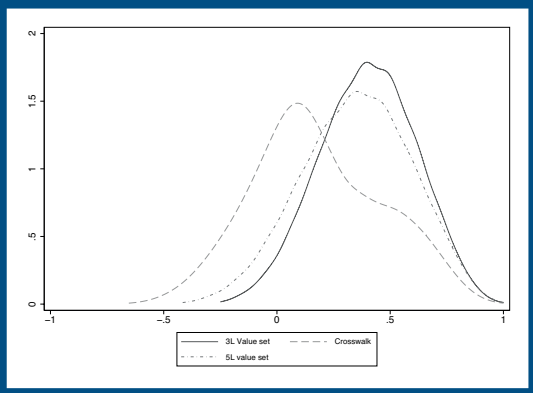

Improving the Valuation of the EQ-5D-5L by Introducing Quality Control and Integrating TTO and DCE

Juan M. Ramos-Goñi

Improving the Valuation of the EQ-5D-5L by Introducing Quality Control and Integrating TTO and DCE

Juan Manuel Ramos-Goñi

Cover design by: Ridderprint

Layout and printed by: Ridderprint

ISBN: 978-94-6299-984-8

Improving the Valuation of the EQ-5D-5L by Introducing Quality Control and Integrating TTO and DCE

Verbeteringen van het waarderen van de EQ-5D-5L door het invoeren van kwaliteitscontrole en het integreren van TTO en DCE

Thesis to obtain the degree of Doctor from the Erasmus University Rotterdam by command of the rector magnificus

Prof.dr. H.A.P. Pols

and in accordance with the decision of the Doctorate Board.
The public defence shall be held on

Friday 8th June 2018, at 11:30 hrs

by
Juan Manuel Ramos-Goñi
born in Tacoronte, Spain

Erasmus University Rotterdam

Doctoral committee:

Promoter:	Prof.dr. J.J. van Busschbach
Other members:	Prof.dr. W.B.F. Brouwer
	Prof.dr. N. Devlin
	Prof.dr. C.D. Dirksen
Copromotores:	Dr. O. Rivero-Arias
	Dr. E.A. Stolk

TABLE OF CONTENTS

1

Chapter 1

Background

Juan M. Ramos-Goñi

Several years ago, the EuroQol Group developed a generic instrument, the EQ-5D, to measure health- related quality of life (HRQoL)[1,2]. The EQ-5D, nowadays called EQ-5D-3L, uses a standardized health state descriptive system consisting of five dimensions: mobility, self-care, usual activities, pain/discomfort and anxiety/depression, each of which has three levels of severity (no problems, some problems, unable/extreme problems) (Figure 1). Together these five dimensions can describe 243 unique health states. Population value sets are available to attach a value to each of these states that reflects how good or bad each health state is according to the general population. These values reflect HRQoL.

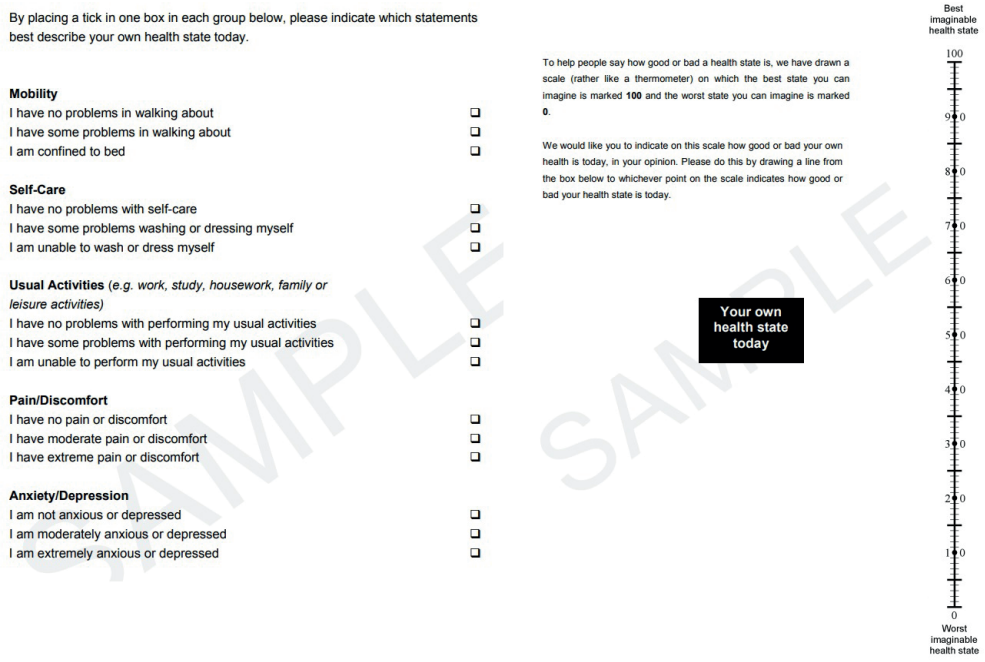

Figure 1.- EQ-5D-3L (Sample of UK version)

In the past two decades, EQ-5D-3L has become one of the most widely-used instrument for measuring health-related quality of life in medical decision-making [3]. Nevertheless, several shortcomings of the EQ-5D-3L have been e noted. Specifically, due to its crude level structure, EQ-5D has suffered from ceiling effects that limit the discriminative power of the instrument. In order to address these problems, in 2009 the EuroQol Group introduced a new version of EQ-5D, namely EQ-5D-5L [4]. This includes the same dimensions as EQ-5D-3L, but the number of severity levels per dimension was increased from three to five (no problems, slight problems, moderate problems, severe problems and unable/extreme problems) (Figure 2).

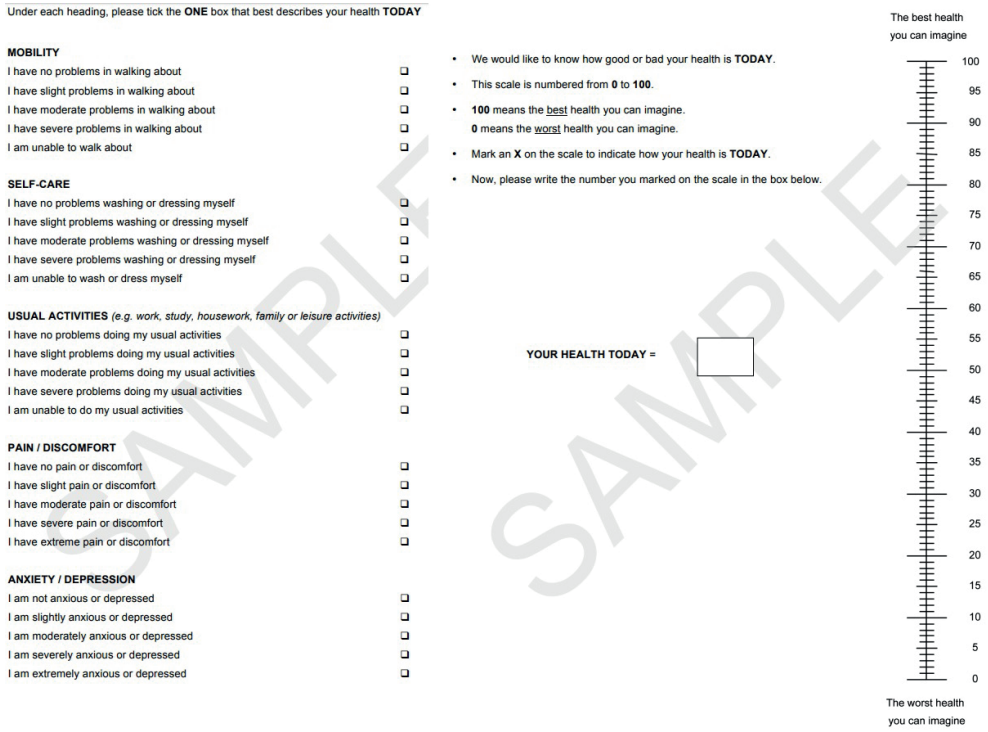

Figure 2.- EQ-5D-5L (Sample of UK version)

Valuation techniques

As a subsequent step, EQ-5D-5L value sets required construction. To harmonize valuation studies across the world and to promote best practice, the EuroQol Group introduced a standardized protocol for the valuation of EQ-5D-5L health states. The protocol developed included two different valuation techniques: Composite Time Trade-Off (C-TTO) and Discrete Choice Experiments (DCE). A detailed description of both techniques is provided in Chapter 2, but in outline, C-TTO is a combination of the traditional Time Trade-Off (TTO) technique for health states considered to be Better Than Dead (BTD) with the Lead-Time TTO for health states considered to be Worse Than Dead (WTD) (Figures 3a and 3b, respectively). To complement the protocol the EuroQol Group also developed a software platform called the EuroQol Valuation Technology (EQ-VT), which embedded the protocol. [5].

