**Modelling EQ-5D-5L health states using the international EQ-5D-5L valuation protocol in Spain**

Juan Manuel Ramos-Goñi[1,2], Jose Luis Pinto-Prades[3], Juan Manuel Cabasés[4], Oliver Rivero-Arias[2,5]

Running title: Spanish valuation of EQ-5D-5L health states

1 HTA Unit of Canary Island Health Service (SESCS)
2 REDISSEC
3 Glasgow Caledonian University
4 Universidad Pública de Navarra
5 Nuffield Department of Population Health, University of Oxford, UK

Corresponding author:
Oliver Rivero-Arias
Senior Health Economist
National Perinatal Epidemiology Unit (NPEU)
Nuffield Department of Population Health
University of Oxford
Old Road Campus
Headington
Oxford
OX3 7LF
Email: oliver.rivero@dph.ox.ac.uk

HESG reference number: A005

Acknowledgments
A research grant with number SESCS 2010/05 was awarded by the "Ministerio de Sanidad, Servicios Sociales e Igualdad" in Spain to conduct this study. The authors are very grateful to the Valuation Methodology Management Team (Frank de Charro, Ben van Hout, Nancy Devlin, Mark Oppe and Paul Krabbe) for the constant and unconditional advice received during this study.

# Introduction

The EQ-5D instrument is the most widely used preference-based health-related quality of life questionnaire in cost-effectiveness analysis of health care technologies. Reimbursement agencies such as the National Institute for Health and Care Excellence (NICE) explicitly recommend the use of the EQ-5D in submissions to the institute and this partly explains the spread use of the instrument in applied studies [1]. To date, the EQ-5D has been translated into more than 100 languages and 9 official country-specific value sets to estimate utility values have been estimated using responses from members of the general public [2].

The original EQ-5D is a relatively simple questionnaire with five domains (mobility, self-care, usual activities, pain/discomfort and anxiety/depression) and three levels in each domain (no problems, some problems and extreme problems) describing 243 possible health states [3]. Extensive body of research supports the use of the instrument in most major disease areas. However, recent studies have shown that the EQ-5D has limited sensitivity to change in mental health and in children [4, 5], and suffers from problems with ceiling effects [6]. To improve and refine the questionnaire, a EuroQol Group task force proposed a new version aimed at correcting these issues. The new version increased the number of severity levels from three to five (no problems, slight, moderate, severe and unable or extreme), and the number of health states to 3125 [7]. This version of the questionnaire received the name of EQ-5D-5L and the original EQ-5D questionnaire is now known as EQ-5D-3L.

Population-based preference EQ-5D-3L value sets estimated using current valuation studies cannot be used directly with five level version responses. At the moment and as temporary solution, an interim scoring algorithm needs be used first [8]. Therefore, new valuation exercises are necessary to obtain preferences from the general public for the health states derived from the EQ-5D-5L. The EuroQol group has developed a valuation study protocol tool to elicit preferences using evidence from a series of pilot studies conducted by research teams worldwide [9]. The methods used to elicit preferences in this first version included a composite time-trade off (C-TTO) method plus a discrete choice experiment (DCE). The additional number of health states in the new version introduced challenges for a stand-alone TTO exercise and the inclusion of a DCE was considered to have the potential to provide additional information to estimate the EQ-5D-5L health states more accurately [10]. A working group with researchers based in Spain and UK have been one of the first teams in implementing this version of the international protocol in members of the general public in Spain.

In this paper we report the results of applying the EuroQol group valuation study protocol to a representative sample of the Spanish general population to obtain a value set representing the 3125 health states of the EQ-5D-5L. The reported EQ-5D-5L value set is expected to assist the conduct of economic evaluations to support decision-making in Spain.

# Methods

## EQ-5D-5L

The EQ-5D instrument contains five dimensions: mobility, self-care, usual activities, pain/discomfort and anxiety/depression. The new EQ-5D-5L extends the original three levels to five (no problems, slight problems, moderate problems, severe problems and unable or extreme problems) describing 3125 ($5^5$) possible health states. In a particular dimension no problems are represented by level 1 whereas extreme problems are represented by level 5. Hence each health state in the EQ-5D-5L can be described using a five-digit number where 11111 indicates perfect health and 55555 indicates the worst possible health state.

## The international EQ-5D-5L valuation protocol

The EuroQol group in collaboration with an international research team conducted a series of pilot studies to evaluate methodological aspects about the use of composite time trade-offs tasks and discrete choice experiments to elicit individual preferences [9]. The evidence obtained from these multinational pilot studies informed the standardised protocol for EQ-5D-5L value sets used in this study. The interview process described in the protocol has five different parts (Box 1) including: 1) a general welcome, 2) introduction to the research and completion of background information, 3) the composite time trade-off task, 4) the discrete choice experiment, and 5) general thank you and goodbye.

The valuation exercise started with welcoming the respondent and explaining the objectives of the research. Next, the respondents were asked to complete the EQ-5D-5L instrument, the visual analog scale (EQ-VAS), and background information related to age, sex and experience with illness. In the next part the respondents were introduced and explained the C-TTO task with an example of being on a wheelchair health state, completed 10 EQ-5D-5L C-TTO valuations and provided feedback about whether they found the task difficult. The respondent then moved to the DCE and was introduced and explained how to carry out the exercise and completed seven paired choices. They were also given the opportunity to provide feedback about difficulties completing the DCE. The respondent could also comment and provide general feedback about the entire interview. At the end of the interview, the respondent was thanked for their time and help completing the survey.

To implement and facilitate this standard protocol, the EuroQol group developed an online system that incorporated all the parts of the protocol described above and collected the data called EuroQol Valuation Technology (EQ-VT). At the time of conducting this study, the Spanish working team used an early version of this system.

**Eliciting preferences methods**

*Composite time trade-off (C-TTO)*

In the traditional time trade-off (TTO) participants select between being in a poor health state $i$ for an amount of time $T$, or being in full health for a shorter period of time $X$. The time $X$ is varied until the respondent is indifferent between the two alternatives and the valuation of the state $i$ is calculated as $X/T$. This method has been widely used in the EQ-5D-3L valuation studies conducted so far (with $T = 10$) and it is appropriate to valuate health states considered better than dead. However, using the TTO method for states worse than dead gives negatives values that are normally transformed to be bounded to -1 and this has been criticised in the literature [11]. Other TTO alternatives to evaluate health states were therefore assessed during the EuroQol pilot studies. Two extensions to the traditional TTO that provide more trading time in full health to respondents valuing poor health states were evaluated: a "lead-time" TTO and a "lag-time" TTO [12, 13]. In the former, the additional trading time is included before the health state whereas in the latter, the trading time is included after the health state to be valued. The pilot studies looked at the potential of using these methods in practice (e.g. ratio of lead/lag time to disease time, framing effects in lead/lag time and the use of a composite TTO method) and concluded that the international protocol should include a composite TTO method. This composite approach involved the use of the traditional TTO approach for states better than dead and lead-time TTO for states worse than dead in a unique task (i.e. the respondent was not asked explicitly to select whether they considered the health state better or worse than dead). For the lead-time TTO, a ratio of lead-time of 1:1 with a 20-year time frame (i.e. 10 years lead time and 10 years in the state to be valued) was used. This lead-time method produced a minimum value of -1 and no transformation of negative values was needed.

