Application of EuroQol Valuation Technology to the EQ Health and Wellbeing Short: A feasibility study

Clara Mukuria[1], Tessa Peasgood[2], Emily McDool[1], Richard Norman[3], Donna Rowen[1], John Brazier[1]

1 University of Sheffield
2 University of Melbourne
3 Curtin University

Abstract

Objective: To test the feasibility of time trade-off (cTTO) and discrete choice experiment (DCE) administered using the EuroQol Valuation Technology (EQ-VT) research protocol to derive a value set for the EQ Health and Wellbeing Short (EQ-HWB-S) and to generate a pilot value set.

Methods: EQ-HWB-S values were elicited using cTTO and DCE tasks with adaptations to fit the new measure. Participants (target n=600) from the UK general population were sampled based on age, sex and ethnicity. Interviews were undertaken using Video-conferencing. Quality control (QC) steps were used to assess interviewers' performance throughout the study. Data were modelled using linear, Tobit, probit and hybrid models. Feasibility was assessed based on the evaluation of the cTTO data, QC assessment and regression modelling results. The pilot value set was selected based on theoretical considerations, monotonicity and statistical significance.

Results: There were 521 participants who provided cTTO and DCE data. The demographic characteristics were broadly representative of the UK general population although participants were more educated and there were slightly more females. Interviewers met quality control requirements. It was feasible to value states described by the EQ-HWB-S using cTTO and DCE . cTTO values ranged between -1 to 1 with increasing disutility associated with more severe states. The hybrid Tobit heteroscedastic model was selected for the pilot value set with values ranging from -0.384 to 1. Pain, mobility, daily activities, sad/depressed had the largest disutilities followed by loneliness, anxiety, exhaustion, control and cognition in the selected model.

Conclusions: EQ-HWB-S can be valued using cTTO and DCE administered using EQ-VT. The pilot value set offers an opportunity to test the validity of the EQ-HWB-S. Further methodological work is recommended to develop a valuation protocol specific to the EQ-HWB-S.

Funding: This study is independent research funded by the UK Medical Research Council (Grant number 170620), and the EuroQol Research Foundation.

Disclaimer: The views expressed in this manuscript are those of the authors and not necessarily of our funders, the National Institute of Health and Care Excellence, the Department of Health and Social Care or those acknowledged.

Acknowledgement: We would like to thank NICE for highlighting the methodological research need to the MRC that resulted in the funding call entitled "Beyond the QALY" which led to this research being funded. We would like to acknowledge the support of the National Institute for Health Research Yorkshire and Humber Applied Research Collaboration (formerly CLAHRC) and the National Institute for Health Research Clinical Research Network (NIHR CRN). We acknowledge the invaluable contributions of members of the project steering group, advisory group and public and patient involvement and engagement groups and Julie Johnson for project administration. We would also like to thank members of the EuroQol Group Association for their input at plenary and academy meetings. We would like to acknowledge the contribution of all the participants who took part in the study and the interviewers who undertook data collection. Finally, we would also like to extend our acknowledgments to the EuroQol Scientists – Aureliano Finch, Bram Roudijk and Elly Stolk for their input.

INTRODUCTION

In the economic evaluation of health interventions, the costs and benefits are considered to inform resource allocation. One method commonly used is to measure outcomes using quality-adjusted life years (QALY) by estimating the incremental cost per QALY gained. A QALY combines the value of health-related quality-of-life (HRQoL) with the length of life in a single number, or index. HRQoL measures scored using preferences are typically used to derive the value element of a QALY, measured in terms of utilities. There are a number of existing generic utility measures such as EQ-5D-5L™ [1]. A new utility measure, the EQ Health and Wellbeing ™ (EQ-HWB™) has been developed to capture a broad range of health and wellbeing outcomes for economic evaluation of interventions in health, public health and social care including for informal carers[2]. There are two versions of the instrument, a longer profile measure with 25-items, and the short version of the measure, the EQ-HWB-S™ which has 9-items[2]. The latter was developed for valuation purposes to generate utilities. The measures cover items related to seven dimensions: activity, relationships, cognition, self-identify, autonomy, feelings and physical sensations. The measures are experimental with further testing and validation being undertaken.

In order to generate utility values for the EQ-HWB-S, a preference elicitation study is required. Different preference elicitation methods can be used including time trade-off (TTO), standard gamble and discrete choice experiment (DCE with or without duration) which have been successfully applied to valuation of other measures[3]. Preference elicitation tasks such as TTO are cognitively demanding and the common use of general population samples means that the states they are asked to imagine they are living in are often hypothetical in nature[3]. Pairwise choice tasks as implemented in a DCE may be simpler to understand for participants, but the amount of information presented can make it difficult for participants to make a choice. Studies have tested overlap in some of the choices across pairs e.g. only varying three out of five dimensions with the other two being the same across the pair and/or highlighting where differences occur in order to minimise the cognitive burden[4, 5]. This approach allows measures with many dimensions to be valued e.g. the EORTC-QLU-C10D (European Organisation for Research and Treatment of Cancer Quality of Life Utility Measure - Core) has 10 dimensions but only five vary during valuation [6, 7]. The EQ-HWB-S was limited to 10 or fewer dimensions in the item selection stage as it was not considered likely that the 25 item questionnaire could be valued using standard methods[2].

A previous small (n=19) mixed-methods pilot tested the feasibility and practicality of applying the EQ Valuation Technology (EQ-VT) vs2 protocol[8] that is used to value the EQ-5D-5L™ to the EQ-HWB-S[9]. The protocol sets out the preference elicitation methods (composite time trade-off (cTTO) and DCE), the EQ-5D states to be valued including the number of states per person, the sample size, interviewer training and a set of quality control steps to ensure high data quality. The protocol has been successfully applied across a number of studies [10] therefore it was useful to apply it to the EQ-HWB-S. In a pilot, we tested whether participants could undertake cTTO and DCE based on the 9-item EQ-HWB-S classifier compared to EQ-5D-5L[9]. Interviews were undertaken using computer assisted personal interviews (CAPI) administered via EQ Portable Valuation Technology (EQ-PVT). EQ-PVT runs via MS PowerPoint which made it possible to adapt for EQ-HWB-S. We found that participants could successfully complete both cTTO and DCE tasks but some changes were required including: reducing the number of states that they would value (seven instead of ten for cTTO due to the cognitive burden of a longer measure) and substituting the autonomy item in the EQ-HWB-S which asked about 'coping' to an item which asked about 'control'[9]. Some participants ignored 'coping' as they considered whether they could cope with the rest of the state while others ignored the rest of the state and focused on coping therefore this needed to be replaced. DCE with overlap and simple formatting was preferred; therefore this was also implemented.

Following the mixed-methods pilot, this study aimed to undertake a larger feasibility study to test the feasibility, practicality and comparability of cTTO and DCE when applied to the EQ-HWB-S as well as to generate an initial pilot value set. The study was initially designed to be undertaken face to face but following ongoing COVID-19 restrictions and the successful use of online interviews for other valuation studies [11-13], online data collection was undertaken instead.

METHODS

Study Design

The research design and data collection adapted methods developed by the EuroQol group for valuation of the EQ-5D-5L. The previous pilot tested feasibility of EQ-PVT in face to face setting rather than online and was designed to compare with EQ-5D-5L[9]. An additional pilot (n=23) was undertaken to test delivering the fully modified EQ-PVT protocol. Ethics was obtained from the University of Sheffield School of Health and Related Research Ethics Committee (038012).

Descriptive System

States were described using the EQ-HWB-S which includes 9 items (mobility, activities, exhaustion, loneliness, concentrating/thinking clearly (cognition), anxiety, sadness/depression, control and physical pain) [2]. Each item described 5 levels of problems: 1) severity (no, mild, moderate, severe and very severe) for physical pain; 2) difficulty (none, slight, some, a lot of, unable) for mobility and activities; 3) frequency (none, only occasionally, sometimes, often, most or all of the time) for the rest of the items. States combine items and levels ranging from having no problems (111111111) to the worst problems in all dimensions (555555555) with a total of 1,953,125 states (5^9). Levels in each state can be summed to present what is referred level sum score (e.g. for state 123111111=1+2+3+1+1+1+1+1+1=12). Presentation of states was tested in the previous pilot to ensure that they made sense to participants (see Supplementary Table 1) [9].

Preference elicitation methods

EQ-VT vs2 valuation protocol employs cTTO where states better than dead are valued using a choice of living for a shorter period in full health (n<10 years) versus living in the impaired state being valued for a longer period (n=10 years), with the time in full health decreased or increased based on the participant's responses until they are indifferent between the two options. For states worse than dead (WTD), the choice is living in full health for a shorter period (n<10) versus living for 10 years in full health followed by 10 years in the impaired state. The 'full health' descriptor was changed to 'full health and quality of life' to reflect the EQ-HWB-S which covers more than health. The protocol also includes DCE (without duration) where respondents are presented with a pair of states and asked to select their preferred state. In the DCE, only five out of nine items had differences in each pair with overlap in four to reduce cognitive burden with colour to highlight the differences (Supplementary Figure 1b). Piloting in the current study also led to some minor modifications including showing all possible levels for each item on the screen for both cTTO and DCE presentation screens (see Supplementary Figure 1a and b).

Health State Selection

Designs were based on the use of a main effects model for regression analysis, predicting utility decrements for each level of each dimensions, and not taking into account interaction effects. For cTTO, an orthogonal array was used to select states resulting in 50 states. One of these was the full health and quality of life state which not included. Following the EQ-VT design approach, 10 mild states were added to the design i.e. with only mild problems in one dimension e.g. state 111111112. Health states were grouped into 10 blocks with seven states per block. All 10 blocks contained the worst possible state (555555555), one mild state, and five health states selected from the orthogonal array that were unique to each block. The number of states in each block was lower than the EQ-5D-5L version (10 states) as the pilot indicated that participants preferred fewer EQ-HWB-S states.

For DCE, an approach that allowed some dimensions to be the same across the pairs within each choice set was considered more feasible in the pilot study. A D-efficient design was implemented using Ngene 1.2.1 [14]. A candidate set (n=5000) of random pairs that had the required constraints across four dimensions (i.e. for each choice set 4 of 9 dimensions were identical, and these dimensions and levels varied across the choice sets) was used to enable overlap in the design with small non-zero priors for the dummy variables for each dimension to denote monotonically increasing severity in levels within each dimension[15]. This resulted in 140 choice sets, with 20 blocks of 7 choice sets. No additional states were added to the DCE design.

Sampling and recruitment

Data collection took place between May and November 2021 for participants in England with expansion to the rest of the UK from October 2021 to November 2021. Participants were recruited based on quota sampling using age and sex combined and ethnicity separately to ensure representativeness with a target sample size of 600 as this was a feasibility study. A blended recruitment approach was used to ensure representativeness across the quota characteristics. First, a recruitment agency sent targeted postal invites to individuals identified via the electoral register and other sources in the public domain. Subsequently, participants were also invited to take part from an online research panel, via social media advertising by the recruitment agency, interviewers and other researchers, via an advert on a research advertising website, and via snowballing from previous participants. In each case, potential participants completed a screener survey and provided their contact details. For postal surveys, the screener survey could be completed on paper or online via a link in the invitation letter. The inclusion criteria were: (1) 18 years of age or older; (2) current UK resident; (3) access to a computer, laptop or large screen tablet with an internet connection and the ability to access Zoom or Google Meet; and (4) the ability to complete the tasks in English. Participants were offered an incentive of a shopping voucher for study participation. The value of the incentive was initially £25 but after the initial postal invitation in June (following completion of 11 interviews), this was increased to £40 to improve recruitment.