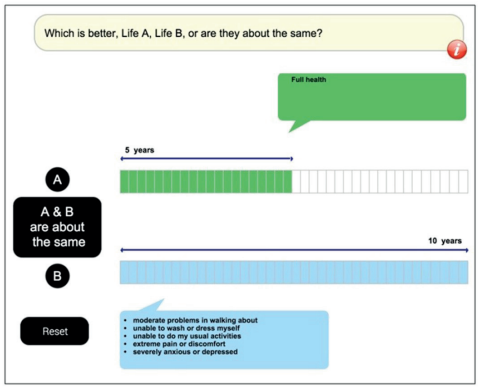

Figure 3a.- C-TTO for health states considered BTD

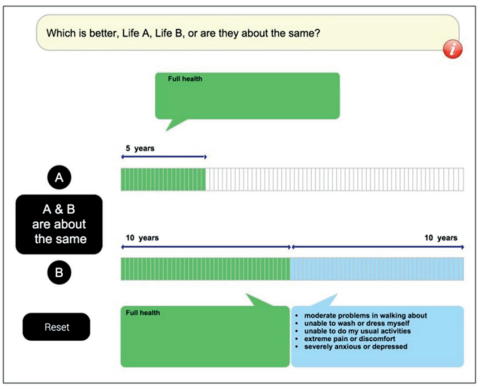

Figure 3b.- C-TTO for health states considered WTD

Figure 3.- Example of C-TTO tasks

In the C-TTO task, respondents are asked questions that aid in understanding their preferences for trade-offs between length of life and quality of life. They are asked to choose which life is better for them, life A or life B, where life B has worse health but an equal or longer lifespan. Whenever the respondent chooses A, life A is made less attractive, i.e. the number of years in life A decreases. Whenever the respondent chooses life B, life A becomes more attractive, i.e., the number of years in life A increases. This process continues until the respondent cannot decide which life is better, hence the indifference point between the two lives is reached. At this point the utility of the health state described in the blue box can be calculated as: U= t/10 where t is the number of years in life A in the case of BTD responses - e.g. in Figure 3a U = 5/10= 0.5; or U= (t-10) / 10 in the case of WTD responses - e.g. in Figure 3b, U = (5 – 10) / 10 = -0.5.

Additional information on people's preferences for health can be collected utilizing a DCE task. This comprises a series of paired comparisons between two EQ-5D-5L health states (Figure 4). The respondent is asked to decide which health state is better for him/her by selecting A or B. Note that no durations are attached to the health states.

The EQ-5D-5L valuation protocol was carefully designed to reflect best practice for the selected valuation methods [5]. The selection of methods was motivated by different considerations. On the one hand, TTO had been the most utilized valuation technique during the EQ-5D-3L era. Hence there was a clear preference for TTO over other techniques such as the standard gamble or the visual analogue scale. However, the TTO version used in 3L studies was criticized due to the arbitrary transformation of WTD values [6]. In order to avoid these transformations C-TTO was identified in an international research programme as the best candidate to replace the traditional TTO method [7]. On the other hand, DCE was an

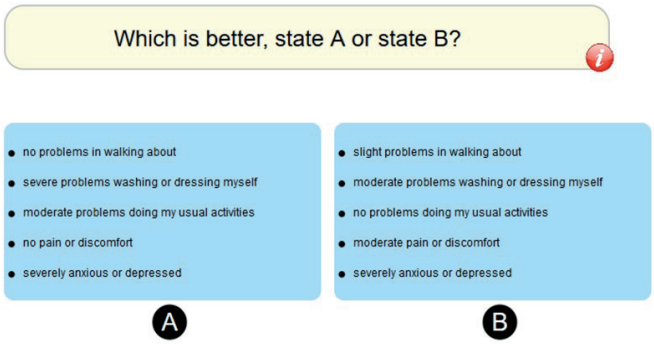

Figure 4.- Example of DCE task

emerging valuation technique at the time that the protocol was developed, and was identified to be a complementary valuation technique to C-TTO. In addition, the theoretical possibility of a hybrid model was proposed [8].

While the EQ-5D-5L valuation protocol was introduced with substantial evidence backing the methodological choices that were made, it was neither officially tested for its intended purpose of constructing value sets, nor was the methodology for combining C-TTO and DCE data in a hybrid model fully developed. As PI of the Spanish EQ-5D-5L valuation study, I was acutely aware of this, because strong interviewer effects were found in the Spanish EQ-5D-5L valuation data, suggesting that data quality was highly variable. These issues were dealt with at two levels: we investigated the scope to improve the EQ-5D-5L valuation protocol to prevent similar issues arising in later studies, and we also explored the consequences of interviewer effects and data quality concerns in ~~for~~ handling ~~of~~ the data.

Specific research questions and thesis structure

This thesis is based on the Spanish valuation study and the methodological research that it inspired. The specific research questions to be addressed are:

1. To what extent is the proposed valuation protocol feasible and are hybrid estimations possible in practice?
2. Is there an explanation for the interviewer effects found in chapter 2? If so, how can the existing protocol be modified to collect better data?
3. What types of techniques are most suitable for modelling the C-TTO and DCE valuation data?

The outline of this thesis is as follows. Chapter 2 addresses research question 1 concerning the feasibility of the protocol. This chapter reports on the national value set study conducted in Spain with the EQ-5D-5L valuation protocol. In addition, this chapter explores the estimation of a hybrid model combining C-TTO and DCE data obtained from the application of the protocol.

As shown in chapter 2, the EQ-5D-5L valuation protocol developed by Oppe et al. seemed to be feasible in terms of producing a value set. However, the first test of the protocol found interviewer effects. Chapters 3, 4 and 5 explore the reasons for these interviewer effects, together with the implementation and testing of protocol modifications (research question 2). In particular, in attempting to explain interviewer effects, there is an exploration of the presence of learning and satisficing effects. Two modifications of the protocol were implemented and tested, namely: (i) the introduction of a ranking task prior to the C-TTO task in order to reduce the impact of learning effects, and (ii) the introduction of a quality control methodology aimed at reducing both satisficing and interviewer effects.

Chapter 3 uses data from six valuation studies conducted in the US, Spain and the Netherlands to explore the presence of learning and satisficing effects on both TTO and DCE data that could explain the interviewer effects found in chapter 2.

Chapter 4 explores the possibility to reintroduce ranking as a warm-up task in the valuation protocol for the EQ-5D-5L with the aim of reducing learning effects. The first valuation study for the EQ-5D-3L instrument was conducted in the UK in 1999. The protocol used included different warm-up tasks employed prior to the administration of the TTO. One of these was a ranking task where participants were asked to rank from best to worst the 10 health states that they valued later using the TTO technique.

Indepth exploration of the interviewer effects reported in chapter 2 showed that protocol violations were present in many interviews. Chapter 5 describes a quality control methodology aimed at reducing these violations and improving interviewer skills. This chapter also illustrates the benefits that can be obtained from quality control by comparing the properties of valuation datasets collected with and without quality control.

The next four chapters deal with the modelling of valuation data (research question 3). Ordinary Least Squares (OLS) has been the preferred method to model TTO data in the past. However, based on the data issues that were recognized in the previous chapters, it was feared that the use of OLS to model valuation data would provide biased estimates. In addition, DCE data cannot be modelled using the traditional OLS approach as no values are observed on DCE tasks. Only preference of one state over another is observed in each task. Thus DCE data has to be modelled using conditional binary response regression methods. Briefly, chapters 6, 7, 8, and 9 focus respectively on testing DCE models, testing C-TTO models, improving the hybrid model to account for intervals and heteroscedasticity, and testing the improved hybrid model, in order to estimate the Spanish EQ-5D-5L value set.

As stated in the introduction, the DCE tasks included in the protocol did not attach duration to the health states. This had the implication that value sets produced by the DCE method were on a latent scale instead of the (0) dead - (1) full health scale required for QALY calculations. Chapter 6 uses data from an EQ-5D-5L valuation pilot study conducted in

Spain in 2011 to explore different modelling techniques to anchor the latent scale value sets produced by the DCE data onto the (0) death - (1) full health scale.

When developing the valuation protocol the EuroQol Group was uncertain about whether the number of health states included in the C-TTO tasks would be enough to make possible value set estimations. Chapter 7 uses data from the national EQ-5D-5L valuation set in Uruguay to explore whether C-TTO valuations can be used alone to generate a national value set.

When dealing with interviewer effects to estimate the Spanish value set for the EQ-5D-5L instrument, the team realized that the required mathematical models were not available. Chapter 8 introduces both the mathematical development and the software implementation which made it possible to extend the initial hybrid model description to allow the inclusion of censored and interval responses. In addition, this chapter introduces hybrid heteroscedastic models to take account of preference heterogeneity when estimating a national value set.

To finalize the research questions and make possible the estimation of a less biased EQ-5D-5L national value set for Spain than the one presented in chapter 2, chapter 9 uses the evidence from chapters 2-5 to construct interval C-TTO responses which aim to correct the interviewer effects and data quality issues encountered in the Spanish valuation study. In addition, this chapter utilizes the processes reported in chapter 8 to incorporate all the information from a hybrid model to estimate a value set for the EQ-5D-5L instrument.

Finally, chapter 10 discusses the findings from the previous chapters and outlines the possible consequences for future research. Chapters 2 to 9 are papers published in peer-reviewed international journals. Hence each can be read independently. Chapter 10 takes text from a paper under the review process in a peer-reviewed international journal (Value in Health).