The final selection of EQ-5D-5L health states in the C-TTO included 86 health states that were selected using efficient design methods [14]. The health states were derived into 10 blocks with similar representation of severity levels and all blocks included the worst health state 55555. In the EQ-VT respondents were randomly allocated to one of the blocks and the order of the health state was also randomly presented to the participants. Each respondent answered 10 C-TTO tasks. The iterative process followed by the respondent to reach the point of indifference in the C-TTO task was adapted from the original UK valuation exercise [15].

*Discrete choice experiment (DCE)*

It is widely accepted that the TTO method to elicit individual preferences is cognitively challenging for respondents. Therefore, alternatives such as the use of ordinal data obtained from DCE studies for health state valuation have received recent attention in the literature [16, 17]. Modelling ordinal data is supported by theoretical foundations based on the random utility theory and can be analysed using robust limited-dependent econometrics tools [18]. Valuations obtained from DCE modelling have shown to replicate similar patterns found in TTO responses but with consistently higher valuations [19]. The valuations obtained from DCE models are expressed in an arbitrary scale and not on the

utility scale of TTO values. Therefore, DCE valuations need to be anchored on the dead (0) - full health (1) scale. Using DCE was also piloted as part of the multinational studies and the results suggested that collecting such information about the ordering and the distance between health states could provide additional useful information to the C-TTO data. Hence, a DCE was included as part of the international protocol and respondents indicated the health state they considered better from EQ-5D-5L state pairs. The DCE included 196 pairs divided in 28 blocks with similar severity representation identified using efficient methods [14]. In the EQ-VT respondents were randomly allocated to one of the blocks and the order of the pairs was also randomly presented to the participants. Each respondent completed 7 pairs.

**Sampling methods**

Sample size calculations for the whole study were based on obtaining 0.01 standard errors in the coefficients of a linear regression model using C-TTO utility values as dependent variable and C-TTO responses as explanatories. The power calculations suggested a sample of 1,000 individuals to achieve that level of statistical efficiency [10]. A two-step stratification procedure was designed to obtain a representative sample of the Spanish population. Spain is divided into autonomous communities and these are divided further into provinces. The first stratification step includes the Spanish provinces. The 50 Spanish provinces were ordered according to population size. The 26 provinces that represented more than 75% of the Spanish population were chosen. Since not all Spanish autonomous communities were selected and some of them were over-represented by many provinces, the provinces with lower population size of the over-represented autonomous communities were replaced for the provinces with higher population size from the autonomous communities that were not included initially.

**Piloting and implementation of international protocol in Spain**

The results of the multinational pilot studies suggested that the valuation exercise (as described in Box 1) had to be conducted face-to-face with the respondent given access to the EQ-VT and the interviewer always available to introduce the tasks and clarify any doubts. We subcontracted an independent market research company, which identified respondents around Spanish provinces and arranged interviews at home or at convenient places during June and July 2011. Respondents did not receive payment for participating in the survey. A total of 33 interviewers were trained by JMRG over a period of one day in Madrid and Barcelona using material also developed by the EuroQol group alongside the EQ-VT. In our case the material was translated into Spanish using a certified translator. A small pilot involving 10 respondents to evaluate the EQ-VT in practice was conducted before the main data collection.

**Quality control**

The early version of the EQ-VT software used in Spain did not provide detailed information about interviewer's compliance to the protocol. Data on timings explaining the C-TTO wheelchair example

was automatically collected in this version but could not be retrieved from the server at the time we started data collection. A second market research company obtained a proxy for this information telephoning a random sample of 15% of respondents. During the phone call we asked the respondent to provide feedback about the EQ-VT, the performance of the interviewer and whether they had any suggestions to improve the interview. The results from this exercise suggested that interviewers seemed to comply with the protocol. However, information on timings became available with the newer version of the software after data collection was completed and the data suggested that on average interviewers spent 1.73 minutes explaining the C-TTO wheelchair example. Given the complexity of the C-TTO this figure was considerably smaller than what we had expected. Figure 1 shows the actual distribution of time explaining the C-TTO example by interviewers. In addition, this new data about interviewer's compliance suggested that 75% of the whole sample did not receive any explanation about the lead-time TTO part of the C-TTO. 30% of this group produced however an average of 2.5 negative values. As a result, the C-TTO data for poorer health states were likely to have been affected by the interviewer's behaviour.

### Excluded observations

We excluded observations using the following two criteria: 1) respondents with a positive slope on a regression between his/her values and the severity of the health states indicating that the participant provided higher utility values for poorer health states on average; and 2) respondents who valued all states equal to death.

## Statistical analyses

### Sample characteristics and C-TTO and DCE responses

Participant's characteristics and responses from the C-TTO and the DCE were first described using descriptive analysis. We calculated a severity index to illustrate C-TTO and DCE responses depending on severity levels. This index was defined as the sum of the levels for all dimensions in a particular state (e.g. state 25431 has a severity index of 2+5+4+3+1=15).

Observations from 27 participants were removed for the reasons included in the previous section. We checked the baseline demographics of the excluded participants against valid responses to ensure they were not different from the cases included in the estimation sample.

### Modelling strategy

Two sources of data were available to estimate the EQ-5D-5L Spanish value set: C-TTO and DCE data. In order to maximise the use of the data available we implemented a hybrid modelling approach that made use of both C-TTO and DCE data to estimate a the potential value set. This hybrid method

estimated a unique set of coefficients from a likelihood function obtained multiplying the likelihood functions of using ordinary least squares (OLS) for the C-TTO data by the likelihood function of a conditional logit for DCE data [20]. This approach combined therefore the utility values elicited in the C-TTO for the 86 health states with information on the ordering and distance between health states elicited in the DCE. The dependent variable in the OLS part was defined as 1 minus de C-TTO value for a given health state to indicate disutility and the coefficients expressing utility decrements. In the conditional logit, the dependent variable was a binary outcome 0/1 indicating the respondent's choice to each EQ-5D-5L pair. The additional information from the DCE helped to correct the possible bias introduced by the interviewer's described above. For a full description and analytical derivation of the hybrid method, the reader is referred to Rowen et al [20].

In this report we also present for comparison purposes statistical models to estimate C-TTO and DCE data separately. We analysed C-TTO data using a random intercept effects model with each respondent contributing with ten observations in the estimation. In a random intercept model between respondents differences are treated as random variables coming from a normal distribution deviating from an overall mean. The dependent variable was defined as 1 minus de C-TTO value for a given health state to indicate disutility and the coefficients expressing utility decrements.

DCE data was analysed using the standard econometric method for ordinal data conditional logit regression [18]. Similar to the conditional logir part of the hybrid model, the dependent variable was a binary outcome 0/1 indicating the respondent's choice to each EQ-5D-5L pair. Coefficients estimated from a conditional logit are expressed on a latent arbitrary utility scale with information on the ordering of the state. To obtain coefficients on the C-TTO utility scale the coefficients were normalised using the C-TTO value of the worst health state. Thus the value of the worst health state in the DCE model was anchored at the C-TTO value of the health state 55555 [20].

For each of the models (hybrid, random-effects and conditional logit) we started using main effects with a 20-parameter model consisting of 4 dummies for each EQ-5D-5L dimensions using level 1 as the reference. We constructed dummies to represent the additional utility decrement of moving from one level to another. For instance for the mobility dimension we created four dummies MO1, MO2, MO3 and MO4 and the coefficient associated to MO1 indicated the utility decrement of moving from no problems (level 1) to slight problems (level 2), MO2 the additional utility decrement of moving from slight (level 2) to moderate (level 3) problems, and so on. Therefore the overall decrement of moving from no to moderate problems could be calculated as the sum of the coefficients of MO1 plus MO2. A similar set of dummies was defined for the other dimensions: self-care (SC), usual activities (UA), pain/discomfort (PD), and anxiety/depression (AD).