Survey administration

Due to the pandemic related restrictions, all interviews used a modified version of EQ-PVT presented via videoconferencing (Zoom). EQ-PVT is the same as EQ-VT but the tasks are presented via MS PowerPoint, with separate programs for cTTO and DCE, rather than a dedicated survey with the responses recorded in MS Excel (Supplementary Figures 1a and 1b). EQ-PVT allowed modifications to be made via PowerPoint including adding the EQ-HWB-S states and it has been used in other contexts for the valuation of EQ-5D-5L [16, 17]. DCE states had colour modifications to show where there was overlap in the state as this has been shown to help with engagement [6, 7] (Supplementary Figure 1b). All EQ-VT studies also employ an interview script and this was modified to match EQ-HWB-S and videoconference presentation.

All participants were asked to consent and complete questions about themselves including completing the EQ-HWB-S via an online survey (no interviewer present) prior to the interview. Those who did not do this were asked to do so at the start of the interview. Following the second pilot to test delivery of EQ-PVT online, participants were initially presented with the EQ-HWB-S including an example of a state that could be considered implausible before proceeding to cTTO. By presenting the measure prior to the interview, participants were made aware of the possible responses to each question and the levels of severity that may be reported by the measure as well as who could complete it (patients, social care users and informal carers). The participants then completed four practice cTTO exercises – a state with mobility problems requiring the use of a wheelchair, a state proposed by the respondent which they considered worse than requiring a wheelchair, a mild EQ-HWB-S state and a severe EQ-HWB-S state. They then completed seven EQ-HWB-S states in random order. Participants were then asked three questions related to understanding and answering the cTTO tasks. This was followed by 7 paired DCE tasks presented in a random order, and randomising of the presentation of the pairs to the left/right of the display. Similar feedback questions were then asked about the DCE component. For all the preference tasks, participants were asked to read each state description aloud but were also offered the option to have the interviewer read the state aloud to increase accessibility. Finally, interviewers completed questions on their views of participants' engagement and understanding, alongside any issues with the interview such as technical problems, difficulties with the participant such as if they became upset and any additional comments.

Interviewers and quality control process

Interviewers were PhD students (n=6) recruited from the University of Sheffield or members of staff (n=1). Interviewers received full training from the study team at the University of Sheffield (who had previously been trained by the EuroQol scientific staff) and EuroQol scientific staff, with adapted materials supplied by the EuroQol group. Quality control criteria were applied during data collection to ensure cTTO data quality was maintained. Interviews received a 'flag' for suspected quality concerns based on four criteria: 1) time spent explaining the first practice state (wheelchair)>3min; 2) demonstration of the lead time part of cTTO (worse than dead) in either of the first two practice states; 3) the participant did not give the worst state (555555555) a cTTO value of 0.5 or more higher than other states they valued; and 4) the participant used at least 5 minutes to complete all 7 EQ-HWB-S states. Interviews that did not meet all 4 criteria were discussed during quality control meetings following discussion with EuroQol scientific staff. Feedback was provided to all interviewers both collectively and individually. The distribution of the data was also reviewed to check for any interviewer effects including clustering at particular values that are easily attained and may indicate disengagement (-1,-0.5, 0, 0.5, 1), differences in distribution and time taken for the cTTO interviews. The protocol dictates that interviewers with more than 40% of their interviews flagged as not meeting protocol compliance are retrained, and their data up to that point is dropped from any analyses. Any flags due to technical errors with the EQ-PVT system (e.g. the system crashing necessitating a restart) were not included as flags of poor protocol compliance, though these were only identified as technical problems part-way through the study.

Data analysis

Descriptive analysis of the respondents was undertaken with distributions for age, gender and ethnicity compared to a representative sample. The cTTO data was assessed in terms of summary statistics (mean, standard deviation, median, minimum and maximum values) for each state, distribution and inconsistencies (where a health state that was better than another health state in all dimensions was assigned a lower cTTO value) and inconsistencies involving the worst state). DCE data was assessed for any evidence of particular patterns e.g. ABABABA/AAAAAAA etc.

The modelling approach used to predict scores for all EQ-HWB-S states was informed by the nature of the data and a recent systematic review of the modelling approaches used in EQ-5D-5L value sets elicited using TTO and DCE administered via EQ-VT[18]. The cTTO data is considered censored because the WTD task only allows participants to go as low as -1 when they may be willing to go lower. Therefore models that take into account this censoring at -1 such as Tobit were preferred. DCE data is binary data where participants make a single choice for each pair of states, therefore conditional logit models or logit/probit models would be appropriate. Each participant completed seven cTTO and DCE tasks therefore accounting for repeated data is important and clustering of standard errors or random effect models were used. Finally, observed variance of cTTO values increased with the severity of the state indicating heteroscedasticity and methods to account this were explored. Finally, a hybrid model could be estimated that combined cTTO and DCE data (see Ramos-Goni et al [19] for details), which has the advantage of using all of the data, particularly in this feasibility dataset.

Combining the data relies on an assumption that the two preference elicitation methods are measuring the same thing (preferences for states described by EQ-HWB-S) with a constant proportional relationship between them which can be modelled jointly [19]. The cTTO and DCE results were compared using Lin's concordance correlation coefficient [20] and a plot of the predictions to assess how they compared.

The dependent variable was the utility value from cTTO tasks which was transformed by subtracting this value from one to generate disutility, or the choice option for DCE data. In each case, dummy variables representing levels 2 to 5 were included for each dimension e.g. mo2, mo3, mo4, mo5 for mobility. For the DCE data, the dummy variables were the difference between the paired choice to facilitate estimation using probit/logit and hybrid models. Only main effects were estimated i.e. no interactions as these were not included in the design. Coefficients were expected to be positive and increasing with severity. Where there were disordered coefficients i.e. where there was an increase in utility as severity increased, levels were merged to constrain them to the same value. No constant term was included as given the broader aspects of health and wellbeing, it was assumed that there is no gap between full health and quality of life and having no problems. Furthermore, the constant term in a linear prediction was small and not statistically significant. DCE models were estimated on a latent scale and are not anchored on the zero to dead utility scale. To allow comparison, the DCE values were anchored using the cTTO value for the worst state from the Tobit heteroscedastic model[21].

Study Sample

The pilot data (n=23) were not included. There were no interviewer quality control issues. Sensitivity analyses were undertaken using the selected model to test the robustness of results including: 1) exclusion of the additional mild states from the cTTO data since these were added onto the orthogonal array; (2) exclusion of participants who did not understand or engage (interviewer-reported); and (3) examination of impact of age, sex and interviewer effects. Interviewer effects were assessed by inclusion of a dummy variable for each interviewer and separately running the models while sequentially excluding one interviewer.

Assessing the feasibility and practicality

To assess how well the EQ-VT methods applied to the EQ-HWB-S, results were assessed on the basis of: 1) the cTTO data including EQ-VT quality control; 2) logical ordering of parameters i.e. larger utility decrements for more severe problems; 3) statistical significance of coefficients across items and for different response levels; and 4) performance of models in terms of predicting values for selected states (three states with one dimension with slight problems as example states) including mean absolute errors by state (MAE).

In order to select a model as the pilot value set, results were compared on the basis of appropriate model for the data (taking into account censoring, heteroscedasticity and multiple data per participant), logical ordering, statistical significance and MAE. Akaike and Bayesian information criterion (AIC and BIC) could not be compared across different model types e.g. Tobit and the hybrid models but could be compared within the same models with lower values preferred. Statistical analyses were performed using Stata MP 17[22].

RESULTS

Recruitment and sample

The response rate was 1.45% (n=29/2000) for postal recruitment (discontinued in June 2021 due to poor response rate). From the panel, 41.6% (n=330/794) participants of those who expressed interest completed interviews. There were 117 participants recruited from social media and 48 participants recruited from the research advertising site. This latter route was prone to suspicious participants e.g. multiple attempts therefore some people who expressed interest were not offered interviews (n=19) although they were all contacted and five interviews were terminated; no data was recorded for these. The final sample consisted of

520 participants with cTTO data and 521 with DCE data. Three participants who had data were not included (2 did not have consent and one participant had two entries in July and November; the latter data was excluded). One participant did not want to complete cTTO and only had DCE data.

Descriptive statistics

The mean age was 48.5 and 45% were male with some differences compared to general population (Table 1). The sample had a smaller proportion of employed individuals (59%) relative to the general population (61%,[23]) and they were more educated than the general population with 66% having a degree compared with 42% of 21–64-year-olds[1] with a degree in the UK [24]. Some participants had caring responsibilities (14%) but few (3%) reported using social care services or support for themselves in the last 6 months.

There were 32% respondents who reported having a longstanding health problem[2] and a broad range of conditions were reported across the sample. A large proportion of the participants reported having no health problems based on EQ-5D-5L (11111= 42%) but only 10% report no health and quality of life problems (111111111) using the EQ-HWB-S.

cTTo and DCE data

Each state was valued approximately 50 times (50-57) with the exception of the worst state which was valued by all participants. Participants took an average of 5.2 (standard deviation = 3.2) iterative steps before they reached their point of indifference in cTTO tasks. The distribution of cTTO values ranged from -1 to 1 (Figure 1) with lower mean cTTO values and larger standard deviation as the misery score increased (Supplementary Table 2). The observed mean cTTO values for the states ranged from 0.982 (state 111111121) to -0.264 (state 555555555) [see Supplementary Table 2]. The proportion of values clustered at -1, -0.5, 0, 0.5 and 1 was 7% (n=257), 3% (n=105), 3% (n=116), 10% (n=380) and 12% (n=446) respectively, with 17.4% (n=632) cTTO responses with a value that was WTD (Figure 1). There were 40 participants with inconsistencies involving the worst state and out of these, 10 participants had inconsistencies that were at least 0.5 higher than that of the state with the lowest value. Each DCE profile was valued an average of 26 times (23-30) with few respondents exhibiting any specific response pattern indicative of poor engagement (n=11).

Interviewers reported only a small proportion of participants who did not understand (cTTO=3% DCE = 0.8%) or engage (cTTO = 1.7% DCE = 1%) with the tasks. Participants self-reported lack of understanding was similar (cTTO=3% and DCE=0.6%). More participants strongly agreed that they understood DCE (81%) compared to cTTO (59%) but they found it more difficult to decide on their answers for DCE (42%) compared to cTTO (23%).

Interviewers completed between 48 and 94 interviews with an average duration of 28.3 minutes (standard deviation = 8.2) for the cTTO part including the practice with some indication that one interviewer (No. 3) was faster than the other interviewers (see Supplementary Figure 2). However, this did not translate into differences in cTTO values (see Supplementary Figure 2). None of the interviewers experienced a flag rate requiring exclusion of data (Supplementary Table 3) and some of the flags are related to technical problems with EQ-PVT although earlier interviews were not flagged in this way.