REFERENCES

1. EuroQol Group. EuroQol: a new facility for the measurement of health-related quality of life. Health Policy. 1990;16(3):199–208.

2. Brooks R. EuroQol Group: the current state of play. Health Policy. 1996;37(1):53–72.

3. Wisløff T, Hagen G, Hamidi V, et al. Estimating QALY gains in applied studies: a review of cost-utility analyses published in 2010. Pharmacoeconomics 2014;32:367–75.

4. Herdman M, Gudex C, Lloyd A, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res 2011;20:1727–36.

5. Oppe M, Devlin NJ, van Hout B, Krabbe PF, de Charro F. A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. Value Health. 2014;17(4):445–53.

6. Craig BM, Oppe M. From a different angle: a novel approach to health valuation. Soc Sci Med. 2010 Jan;70(2):169-74.

7. Devlin N, Krabbe P. The development of new research methods for the valuation of EQ-5D-5L. Eur J Health Econ. 2013;14(Suppl 1):S1–3.

8. Oppe M, van Hout B. The optimal hybrid: experimental design and modeling of a combination of TTO and DCE. EuroQol Group Proceedings. 2013. Available at:

9. https://eq-5dpublications.euroqol.org/download?id=0_53738&fileId=54152. Accessed December 20, 2017.

2

Chapter 2

Valuation and modelling of EQ-5D-5L health states using a hybrid approach

Juan M. Ramos-Goñi,
José L. Pinto-Prades,
Mark Oppe,
Juan M. Cabasés,
Pedro Serrano-Aguilar,
Oliver Rivero-Arias

Med Care. 2017 Jul;55(7):e51-e58

ABSTRACT

Background: 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. Recently, a version called EQ-5D-5L with 5 levels on each dimension was developed. This manuscript explores the performance of a hybrid approach for the modeling of EQ-5D-5L valuation data.

Methods: Two elicitation techniques, the composite time trade-off, and discrete choice experiments, were applied to a sample of the Spanish population (n = 1000) using a computer-based questionnaire. The sampling process consisted of 2 stages: stratified sampling of geographic area, followed by systematic sampling in each area. A hybrid regression model combining composite time trade-off and discrete choice data was used to estimate the potential value sets using main effects as starting point. The comparison between the models was performed using the criteria of logical consistency, goodness of fit, and parsimony.

Results: Twenty-seven participants from the 1000 were removed following the exclusion criteria. The best-fitted model included 2 significant interaction terms but resulted in marginal improvements in model fit compared to the main effects model. We therefore selected the model results with main effects as a potential value set for this methodological study, based on the parsimony criteria. The results showed that the main effects hybrid model was consistent, with a range of utility values between 1 and -0.224.

Conclusion: This paper shows the feasibility of using a hybrid approach to estimate a value set for EQ-5D-5L valuation data.

Key Words: utility theory, quality of life, maximum likelihood estimation, time trade-off, discrete choice experiment

BACKGROUND

The EQ-5D instrument is the most widely used preference- based health-related quality of life questionnaire in cost- effectiveness analysis. Reimbursement agencies such as the UK National Institute for Health and Care Excellence (NICE) 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].

The original EQ-5D (EQ-5D-3L) is a questionnaire with 5 dimensions (mobility, self-care, usual activities, pain/ discomfort, and anxiety/depression) and 3 levels in each dimension (no problems, some problems, and extreme problems) [2]. Extensive research supports the use of the instrument in many disease areas but recent studies have shown ceiling effects issues, particularly in general population samples [3,4]. In response to this, the EuroQol Group proposed a new version of the instrument the EQ-5D-5L. This new version increased the number of severity levels from 3 to 5 (no problems, slight, moderate, severe, and unable or extreme) describing 3125 (5^5) possible health states [3]. Each health state is usually represented using a 5-digit number (profile) where 11111 indicates perfect health and 55555 the worst health state or pits state.

Available EQ-5D-3L value sets cannot be used directly with 5-level version responses. As a temporary solution, an interim scoring algorithm needs to be used [5]. Therefore, new valuation studies are necessary to obtain preferences from the general public for EQ-5D-5L health states. The EuroQol Group has developed a valuation protocol to elicit preferences after a series of pilot studies conducted by research teams worldwide [6]. A group of researchers based in Spain, the UK, and the Netherlands, has been one of the first teams in implementing this protocol. This manuscript explores the feasibility of a hybrid method to estimate a potential value set for EQ-5D-5L valuation data.

METHODS

Protocol

The results obtained from the pilot studies [6] informed the standardized protocol for EQ-5D-5L value sets used in this study [7]. The interview process described in the protocol has 5 sections. First, a general welcome and an introduction to the research were given. Next, respondents were asked to provide background information, including their own health using the EQ-5D-5L, age, sex, and experience with illness. This was followed by the composite time trade-off (C-TTO) task, which was administered after giving an explanation of the task, and included 10 EQ-5D-5L C-TTO valuations. The next part was a discrete choice (DC) experiment, which consisted of 7 paired comparisons. Finally, there was a general thank you and goodbye. After each block of tasks (C- TTO and DC experiments) and at the end of the

interview, participants were given the opportunity to clarify whether they found difficulties completing the tasks and the overall survey. The EuroQol Group developed the online system to carry out the survey called EuroQol Valuation Technology (EQ-VT).

Eliciting Preferences Methods

C-TTO

The traditional time trade-off (TTO) has been widely used in the EQ-5D-3L valuation studies conducted so far and it is appropriate to value health states considered better than dead [8,9]. However, using the traditional TTO method for states worse than dead gives negative values that are normally transformed to be bounded to -1, which has been criticized in the literature [10]. Other TTO alternatives to evaluate health states were therefore assessed during the EuroQol pilot studies including lead and lag time [11,12]. In the former, additional trading time is included before the health state, whereas in the latter, trading time is included after the health state to be valued. The pilot studies looked at the potential of using these methods in practice and concluded that the 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 single task [13]. For the lead-time TTO, 10 years lead-time and 10 years in the state were used. This lead-time method produces a minimum value of -1 and no transformation of negative values is needed. The iterative process used in the original UK valuation exercise [8] was adapted to be used in the C-TTO task. The C-TTO design included 86 health states selected using Monte Carlo simulation. The health states were distributed over 10 blocks and each block contained 1 very mild state (1 dimension at level 2, the remaining dimensions at level 1), the pits state 55555, and a balanced set of intermediate states. The EQ-VT randomly assigned respondents to one of the blocks and presented the states in random order.

DC Experiment

The use of DC experiments for health state valuation has received recent attention in the literature [14,15]. Modeling ordinal data follows the theoretical foundations of random utility theory [16]. Values obtained with DC models have been shown to have patterns similar to those obtained with TTO models [17]. The values obtained from DC models are expressed on an arbitrary scale and need to be rescaled on the dead (0) full health (1) scale [17,18]. Using DC experiments was also piloted and the results suggested that collecting such information could provide additional useful information to the C-TTO data. Hence, a DC experiment was included as part of the protocol. The DC experiment design included 196 pairs divided in 28 blocks with similar severity representation identified using Bayesian design[19]. The EQ-VT randomly assigned respondents to one of the blocks, presented the pairs in random order, and randomized the location of the states within the pair (i.e., left and right).

Sampling and Data Collection

Our power calculations estimated that to obtain a 0.01 SE of the observed mean C-TTO, we needed 9735 C-TTO responses. We therefore recruited 1000 participants that after completing the valuations tasks provided 10,000 C-TTO and 7000 DC responses to estimate the models.

A 2-stage sampling strategy was designed to obtain a representative sample of the Spanish population. In a first stage, we stratified geographically by Spanish provinces, whereas in a second stage we systematically sample individuals from a panel until an accurate age and sex distribution for that province was achieved. We contracted an independent market research company, which identified respondents and arranged interviews at convenient places. Interviews were conducted face-to-face during June and July 2012 by 33 trained interviewers. Respondents did not receive payment for participating in the survey. A different market research company was contracted to call a random sub- sample of 15% of respondents as quality control of the process.

Statistical Analyses

Descriptive statistics were used to summarize respondent's characteristics and responses to the C-TTO and DC experiments.

Two sources of data were available to estimate the EQ-5D-5L value set: C-TTO and DC data. To maximize the use of the available data, we implemented a hybrid modeling approach that made use of both C-TTO and DC data to estimate the potential value sets. This hybrid method estimated a unique set of coefficients from a likelihood function obtained multiplying the likelihood functions of a normal distribution for the C-TTO data by the likelihood function of a conditional logit distribution for DC data [20]. As the coefficients estimated from a conditional logit are expressed on a latent arbitrary utility scale, we used a rescaled parameter θ, which assumes that the C-TTO model coefficients are proportional to DC model coefficients. See the Appendix for a full description and analytical derivation of the hybrid method. This method combines the utility values elicited in the C-TTO for the 86 health states with utility values elicited in the DC experiment for 196 pairs of states. The dependent variable in the C-TTO part of the model was defined as 1 minus the C-TTO observed values for a given health state to indicate disutility and therefore coefficients expressed utility decrements. In the DC part of the model, the dependent variable was a binary outcome 0/1 indicating the respondent's choice for each pair of EQ-5D-5L states. We used cluster estimation to acknowledge that for each participant included in the models, 10 C-TTO and 7 DC responses were available.