Our starting point for the selection of additional covariates for the models was the US valuation study [21]. In that research, several variables were defined to explain the effect of marginal disutility. The increase in the number of levels in the EQ-5D-5L indicated that this effect might be even more

marked in this version of the questionnaire. We thus explored the impact on the estimation results of marginal disutility using the terms described in Box 2. Squared of all terms were also introduced to assess non-linear effects on the dependent variable. There is currently no guidance or rules to follow about the order and inclusion of terms in Box 2. Hence we included all terms first, and use a stepwise approach removing non-significant terms and collinearity problems until a consistent model was obtained (see next section for more details on model selection).

In this paper, we present the results of the regression with the main effects and the best-fitted model with significant terms using random effects, conditional logit or hybrid method. Statistical analysis and regression modelling were conducted in Stata MP 11 [22]. The hybrid model was not available in any standard package and was programmed in Stata specifically for this study.

### Evaluation of model performance

We evaluated model performance using 1) logical consistency of parameters, 2) goodness of fit, 3) prediction accuracy and 4) parsimony. Estimated coefficients are said to be logically consistent if magnitude values from worse health states are lower than those from better health states. In our estimated results this translated to all main effects coefficients being positive. Goodness of fit was assessed using the Akaike (AIC) and the Bayesian information criteria (BIC). Prediction accuracy was evaluated using mean square error (MSE) and mean absolute error (MAE). For the hybrid model and the conditional logit, we used the responses from the C-TTO exercise as actual values to calculate these values given the lack of an appropriate counterfactual. Finally, the principle of parsimony stated that if competing models were similar in logical consistency and goodness of fit, the model with fewer parameters was preferred.

These four criteria were used to compare each model (hybrid, random-effect and conditional logit) separately. However, comparison between models was only possible using logical consistency of parameters and parsimony. AIC and BIC cannot be used to compare models estimated using different data. MSE and MAE were not appropriate to compare the random-effect model with the hybrid method or the re-scale conditional logit. Given that the random-effect fitted C-TTO data directly, it provided more accurate predictions for such data (i.e. lower MSE and MAE) compared to the hybrid model and the conditional logit by definition.

## Results

### Sample characteristics and C-TTO and DCE responses

Background characteristics of the estimation sample compared with the Spanish general population are presented in Table 1. Overall both samples were similar in the distribution of employment status;

mean age and gender distribution but the estimation sample had a larger number of respondents in age group 25-34 and fewer participants over 75 than the general population.

Table 1 also presents the sample characteristics of those respondents excluded from the estimation sample. Overall the excluded observations were older with no studies or primary school studies than the estimation sample.

Table 2 presents the distribution of EQ-5D-5L responses of participants in the estimation sample. Most individuals reported no problems in mobility, self-care, and usual activities with 29.4% reporting slight or moderate problems in pain/discomfort and 18.90% reporting slight or moderate problems in anxiety/depression.

The observed mean C-TTO utility values for the 86 health states ordered by severity index are reported in Table 3. On average, mean utility values decreased as severity scores increased with larger variability in responses for poorer health states. Figure 2 shows the distribution of observed C-TTO values across all respondents. The distribution shows spikes at 1, 0.5, 0, -0.5 and -1 with most of the negative values concentrated between -1 and -0.5. Figure 3 shows the same information but reported across the severity index. The spikes were not present for lower values of severity index (<12) but seemed to be available for values of severity index greater or equal to 12.

**Modelling results**

The estimation results for the three models using main effects only are presented in Table 5. The hybrid method with main effects was a consistent model with a predicted utility for the worst health state of -0.225. Both the random effects and the re-scaled conditional logit presented logical inconsistencies in the usual activities dimension. In the case of the random effects the inconsistency appeared in the coefficients of levels 4 and 5 whereas in the re-scaled conditional logit the inconsistency was present in level 2 and level 3.

The best-fitted estimation results using terms and interactions for the three models are presented in Table 6. These final models presented marginal improvements in AIC, BIC, MSE and MAE compared to the main effects models in Table 5. We could not estimate a consistent model with any of the combinations of terms in the re-scaled conditional logit, and was therefore excluded from further evaluation.

In the hybrid and the random-effects models, the best-fitted model included the terms $D1^2$ and $K45^2$. D1 was the number of dimensions at level 2, 3, 4, and 5 beyond the first whereas K45 indicated the number of dimensions at level 4 or 5. The sign associated to these coefficients was negative indicating a declining marginal disutility. The square of the terms suggested that the effect of the terms was not linear. The inclusion of these terms also corrected the inconsistency from the main effects model in the random effect model.

**Selected valuation model for EQ-5D-5L health states**

Table 6 shows the mean actual C-TTO values versus the predicted utility values by severity index for the main effects and best-fitted hybrid model and best-fitted random-effects. As expected, the random-effects model produced predictions similar to the actual C-TTO values. Both hybrid models produced wider utility values at the upper and lower end of the scale compared to the actual C-TTO values. This result was also expected as the DCE introduced additional information about the distance between health states and the ordering of health states. However, including the marginal disutility terms $D1^2$ and $K45^2$ in the hybrid model translated into predicted utility values closer to the actual C-TTO than the main effects. Given that the magnitude of the coefficients associated to $D1^2$ and $K45^2$ is close to zero and that the improvement in fitness between the main effects and the best-fitted model was marginal, we have selected the estimation results from the hybrid model with main effects for the Spanish EQ-5D-5L value set based on the parsimony criteria.

# Discussion

In this manuscript we have reported the first valuation study estimating a value set for the EQ-5D-5L questionnaire in Spain using the recent international valuation protocol developed and endorsed by the EuroQol group. C-TTO and DCE data were collected during the study and we implemented a hybrid method that combines both sources of data to estimate the final value set.

The C-TTO data presented clear spikes at -1, -0.5, 0, 0.5 and 1 and 40% of these values were concentrated around these figures. It is likely that these are valid responses but it was possible that the iterative process in the C-TTO exercise played a role since those values could have been produced by a strategic behaviour from the respondent (e.g. to short the duration of the interview). We checked the number of steps needed to reach these values in the iterative process and except for 0 and 1, all the remaining values needed at least three steps. As a result it is unlikely that these spikes represented strategic behaviour. We did observed however a small proportion of negative values for health states considered worse than death. In particular few valuations were placed between 0 and -0.5. One possible explanation is the bias introduced by interviewers not complying with the protocol. Nevertheless, we cannot completely verify whether this bias is completely attributable to interviewer's behaviour or whether there are other aspects of the lead-time TTO or the iterative process that have contributed to this.