Modelling results

Nine models were estimated including linear random effect models (models 1 and 2) and Tobit models (models 3 and 4) for cTTO data both with and without controlling for heteroscedasticity, a probit random effects (model 5) and probit model controlling for heteroscedasticity (model 6) for DCE data and a hybrid Tobit

[1] This statistic is based on a younger population than our sample (which includes aged 66+), however in this age group the proportion holding a degree in the general population is relatively low.

[2] Participants are asked if they have any long-standing physical or mental impairment, illness or disability which has been diagnosed by a doctor and has, or is likely to, trouble them for a period of at least 12 months.

model for the combined data with and without heteroscedasticity and merging inconsistent levels (models 7-9) (Table 2 and 3).

Pain had the biggest disutility across all the models (Tables 2 and 3). This was followed by mobility or daily activities then sadness/depression or loneliness depending on the model. The order for the rest of the dimensions varied depending on the model, with anxiety and control having the least impact. Pain dominated the other physical health dimensions (mobility, activities, exhaustion) with a large difference at the worst level e.g. the disutility at level 5 for mobility was 0.20 (hybrid heteroscedastic tobit model, Table 3) and it was 0.357 for pain. Exhaustion had a low impact relative to the other physical health dimensions. In the other dimensions, sadness/depression and loneliness had larger impact than cognition, anxiety and control.

Coefficients for levels 2 and 3 were less than 0.1 in all the models for all dimensions (Tables 2 and 3). Some of the wellbeing and mental health related dimensions (anxiety, control, cognition) also had coefficients that were less than 0.1 for levels 4 and 5 in all models. Moving between levels was associated with incremental changes that were below 0.1 apart from in the pain dimension where moving from level 3 to 4 and 4 to 5 was associated with changes greater than 0.2 and 0.1 respectively (Tables 2 and 3). There were differences across the coefficient sizes depending on the model especially with differences between models that took into account heteroscedasticity and those that did not (Tables 2 and 3). All the models had some logical inconsistencies or negative coefficients (Tables 2 and 3). There were logical inconsistencies in the cognition (level 5) and anxiety (level 3) dimensions across all the models and for exhaustion (level 5) and control (level 2) for at least three or more models. In addition, there was disordering in other anxiety levels (2 and 4), loneliness (level 4) and sadness/depression (level 3) in at least one model. Merging variables where there was disordering led to new disordered variables, sometimes on different dimensions across models apart from in the hybrid models.

Most of the coefficients were statistically significant at the 5% level (Tables 2 and 3). Statistical significance varied by the level (not statistically significant at level 2 or 3), whether or not heteroscedasticity was taken into account (e.g. level 2 exhaustion was statistically significant in the linear and Tobit heteroscedastic models) and the dimension (cognition, anxiety, sadness/depression and control). Hybrid models had fewer coefficients that were not statistically significant compared to the other models. Some of the logically inconsistent variables were not statistically significant.

Overall MAE for predicting states ranged between 0.027 to 0.066 while for the mild states this ranged between 0.005 to 0.022 (Tables 2 and 3). There was slight variation in the predicted values for three selected mild states depending on the model e.g. for state 1111111112 (mild pain), predictions ranged between 0.952 and 0.963. For the worst state, predictions ranged from -0.263 to -0.368. Predictions from the Tobit heteroscedastic model were correlated with those from the Probit heteroscedastic model (Lin's concordant correlation coefficient = 0.98, see Supplementary Figure 3).

Selected model

The hybrid Tobit model which controls for heteroscedasticity was selected as the pilot value set with values ranging from -0.384 for the worst state to 0.997 for the mildest state (Table 3, Figure 2). This model takes into account the nature of the data, combines data from cTTO and DCE which maximizes the data, and also had a low number of coefficients that were not statistically significant or disordered. It also had a lower AIC and BIC than the hybrid Tobit model but higher MAE.

Sensitivity analysis

Excluding those who did not understand or engage with either TTO or DCE had an impact on the coefficients (differences range = -0.0064 to 0.0096) (Table 3). The overall range when these respondents were excluded was -0.4 to 0.997. Excluding the mild states that were added to the design also had an impact on the coefficients (differences range= -0.0129 to 0.0224) with the overall range going from -0.395 to 1.019 (Table 3) and negative coefficients in the cognition and control level 2 dimensions.

Respondent gender was not a significant predictor and there was a u-shaped relationship between utility values and age (Supplementary Table 4).There was some evidence of interviewer effects on coefficient size and statistical significance. Therefore estimated range varied when individual interviewers were excluded sequentially (Supplementary Table 4).

DISCUSSION

The primary aim of this study was to assess the feasibility of valuing a new measure, the EQ-HWB-S, using TTO and DCE. The EQ-VTv2 international protocol involving TTO and DCE for valuing the EQ-5D-5L was modified and administered in a general population sample. The modifications accounted for the new measure including changing the upper anchor (health and quality of life), the number of tasks (less for cTTO) and the presentation of tasks (levels shown in the task and overlap for DCE). The need for less tasks, overlap and presentation of levels on screen reflect the cognitive burden of valuing a longer measure. Participants in the pilot recommended that all the severity levels were presented on screen to reduce the cognitive burden of considering states. This may also indicate that the levels are not distinct enough for participants to distinguish between them independently. Conducting interviews online via videoconference enabled the study to proceed during a period when face-to-face interactions were more restricted due to COVID-19.

The results indicate that applying the modified protocol to the new measure was feasible. The cTTO data covered the full TTO range from -1 to 1 with a proportion of values below zero with lower mean and larger standard deviations observed for states as the level sum score increased. However, a number of states had the same level sum score due to the mix of response levels which may impact on the comparison of states by level sum scores. The proportion of states valued at selected points (e.g. -1, 0 and 1) was 7%, 3% and 12% which was reasonable compared to other EQ-VT studies e.g. US [25] had 15%, 5% and 21% while for Italy [11] (that was also conducted online via videoconferencing) this was 8%, 2% and 11% respectively.

There was a small proportion of participants whom interviewers thought did not engage or understand the preference-elicitation task, which is similar to the figure observed in other studies [11, 26]. The video conferencing added other challenges related to who was recruited and how they engaged which may have impacted on this. As with other studies, participants found it easier to understand the DCE task but they found it harder to decide between the pairs even though we had variation in only five out of the nine items. There were no major concerns raised with the quality control including in the number of inconsistencies or non-traders.

A hybrid Tobit model which takes into account heteroscedasticity was selected as a pilot value set on the basis of appropriate models, maximizing use of data, the number of inconsistencies and statistical significance. The values for this model range from -0.384 for the worst state to 0.997 for the mildest state. Although direct comparison of disutilities associated with other measures is difficult due to differences in questions, responses and states (due to additional dimensions), it is useful for contextualise the EQ-HWB-S disutilities. The disutility associated with being at level 5 were generally smaller in EQ-HWB-S compared to EQ-5D-5L in overlapping dimensions (mobility: 0.200 vs. 0.22 to 0.613; usual activities: 0.195 vs. 0.153 to 0.385; pain: 0.355 vs 0.246 to 0.612; anxiety/depression: 0.088/0.171 vs. 0.19 to 0.646) based on results from 25 EQ-5D-5L value sets [27].

In terms of ranking the size of utility decrement across dimensions, pain, daily activities, mobility, sadness/depression and loneliness tended to have the largest overall impact with anxiety, exhaustion, control and cognition ranking lower. EQ-5D-5L studies tend to have pain, mobility and anxiety/depression as the worst dimensions [10]. The EQ-5D-3L UK value set had pain, mobility and anxiety/depression as the dimensions with the biggest impact [28] while this was pain, anxiety/depression then mobility for the English EQ-5D-5L value set [29]. Anxiety is combined with depression in the EQ-5D studies but is a separate item in the EQ-HWB-S and it did not get a large weight in this feasibility study. This may indicate that the weight given to anxiety in EQ-5D reflects the weight of depression. McDonald et al [30] found that in valuation, depression was perceived as worse than anxiety at the same level but the values associated with the composite anxiety/depression were between depression and anxiety values. In the context of additional dimensions, the relative weight given to anxiety may be even less although direct comparisons are not possible as the responses are severity for EQ-5D-5L while they are frequency for EQ-HWB-S.

Ranking of the remaining dimensions, reflecting aspects of wellbeing, showed that loneliness is important but the other dimensions had relatively low weight even at the most severe levels. Fatigue can have low utility decrements relative to physical, emotional and social functioning and pain e.g. in the EORTC QLU-C10D disutility ranges between -0.036 and -0.058 for UK weights [6] and -0.023 to -0.037 for the Australian weights [7]. In an exploratory TTO study [31], addition of a tiredness bolt-on dimension to the EQ-5D-3L states did not have a statistically significant impact on the values for these states which may indicate that relative to the other dimensions, it did not matter as much. Pairwise studies show that EQ-5D-5L dimensions are ranked higher than social care related dimensions such as control [32] while cognition ranks high relative to other potential bolt-on dimensions such as relationships, energy (tiredness) and sleep [33]. These pairwise studies do not provide weights to assess the actual weight – it may be that highly ranked dimensions do not attract large utility decrements relative to the other dimensions. In the Adult Social Care Outcomes Toolkit measures for social care users [34] and for informal carers [35] having control over daily life was weighted highly using best-worst scaling and not having control was considered to be the worst aspect relative to other social care and informal care dimensions (which did not include any health dimensions), indicating that control is important.

Small utility decrements for some dimensions and levels, inconsistencies and lack of statistical significance may raise the question about the relevance of these dimensions or severity EQ-HWB-S. However, other studies also have inconsistencies e.g. US **[25]** and lack of statistical significance in some models e.g. Italy [11], meaning that this may not be indicative in its own right. Further, these may also be explained by the relatively small sample and size of the design relative to the size of the classification system. In addition, a single study may not be sufficient to make a judgement. Information may be required regarding the applicability of the selected items e.g. Finch et al **[33]** used 'remembering' as the cognition test bolt-on dimension while we used 'concentrating and thinking clearly'. Alternative questions are available in the EQ-HWB which has 25 questions. There are also questions regarding the impact of experience. The additional dimensions in the EQ-HWB-S are aspects of quality of life that participants may think they understand e.g. feeling anxious or exhausted because they have experienced these and think they can cope. On the other hand, although few participants may have experienced complete immobility or very severe pain, they imagine that this would be difficult to cope with. Crocker et al **[36]** found that individuals who had a disability placed relatively higher importance on broader quality of life dimensions such as control relative to health status focused dimensions based on a ranking exercise. Qualitative evidence **[37]**, psychometric evidence **[38]** and stakeholders views [2] highlight the importance and validity of these additional dimensions and it is important to ensure they also reflect the values of those who will be impacted by decisions-based on the EQ-HWB. It is important to note that the small size of utility decrements is indicative of relative importance, not that these dimensions are unimportant. We used an international protocol that has been tested and refined over several stages to ensure high data quality for EQ-5D-5L valuation [8]. However, it would be useful to develop and test different approaches to minimise the burden of valuing a longer measure such as EQ-HWB-S in order to inform an international protocol for future valuation studies. This includes selecting states, presentation of TTO and DCE states and potentially separation of the TTO and DCE into different samples/modes of administration as has been done for the EQ-5D-Y [17] or in 'Lite' [39] protocol studies that offer a cost effective valuation option.