We also present models to estimate C-TTO and DC data separately, to illustrate how the hybrid model combined both types of data. We analysed C-TTO data using a linear regression model assuming normal distribution in its errors, as it is the C-TTO part of hybrid model. We analysed DC data using the standard econometric method for ordinal data conditional logit

regression [16]. To make model coefficients comparable, we rescaled the DC model coefficients using the same rescaling parameter y that was estimated in the hybrid model.

We started exploring the hybrid 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 4 dummies MO1 to 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. The same set of dummy variables was defined for each of the remaining dimensions: self-care (SC), usual activities (UA), pain/discomfort (PD), and anxiety/depression (AD). We also estimated the model using the definition of dummies implemented in most previous EQ-5D-3L valuation exercises [21] and such analyses are available from the authors upon request.

Our starting point for the selection of additional co-variates for the models was the US valuation study.[9] Several variables were defined. For example, $D1$ as the number of dimensions at levels 2, 3, 4, or 5 beyond the first; IJ as the number of dimensions at level J beyond the first; $K45$ as the number of dimensions at level 4 or 5, and others. Squared of all terms were also introduced to assess nonlinear effects on the dependent variable. We included all terms first, and use a stepwise approach removing non-significant terms and ensuring model consistency.

Exclusion Criteria and Interviewer Assessment

We excluded observations using the following 2 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.

We used the Kruskal-Wallis test to assess the differences among mean values by interviewer in the C-TTO responses. We further assess this including dummies that identified interviewers in the main effects model and using an F test among the dummy coefficients.

Evaluation of Model Performance

We evaluated model performance using (1) logical consistency of parameters; (2) goodness of fit; and (3) parsimony. Estimated coefficients are said to be logically consistent if magnitude values from logically worse health states are lower than those from logically better health states. In our estimated results this is translated to all main effects coefficients being positive. Goodness of fit was assessed using the Akaike (AIC) and the Bayesian information criteria

(BIC). 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 3 criteria were used to compare different hybrid model specifications using different interaction terms. However, prediction accuracy evaluated using mean square error or mean absolute error are not appropriate measures in this case, given the lack of an appropriate counterfactual for hybrid model predictions.

We present the results of the regression with the main effects and the best-fitted model with significant terms. Statistical analysis and regression modeling 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.

Comparison with EQ-5D-3L Value Set

We calculated and compared predictions for the 3,125 health states using the final selected EQ-5D-5L value set and the interim solution to calculate EQ-5D-3L values [5] presented for a selected set of health states covering mild, moderate, and severe states. In addition, we compared the kernel density functions for the index values of the 243 states of the Spanish EQ-5D-3L value set[23] and for the 3,125 states of the final selected EQ-5D-5L value set.

RESULTS

Descriptive Statistics

Twenty-seven participants from the 1000 were removed following the exclusion criteria: 18 respondents with a positive slope on a regression between his/her values and the severity of the health states and 9 respondents who valued all states equal to death. Overall the excluded observations were older with no studies or primary school studies than the estimation sample (Table 1). The estimation sample was similar in the distribution of employment status; mean age and sex distribution than Spanish population, but the estimation sample had a larger number of respondents in age group 25–34 and fewer participants over 75 (Table 1). The self-reported health using the EQ-5D-5L of respondents showed that 18.90% reported problems in usual activities and 30.8% reported problems in anxiety or depression dimension (Table 1). For the remaining dimensions, proportions of respondents with problems were <10% (Table 1).

The outcome of the quality control reported no incidences, but we observed significant differences between interviewers in the valuations obtained with Kruskal-Wallis ($P < 0.0001$) and F tests ($P < 0.0001$). We report further descriptive information about the C-TTO and the DC data in the online supplemental digital content (Tables 1 and 2 and SDC Figures 1 and 2, Supplemental Digital Content, http://links.lww.com/MLR/A839).

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	%	%	
Age groups						
- 18-24	3	11.2	114	11.7	9.0	
- 25-34	4	14.8	270	27.8	18.3	
- 35-44	5	18.5	170	17.5	19.6	
- 45-54	5	18.5	148	15.2	17.9	
- 55-64	4	14.8	111	11.4	13.5	
- 65-74	2	7.4	108	11.1	10.2	
- 75+	4	14.8	52	5.3	11.0	
Gender						
- Male	12	44.4	463	47.6	49.3%	
- Female	15	55.6	510	52.4	50.7%	
Employment status						
- Housewife/house husband	1	3.7	70	7.2	10.51	
- Employed or freelance	11	40.8	529	54.4	44.98	
- Student	2	7.4	89	9.1	6.33	
- Retired	8	29.6	132	13.6	20.12	
- Unemployed	5	18.5	139	14.3	15.01	
- Disabled	0	0	8	0.8	3.03	
- Missing	-	-	6	0.6	-	
Education						
- Higher education	10	37.0	314	32.47	17.70	
- High school	2	7.4	374	38.68	53.90	
- Primary school	10	37.0	234	24.20	26.30	
- No studies	5	18.5	45	4.65	2.10	
- Missing	-	-	6	0.6		
Experience with illness						

Variables	Excluded sample (n =27)		Estimation sample (n = 973)		Spanish General Population*
- Personal (%YES)	4	14.8	140	14.4	n/a
- Relatives (%YES)	17	62.96	616	63.3	n/a
- Other (%YES)	9	33.3	338	34.7	n/a
Self-reported EQ-5D-5L					
Mobility					
- No problems	22	81.48%	864	88.80%	86.1%
- Slight problems	4	14.81%	69	7.09%	6.1%
- Moderate problems	1	3.70%	32	3.29%	4.7%
- Severe problems	0	0%	7	0.72%	2.4%
- Unable/extreme problems	0	0%	1	0.10%	0.8%
Self-care					
- No problems	24	88.89%	933	95.89%	93.9%
- Slight problems	2	7.41%	30	3.08%	2.5%
- Moderate problems	1	3.70%	9	0.92%	1.7%
- Severe problems	0	0%	1	0.10%	0.9%
- Unable/extreme problems	0	0%	0	0%	1.0%
Usual activities					
- No problems	22	81.48%	891	91.57%	89.2%
- Slight problems	3	11.11%	57	5.86%	4.7%
- Moderate problems	2	7.41%	20	2.06%	3.2%
- Severe problems	0	0%	4	0.41%	1.5%
- Unable/extreme problems	0	0%	1	0.10%	1.4%
Pain					
- No problems	20	74.07%	772	79.34%	75.2%
- Slight problems	5	18.52%	149	15.31%	12.3%
- Moderate problems	1	3.70%	37	3.80%	8.7%
- Severe problems	0	0%	10	1. 03%	3.5%
- Unable/extreme problems	1	3.70%	5	0.51%	0.4%
Anxiety/Depression					
- No problems	15	55.56%	673	69.17%	85.4%
- Slight problems	8	29.63%	214	21.99%	8.4%
- Moderate problems	2	7.41%	71	7.30%	4.2%
- Severe problems	1	3.70%	15	1.54%	1.6%
- Unable/extreme problems	1	3.70%	0	0%	0.4%

n/a: not available; *Data extracted from the 2012-2013 National Spanish Health Survey

Modeling Results

The hybrid model with main effects was a consistent model predicting utilities with a range between 1 and 0.224 (Table 2). Both, the C-TTO and the DC models derived logical inconsistencies. It is shown how the hybrid model corrects the inconsistencies in the C-TTO model by using DC information and the DC model inconsistencies with C-TTO information. As described in the Appendix, the log likelihood in the hybrid model was approximately the sum of the log likehood of both C-TTO and DC models separately.

Table 2: Estimation results for hybrid model using main effects only

	Hybrid (C-TTO+DCE) model			C-TTO model			Re-scaled DCE model		
	Coeff.	**SE**	**p-value**	**Coeff.**	**SE**	**p-value**	**Coeff.**	**SE**	**p-value**
MO1	0.084	0.008	0.000	0.015	0.015	0.293	0.088	0.010	0.000
MO2	0.014	0.009	0.101	0.053	0.015	0.000	0.012	0.011	0.272
MO3	0.130	0.010	0.000	0.152	0.018	0.000	0.115	0.012	0.000
MO4	0.060	0.010	0.000	0.023	0.019	0.230	0.081	0.013	0.000
SC1	0.056	0.008	0.000	0.038	0.015	0.009	0.030	0.011	0.008
SC2	0.000	0.009	0.989	**-0.001**	0.017	0.964	0.017	0.011	0.126
SC3	0.097	0.011	0.000	0.131	0.020	0.000	0.079	0.013	0.000
SC4	0.016	0.009	0.090	0.012	0.018	0.506	0.022	0.011	0.047
UA1	0.053	0.008	0.000	0.040	0.014	0.006	0.037	0.011	0.000
UA2	0.005	0.010	0.572	0.035	0.019	0.069	**-0.008**	0.011	0.485
UA3	0.072	0.010	0.000	0.085	0.021	0.000	0.069	0.011	0.000
UA4	0.004	0.010	0.705	**-0.030**	0.017	0.082	0.024	0.013	0.056
PD1	0.078	0.008	0.000	0.049	0.014	0.000	0.066	0.011	0.000
PD2	0.024	0.009	0.007	0.044	0.019	0.024	0.019	0.012	0.093
PD3	0.115	0.011	0.000	0.100	0.019	0.000	0.130	0.013	0.000
PD4	0.105	0.010	0.000	0.090	0.023	0.000	0.118	0.014	0.000
AD1	0.085	0.008	0.000	0.057	0.018	0.002	0.065	0.011	0.000
AD2	0.044	0.010	0.000	0.038	0.018	0.040	0.049	0.012	0.000
AD3	0.121	0.010	0.000	0.116	0.019	0.000	0.119	0.013	0.000
AD4	0.053	0.010	0.000	0.049	0.017	0.005	0.063	0.013	0.000
Const.	0.007	0.004	0.087	0.098	0.018	0.000			
LogL	-10292.97			-6565.7			-3675.81		
AIC	20631.95			13173.41			7391.62		
BIC	20809.62			13324.25			7528.69		
U(55555)	-0.224			-0.194			-0.196		
Lowest prediction (state)	-0.224 (55555)			-0.224 (55455)			-0.196 (55555)		