The choice of the hybrid method for the statistical modelling was not only based on practical grounds. The model is based on the assumption that subjects have a unique utility function that generates both sets of responses. If responses were totally unbiased there would be no need of combining the two methods, since they would produce the same results. However, this is not often the case. If results are different using two theoretically equivalent methods, researchers have tried to find arguments in

favour of one method or the other. We believe that the evidence suggests that neither method based on matching (like C-TTO) nor choices (DCE) are unbiased. Matching methods are influenced by scale compatibility and, in the case of C-TTO, by loss aversion. Choices are not without problems since it has been shown that responses to choices are more lexicographic (the prominence effect). Finally, it has also been observed that subjects perceive differently distances between outcomes when comparisons are conducted in a separate or in a joint model, again without clear evidence that one method is better than another. C-TTO can be conceived as a method based on a separate evaluation mode while DCE is a method based on a joint evaluation mode. We then do not think that the "true" values can be inferred from one single method and for this reason we suggest that it can make sense to use a hybrid model. We are not claiming that the biases present in one method compensate the biases present in the other so that adding up the two methods we get unbiased results. There is no evidence to suggest this is the case. In the absence of such empirical evidence, we think that there are reasons to suspect that, at least, the potential biases present in the C-TTO are not enhanced by choices of the DCE, rather the opposite. For example, while C-TTO may give too much weight to the time dimension (scale compatibility), this clearly does not happen in our DCE since subjects did not see time attached to the health states. Also, given that in DCE subjects compare different combinations of EQ-5D-5L directly (in a joint mode), this can make people to focus more on the relative value of different quality of life dimensions. In the absence of clear empirical evidence we think that it can make sense to have a hypothesis implying that in C-TTO people focus more in the time dimension while in DCE they focus more on quality of life. For this reason, the hybrid model can help to solve some of the limitations present in DCE and C-TTO. In our study the DCE provided additional data about the ordering and distance between health states, this helping to control for any possible interviewer's bias.

The effect of marginal disutility was assessed using a series of ordinal variables. The best fitted hybrid model included two marginal disutility terms $D1^2$ and $K45^2$.The fact that the introduction of these terms provided a better fit to the data seems to suggest that the final value set should have included such effect. However, using these terms in the model raised some concerns. It did not seem logical to suggest that an improvement from 44344 to 44444 has little or no value. In order to address this issue we believe it was crucial to make a distinction between the descriptive and the normative interpretations of these terms. From a descriptive point of view it seemed clear that the variables should have been included. The effect was there. However, from a normative perspective it was not clear to us what should have been done. It depends on what we think was the origin of this effect. It might be the case that this was the way that people value health. That is, a health problem was not so bad if subjects were already in bad health. That is, a health improvement would not be so relevant if other concomitant health problems remain. If this is how people value health, we can say that $D1^2$ and $K45^2$ identified true preferences and these variables should (from a normative point of view) be included in the valuation of a health state. However, looking at the data we observed some results that were compatible with an alternative interpretation of D1 and K45 as parameters that identified the effect of a bias generated by psychological effects (diminishing sensitivity). In this case the terms

could be thought of as a debasing tool and it should not be used to estimate the final utilities. That is, the real difference in utilities between 44344 and 4444 is not -0.054 (with $D1^2$ and $K45^2$) but -0.095 (the additional utility decrement of UA3 in table 5). This would be so if the reduction in this distance (from 0.095 to 0.054) generated by $D1^2$ and $D45^2$ was generated by limitations of subjects in perceiving real differences. There are some results in the data that suggests this interpretation was not implausible. For example, we observed that the predicted utilities using Table 5 estimations are U(12244)=0.381 and U(34244)=0.347, a very small difference for two very different health states. Given that in our case, the improvement in fitness of using marginal disutility variables was marginal as suggested by the AIC, BIC, MSE and MAE, we based the selection of our final model without $D1^2$ and $D45^2$ using the parsimony criterion.

The experience from this study has provided valuable insight about the training and experience of interviewers conducting interviews for the valuation exercise. Unfortunately, we could not have access to the timings to assess protocol compliance until the data was completed. This is a limitation of the current study and we have used the hybrid method to help correcting any possible bias on this. Our experience has been shared with the EuroQol group who will incorporate in the next version of the protocol three additional examples before participants start the C-TTO task. In addition, future studies should closely monitor interviewers through timings who at least do not spend 3-4 minutes with the examples. Our work also questioned whether the training received by interviewers was complete enough to communicate properly to participants and whether the interviews should have been conducted by researchers with solid experience in outcomes research. This should be evaluated in future studies.

The final selected model has a wider range of values in the upper scale and narrower range of values in the lower range compared with the EQ-5D-3L valuation study conducted in Spain [23]. In that study the mean utility value across health states with a severity index of 6 was estimated to be 0.838 whereas in our study this mean was estimated to be 0.926. We expect that our new value set will be help to overcome some of the ceiling effect issues associated to the EQ-5D-3L. A different story however was observed in the lower scale. In the paper by Badia and colleagues, the mean lowest utility value was associated to health state 33333 with a mean estimate of -0.654. In our case this mean translated into -0.225 for the health state 55555. The issue of the possible interviewer bias applies here. However, the change in the wording of the mobility level "confined to bed" in EQ-5D-3L to "severe problems in walking about" and "unable to walk about" in EQ-5D-5L have transformed the worst possible health state. Given that these two new levels are not as severe as "confined to bed" (who also had the largest decrement from all dimensions in the Spanish study) it is expected to obtain higher valuations for 55555 than 33333.

# References


1. National Institute for Health and Care Excellence. Guide to the methods of technology appraisal. 2013, London: National Institute for Health and Care Excellence.

2. Szende, A., M. Oppe, and N. Devlin. EQ-5D value sets: inventory, comparative review and user guide, ed. A. Szende, M. Oppe, and N. Devlin. 2007, Dordrecht: Springer.

3. EuroQol, G. EuroQol - a new facility for the measurement of health-related quality of life. Health Policy, 1990. **16**: 199-208.

4. Brazier, J. Is the EQ-5D fit for purpose in mental health? Br J Psychiatry, 2010. **197**: 348-9.

5. Canaway, A.G. and E.J. Frew. Measuring preference-based quality of life in children aged 6-7 years: a comparison of the performance of the CHU-9D and EQ-5D-Y--the WAVES pilot study. Qual Life Res, 2013. **22**(1): 173-83.

6. Kontodimopoulos, N., E. Pappa, D. Niakas, J. Yfantopoulos, C. Dimitrakaki, and Y. Tountas. Validity of the EuroQoL (EQ-5D) instrument in a Greek general population. Value Health, 2008. **11**(7): 1162-9.

7. Herdman, M., C. Gudex, A. Lloyd, M. Janssen, P. Kind, D. Parkin, G. Bonsel, and X. Badia. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res, 2011. **20**(10): 1727-36.

8. van Hout, B., M.F. Janssen, Y.S. Feng, T. Kohlmann, J. Busschbach, D. Golicki, A. Lloyd, L. Scalone, P. Kind, and A.S. Pickard. Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. Value Health, 2012. **15**(5): 708-15.

9. Devlin, N. and P. Krabbe. The development of new research methods for the valuation of EQ-5D-5L. European Journal of Health Economics, 2013. **14**(Issue 1 Supplement).

10. Oppe, M., N. Devlin, B. Van Hout, P. Krabbe, and F. de Charro. A programme of methodological research to arrive at the new international EQ-5D-5L valuation protocol. Value Health, 2013. **Submitted**.

11. Craig, B.M. and M. Oppe. From a different angle: a novel approach to health valuation. Soc Sci Med, 2010. **70**(2): 169-74.

12. Augustovski, F., L. Rey-Ares, V. Irazola, M. Oppe, and N.J. Devlin. Lead versus lag-time trade-off variants: does it make any difference? Eur J Health Econ, 2013. **14 Suppl 1**: S25-31.

13. Robinson, A. and A. Spencer. Exploring challenges to TTO utilities: valuing states worse than dead. Health Econ, 2006. **15**(4): 393-402.

14. Bliemer, M.C.J., J.M. Rose, and S. Hess. Approximation of bayesian efficiency in experimental choice designs. Journal of Choice Modelling, 2008. **1**(1): 98-126.

15. Dolan, P. Modeling valuations for EuroQol health states. Medical Care, 1997. **35**(11): 1095-1108.

16. Salomon, J. Reconsidering the use of rankings in the valuation of health states: a model for estimating cardinal values from ordinal data. Population Health Metrics, 2003. **1**(1): 12.