Limitations

This study was a feasibility study with a relatively small sample. Due to the pandemic related restrictions, interviews were conducted online. Online participants may not be representative – our sample was highly educated and there were more females. We also encountered problems with recruitment, and engagement can be more difficult in an online environment. Although the quality control criteria was met by all the interviewers, interviewers did varying numbers of interviews and there was some evidence of interviewer effects including time taken. The mixed states from the TTO design may have been more difficult for participants to engage with. This was mitigated by reducing the number of states participants valued but this reduced the number of times each state was valued. Coverage across different level sum scores was low in the TTO health state selection. Future studies should consider using the existing study data to optimize the design and selection of the states for valuation of the EQ-HWB-S.

Conclusion

This feasibility study demonstrated that EQ-VT could be applied to the EQ-HWB-S and an initial pilot value set has been generated. Future work to develop a valuation protocol for EQ-HWB as well as to test whether all the dimensions of the EQ-HWB-S should be retained is recommended.

REFERENCES

1. Herdman, M., et al., *Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L)*. 2011. p. 1727-1736.
2. Brazier, J., et al., *The EQ Health and wellbeing: overview of the development of a measure of health and wellbeing and key results.* Value in Health, 2022.
3. Brazier, J., et al., *Measuring and valuing health benefits for economic evaluation*. 2017: OXFORD university press.
4. Jonker, M.F., et al., *Attribute level overlap (and color coding) can reduce task complexity, improve choice consistency, and decrease the dropout rate in discrete choice experiments.* Health economics, 2019. **28**(3): p. 350-363.
5. Mulhern, B., et al., *How should discrete choice experiments with duration choice sets be presented for the valuation of health states?* Medical Decision Making, 2018. **38**(3): p. 306-318.
6. Norman, R., et al., *UK utility weights for the EORTC QLU ‐C10D.* Health Economics, 2019. **28**(12): p. 1385-1401.
7. 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.
8. 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 in Health, 2019. **22**(1): p. 23-30.
9. Mukuria, C., T. Peasgood, and J. Brazier, *Applying EuroQol Portable Valuation Technology to the EQ Health and Wellbeing Short (EQHWB-S): a pilot study.* School of Health and Related Research, University of Sheffield Discussion Paper Series, 2021.
10. Roudijk, B., K. Ludwig, and N. Devlin, *EQ-5D-5L Value Set Summaries*, in *Value Sets for EQ-5D-5L: A Compendium, Comparative Review & User Guide*, N. Devlin, B. Roudijk, and K. Ludwig, Editors. 2022, Springer International Publishing: Cham. p. 55-212.
11. Finch, A.P., et al., *An EQ-5D-5L value set for Italy using videoconferencing interviews and feasibility of a new mode of administration.* Social Science & Medicine, 2022. **292**: p. 114519.

12. Estévez-Carrillo, A., et al., *Exploring the Comparability of Face-to-Face Versus Video Conference-Based Composite Time Trade-Off Interviews: Insights from EQ-5D-Y-3L Valuation Studies in Belgium and Spain.* The Patient-Patient-Centered Outcomes Research, 2022: p. 1-15.
13. Lipman, S.A., *Time for tele-TTO? Lessons learned from digital interviewer-assisted time trade-off data collection.* The Patient-Patient-Centered Outcomes Research, 2021. **14**(5): p. 459-469.
14. Choice Metrics, *Ngene [software for experimental design]*. 2018, Choice Metrics: Sydney.
15. Cook, R.D. and C.J. Nachtrheim, *A comparison of algorithms for constructing exact D-optimal designs.* Technometrics, 1980. **22**(3): p. 315-324.
16. 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.
17. Yang, F., et al., *Developing the EQ-5D-5L Value Set for Uganda Using the 'Lite'Protocol.* PharmacoEconomics, 2022. **40**(3): p. 309-321.
18. Rowen, D., C. Mukuria, and E. McDool, *A systematic review of the methodologies and modelling approaches used to generate international EQ-5D-5L value sets.* PharmacoEconomics, 2022.
19. 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.
20. Lawrence, I. and K. Lin, *A concordance correlation coefficient to evaluate reproducibility.* Biometrics, 1989: p. 255-268.
21. Rowen, D., J. Brazier, and B. Van Hout, *A comparison of methods for converting DCE values onto the full health-dead QALY scale.* Medical Decision Making, 2015. **35**(3): p. 328-340.
22. StataCorp, *Stata Statistical Software: Release 17*. 2021, StataCorp LLC: College Station, TX.
23. Office for National Statistics, *Dataset: A05 SA: Employment, unemployment and economic inactivity by age group (seasonally adjusted)*, O.f.N. Statistics, Editor. 2022.
24. Office for National Statistics. *Graduates in the UK labour market: 2017*. 2017; Available from: https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/articles/graduatesintheuklabourmarket/2017.
25. Pickard, A.S., et al., *United States valuation of EQ-5D-5L health states using an international protocol.* Value in Health, 2019. **22**(8): p. 931-941.
26. 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.
27. Roudijk, B., B. Janssen, and J.A. Olsen, *How Do EQ-5D-5L Value Sets Differ?*, in *Value Sets for EQ-5D-5L: A Compendium, Comparative Review & User Guide*, N. Devlin, B. Roudijk, and K. Ludwig, Editors. 2022, Springer International Publishing: Cham. p. 235-258.
28. Dolan, P., *Modeling valuations for EuroQol health states.* Med Care, 1997. **35**(11): p. 1095-108.
29. 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.
30. McDonald, R., T.L. Mullett, and A. Tsuchiya, *Understanding the composite dimensions of the EQ-5D: an experimental approach.* Social Science & Medicine, 2020. **265**: p. 113323.
31. Yang, Y., et al., *An exploratory study to test the impact on three "bolt-on" items to the EQ-5D.* Value in Health, 2015. **18**(1): p. 52-60.
32. Mulhern, B., et al., *Investigating the relative value of health and social care related quality of life using a discrete choice experiment.* Social science & medicine, 2019. **233**: p. 28-37.
33. Finch, A.P., J. Brazier, and C. Mukuria, *Selecting bolt-on dimensions for the EQ-5D: testing the impact of hearing, sleep, cognition, energy, and relationships on preferences using pairwise choices.* Medical Decision Making, 2021. **41**(1): p. 89-99.
34. Netten, A., et al., *Outcomes of social care for adults: developing a preference-weighted measure.* 2012. **16**: p. 16.
35. Batchelder, L., et al., *Carer Social Care-Related Quality of Life Outcomes: Estimating English Preference Weights for the Adult Social Care Outcomes Toolkit for Carers.* Value in Health, 2019. **22**(12): p. 1427-1440.
36. Crocker, M., et al., *Assessing the relative importance of key quality of life dimensions for people with and without a disability: an empirical ranking comparison study.* Health and Quality of Life Outcomes, 2021. **19**(1): p. 1-12.
37. Mukuria, C., et al., *Qualitative Review on Domains of Quality of Life Important for Patients, Social Care Users, and Informal Carers to Inform the Development of the EQ-HWB.* Value in Health, 2022. **25**(4): p. 492-511.
38. Peasgood, T., et al., *Developing a New Generic Health and Wellbeing Measure: Psychometric Survey Results for the EQ Health and Wellbeing.* Value in Health, 2022.
39. Ramos-Goñi, J.M., et al., *International valuation protocol for the EQ-5D-Y-3L.* Pharmacoeconomics, 2020. **38**(7): p. 653-663.
40. Office for National Statistics, *Zipped population projections data files, UK* O.f.N. Statistics, Editor. 2022.

TABLES

Table 1: Respondent characteristics

	(1) Full sample (%) n=521[b]	(2) General population UK (%)
Age	48.45	
18-30	20.15	20.53
31-50	31.67	32.61
51-65	29.94	24.33
66+	18.23	22.53
Sex		
Male	45.30	49.4
Female	54.13	50.6
Other	0.58	
Ethnicity		
White	81.96	86
Black	5.95	4
Asian	9.40	8
Mixed / Other	2.69	2
Employment/activity status		
Employed	59.42	
Unemployed	2.88	
Caring for family	3.08	
Looking after home	4.62	
Student	4.62	
Retired	22.50	
Long-term sick	2.69	
Other inc. volunteer	0.19	
Working status		
Usual place	47.87	
Working from home	50.49	
Furloughed	0.66	
Leave of absence	0.98	
Degree	66.35	
Social care services used	3.27	
Caring responsibilities (elderly/disabled)	14.04	
1 to 19 hours	61.64	
20 to 49 hours	19.18	
50 or more hours	13.70	
Don't know – caring hours	5.48	
Additional caring responsibilities e.g. children	11.73	
Long term health condition	31.54	
Experience of serious illness		
Serious illness – self	21.92	
Serious illness – family	52.69	
Serious illness – others	15.77	
Health satisfaction	7.01	
Life satisfaction	7.18	
EQ-VAS (n=520)	76.29	
EQ-5D-5L index value (n=429)[a]	0.80	
General health		
Excellent/very good/good	84.42	
Fair/poor	15.58	

Notes: The age group percentages are calculated as the percentage of the adult UK population based on population projection data from the ONS [40]. Ethnicity figures are sourced from the 2011 census. [a]All participants were asked to complete the EQ-5D-5L but these questions were not compulsory to complete. [b]Sample consists of 521 individuals but we only have the full characteristics of 520 individual – for one respondent we only have age sex and ethnicity