Bold values indicate logical inconsistencies

After exploring many interactions terms, the best-fitted estimation model we found was using the interaction terms $D1^2$ and $K45^2$ (Table 3). The constant term of this model was suppressed as the $D1^2$ term captures the effect of the constant. The reduction of the hybrid log likehood estimation for those terms inclusion only reduces the AIC and BIC by 0.4%. About 3/4 of this reduction was produced by a reduction in the C-TTO part of the model.

Table 3: Estimation results using best-fitted model

	Hybrid (C-TTO+DCE) model			C-TTO model			Re-scaled DCE model		
	Coeff.	**SE**	**p-value**	**Coeff.**	**SE**	**p-value**	**Coeff.**	**SE**	**p-value**
MO1	0.112	0.009	0.000	0.065	0.015	0.000	0.119	0.017	0.000
MO2	0.020	0.008	0.018	0.036	0.015	0.014	0.016	0.011	0.138
MO3	0.143	0.014	0.000	0.199	0.022	0.000	0.134	0.018	0.000
MO4	0.070	0.009	0.000	0.035	0.020	0.079	0.076	0.012	0.000
SC1	0.079	0.009	0.000	0.090	0.013	0.000	0.064	0.019	0.001
SC2	0.006	0.009	0.518	**-0.007**	0.017	0.680	0.018	0.011	0.093
SC3	0.115	0.013	0.000	0.180	0.024	0.000	0.101	0.019	0.000
SC4	0.024	0.009	0.008	0.032	0.020	0.097	0.021	0.010	0.038
UA1	0.081	0.009	0.000	0.102	0.015	0.000	0.073	0.021	0.000
UA2	0.005	0.009	0.607	0.021	0.020	0.297	**-0.005**	0.011	0.622
UA3	0.095	0.013	0.000	0.125	0.025	0.000	0.090	0.017	0.000
UA4	0.012	0.009	0.198	**-0.008**	0.017	0.620	0.024	0.012	0.037
PD1	0.104	0.009	0.000	0.095	0.011	0.000	0.100	0.019	0.000
PD2	0.022	0.008	0.008	0.040	0.020	0.045	0.016	0.011	0.130
PD3	0.146	0.014	0.000	0.159	0.022	0.000	0.153	0.020	0.000
PD4	0.106	0.010	0.000	0.097	0.025	0.000	0.109	0.012	0.000
AD1	0.105	0.008	0.000	0.110	0.012	0.000	0.100	0.020	0.000
AD2	0.043	0.009	0.000	0.016	0.018	0.376	0.044	0.011	0.000
AD3	0.133	0.013	0.000	0.153	0.020	0.000	0.139	0.019	0.000
AD4	0.061	0.010	0.000	0.070	0.019	0.000	0.056	0.012	0.000
$D1^2$	-0.009	0.001	0.000	-0.013	0.003	0.000	-0.007	0.003	0.020
$K45^2$	-0.006	0.002	0.001	-0.009	0.003	0.001	-0.007	0.003	0.031
LogL	-10247.2			-6550.7			-3670.1		
AIC	20542.4			13145.3			7384.3		
BIC	20727.8			13303.4			7535.1		
U(55555)	-0.175			-0.173			-0.160		
Lowest prediction (state)	-0.1.75 (55555)			-0.181 (55455)				-0.160 (55555)	

Bold values indicate logical inconsistencies

The main effects hybrid model produced a wider range of utility values at the upper and lower end of the scale compared to the hybrid model including the terms $D1^2$ and $K45^2$ (Table 4). Given that the improvement in goodness of fit between the main effects and the best-fitted model was marginal (0.4%), we have selected the estimation results from the hybrid model with main effects as the value set for this methodological study based on the parsimony criteria.

Table 4: Predicted utility values by health state for estimated models and from the interim EQ-5D-3L solution

State	Hybrid ME	Hybrid ME + D1² + K45²	Interim solution from EQ-5D-3L
11112	0.9072	0.8945	0.9320
21111	0.9081	0.8877	0.8930
11121	0.9146	0.8958	0.9100
12111	0.9363	0.9207	0.8680
11211	0.9392	0.9190	0.9240
42114	0.4574	0.4259	0.4790
33511	0.7030	0.6342	0.2500
25331	0.5788	0.5418	0.1290
35411	0.5942	0.5265	0.1580
34511	0.6064	0.5383	0.2080
35412	0.5088	0.4676	0.1090
33531	0.6012	0.5541	0.1610
55512	0.3160	0.2747	-0.2980
52533	0.2829	0.2822	-0.2030
34544	0.1389	0.1716	-0.0530
34553	0.1551	0.1541	-0.1160
55433	0.1740	0.1817	-0.3170
35552	0.1831	0.1738	-0.2350
54454	-0.1517	-0.0777	-0.3970
55444	-0.0625	0.0045	-0.4260
55552	-0.0060	0.0057	-0.5590
54455	-0.2046	-0.1391	-0.4380
55554	-0.1712	-0.1134	-0.6120
55545	-0.1192	-0.0689	-0.5510
55555	-0.2242	-0.1748	-0.654

The probability density functions of the Spanish EQ-5D-3L value set and the EQ-5D-5L value set presented here (Fig. 1) show a symmetric distribution for EQ-5D-5L, whereas the

EQ-5D-3L has a bimodal distribution. The proportion of states considered worse than death is lower in the EQ-5D-5L value set.

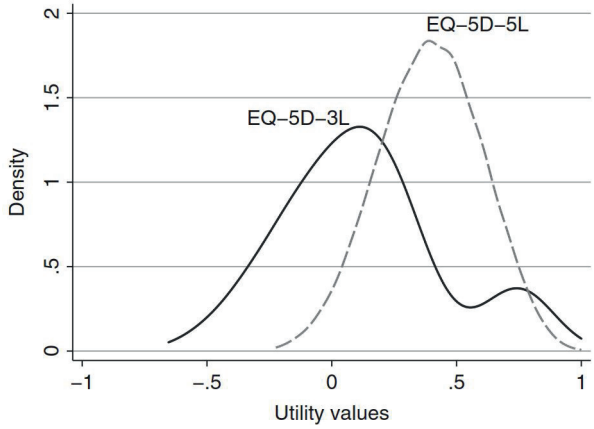

Figure 1 - Probability density function of EQ-5D-3L and EQ-5D-5L value sets.

DISCUSSION

In this manuscript we have reported the performance of a hybrid approach to estimate a value set for the EQ-5D-5L questionnaire. The choice of the hybrid approach is based on the assumption that subjects have a unique utility function that generates both the sets of responses. If utilities were the same in the C-TTO and DC methods, there would be no need of combining them except for having more precise estimates. Our hypothesis is that this disparity is related to the choice versus matching discrepancy as it is one of the most replicated effects in preference elicitation literature [24-27]. Some researchers have tried to find arguments in favour of one method or the other [28]. We believe that neither matching-based (like C-TTO) nor choices (DC) methods are unbiased [29]. Matching methods are influenced by scale compatibility and, in the case of C-TTO, by loss aversion [30]. Choices are also subject to problems as it has been shown that responses are more lexicographic in choice than in matching. Evidence on the prominence effect suggests that in choices, subjects tend to choose the alternative that is better with respect to the more important attribute without paying enough attention to how much better the option is [31]. Finally, it has also been observed that subjects perceive the distances between outcomes differently when comparisons are conducted in a separate or in a joint model, again without clear evidence that one method is better than another [32]. We then do not think that the "true" values can be inferred from 1 single method and for this reason we suggest that it can make sense to use a hybrid approach. We are not claiming that the biases present in 1 method compensate the biases present in the other so that adding up the 2 methods we get unbiased results. There is no evidence to suggest this is the case. Even 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 DC experiment, rather the opposite.

In our results, introducing the $D1^2$ and $K45^2$ terms provided a better fit to the data suggesting that the selected value set should have included such effect. However, the improvement in fit is mostly captured by the C-TTO part of the model. Given that the improvement in the goodness of fit of using $D1^2$ and $K45^2$ variables was marginal as suggested by the AIC and the BIC, we selected the main effects model using the parsimony criterion.