17. McCabe, C., J. Brazier, P. Gilks, A. Tsuchiya, J. Roberts, A. O'Hagan, and K. Stevens. Using rank data to estimate health state utility models. Journal of Health Economics, 2006. **25**(3): 418-431.

18. Hensher, D.A., J.M. Rose, and W.H. Greene. Applied Choice Analysis: A Primer. 2005, Cambridge: Cambridge University Press.

19. Stolk, E.A., M. Oppe, L. Scalone, and P.F. Krabbe. Discrete choice modeling for the quantification of health states: the case of the EQ-5D. Value Health, 2010. **13**(8): 1005-13.

20. Rowen, D., J. Brazier, and B. Van Hout. A comparison of methods for converting DCE values onto the full health - dead QALY scale. HEDS Discussion Paper Series, 2011. **11/15**.

21. Shaw, J.W., J.A. Johnson, and S.J. Coons. US valuation of the EQ-5D health states: development and testing of the D1 valuation model. Med Care, 2005. **43**(3): 203-220.

22. StataCorp. Stata Statistical Software. 2011, Stata Press: College Station, TX: StataCorp LP.

23. Badia, X., M. Roset, M. Herdman, and P. Kind. A Comparison of United Kingdom and Spanish General Population Time Trade-off Values for EQ-5D Health States. Medical Decision Making, 2001. **21**(1): 7-16.

Box 1: Elements of the EQ-5D-5L international valuation protocol [10]

1. General welcome
2. Introduction
    a. Self-reported health on the EQ-5D-5L descriptive system
    b. Self-reported health on the EQ-VAS
    c. Background questions
3. Composite Time Trade-Off
    a. Instructions and example of TTO task
    b. TTO valuation of 10 EQ-5D-5L states
    c. TTO debriefing/structured feedback
4. Discrete Choice Experiment
    a. Instructions and example of DC exercise
    b. DC valuation of 10 pairs of EQ-5D-5L states
    c. DC debriefing/structured feedback
5. General thank you and goodbye
    **End of interview**

Box 2: Terms included in the models

D1: the number of dimensions at levels 2, 3, 4, and 5 beyond the first

I2: the number of dimensions at level 2 beyond the first

I3: the number of dimensions at level 3 beyond the first

I4: the number of dimensions at level 4 beyond the first

I5: the number of dimensions at level 5 beyond the first

D45: the number of dimensions at levels 4 and 5 beyond the first

K45: the number of dimensions at level 4 or 5

Squared of all terms were also introduced to assess non-linear effects on the dependent variable.

Figure 1: Mean time duration spent explaining the C-TTO wheelchair example by interviewers

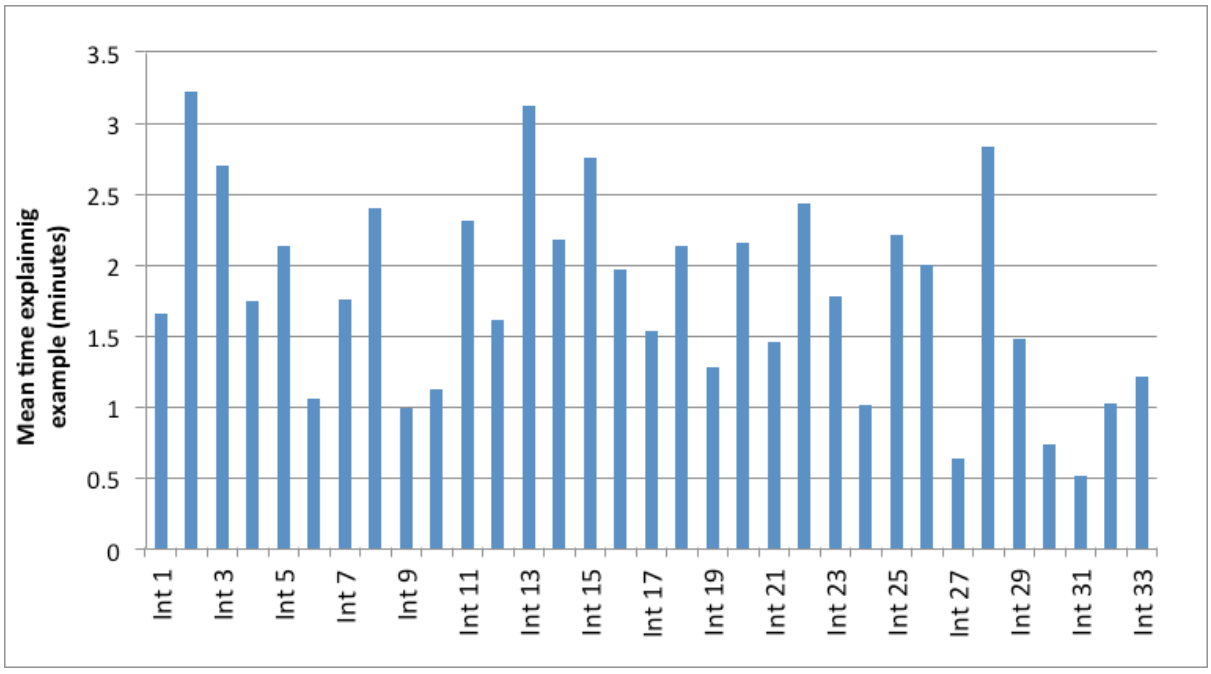

Table 1: Background characteristics of excluded sample, estimation sample and comparison against Spanish general population

| **Variables** | **Excluded sample (n =27)** | | **Estimation sample (n = 973)** | | **Spanish General Population** | |
|---|---|---|---|---|---|---|
| | **Mean** | **SD** | **Mean** | **SD** | **Mean** | **SD** |
| Age | 49.26 | 18.2 | 43.62 | 17.2 | 40.2 | n/a |
| | **n** | **%** | **n** | **%** | **n** | **%** |
| Age groups | | | | | | |
| - 18-24 | 3 | 11.2 | 114 | 11.7 | 3,439,383 | 9.0 |
| - 25-34 | 4 | 14.8 | 270 | 27.8 | 6,981,335 | 18.3 |
| - 35-44 | 5 | 18.5 | 170 | 17.5 | 7,490,547 | 19.6 |
| - 45-54 | 5 | 18.5 | 148 | 15.2 | 6,829,081 | 17.9 |
| - 55-64 | 4 | 14.8 | 111 | 11.4 | 5,169,932 | 13.5 |
| - 65-74 | 2 | 7.4 | 108 | 11.1 | 3,899,962 | 10.2 |
| - 75+ | 4 | 14.8 | 52 | 5.3 | 4,216,388 | 11.0 |
| Gender | | | | | | |
| - Male | 12 | 44.4 | 463 | 47.6 | 23,283,187 | 49.3% |
| - Female | 15 | 55.6 | 510 | 52.4 | 23,907,306 | 50.7% |
| Employment status | | | | | | |
| - Housewife/house husband | 1 | 3.7 | 70 | 7.2 | 4,038,800 | 10.51 |
| - Employed or freelance | 11 | 40.8 | 529 | 54.4 | 17,282,000 | 44.98 |
| - Student | 2 | 7.4 | 89 | 9.1 | 2,433,300 | 6.33 |
| - Retired | 8 | 29.6 | 132 | 13.6 | 7,732,700 | 20.12 |
| - Unemployed | 5 | 18.5 | 139 | 14.3 | 5,769,000 | 15.01 |
| - Disabled | 0 | 0 | 8 | 0.8 | 1,167,400 | 3.03 |
| - Missing | - | - | 6 | 0.6 | - | - |
| Education | | | | | | |
| - Higher education | 10 | 37.0 | 314 | 32.47 | 6,835,700 | 17.70 |
| - High school | 2 | 7.4 | 374 | 38.68 | 20,520,000 | 53.90 |
| - Primary school | 10 | 37.0 | 234 | 24.20 | 10,116,300 | 26.30 |
| - No studies | 5 | 18.5 | 45 | 4.65 | 809,000 | 2.10 |
| - Missing | - | - | 6 | 0.6 | | |
| Experience with illness | | | | | | |
| - Personal (%YES) | 4 | 14.8 | 140 | 14.4 | n/a | n/a |
| - Relatives (%YES) | 17 | 62.96 | 616 | 63.3 | n/a | n/a |
| - Other (%YES) | 9 | 33.3 | 338 | 34.7 | n/a | n/a |