Table 2: Parameter estimates for main effects models – Linear, Tobit, Probit

	Linear RE		Linear Het		Tobit RE		Tobit Het		Probit RE		Rescaled	Probit Het		Rescaled
	1		2		3		4		5			6		
Mobility2	0.0114	(0.019)	0.0494***	(0.012)	0.0077	(0.021)	0.0489***	(0.012)	0.3488***	(0.083)	0.0552	0.2574**	(0.105)	0.0415
Mobility3	0.0521**	(0.021)	0.0977***	(0.024)	0.0515**	(0.023)	0.1004***	(0.027)	0.4562***	(0.067)	0.0721	0.5040***	(0.097)	0.0813
Mobility4	0.1155***	(0.021)	0.1171***	(0.027)	0.1158***	(0.022)	0.1153***	(0.028)	0.8224***	(0.092)	0.1300	0.8411***	(0.121)	0.1357
Mobility5	0.1622***	(0.020)	0.1841***	(0.022)	0.1672***	(0.021)	0.1947***	(0.025)	1.3004***	(0.088)	0.2056	1.2709***	(0.122)	0.2050
Activity2	0.0438**	(0.021)	0.0468***	(0.010)	0.0438**	(0.022)	0.0458***	(0.010)	0.1751*	(0.091)	0.0277	0.0921	(0.124)	0.0149
Activity3	0.0804***	(0.022)	0.0879***	(0.026)	0.0833***	(0.024)	0.0890***	(0.029)	0.3702***	(0.072)	0.0585	0.3138***	(0.105)	0.0506
Activity4	0.1364***	(0.021)	0.1298***	(0.021)	0.1395***	(0.022)	0.1314***	(0.022)	0.8700***	(0.080)	0.1376	0.9122***	(0.102)	0.1471
Activity5	0.1720***	(0.022)	0.1723***	(0.027)	0.1790***	(0.024)	0.1826***	(0.030)	1.1283***	(0.094)	0.1784	1.1915***	(0.108)	0.1922
Exhaustion2	0.0202	(0.021)	0.0292***	(0.009)	0.0206	(0.022)	0.0286***	(0.009)	0.1676**	(0.069)	0.0265	0.1159	(0.091)	0.0187
Exhaustion3	0.0405**	(0.020)	0.0608***	(0.022)	0.0404*	(0.022)	0.0589**	(0.024)	0.2060***	(0.070)	0.0326	0.1569**	(0.080)	0.0253
Exhaustion4	0.0663***	(0.022)	0.0734***	(0.027)	0.0660***	(0.023)	0.0709**	(0.030)	0.4155***	(0.070)	0.0657	0.3818***	(0.083)	0.0616
Exhaustion5	0.0496**	(0.022)	0.0565**	(0.026)	0.0527**	(0.024)	0.0590**	(0.028)	0.5966***	(0.069)	0.0943	0.5744***	(0.095)	0.0926
Loneliness2	0.0353*	(0.020)	0.0230***	(0.006)	0.0379*	(0.022)	0.0224***	(0.007)	0.0814	(0.077)	0.0129	0.0980	(0.092)	0.0158
Loneliness3	0.0619***	(0.021)	0.0735***	(0.026)	0.0626***	(0.023)	0.0703**	(0.029)	0.2859***	(0.073)	0.0452	0.3380***	(0.097)	0.0545
Loneliness4	0.0873***	(0.022)	0.0536**	(0.027)	0.0894***	(0.023)	0.0507*	(0.029)	0.5935***	(0.085)	0.0938	0.5669***	(0.102)	0.0914
Loneliness5	0.1271***	(0.021)	0.0969***	(0.024)	0.1309***	(0.023)	0.0991***	(0.026)	0.7222***	(0.091)	0.1142	0.7747***	(0.120)	0.125
Cognition2	0.0123	(0.019)	0.0212***	(0.006)	0.0088	(0.020)	0.0204***	(0.006)	0.2519***	(0.076)	0.0398	0.1780*	(0.092)	0.0287
Cognition3	0.0440**	(0.021)	0.0634***	(0.022)	0.0422*	(0.022)	0.0610***	(0.024)	0.3219***	(0.080)	0.0509	0.1812**	(0.091)	0.0292
Cognition4	0.0803***	(0.021)	0.0783***	(0.026)	0.0788***	(0.022)	0.0774***	(0.029)	0.6453***	(0.082)	0.102	0.5019***	(0.094)	0.0809
Cognition5	0.0281	(0.021)	0.0433**	(0.022)	0.0264	(0.022)	0.0434*	(0.023)	0.5977***	(0.085)	0.0945	0.3704***	(0.091)	0.0597
Anxiety2	0.0319	(0.022)	0.0283***	(0.010)	0.0307	(0.023)	0.0270***	(0.010)	-0.0616	(0.082)	-0.0097	-0.0737	(0.096)	-0.0119
Anxiety3	0.0241	(0.022)	0.0356*	(0.019)	0.0231	(0.023)	0.0339*	(0.020)	-0.0109	(0.082)	-0.0017	0.0482	(0.087)	0.0078
Anxiety4	0.0478**	(0.021)	0.0318	(0.024)	0.0491**	(0.022)	0.0304	(0.026)	0.3141***	(0.078)	0.0497	0.5137***	(0.094)	0.0828
Anxiety5	0.0603***	(0.022)	0.0561**	(0.025)	0.0651***	(0.023)	0.0624**	(0.027)	0.5057***	(0.088)	0.08	0.5666***	(0.109)	0.0914
Sad/depress2	0.0344	(0.021)	0.0345***	(0.009)	0.0343	(0.023)	0.0333***	(0.009)	0.1396*	(0.084)	0.0221	0.0855	(0.098)	0.0138
Sad/depress3	0.0400*	(0.021)	0.0266	(0.018)	0.0413*	(0.022)	0.0273	(0.019)	0.2588***	(0.077)	0.0409	0.1178	(0.097)	0.019
Sad/depress4	0.0933***	(0.022)	0.0528*	(0.030)	0.0946***	(0.024)	0.0520	(0.033)	0.6941***	(0.085)	0.1098	0.5462***	(0.105)	0.0881
Sad/depress5	0.1672***	(0.021)	0.1466***	(0.024)	0.1731***	(0.022)	0.1545***	(0.027)	1.0400***	(0.089)	0.1644	0.9370***	(0.112)	0.1511
Control2	0.0073	(0.021)	0.0108*	(0.006)	0.0078	(0.022)	0.0103*	(0.006)	-0.0011	(0.075)	-0.0002	-0.1303	(0.099)	-0.021
Control3	0.0589***	(0.023)	0.0541*	(0.030)	0.0596**	(0.024)	0.0543*	(0.033)	0.2599***	(0.078)	0.0411	0.1279	(0.103)	0.0206
Control4	0.0752***	(0.022)	0.0664**	(0.029)	0.0778***	(0.024)	0.0703**	(0.032)	0.3876***	(0.087)	0.0613	0.2105	(0.147)	0.0339

	Linear RE		**Linear Het**		**Tobit RE**		**Tobit Het**		**Probit RE**		**Rescaled**	**Probit Het**		**Rescaled**
	1		**2**		**3**		**4**		**5**			**6**		
Control5	0.0779***	(0.021)	0.0780***	(0.022)	0.0819***	(0.022)	0.0827***	(0.024)	0.4986***	(0.081)	0.0788	0.4223***	(0.107)	0.0681
Pain2	0.0481**	(0.020)	0.0409***	(0.014)	0.0460**	(0.021)	0.0398***	(0.014)	0.2568***	(0.085)	0.0406	0.2772***	(0.094)	0.0447
Pain3	0.0610***	(0.021)	0.0827***	(0.025)	0.0600***	(0.022)	0.0822***	(0.027)	0.4845***	(0.079)	0.0766	0.5852***	(0.102)	0.0944
Pain4	0.3170***	(0.022)	0.3149***	(0.028)	0.3241***	(0.023)	0.3250***	(0.031)	1.4529***	(0.090)	0.2297	1.4794***	(0.127)	0.2386
Pain5	0.4184***	(0.021)	0.4290***	(0.031)	0.4336***	(0.023)	0.4566***	(0.035)	2.0528***	(0.114)	0.3246	2.1692***	(0.172)	0.3499
Observations	3,640		3,640		3,640		3,640		3,647			3,647		
AIC	3554		3616		4359		4280		2997			2977		
BIC	3790		4068		4595		4732		3221			3424		
Disordered	Ex5,Cg5,An3		Ex5,Cg5,Ln4,An4,Sd3		Ex5,Cg5,An3		Ex5,Cg5,Ln4,An4,Sd3		Cg5,An3,Cl2			Cg5,An3,Cl2		
MAE	0.030		0.027		0.036		0.039				0.059			0.066
MAE mild states	0.011		0.005		0.012		0.006				0.016			0.022
Ranking	PN>AC>SD>		PN>MO>AC>		PN>AC>SD>		PN>MO>AC>		PN>MO>AC>			PN>MO>AC>		
	MO>LN>GG>		SD>LN>CG>		MO>LN>CL>		SD>LN>CL>		SD>LN>CG>			SD>LN>EX>		
	CL>EX>AN		CL>EX>AN		CG>EX>AN		CG>EX>AN		EX>AN>CL			AN>CL>CG		
Range of values	-0.263 to 0.993		-0.263 to 0.989		-0.310 to 0.992		-0.335 to 0.990		-0.335 to 1.010			-0.335 to 1.021		
Estimated utility by selected states														
111111112	0.952		0.959		0.954		0.960				0.959			0.955
211111111	0.989		0.951		0.992		0.951				0.945			0.958
111121111	0.988		0.979		0.991		0.980				0.960			0.971
555555555	-0.263		-0.263		-0.310		-0.335				-0.335			-0.335

RE – random effects; Het – Heteroscedasticity

Standard errors in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

MAE Mean absolute error by state; mild states – states with only one or two dimensions at level 2.

Heteroscedastic models estimated with clustered standard errors to account for repeated data

Ranking based on largest decrement for each dimension

MO mobility AC activity EX exhaustion LN loneliness CG cognition AN anxiety SD sadness/depression CL control PN pain

Rescaled - the DCE values were anchored using the cTTO value for the worst state from the Tobit heteroscedastic model