As far as we are aware there is no EQ-5D-5L value set available in the literature for direct comparison. Given the lack of such information, we compared our model with the Spanish value set for the 3L version of EQ-5D [23]. Our model has higher values in the upper scale compared to the EQ-5D- 3L valuation study conducted in Spain. This was expected as the label for level 2 in the 3L version is "some problems" and the label for level 2 in the 5L version is "slight problems." However, the utility decrement of level 2 for AD dimension is higher in our study than in the 3L study. A possible ex planation for this is the fact that the self-reported health results in our sample showed a high rate of people reporting problems in the AD dimension, causing them to put more weight on this dimension in the valuation tasks. On the other side of the scale the pits state prediction was higher in our study. Something expected as well, as the change in the wording of the mobility level "confined to bed" in EQ-5D-3L to "unable to walk about" in EQ-5D-5L has changed the definition of the worst possible health state. Given that this new level is not as severe as "confined to bed" (which had the largest decrement of all dimensions in the Spanish 3L study) it is expected to obtain higher valuations for 55555 than for 33333. We observed a lower proportion of negative values in our study in comparison with the Spanish EQ-5D-3L value set. The number of non-extreme health states has increased >10-fold in the EQ-5D-5L compared with the 3L version reducing the proportion of the extreme health states, and partly explaining why the kernel density distribution of the 5L value set shows a smaller area below 0 than the 3L value set.

The hybrid model is not exempt of limitations. The assumption of normal distribution for the errors in the C-TTO part suffers from problems related to the robustness of the estimation of SE and related to the violation of the homoscedasticity condition. In addition, the use of conditional logit model for DC data does not explicitly consider within respondents correlations. We try to limit the impact of these limitations by using cluster estimations of the SEs of the estimated coefficients. However, further exploration of more sophisticated hybrid models for both types of data is needed. For example, the use of random coefficient models for the C-TTO part and mixed (conditional) logit models for the DC part of the model.

We have observed significant differences in the valuations observed by interviewers that lead us to be cautious about suggesting a final value set to use in practice in Spain. We are now

trying to understand the nature of these differences, which could be attributable to several factors including issues with the EQ-VT software, the use of C-TTO, or noncompliance of the protocol by the interviewer.

We present here a novel methodological approach to obtain an EQ-5D-5L value set. Our results show the feasibility of using a hybrid model to estimate a value set for EQ- 5D-5L valuation data.

ACKNOWLEDGMENTS

The authors are very grateful to the EuroQol Valuation Methodology Management Team (Frank de Charro, Ben van Hout, Nancy Devlin, and Paul Krabbe) and the support team (Arnd Jan Prause, Gerben Bakker, and Job A. de Bruyne) for the constant and unconditional advice and support received during the conduct of this study.

Appendix: The hybrid model

There are several methods that enable the combination of both sets of data in a single model. The hybrid model we present here uses a maximum likelihood approach. It builds on the notion that both linear regression (as applied to the **C-**TTO data) and logistic regression (as applied to the **DC** data) can be obtained by maximum likelihood estimation and that both models contain a similar linear component βx underlying the values and choices. If one assumes that this component, which reflects the weight given to the dimensions and labels, is identical between both approaches, one can find the optimal parameters for the combination of the data. This is done by creating a single likelihood function for the joined data by multiplying the likelihoods of the C-TTO data and the **DC** data (or-equivalently- by adding the log likelihoods).

However, we know that the C-TTO model and the **DC** model are anchored on a different scale. We can take this into account in the combined likelihood by including (an) additional parameter(s) relating both linear functions with each other. In the model presented here, we assume that the weights (i.e. the β's) in both models differ up to a monotonic transformation θ.

The likelihood of the C-TTO data is expressed as follows:

$$like_{\boldsymbol{C\text{-}TTO}} = Normal\_pdf\left(\sum_{j=1}^{J}\beta_j d_{ij}\right)$$

where, β_j is the vector of C-TTO regression coefficients, d_{ij} the vector of dummy variables for state *i*, *J* the number of dummies, and pdf the probability density function.

The likelihood of the **DC** data is defined as:

$$like_{\boldsymbol{DC}} = ConditionalLogit\_\boldsymbol{cdf}\left(\sum_{j=1}^{J}\beta_j^{'}\left(d_{ij}^{A}\text{-}d_{ij}^{B}\right)\right)$$

where $\beta_j^{'}$ is the vector of **DC** regression coefficients, d_{ij}^{A} the vector of dummy variables for state A of pair *i*, d_{ij}^{B} the vector of dummy variables for state B of pair *i*, and cdf the cumulative density function.

Finally: $Loglike_{hybrid} = Log(like_{\boldsymbol{C\text{-}TTO}} * like_{\boldsymbol{DC}}) = Loglike_{\boldsymbol{C\text{-}TTO}} + Loglike_{\boldsymbol{DC}}$, and the relation between β and $\beta^{'}$ is assumed in the estimation to be: $\theta\beta^{'} = \beta$.

REFERENCES

1. National Institute for Health and Care Excellence. *Guide to the Methodsof Technology Appraisal*. London: National Institute for Health and Care Excellence; 2013.
2. EuroQol G. EuroQol—a new facility for the measurement of health- related quality of life. *Health Policy*. 1990;16:199–208.
3. Herdman M, Gudex C, Lloyd A, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res*. 2011;20:1727–1736.
4. Bharmal M, Thomas J. Comparing the EQ-5D and the SF-6D descriptive systems to assess their ceiling effects in the US general population. *Value Health*. 2006;9:262–271.
5. van Hout B, Janssen MF, Feng YS, et al. Interim scoring for the EQ-5D- 5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. *Value Health*. 2012;15:708–715.
6. Devlin N, Krabbe P. The development of new research methods for the valuation of EQ-5D-5L. *Eur J Health Econ*. 2013;14(suppl):1–3.
7. Oppe M, Devlin NJ, van Hout B, et al. A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. *Value Health*. 2014;17:445–453
8. Dolan P. Modeling valuations for EuroQol health states. *Med Care*. 1997;35:1095–1108.
9. Shaw JW, Johnson JA, Coons SJ. US valuation of the EQ-5D health states: development and testing of the D1 valuation model. *Med Care*. 2005;43:203–220.
10. Craig BM, Oppe M. From a different angle: a novel approach to health valuation. *Soc Sci Med*. 2010;70:169–174.
11. Augustovski F, Rey-Ares L, Irazola V, et al. Lead versus lag-time trade- off variants: does it make any difference? *Eur J Health Econ*. 2013; 14(suppl 1):S25–S31.
12. Robinson A, Spencer A. Exploring challenges to TTO utilities: valuing states worse than dead. *Health Econ*. 2006;15:393–402.
13. Janssen BM, Oppe M, Versteegh MM, et al. Introducing the composite time trade-off: a test of feasibility and face validity. *Eur J Health Econ*. 2013;14(suppl 1):S5–S13.
14. Salomon J. Reconsidering the use of rankings in the valuation of health states: a model for estimating cardinal values from ordinal data. *Popul Health Metr*. 2003;1:12.
15. McCabe C, Brazier J, Gilks P, et al. Using rank data to estimate health state utility models. *J Health Econ*. 2006;25:418–431.
16. Hensher DA, Rose JM, Greene WH. *Applied Choice Analysis: A Primer*. Cambridge: Cambridge University Press; 2005.
17. Stolk EA, Oppe M, Scalone L, et al. Discrete choice modeling for the quantification of health states: the case of the EQ-5D. *Value Health*. 2010;13:1005–1013.
18. Ramos-Goni JM, Rivero-Arias O, Errea M, et al. Dealing with the health state 'dead' when using discrete choice experiments to obtain values for EQ-5D-5L heath states. *Eur J Health Econ*. 2013;14(suppl 1):33–42.

19. Bliemer MCJ, Rose JM, Hess S. Approximation of bayesian efficiency in experimental choice designs. *J Choice Model*. 2008;1:98–126.

20. Oppe M, van Hout B. The optimal hybrid: experimental design and modeling of a combination of TTO and DCE. EuroQol Group Proceedings. 2013. Available at: http://www.euroqol.org/uploads/media/EQ2010_-_CH03_-_Oppe_-_The_optimal_hybrid_-_Experimental_ design_and_modeling_of_a_combination_of_TTO_and_DCE.pdf. Accessed October 11, 2014.

21. EQ-5D value sets: inventory, comparative review and user guide. In: Szende A, Oppe M, Devlin N, eds. *EuroQol Group Monographs*. Vol. 2. Dordrecht: Springer; 2007.

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

23. Badia X, Roset M, Herdman M, et al. A comparison of United Kingdom and Spanish general population time trade-off values for EQ-5D health states. *Med Decis Making*. 2001;21:7–16.

24. Hsee CK, Dube J-P, Zhang Y. The prominence effect in Shanghai apartment prices. *J Market Res*. 2008;45:133–144.

25. Fischer GW, Hawkins SA. Strategy compatibility, scale compatibility, and the prominence effect. *J Exp Psychol*. 1993;19:580–597.