n/a: not available

Table 2: Distribution of EQ-5D-5L responses of participants in the estimation sample

| | MOBILITY | | SELF CARE | | USUAL ACTIVITIES | | PAIN/ DISCOMFORT | | ANXIETY/ DEPRESSION | |
|---|---|---|---|---|---|---|---|---|---|---|
| | N | % | N | % | n | % | N | % | n | % |
| **NO PROBLEMS** | 864 | 88.80% | 933 | 95.89% | 673 | 69.17% | 772 | 79.34% | 891 | 91.57% |
| **SLIGHT PROBLEMS** | 69 | 7.09% | 30 | 3.08% | 214 | 21.99% | 149 | 15.31% | 57 | 5.86% |
| **MODERATE PROBLEMS** | 32 | 3.29% | 9 | 0.92% | 71 | 7.30% | 37 | 3.80% | 20 | 2.06% |
| **SEVERE PROBLEMS** | 7 | 0.72% | 1 | 0.10% | 15 | 1.54% | 10 | 1. 03% | 4 | 0.41% |
| **UNABLE/EXTREME** | 1 | 0.10% | 0 | 0% | 0 | 0% | 5 | 0.51% | 1 | 0.10% |

Table 3: Observed mean C-TTO values by health state and severity index

| STATE | SEVERITY | Mean | SD | STATE | SEVERITY | Mean | SD | STATE | SEVERITY | Mean | SD |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **11112** | 6 | 0.90 | 0.19 | **23242** | 13 | 0.60 | 0.39 | **45133** | 16 | 0.37 | 0.48 |
| **11121** | 6 | 0.90 | 0.18 | **25222** | 13 | 0.62 | 0.33 | **51451** | 16 | 0.18 | 0.58 |
| **11211** | 6 | 0.90 | 0.18 | **32314** | 13 | 0.56 | 0.46 | **24443** | 17 | 0.26 | 0.52 |
| **12111** | 6 | 0.87 | 0.27 | **35311** | 13 | 0.50 | 0.54 | **34244** | 17 | 0.27 | 0.57 |
| **21111** | 6 | 0.90 | 0.18 | **42115** | 13 | 0.27 | 0.59 | **43514** | 17 | 0.30 | 0.56 |
| **11122** | 7 | 0.84 | 0.23 | **53221** | 13 | 0.49 | 0.42 | **45233** | 17 | 0.31 | 0.54 |
| **11212** | 7 | 0.83 | 0.25 | **12344** | 14 | 0.35 | 0.53 | **45413** | 17 | 0.22 | 0.56 |
| **11221** | 7 | 0.84 | 0.25 | **25331** | 14 | 0.54 | 0.43 | **53243** | 17 | 0.31 | 0.48 |
| **12112** | 7 | 0.81 | 0.28 | **31514** | 14 | 0.53 | 0.37 | **34155** | 18 | 0.22 | 0.55 |
| **12121** | 7 | 0.86 | 0.18 | **34232** | 14 | 0.53 | 0.49 | **34515** | 18 | 0.22 | 0.56 |
| **21112** | 7 | 0.82 | 0.26 | **51152** | 14 | 0.17 | 0.58 | **43542** | 18 | 0.31 | 0.47 |
| **11421** | 9 | 0.72 | 0.33 | **12543** | 15 | 0.45 | 0.42 | **45144** | 18 | 0.09 | 0.57 |
| **13122** | 9 | 0.76 | 0.31 | **21345** | 15 | 0.32 | 0.57 | **52335** | 18 | 0.24 | 0.56 |
| **14113** | 10 | 0.67 | 0.36 | **21444** | 15 | 0.34 | 0.45 | **53244** | 18 | 0.19 | 0.51 |
| **11414** | 11 | 0.56 | 0.40 | **22434** | 15 | 0.37 | 0.52 | **54153** | 18 | 0.09 | 0.54 |
| **13313** | 11 | 0.69 | 0.37 | **23514** | 15 | 0.55 | 0.33 | **54342** | 18 | 0.05 | 0.56 |
| **11235** | 12 | 0.40 | 0.58 | **24342** | 15 | 0.48 | 0.42 | **55233** | 18 | 0.16 | 0.58 |
| **12513** | 12 | 0.61 | 0.36 | **31524** | 15 | 0.43 | 0.49 | **14554** | 19 | 0.07 | 0.57 |
| **13224** | 12 | 0.51 | 0.53 | **52215** | 15 | 0.37 | 0.48 | **24445** | 19 | 0.10 | 0.59 |
| **21315** | 12 | 0.48 | 0.48 | **52431** | 15 | 0.36 | 0.53 | **24553** | 19 | 0.15 | 0.56 |
| **25122** | 12 | 0.57 | 0.45 | **53412** | 15 | 0.41 | 0.54 | **35245** | 19 | 0.08 | 0.57 |
| **42321** | 12 | 0.50 | 0.53 | **54231** | 15 | 0.32 | 0.55 | **55225** | 19 | 0.19 | 0.57 |
| **11425** | 13 | 0.37 | 0.54 | **31525** | 16 | 0.33 | 0.56 | **44345** | 20 | 0.05 | 0.53 |
| **12244** | 13 | 0.33 | 0.55 | **32443** | 16 | 0.40 | 0.40 | **55424** | 20 | 0.03 | 0.56 |
| **12334** | 13 | 0.61 | 0.40 | **33253** | 16 | 0.41 | 0.50 | **44553** | 21 | 0.03 | 0.58 |
| **12514** | 13 | 0.41 | 0.54 | **35143** | 16 | 0.38 | 0.44 | **52455** | 21 | -0.01 | 0.52 |
| **15151** | 13 | 0.37 | 0.52 | **35332** | 16 | 0.34 | 0.59 | **43555** | 22 | -0.06 | 0.57 |
| **21334** | 13 | 0.56 | 0.45 | **43315** | 16 | 0.33 | 0.54 | **55555** | 25 | -0.17 | 0.53 |
| **23152** | 13 | 0.42 | 0.57 | **44125** | 16 | 0.16 | 0.53 | | | | |