Table 3: Parameter estimates for main effects models – Hybrid models

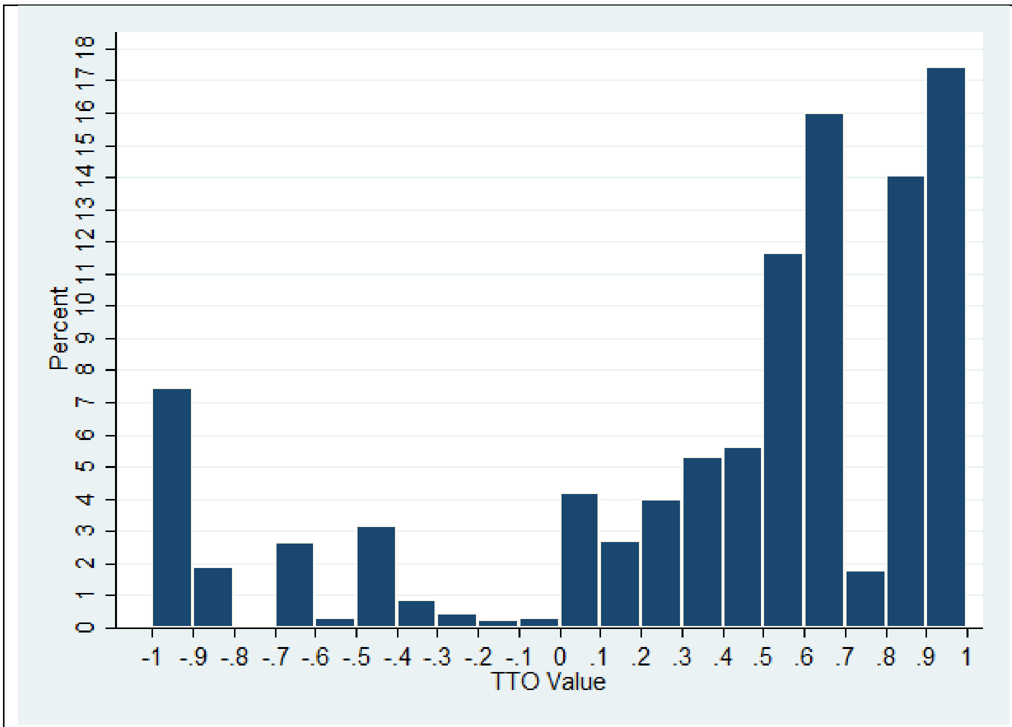
	Hybrid Tobit		Hybrid Tobit with control for Heteroscedasticity					
	7		8		9		Exclusions from 9	
							Understand/engage	No Mild
Mobility2	0.0474***	(0.011)	0.0496***	(0.009)	0.0534***	(0.009)	0.0521***	0.0489***
Mobility3	0.0720***	(0.010)	0.0705***	(0.009)	0.0699***	(0.010)	0.0702***	0.0742***
Mobility4	0.1359***	(0.013)	0.1315***	(0.013)	0.1364***	(0.013)	0.1300***	0.1493***
Mobility5	0.1966***	(0.012)	0.2002***	(0.012)	0.2071***	(0.013)	0.2004***	0.2151***
Activity2	0.0400***	(0.014)	0.0379***	(0.008)	0.0409***	(0.008)	0.0403***	0.0441***
Activity3	0.0627***	(0.011)	0.0593***	(0.010)	0.0627***	(0.010)	0.0636***	0.0691***
Activity4	0.1451***	(0.012)	0.1414***	(0.011)	0.1498***	(0.011)	0.1499***	0.1576***
Activity5	0.1924***	(0.014)	0.1952***	(0.014)	0.1985***	(0.014)	0.2045***	0.2030***
Exhaustion2	0.0219**	(0.011)	0.0226***	(0.007)	0.0187**	(0.007)	0.0215***	0.0079
Exhaustion3	0.0336***	(0.010)	0.0290***	(0.009)	0.0273***	(0.009)	0.0291***	0.0228**
Exhaustion4	0.0736***	(0.011)	0.0713***	(0.011)	0.0664***	(0.011)	0.0653***	0.0651***
Exhaustion5	0.0855***	(0.010)	0.0838***	(0.010)	0.0820***	(0.011)	0.0820***	0.0796***
Loneliness2	0.0197*	(0.011)	0.0190***	(0.006)	0.0207***	(0.006)	0.0219***	0.0190
Loneliness3	0.0484***	(0.012)	0.0443***	(0.010)	0.0515***	(0.011)	0.0528***	0.0523***
Loneliness4	0.0946***	(0.012)	0.0921***	(0.011)	0.1010***	(0.011)	0.1053***	0.1050***
Loneliness5	0.1162***	(0.013)	0.1159***	(0.012)	0.1201***	(0.012)	0.1239***	0.1258***
Cognition2	0.0239**	(0.011)	0.0222***	(0.006)	0.0033	(0.007)	0.0027	-0.0191*
Cognition3	0.0425***	(0.012)	0.0385***	(0.011)	0.0158	(0.011)	0.0157	0.0112
Cognition4	0.0944***	(0.012)	0.0925***	(0.011)	0.0569***	(0.009)	0.0583***	0.0506***
Cognition5	0.0771***	(0.012)	0.0765***	(0.011)	0.0569***	(0.009)	0.0583***	0.0506***
Anxiety2	0.0103	(0.012)	0.0134*	(0.007)	0.0219***	(0.006)	0.0225***	0.0246***
Anxiety3	0.0062	(0.012)	0.0079	(0.009)	0.0219***	(0.006)	0.0225***	0.0246***
Anxiety4	0.0601***	(0.012)	0.0627***	(0.011)	0.0688***	(0.011)	0.0710***	0.0729***
Anxiety5	0.0840***	(0.013)	0.0877***	(0.011)	0.0924***	(0.011)	0.0923***	0.0999***
Sad/depress2	0.0297**	(0.012)	0.0314***	(0.007)	0.0311***	(0.007)	0.0342***	0.0261**
Sad/depress3	0.0418***	(0.012)	0.0457***	(0.010)	0.0338***	(0.009)	0.0368***	0.0330***
Sad/depress4	0.1127***	(0.013)	0.1133***	(0.012)	0.1130***	(0.012)	0.1158***	0.1134***
Sad/depress5	0.1671***	(0.012)	0.1717***	(0.011)	0.1727***	(0.011)	0.1791***	0.1727***
Control2	-0.0039	(0.011)	0.0042	(0.006)	0.0038	(0.006)	0.0055	-0.0101

	Hybrid Tobit		**Hybrid Tobit with control for Heteroscedasticity**					
	7		**8**		**9**		**Exclusions from 9**	
							Understand/engage	**No Mild**
Control3	0.0451***	(0.012)	0.0479***	(0.011)	0.0447***	(0.011)	0.0457***	0.0420***
Control4	0.0633***	(0.013)	0.0650***	(0.011)	0.0653***	(0.012)	0.0674***	0.0605***
Control5	0.0754***	(0.011)	0.0797***	(0.010)	0.0820***	(0.011)	0.0844***	0.0746***
Pain2	0.0361***	(0.012)	0.0368***	(0.010)	0.0383***	(0.010)	0.0287***	0.0267**
Pain3	0.0783***	(0.012)	0.0753***	(0.011)	0.0802***	(0.011)	0.0774***	0.0752***
Pain4	0.2471***	(0.014)	0.2476***	(0.014)	0.2575***	(0.014)	0.2556***	0.2578***
Pain5	0.3548***	(0.019)	0.3567***	(0.018)	0.3718***	(0.018)	0.3749***	0.3738***
Observations	7,287		7,287		7,287		7,042	6,767
AIC	8746		7248		7301		6910	8277
BIC	9008		7758		7784		7390	8754
Disordered	Cg5,An3,Cl2		Cg5,An3,Cl2		-		-	Cg2, Cl2
MAE	0.049		0.056		0.056		0.058	0.058
MAE mild	0.009		0.010		0.010		0.010	0.016
Ranking	PN>MO>AC> SD>LN>CG> EX>AN>CL		PN>MO>AC> SD>LN>CG> EX>AN>CL		PN>MO>AC> SD>LN>AN> EX>CL>CG		PN>AC>MO> SD>LN>AN> CL>EX>CG	PN>AC>MO> SD>LN>AN> EX>CL>CG
Range of values	-0.349 to 1.004		-0.368 to 0.996		-0.384 to 0.997		-0.4 to 0.997	-0.395 to 1.019
Estimated utility by selected states								
111111112	0.962		0.963		0.962		0.971	0.973
211111111	0.950		0.950		0.947		0.948	0.951
111121111	0.976		0.978		0.997		0.977	1.019
555555555	-0.305		-0.368		-0.384		-0.4	-0.395

Standard errors in parenthesis. Models estimated with clustered standard errors to account for repeated data. *** p<0.01, ** p<0.05, * p<0.1
MAE Mean absolute error by state; mild states – states with only one or two dimensions at level 2. Ranking based on largest decrement for each dimension
MO mobility AC activity EX exhaustion LN loneliness CG cognition AN anxiety SD sadness/depression CL control PN pain

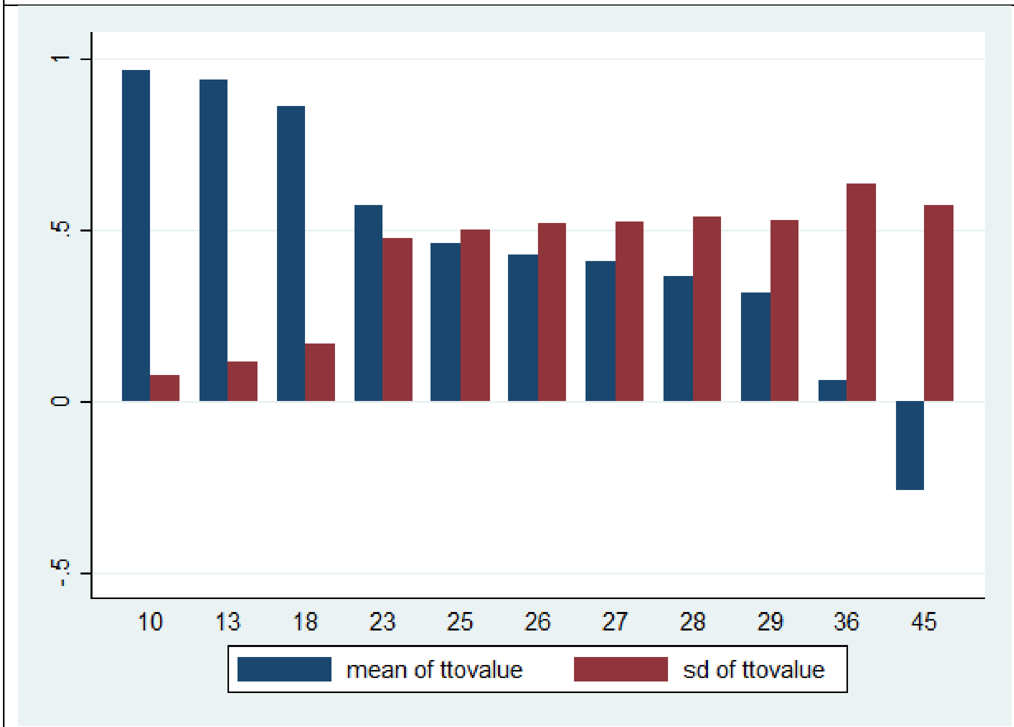
FIGURES

Figure 1 Distribution of TTO values



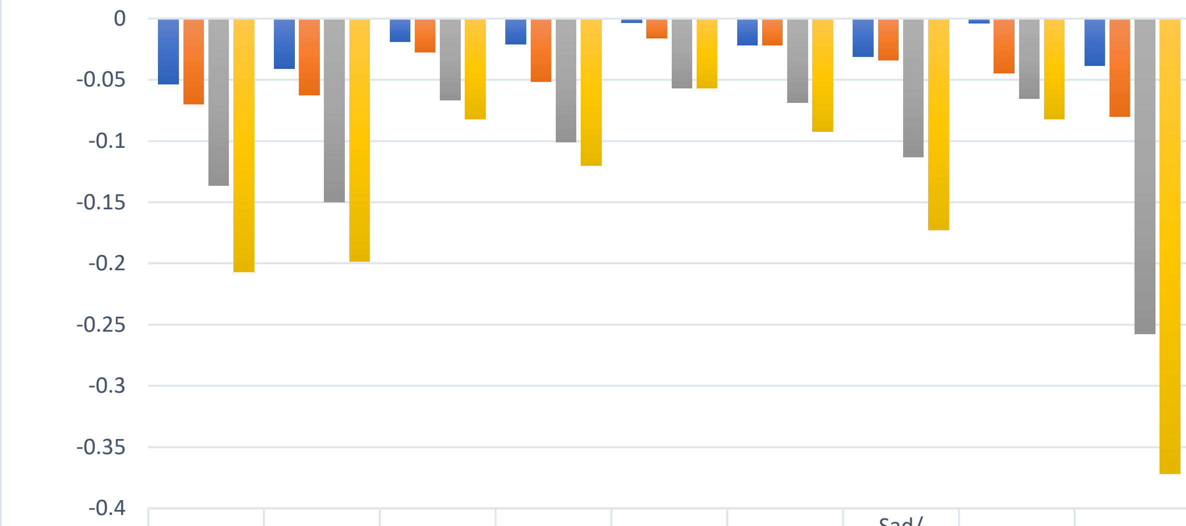
1a Distribution of TTO values

For example, the rightmost bar shows the proportion of observations of values greater than 0.90 and less than or equal to 1.0. This excludes practice TTO tasks.