26. Tversky A, Sattath S, Slovic P. Contingent weighting in judgment and choice. *Psychol Rev*. 1988;95:371–384.

27. Carmon Z, Simonson I. Price-quality trade-offs in choice versus matching: new insights into the prominence effect. 1998. *J Consum Psychol*. 1998;7:323–343.

28. Oliver A. Further evidence of preference reversals: choice, valuation and ranking over distributions of life expectancy. *J Health Econ*. 2006; 25:803–820.

29. Sumner W II, Nease RF Jr. Choice-matching preference reversals in health outcome assessments. *Med Decis Making*. 2001;21:208–218.

30. Bleichrodt H, Pinto JL. Loss aversion and scale compatibility in two- attribute trade-offs. *J Math Psychol*. 2002;46:315–337.

31. Hawkins SA. Information processing strategies in riskless preference reversals: the prominence effect. *Organ Behav Hum Decis Proces*. 1994;59:1–26.

32. Hsee CK, Loewenstein GF, Blount S, et al. Preference reversals between joint and separate evaluations of options: a review and theoretical analysis. *Psychol Bull*. 1999;125:576–590.

3

Chapter 3

Learning and Satisficing: An Analysis of Sequence Effects in Health Valuation

Benjamin M. Craig,
Shannon K. Runge,
Kim Rand-Hendriksen,
Juan M. Ramos-Goñi,
Mark Oppe

Value Health. 2015 Mar;18(2):217-23

ABSTRACT

Objective: To estimate the effect of sequence on response precision and response behavior in health valuation studies.

Methods: Time trade-off (TTO) and paired comparison responses from six health valuation studies—four US, one Spanish, and one Dutch—were examined (22,225 respondents) to test whether task sequence influences response precision (e.g., rounding), response changes, and median response times. Each study used a computer-based instrument that randomized task sequence among a national sample of adults, age 18 years or older, from the general population.

Results: For both TTO and paired comparisons, median response times decreased with sequence (i.e., learning), but tended to flatten after the first three tasks. Although the paired comparison evidence demonstrated that sequence had no effect on response precision, the frequency of rounded TTO responses (to either 1- year or 5-year units) increased with sequence.

Conclusions: Based on these results, randomizing or reducing the number of paired comparison tasks does not appear to influence response precision; however, generalizability, practicality, and precautionary considerations remain. Overall, participants learned to respond efficiently within the first three tasks and did not resort to satisficing, but may have rounded their TTO responses.

Keywords: health valuation, para-data, preferences, QALY, response precision, sequence effects, time trade-off.

INTRODUCTION

Most economic evaluations summarize effectiveness using preference weights on a quality-adjusted life-year (QALY) scale, as recommended by numerous health technology assessment agencies. Such QALY weights may be from societal or patient perspectives and derived using a wealth of preference elicitation tasks (e.g., best-worst scaling). Although valuation research has a well-established history, the use of online computer-based surveys for health valuation offers an array of new capabilities, such as quota-sampling at the task level; para-data on respondent behaviour, device, and browser; and other interactive technologies. Compared with interview, postal, or telephone surveys, online computer-based experiments increase control in the randomization of tasks, while reducing cognitive burden and minimizing missing data and other data collection errors and biases.

Although online instruments typically randomize the order of presentation of tasks, response precision and behaviour may change with sequence. For example, when a respondent is shown two alternatives and asked, "Which do you prefer?" he or she may take longer or change his or her responses on initial pairs while becoming acquainted with the valuation task as compared with later pairs. Furthermore, a respondent's attention may wane in later pairs, leading to satisficing (i.e., expediting selection among alternatives to minimize effort), reducing response precision [1,2]. This article examines whether response precision and response behaviour vary with the number of tasks completed (i.e., sequence effect) in health valuation studies for two types of valuation tasks, time trade-off (TTO) and paired comparisons.

Understanding the relationship between response precision and task sequence guides the number of tasks to be included in a valuation study, informs weights that place a greater emphasis on earlier or later tasks, and justifies the randomization of task sequence. Although studies have attempted to identify respondents who randomize all responses (i.e., shufflers and satisficers) [3], few studies to date have examined the effect of sequence on response precision in health valuation [4].

Sequence effects have been identified in other forms of discrete choice experiments (DCEs) as a type of ordering effect specifically related to the order in which choice sets are presented (i.e., position-dependent order effects) [5]. This type of order effect differs from those related to the order or position of attributes within a choice set [5–7]. Experimental design, such as the layout of questions, the number of attributes, and the number of tasks, can influence ordering effects and response time [8–10]. A key example in survey research is the primacy effect or the tendency for respondents to choose the first reasonable answer to a survey question (e.g., first response option in a list of potential answers) [6,11]. This weak form of satisficing leads to nonrandom response; expedites response with minimum effort; reduces response quality and time; and is commonly cited by experimenters to justify randomization and reduction in the number of attributes, scenarios, and tasks [12].

A wealth of studies have examined order effects in terms of perception and salience [5,7,9,10,13–17], although the results have been somewhat inconsistent. For example, some evidence suggests that the order of attributes affects choice [5,7], yet other studies did not find this effect [9,14,18]. In addition, the number and complexity of task sets within an experiment may induce order effects through respondent fatigue or boredom [19]. Evaluating the association between participant response behaviours (i.e., response times and changes) and task sequence has the potential to provide valuable insight regarding the influence of study design.

In complement to evidence on response precision, we examine response behaviours (i.e., response times and changes) that may indicate learning and added deliberative effort beyond that which is needed to satisfy the task requirements. Typically, response behaviour is examined at the questionnaire level (e.g., the amount of time it takes a respondent to complete all tasks). In addition to evaluating response behaviour at the questionnaire level, computerized software offers a unique opportunity to examine response behaviours at the level of individual questions (e.g., the amount of time it takes to complete a single task set or a series of different task sets). A better understanding of response behaviour at each of these levels can aid in the interpretation of the empirical association between sequence and response precision and in the improvement of survey design (e.g., cognitive burden).

The present study contributes to an innovative evaluation of client-side paradata. Client-side paradata is the information recorded in Web surveys by the respondent's computer (e.g., the number of times and locations of mouse clicks on a computer screen). Unlike server-side paradata, which refers to data management processes, client-side information allows researchers to interpret participant response behaviours in terms of changed responses (CRs) and response time at the level of individual questions [20]. Evaluating response behaviour patterns at such a specific level contributes to our knowledge of how sequence influences preferences. In this secondary analysis of health valuation data, we examine sequence effects, specifically whether response precision and response behaviour vary with the number of tasks completed.

METHODS

Preference Elicitation

In a paired comparison, respondents are asked, "Which do you prefer?" given two health episodes, and their choices define the relative value between these episodes. An original TTO task is more involved, using an adaptive series of paired comparisons based on either time with no health problems or "immediate death." Specifically, each TTO begins with a paired comparison in which the respondent must first decide whether the health episode is

preferred to immediate death. If so, an adaptive series of paired comparisons is presented to determine the number of years with no health problems that is equivalent to the health episode (i.e., better-than-dead indifference statement). If the respondent prefers immediate death, an alternative series of paired comparisons is completed to identify a worse-than-dead indifferent statement. The original adaptation procedure [21–23] is like a dose-response study in that it increases the duration of problems within an episode until it is equivalent to immediate death (e.g., how much poison is needed until it kills you). Thus, the TTO exercise is a matching task that produces an equivalence statement regardless of whether the original paired comparison response is better or worse than death.

Data

To test the effect of sequence on response precision and behaviour, we examined paired comparisons and TTO responses from six health valuation studies—four US, one Spanish, and one Dutch—totalling 259,318 responses from 22,225 respondents who completed 17 to 37 tasks [2,24–27]. Table 1 summarizes the characteristics of these six studies. All studies used a computerized instrument that randomized task sequence using national samples of adults from the general population. For the US-based studies, respondents completed a set of paired comparisons trading improvements in health-related quality of life (HRQOL) for reduced lifespan (i.e., lifespan pairs) before completing a second set that traded alternative HRQOL scenarios of a common duration (i.e., health pairs). For the valuation of the five-level EuroQol five-dimensional questionnaire, respondents completed a set of TTOs before completing a set of paired comparisons that traded alternative HRQOL scenarios without a description of duration (i.e., health state pairs). Further description of the protocol of each study is provided online [2,24–27].

The TTO task in the Spanish and Dutch studies was an adaptive hierarchy of steps known as the composite TTO (Fig. 1) [27]. The composite TTO is derived from both the original and lead-time TTO [21–23]. Each step displayed two scenarios, and the respondent was asked, “Which is better?” If the respondent did not wish to choose, the respondent may instead state indifference (i.e., the scenarios were “about the same”).

In this adaptive process, the task began with the choice between 10 years in full health and 10 years in the health state (i.e., step 1). If the respondent chose the health state scenario or stated indifference, the TTO response was þ10 and the task ended.

If the respondent chose the full health scenario in step 1, the task continue on to step 2 and displayed 0 years in full health (i.e., immediate death) instead of 10 years in full health.

If the respondent chose the full health scenario in step 2, the task continued to step 3 and displayed 5 years in full health instead of 0 years in full health. If the respondent chose

the health state scenario in step 2, the task continued to step 3 and displayed 5 years in full health instead of 0 years in full health. If the respondent stated indifference in step 2, the TTO response was 0 and the task ended. This task continued for up to nine steps until the respondent expressed indifference between the two scenarios (Fig. 1).