Figure 2: Distribution of observed C-TTO utility values across respondents

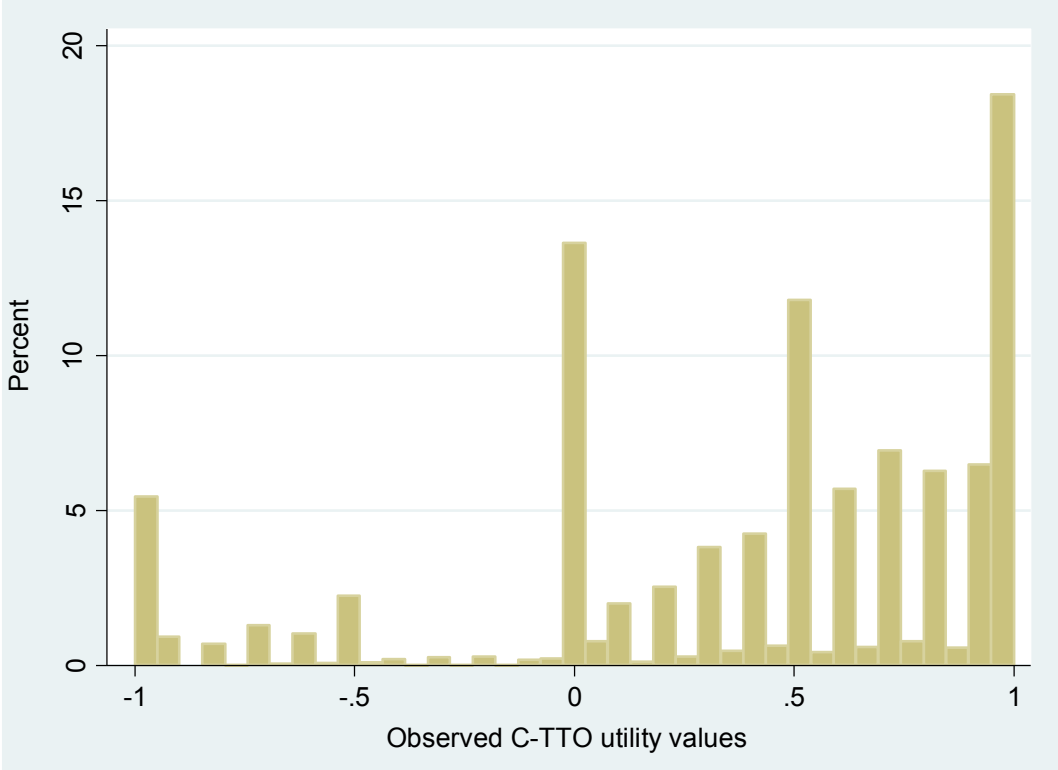


Figure 3: Distribution of observed C-TTO utility values by severity index

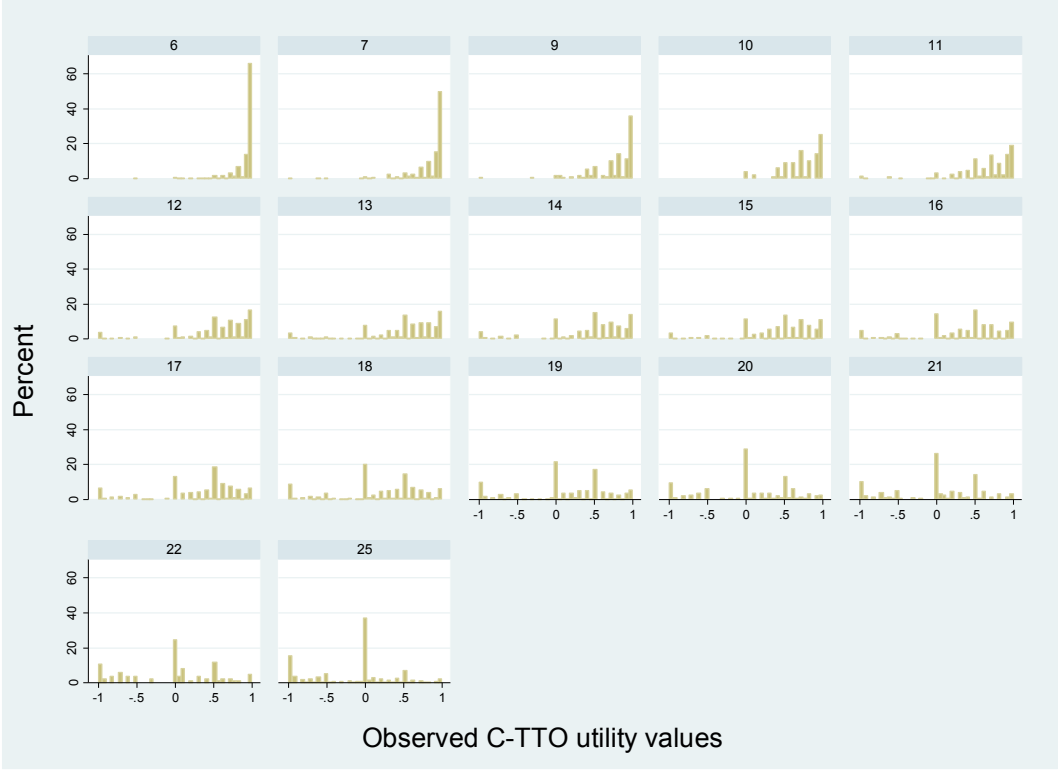

Table 4: Estimation results for random effects, conditional logit and hybrid model using main effects only

| | Hybrid (C-TTO+DCE data) | | | Random-effects (C-TTO data) | | | Re-scaled conditional Logit (DCE data) | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Coeff.** | **SE** | **p-value** | **Coeff.** | **SE** | **p-value** | **Coeff.** | **SE** | **p-value** |
| **MO1** | 0.086 | 0.008 | 0.000 | 0.033 | 0.012 | 0.005 | 0.088 | 0.010 | 0.000 |
| **MO2** | 0.014 | 0.009 | 0.120 | 0.054 | 0.012 | 0.000 | 0.012 | 0.012 | 0.292 |
| **MO3** | 0.131 | 0.010 | 0.000 | 0.125 | 0.013 | 0.000 | 0.115 | 0.011 | 0.000 |
| **MO4** | 0.059 | 0.010 | 0.000 | 0.044 | 0.015 | 0.003 | 0.081 | 0.012 | 0.000 |
| **SC1** | 0.058 | 0.008 | 0.000 | 0.043 | 0.011 | 0.000 | 0.030 | 0.011 | 0.007 |
| **SC2** | 0.000 | 0.009 | 0.975 | 0.010 | 0.015 | 0.516 | 0.017 | 0.012 | 0.151 |
| **SC3** | 0.097 | 0.011 | 0.000 | 0.094 | 0.015 | 0.000 | 0.079 | 0.013 | 0.000 |
| **SC4** | 0.015 | 0.009 | 0.107 | 0.032 | 0.013 | 0.015 | 0.022 | 0.011 | 0.052 |
| **UA1** | 0.055 | 0.008 | 0.000 | 0.041 | 0.011 | 0.000 | 0.037 | 0.011 | 0.001 |
| **UA2** | 0.005 | 0.010 | 0.638 | 0.039 | 0.014 | 0.004 | **-0.008** | 0.012 | 0.494 |
| **UA3** | 0.072 | 0.010 | 0.000 | 0.070 | 0.015 | 0.000 | 0.068 | 0.012 | 0.000 |
| **UA4** | 0.004 | 0.010 | 0.685 | **-0.022** | 0.014 | 0.110 | 0.024 | 0.012 | 0.044 |
| **PD1** | 0.080 | 0.008 | 0.000 | 0.049 | 0.011 | 0.000 | 0.065 | 0.011 | 0.000 |
| **PD2** | 0.024 | 0.009 | 0.008 | 0.042 | 0.014 | 0.002 | 0.019 | 0.012 | 0.104 |
| **PD3** | 0.114 | 0.011 | 0.000 | 0.113 | 0.014 | 0.000 | 0.129 | 0.012 | 0.000 |
| **PD4** | 0.106 | 0.010 | 0.000 | 0.078 | 0.016 | 0.000 | 0.117 | 0.012 | 0.000 |
| **AD1** | 0.088 | 0.008 | 0.000 | 0.067 | 0.013 | 0.000 | 0.064 | 0.012 | 0.000 |
| **AD2** | 0.043 | 0.010 | 0.000 | 0.045 | 0.013 | 0.001 | 0.048 | 0.012 | 0.000 |
| **AD3** | 0.123 | 0.010 | 0.000 | 0.112 | 0.014 | 0.000 | 0.118 | 0.013 | 0.000 |
| **AD4** | 0.052 | 0.010 | 0.000 | 0.036 | 0.013 | 0.005 | 0.063 | 0.012 | 0.000 |
| **Const.** | | | | 0.081 | 0.014 | 0.000 | | | |
| **LogL** | -10294.2 | | | -3093.07 | | | -3675.81 | | |
| **AIC** | 20632.44 | | | 6226.14 | | | 7391.62 | | |
| **BIC** | 20802.39 | | | 6369.8 | | | 7528.69 | | |
| **MSE** | 0.0048 | | | 0.0034 | | | 0.0108 | | |
| **MAE** | 0.0553 | | | 0.0479 | | | 0.0872 | | |
| **U(55555)** | -0.225 | | | -0.188 | | | -0.188 | | |

