlmann T, Busschbach J, Golicki D, et al. Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. Value Health. 2012;15:708–15.
7. Oppe M, Devlin NJ, van Hout B, Krabbe PFM, de Charro F. A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. Value Health. 2014;17:445–53.
8. Shah K, Rand-Hendriksen K, Ramos-Goñi JM, Prause AJ, Stolk E. Improving the quality of data collected in EQ-5D-5L valuation studies: a summary of the EQ-VT research methodology programme. In: Proceedings of the 31st Scientific Plenary Meeting of the EuroQol Group; Stockholm, Sweden, September 25–26, 2014. 2015: p. 1–18. http://eq-5dpublications.euroqol.org/download?id=0_53918&fileId=54332. Accessed 21 Aug 2017.
9. Stolk E, Ludwig K, Rand K, van Hout B, Ramos-Goñi JM. Overview, update and lessons learned from the international EQ-

5D-5L valuation work: version 2 of the EQ-5D-5L valuation protocol. Value Health. (Submitted).

10. Ludwig K, Graf von der Schulenburg JM, Greiner W. Valuation of the EQ-5D-5L with composite time trade-off for the German population: an exploratory study. Health Qual Life Outcomes. 2017;15:39. https://doi.org/10.1186/s12955-017-0617-9.

11. Oppe M, van Hout B. The "power" of eliciting EQ-5D-5L values: the experimental design of the EQ-VT. EuroQol working paper series. 2017;17003. http://euroqol.org/wp-content/uploads/2016/10/EuroQol-Working-Paper-Series-Manuscript-17003-Mark-Oppe.pdf. Accessed 26 Oct 2017.

12. Federal Statistical Office. Mikrozensus. Bevölkerung und Erwerbstätigkeit. Stand und Entwicklung der Erwerbstätigkeit in Deutschland 2014. http://www.destatis.de/DE/Publikationen/Thematisch/Arbeitsmarkt/Erwerbstaetige/StandEntwicklungErwerbstaetigkeit.html. Accessed 10 Aug 2017.

13. Oppe M, Rand-Hendriksen K, Shah K, Ramos-Goñi JM, Luo N. EuroQol protocols for time trade-off valuation of health outcomes. PharmacoEconomics. 2016;34:993–1004.

14. Devlin NJ, Tsuchiya A, Buckingham K, Tilling C. A uniform time trade off method for states better and worse than dead: feasibility study of the 'lead time' approach. Health Econ. 2011;20:348–61.

15. Janssen BMF, Oppe M, Versteegh MM, Stolk EA. Introducing the composite time trade-off: a test of feasibility and face validity. Eur J Health Econ. 2013;14(Suppl. 1):5–13.

16. Ramos-Goñi JM, Oppe M, Slaap B, Busschbach JJV, Stolk E. Quality control process for EQ-5D-5L valuation studies. Value Health. 2016;20:466–73.

17. Lin HW, Li CI, Hsu CN, Tang CH. Does time of full health in worse-than-death (WTD) tasks matter? Noordwijk; 2nd EuroQol Academy Meeting. Noordwijk, The Netherlands, 7–8 March, 2017; 2017.

18. Feng Y, Devlin NJ, Shah KK, Mulhern B, van Hout B. New methods for modelling EQ-5D-5L value sets: an application to English data. Health Econ. 2017. https://doi.org/10.1002/hec.3560.

19. Ramos-Goñi JM, Pinto-Prades JL, Oppe M, Cabasés JM, Serrano-Aguilar P, Rivero-Arias O. Valuation and modeling of EQ-5D-5L health states using a hybrid approach. Med Care. 2014;55:51–8.

20. Ramos-Goñi JM, Craig B, Oppe M, van Hout B. Combining continuous and dichotomous responses in a hybrid model. EuroQol working paper series. 2016; 16002. https://euroqol.org/wp-content/uploads/working_paper_series/EuroQol_Working_Paper_Series_Manuscript_16002_-_Juan_Ramos-Goni.pdf. Accessed 21 Aug 2017.

21. Cameron AC, Trivedi PK. Microeconometrics using Stata. Revised ed. College Station (TX): Stata Press; 2010.

22. Greiner W, Claes C, Busschbach JJV, Graf von der Schulenburg JM. Validating the EQ-5D with time trade off for the German population. Eur J Health Econ. 2005;6:124–30.

23. Devlin NJ, Shah KK, Feng Y, Mulhern B, van Hout B. Valuing health-related quality of life: an EQ-5D-5L value set for England. Health Econ. 2017. https://doi.org/10.1002/hec.3564.

24. Versteegh MM, Vermeulen KM, Evers SMAA, de Wit GA, Prenger R, Stolk EA. Dutch tariff for the five-level version of EQ-5D. Value Health. 2016;19:343–52.

25. Bleichrodt H. A new explanation for the difference between time trade-off utilities and standard gamble utilities. Health Econ. 2002;11:447–56.

26. Lancsar E, Louviere J. Deleting 'irrational' responses from discrete choice experiments: a case of investigating or imposing preferences? Health Econ. 2006;15:797–811.

27. Alemu MH, Mørkbak MR, Olsen SB, Jensen CL. Attending to the reasons for attribute non-attendance in choice experiments. Environ Resource Econ. 2013;54:333–59.