1b Distribution by level sum score

Figure 2: Selected model – hybrid Tobit with heteroscedasticity



	Mobility	Activity	Exhaustion	Loneliness	Cognition	Anxiety	Sad/ depressed	Control	Pain
Level 2	-0.0534	-0.0409	-0.0187	-0.0207	-0.0033	-0.0219	-0.0311	-0.0038	-0.0383
Level 3	-0.0699	-0.0627	-0.0273	-0.0515	-0.0158	-0.0219	-0.0338	-0.0447	-0.0802
Level 4	-0.1364	-0.1498	-0.0664	-0.101	-0.0569	-0.0688	-0.113	-0.0653	-0.2575
Level 5	-0.2071	-0.1985	-0.082	-0.1201	-0.0569	-0.0924	-0.1727	-0.082	-0.3718

SUPPLEMENTARY MATERIAL

Supplementary Table 1: Sample of states described by the EQ-HWB

111111111	223322233	423142545	555555555
No difficulty getting around inside and outside	Slight difficulty getting around inside and outside	A lot of difficulty getting around inside and outside	Unable to get around inside and outside
No difficulty doing day to day activities	Slight difficulty doing day to day activities	Slight difficulty doing day to day activities	Unable to do day to day activities
Never exhausted	Sometimes exhausted	Sometimes exhausted	Exhausted most or all of the time
Never lonely	Sometimes lonely	Never lonely	Lonely most or all of the time
Never have trouble concentrating/thinking clearly	Only occasionally have trouble concentrating/thinking clearly	Often have trouble concentrating/thinking clearly	Trouble concentrating/thinking clearly most or all of the time
Never anxious	Only occasionally anxious	Only occasionally anxious	Anxious most or all of the time
Never sad/depressed	Only occasionally sad/depressed	Sad/depressed most or all of the time	Sad/depressed most or all of the time
Never feel you have no control over your day to day life	Sometimes feel you have no control over your day to day life	Often feel you have no control over your day to day life	Feel you have no control over your day to day life most or all of the time
No physical pain	Moderate physical pain	Very severe physical pain	Very severe physical pain

Supplementary Table 2: Observed TTO values by combined level sum score and state (e.g. 10_111111112)

State	mean	Sd	min	max	State	mean	sd	min	Max
10111111112	0.953	0.128	0.1	1	27124213545	0.144	0.627	-1	0.95
10111111121	0.982	0.044	0.8	1	27155134224	0.269	0.509	-1	1
10111111211	0.963	0.062	0.7	1	27242115453	0.534	0.459	-1	1
10111112111	0.975	0.065	0.6	1	27254452311	0.565	0.432	-0.8	1
10111121111	0.974	0.043	0.8	1	27321542154	0.226	0.589	-1	0.95
10111211111	0.973	0.047	0.75	1	27333333333	0.575	0.43	-1	1
10112111111	0.965	0.069	0.65	1	27413525241	0.587	0.414	-1	1
10121111111	0.955	0.077	0.65	1	27445251132	0.564	0.367	-0.6	1
10211111111	0.938	0.1	0.5	1	27512344512	0.445	0.49	-1	1
13111113311	0.933	0.113	0.5	1	27531421425	0.167	0.561	-1	0.95
18222222222	0.859	0.166	0.1	1	28131554342	0.59	0.459	-1	1
23125332421	0.688	0.368	-1	1	28143341255	0.118	0.573	-1	0.8
23313222541	0.462	0.54	-1	1	28215431543	0.552	0.404	-1	1
25112423354	0.475	0.485	-1	1	28334145521	0.499	0.483	-1	1
25135242413	0.515	0.483	-1	1	28352414135	0.153	0.565	-1	0.8
25234511234	0.315	0.571	-1	1	28425315314	0.228	0.567	-1	1
25253143142	0.52	0.476	-1	0.95	28451233451	0.465	0.5	-1	1
25322351441	0.57	0.421	-1	1	28513152434	0.29	0.549	-1	1
25341435212	0.5	0.364	-1	0.95	28544523113	0.353	0.532	-1	0.95
25414332125	0.198	0.61	-1	1	29154325432	0.407	0.539	-1	1
25421124533	0.499	0.437	-1	1	29223534415	0.152	0.548	-1	0.9
25543214321	0.559	0.502	-1	1	29241353524	0.24	0.444	-1	0.8
26123455123	0.628	0.38	-1	1	29314254253	0.532	0.427	-1	1
26142532531	0.379	0.576	-1	1	29345122345	0.158	0.609	-1	0.95
26211245335	0.319	0.599	-1	1	29433412552	0.31	0.519	-1	1
26235324151	0.612	0.365	-1	1	29452541323	0.407	0.474	-1	1
26315513422	0.642	0.338	-0.6	1	29525443231	0.398	0.481	-1	0.9
26353221514	0.209	0.574	-1	0.9	29532235144	0.219	0.607	-1	0.9
26432153215	0.205	0.564	-1	0.95	36444444444	0.058	0.634	-1	0.95
26524131352	0.482	0.44	-1	1	45555555555	-0.264	0.569	-1	1
26551312243	0.342	0.519	-1	1	Total	0.388	0.594	-1	1

Supplementary Table 3. Flagged interviews

		1		2		3		4		5		6		7
Interviewer	**Total interviews**	**Flagged interviews**		**Total flags**		**No lead time in Wheelchair (WC)**		**Inconsistency on worst state**		**WC time less than 3 min**		**TTO time less than 5 min**		**Technical problems**
	N	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**
1	50	2	4	2	4	0	0	2	4	0	0	0	0	0
2	80	2	3	2	3	2	3	0	0	0	0	0	0	0
3	68	4	6	7	10	2	3	1	1	2	3	2	3	2
4	87	4	5	4	5	0	0	0	0	2	2	2	2	2
5	94	2	2	4	4	2	2	0	0	2	2	0	0	2
6	48	6	13	7	15	3	6	3	6	1	2	0	0	0
7	93	9	10	10	11	1	1	4	4	1	1	4	4	0

This table shows how many times each interviewer's TTO data have been flagged for data quality reasons.

1) Flagged interviews – shows the number of interviews that were flagged.
2) Total number of flags – a given interview may be flagged more than once
3) No lead time in wheelchair - Interview is flagged if the interviewer does not enter the worse-than-dead element of one of the wheelchair examples
4) Inconsistency- Interview is flagged if the respondent has a clear inconsistency in their TTO ratings (the value for 55555 is not the lowest and is at least 0.5 higher than that of the state with the lowest value).
5) WC time - Interview is flagged if the interviewer does not spend at least 180 seconds (3 minutes) on the wheelchair example.
6) TTO time - Interview is flagged if the respondent does not spend at least 5 minutes on the 7 TTO tasks. Note this is a conservative estimate based on the longer EQ-HWB relative to EQ-5D-5L (5 min for 10 tasks).
7) Technical problems – number of interviews that were flagged and also had technical problems i.e. PowerPoint stopped and needed to be restart

Supplementary Table 4: Inclusion of age, gender and interviewer effects (hybrid Tobit controlling for heteroscedasticity)

		Age,gender	Interviewer Effects Excluding Individual Interviewers							
			Dummy	a	b	c	d	e	f	g
Mobility2	0.0534***	0.0504***	0.0526***	0.0540***	0.0556***	0.0527***	0.0544***	0.0590***	0.0535***	0.0432***
Mobility3	0.0699***	0.0702***	0.0703***	0.0753***	0.0724***	0.0616***	0.0672***	0.0705***	0.0723***	0.0667***
Mobility4	0.1364***	0.1354***	0.1362***	0.1404***	0.1484***	0.1324***	0.1286***	0.1288***	0.1392***	0.1332***
Mobility5	0.2071***	0.2078***	0.2076***	0.2099***	0.2050***	0.2069***	0.2099***	0.2028***	0.2154***	0.1956***
Activity2	0.0409***	0.0395***	0.0403***	0.0420***	0.0381***	0.0374***	0.0403***	0.0416***	0.0434***	0.0431***
Activity3	0.0627***	0.0636***	0.0630***	0.0645***	0.0600***	0.0607***	0.0648***	0.0616***	0.0676***	0.0578***
Activity4	0.1498***	0.1503***	0.1497***	0.1580***	0.1537***	0.1480***	0.1464***	0.1533***	0.1503***	0.1353***
Activity5	0.1985***	0.2001***	0.1991***	0.2007***	0.1935***	0.2065***	0.1984***	0.1935***	0.2015***	0.1967***
Exhaustion2	0.0187**	0.0185**	0.0180**	0.0213***	0.0183***	0.0192**	0.0156**	0.0244***	0.0175**	0.0139*
Exhaustion3	0.0273***	0.0261***	0.0266***	0.0254***	0.0296***	0.0266***	0.0266***	0.0271***	0.0287***	0.0257***
Exhaustion4	0.0664***	0.0666***	0.0662***	0.0646***	0.0705***	0.0616***	0.0647***	0.0711***	0.0662***	0.0655***
Exhaustion5	0.0820***	0.0818***	0.0815***	0.0773***	0.0819***	0.0803***	0.0817***	0.0928***	0.0816***	0.0773***
Loneliness2	0.0207***	0.0194***	0.0202***	0.0197***	0.0168***	0.0183***	0.0255***	0.0199***	0.0201***	0.0242***
Loneliness3	0.0515***	0.0525***	0.0520***	0.0538***	0.0421***	0.0478***	0.0543***	0.0467***	0.0541***	0.0602***
Loneliness4	0.1010***	0.1016***	0.1014***	0.1022***	0.0973***	0.1023***	0.1004***	0.0995***	0.0987***	0.1045***
Loneliness5	0.1201***	0.1216***	0.1206***	0.1198***	0.1145***	0.1226***	0.1148***	0.1158***	0.1179***	0.1334***
Cognition2	0.0033	0.0015	0.0024	0.0006	0.0033	0.0050	0.0087	0.0006	0.0062	-0.0014
Cognition3	0.0158	0.0151	0.0153	0.0139	0.0237*	0.0196*	0.0087	0.0139	0.0152	0.0149
Cognition4	0.0569***	0.0553***	0.0561***	0.0553***	0.0599***	0.0580***	0.0505***	0.0568***	0.0589***	0.0584***
Cognition5	0.0569***	0.0553***	0.0561***	0.0553***	0.0599***	0.0580***	0.0505***	0.0568***	0.0589***	0.0584***
Anxiety2	0.0219***	0.0190***	0.0208***	0.0212***	0.0219***	0.0232***	0.0212***	0.0212***	0.0229***	0.0211***
Anxiety3	0.0219***	0.0190***	0.0208***	0.0212***	0.0219***	0.0232***	0.0212***	0.0212***	0.0229***	0.0211***
Anxiety4	0.0688***	0.0666***	0.0677***	0.0706***	0.0689***	0.0672***	0.0720***	0.0676***	0.0664***	0.0657***
Anxiety5	0.0924***	0.0910***	0.0917***	0.0919***	0.0930***	0.0891***	0.0948***	0.0892***	0.0904***	0.0957***
Sad/depress2	0.0311***	0.0284***	0.0303***	0.0321***	0.0312***	0.0322***	0.0361***	0.0294***	0.0290***	0.0265***
Sad/depress3	0.0338***	0.0321***	0.0335***	0.0299***	0.0294***	0.0345***	0.0363***	0.0320***	0.0314***	0.0428***
Sad/depress4	0.1130***	0.1106***	0.1125***	0.1102***	0.1061***	0.1146***	0.1209***	0.1172***	0.1120***	0.1068***
Sad/depress5	0.1727***	0.1721***	0.1728***	0.1760***	0.1689***	0.1784***	0.1676***	0.1763***	0.1727***	0.1663***
Control2	0.0038	0.0039	0.0032	0.0028	0.0040	0.0013	0.0066	0.0018	0.0024	0.0070
Control3	0.0447***	0.0441***	0.0441***	0.0456***	0.0479***	0.0402***	0.0429***	0.0490***	0.0448***	0.0410***
Control4	0.0653***	0.0651***	0.0649***	0.0682***	0.0696***	0.0625***	0.0532***	0.0711***	0.0615***	0.0674***
Control5	0.0820***	0.0830***	0.0819***	0.0854***	0.0800***	0.0786***	0.0846***	0.0898***	0.0804***	0.0735***
Pain2	0.0383***	0.0361***	0.0375***	0.0342***	0.0409***	0.0426***	0.0331***	0.0397***	0.0357***	0.0404***
Pain3	0.0802***	0.0791***	0.0798***	0.0803***	0.0760***	0.0836***	0.0762***	0.0804***	0.0784***	0.0836***