Bold values indicate logical inconsistencies

Table 5: Estimation results using best-fitted model for random effects, conditional logit and hybrid model

| | Hybrid (C-TTO+DCE data) | | | Random effects (C-TTO data) | | | Re-scaled conditional Logit (DCE data) | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Coeff.** | **SE** | **p-value** | **Coeff.** | **SE** | **p-value** | **Coeff.** | **SE** | **p-value** |
| **MO1** | 0.112 | 0.009 | 0.000 | 0.039 | 0.019 | 0.037 | 0.120 | 0.018 | 0.000 |
| **MO2** | 0.020 | 0.008 | 0.018 | 0.040 | 0.012 | 0.001 | 0.016 | 0.011 | 0.149 |
| **MO3** | 0.143 | 0.014 | 0.000 | 0.186 | 0.018 | 0.000 | 0.135 | 0.018 | 0.000 |
| **MO4** | 0.070 | 0.009 | 0.000 | 0.060 | 0.015 | 0.000 | 0.077 | 0.011 | 0.000 |
| **SC1** | 0.079 | 0.009 | 0.000 | 0.055 | 0.017 | 0.001 | 0.065 | 0.018 | 0.000 |
| **SC2** | 0.006 | 0.009 | 0.518 | 0.010 | 0.015 | 0.492 | 0.018 | 0.011 | 0.115 |
| **SC3** | 0.115 | 0.013 | 0.000 | 0.155 | 0.019 | 0.000 | 0.102 | 0.019 | 0.000 |
| **SC4** | 0.024 | 0.009 | 0.008 | 0.058 | 0.014 | 0.000 | 0.022 | 0.011 | 0.044 |
| **UA1** | 0.081 | 0.009 | 0.000 | 0.058 | 0.018 | 0.001 | 0.073 | 0.019 | 0.000 |
| **UA2** | 0.005 | 0.009 | 0.607 | 0.031 | 0.015 | 0.039 | **-0.005** | 0.011 | 0.627 |
| **UA3** | 0.095 | 0.013 | 0.000 | 0.127 | 0.018 | 0.000 | 0.091 | 0.017 | 0.000 |
| **UA4** | 0.012 | 0.009 | 0.198 | 0.005 | 0.014 | 0.742 | 0.025 | 0.011 | 0.029 |
| **PD1** | 0.104 | 0.009 | 0.000 | 0.058 | 0.015 | 0.000 | 0.100 | 0.019 | 0.000 |
| **PD2** | 0.022 | 0.008 | 0.008 | 0.043 | 0.014 | 0.002 | 0.016 | 0.011 | 0.140 |
| **PD3** | 0.146 | 0.014 | 0.000 | 0.184 | 0.019 | 0.000 | 0.154 | 0.019 | 0.000 |
| **PD4** | 0.106 | 0.010 | 0.000 | 0.084 | 0.017 | 0.000 | 0.110 | 0.012 | 0.000 |
| **AD1** | 0.105 | 0.008 | 0.000 | 0.077 | 0.015 | 0.000 | 0.101 | 0.020 | 0.000 |
| **AD2** | 0.043 | 0.009 | 0.000 | 0.031 | 0.014 | 0.022 | 0.044 | 0.011 | 0.000 |
| **AD3** | 0.133 | 0.013 | 0.000 | 0.165 | 0.019 | 0.000 | 0.140 | 0.019 | 0.000 |
| **AD4** | 0.061 | 0.010 | 0.000 | 0.059 | 0.013 | 0.000 | 0.057 | 0.011 | 0.000 |
| **$D1^2$** | -0.009 | 0.001 | 0.000 | -0.006 | 0.003 | 0.045 | -0.007 | 0.003 | 0.017 |
| **$K45^2$** | -0.006 | 0.002 | 0.001 | -0.012 | 0.003 | 0.000 | -0.007 | 0.003 | 0.030 |
| **Const.** | | | | 0.047 | 0.019 | 0.015 | | | |
| **LogL** | -10247.2 | | | -3067.23 | | | -3670.13 | | |
| **AIC** | 20542.37 | | | 6178.46 | | | 7384.27 | | |
| **BIC** | 20727.77 | | | 6336.48 | | | 7535.05 | | |
| **MSE** | 0.0035 | | | 0.0027 | | | 0.0043 | | |
| **MAE** | 0.0472 | | | 0.0419 | | | 0.0531 | | |
| **U(55555)** | -0.175 | | | -0.167 | | | -0.167 | | |

Bold values indicate logical inconsistencies

Table 6: Actual C-TTO values versus the predicted utility values by severity for the hybrid model main effects and best-fitted, and best-fitted random-effects

| | Hybrid ME | Best-fitted Hybrid | Best fitted random -effects | Actual C-TTO utility values |
|---|---|---|---|---|
| **Severity index** | **Mean predicted utility** | **Mean predicted utility** | **Mean predicted utility** | **Mean predicted utility** |
| **6** | 0.926 | 0.903 | 0.896 | 0.892 |
| **7** | 0.850 | 0.815 | 0.835 | 0.832 |
| **9** | 0.781 | 0.737 | 0.735 | 0.743 |
| **10** | 0.714 | 0.667 | 0.644 | 0.670 |
| **11** | 0.683 | 0.645 | 0.618 | 0.623 |
| **12** | 0.577 | 0.541 | 0.552 | 0.510 |
| **13** | 0.505 | 0.474 | 0.475 | 0.472 |
| **14** | 0.462 | 0.427 | 0.419 | 0.425 |
| **15** | 0.391 | 0.366 | 0.368 | 0.400 |
| **16** | 0.345 | 0.320 | 0.316 | 0.321 |
| **17** | 0.285 | 0.285 | 0.270 | 0.281 |
| **18** | 0.185 | 0.179 | 0.172 | 0.171 |
| **19** | 0.130 | 0.141 | 0.139 | 0.118 |
| **20** | 0.053 | 0.094 | 0.074 | 0.037 |
| **21** | -0.042 | -0.010 | 0.005 | 0.007 |
| **22** | -0.055 | -0.024 | -0.007 | -0.059 |
| **25** | -0.225 | -0.175 | -0.167 | -0.166 |