		Age,gender	**Interviewer Effects Excluding Individual Interviewers**							
			Dummy	**a**	**b**	**c**	**d**	**e**	**f**	**g**
Pain4	0.2575***	0.2580***	0.2576***	0.2557***	0.2644***	0.2578***	0.2442***	0.2599***	0.2584***	0.2594***
Pain5	0.3718***	0.3737***	0.3724***	0.3818***	0.3652***	0.3828***	0.3625***	0.3564***	0.3760***	0.3747***
Age		-0.0007**								
Age squared		0.0000**								
Female		0.0060								
interv1			0.0084							
interv2			0.0014							
interv3			-0.0019							
interv5			0.0011							
interv4			0.0063							
interv6			-0.0030							
MAE		0.056	0.056	0.058	0.053	0.059	0.056	0.056	0.057	0.055
MAE mild states		0.008	0.009	0.011	0.011	0.010	0.009	0.010	0.010	0.011
Range of utility		-0.423 to	-0.392 to	-0.398 to	-0.362 to	-0.403 to	-0.365 to	-0.374 to	-0.395 to	-0.372 to
values		1.007	1.001	0.999	0.997	0.999	0.993	0.999	0.998	1.001
Observations	7,287	7,287	7,287	6,587	6,167	6,335	6,069	5,964	6,615	5,985

*** p<0.01, ** p<0.05, * p<0.1

Supplementary Figure 1: TTO and DCE tasks

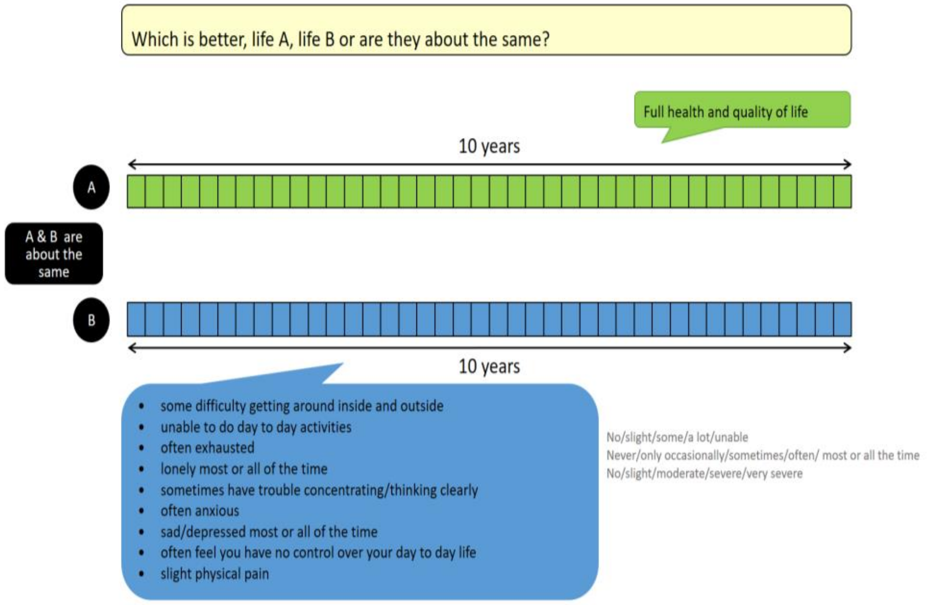


Figure 1a: TTO

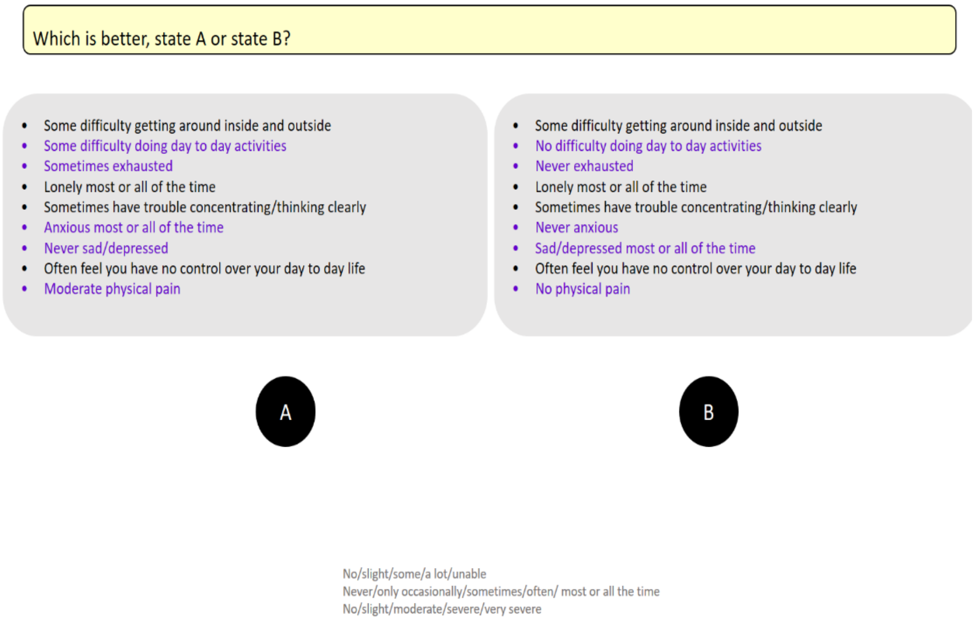
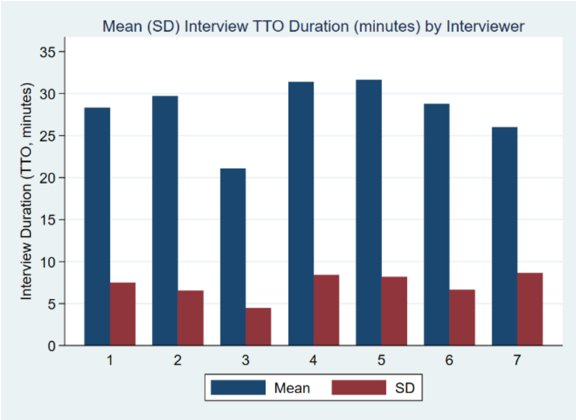


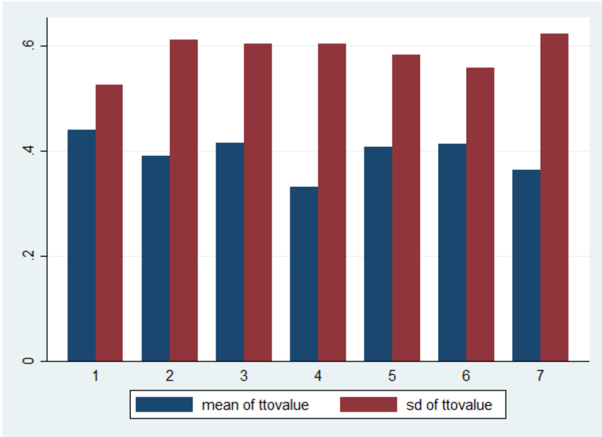
Figure 1b: DCE

Supplementary Figure 2: TTO duration and mean values by interviewer

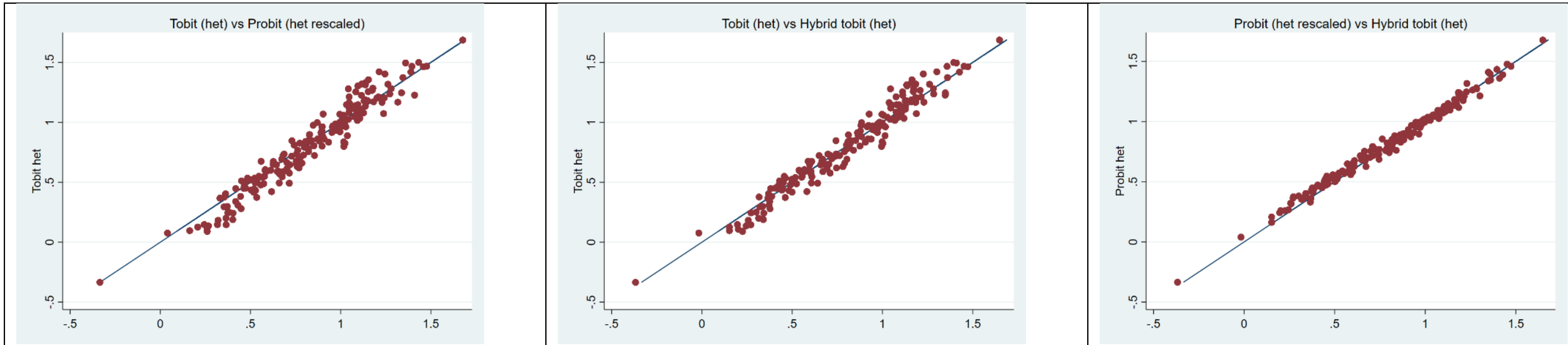
This figure shows the mean (and standard deviation) amount of time taken (in minutes) to complete the cTTO questionnaire, by interviewer. This excludes any time taken to complete DCE (not recorded in EQ-PVT) and any additional questions. X axis: Interviewers 1 to 7



This figure shows the mean (and standard deviation) value observed across all TTO tasks, by interviewer. This excludes the wheelchair example and practice TTO tasks.
X axis: Interviewers 1 to 7



Supplementary Figure 3: Relationship between Tobit, Probit and Hybrid predictions



Het – heteroscedasticity