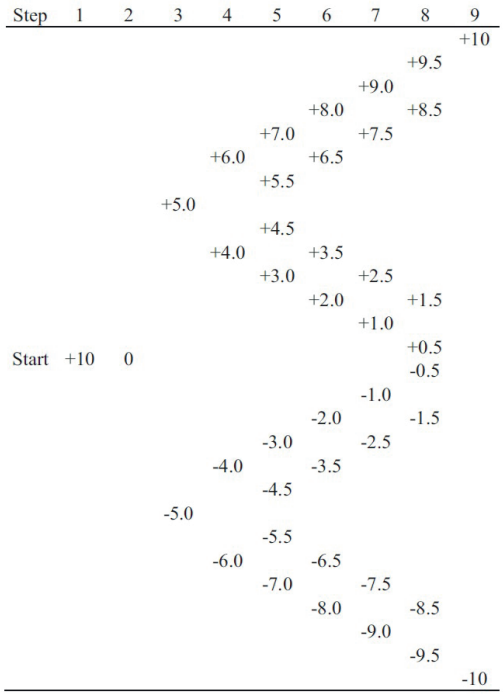

Figure 1. -Minimum number of steps involved in each composite time trade-off response. Numbers in the time trade-off represent the value of 10 years in health state on a quality-adjusted life-year (QALY) scale based on a statementof indifferences (e.g., 10 years in health state ¼þ5 QALYs).

Aside from the highest possible response (+10), which required either one or nine steps, each TTO response required a minimum number of steps (i.e., some TTO responses required more steps than did others). The lowest possible response (-10) required the most effort (i.e., nine steps). By construction, about half of any TTO sample should have been in half-year units.

Paired comparison tasks differed by the studies. The US-based paired comparisons began with three examples and asked “Which do you prefer?” showing two health scenarios with only two attributes and their durations. The Spanish and Dutch paired comparisons had no examples. Respondents completed between 7 and 37 paired comparisons. Unlike the TTO task, indifference was not allowed in any of these paired comparison tasks.

Econometrics

For each study, we graphed the median response time and the relative risk of a CR and a modal response (MR) by sequence. Response time was measured in seconds from the time that the task was first shown until the final response to the task.

A CR is when multiple responses were registered in the paradata for the task (e.g., a respondent may choose the first scenario in a paired comparison as the preferred scenario and then change his or her response to the second scenario). Changing a response may be related to the difficulty of the choice. For example, if two scenarios seemed similar, the probability of changing a response is higher than for a pair with dissimilar scenarios. Nevertheless, we hypothesize that sequence is unrelated to CRs when pairs are randomly sequenced. Specifically, the relative risk of a CR is the risk of a CR at the location in the sequence divided by the overall risk of a CR. We did, however, investigate the impact of the difficulty of the choice on the CR in a sensitivity analysis.

To identify half-year unit responses in the TTO tasks, respondents may be required to complete additional steps to achieve the final response. These steps include overshooting the point of indifference by half a year and backtracking half a year. For example, for a respondent to achieve a final TTO of 6.5, he or she would first be presented with additional scenarios comparing 7 years in a health state (overshooting) and 6 years in a health state (backtracking). Therefore, a TTO CR requires added steps and responses, and a DCE CR implies just added responses. In either case, we hypothesize that sequence is unrelated to the relative risk of CR.

An MR is whether the respondent provided the same response as the modal response for the task. For example, in a choice between mild pain and mild depression, 80% may choose mild pain and this MR should not vary by sequence. If respondent attention waned, however, the frequency of MRs should diminish until just 50% prefer mild pain. Specifically, the relative risk of an MR is the risk of an MR at the location in the sequence divided by the overall risk of an MR.

For a TTO task, the responses are not binary but are integer and half-integer values on a scale ranging from þ10 to 10. Therefore, the risk of a TTO MR may be lower than a risk of a DCE MR. In either case, we hypothesize that sequence is unrelated to the relative risk of MR, the relative risk of CR, or median response times.

As ancillary measures of TTO response precision, we illustrated the frequency of 5-year and half-year unit TTO responses by sequence. A half-year unit response requires that the respondent complete at least one more step than a 1-year unit response. The frequency of half-unit responses represented a trade-off between added effort and greater precision, which may have varied by sequence. Likewise, a respondent may have stopped the task early (i.e., within three steps: þ10, 0, þ5 or 5) and responded in 5-year units. Rounding to 1-year or to 5-year units was a tacit way to avoid added effort in the TTO task (i.e., satisficing).

All analyses were repeated using varying levels of difficulty (i.e., comparing different levels of severe health states) on the basis of the assumption that decision difficulty increases as respondents compare health scenarios with similar levels of severity. For the TTO tasks, decision difficulty is assumed to peak at the point of respondent indifference between health scenarios. For the DCE tasks, this point occurs when the choice probability of two health scenarios is approximately 50%. Subsequently, we used posterior information about DCE pair probabilities to describe subgroups.

RESULTS

Figure 2 illustrates median response times by sequence. At the beginning of each sequence, response times were reduced substantially. Each line exhibits the same downward sloping shape (i.e., learning) and shows a flattening out. Dutch respondents had a higher median time than did Spanish and US respondents, regardless of task. Spanish paired comparisons had a higher response time than US tasks, possibly due to differences in the number of attributes of each alternative (5 vs. 2). This pattern was also observed in the subgroup analysis, which confirmed that more time was needed when the task was more difficult. We examined, however, whether sequence effects (i.e., median response times, CR, MR, and rounding) were similar among tasks with different levels of difficulty (e.g., greater effect seen in easier tasks) and found no differences.

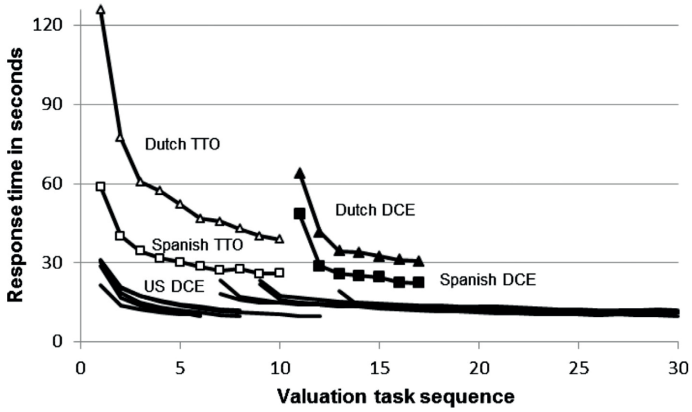

Figure 2. - Median response time by sequence. DCE, discrete choice experiment; TTO, time trade-off.

Figure 3 illustrates the relative risk of CR by sequence, which decreases over the initial tasks. Figure 4 illustrates the relative risk of MR by sequence, and the MR lines appear flat (i.e., relative risks range from 1.1 to 0.9) aside from some wavering.

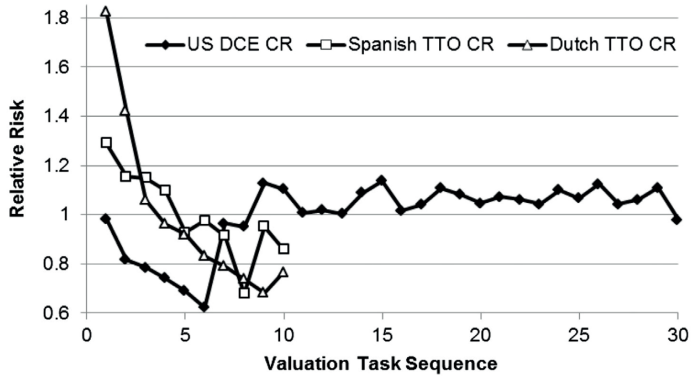

Figure 3. - Relative risk of changed responses (CRs) by sequence. The Dutch and Spanish studies did not collect CR data on paired comparisons. DCE, discrete choice experiment; TTO, time trade-off.

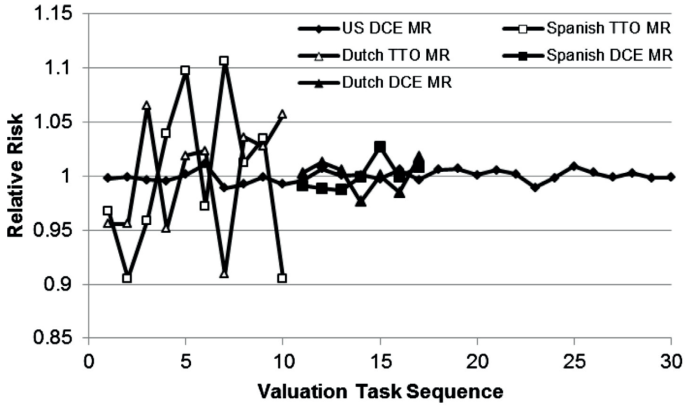

Figure 4. - Relative risk of modal response (MR) by sequence. DCE, discrete choice experiment; TTO, time trade-off.

Unlike the paired comparison responses, TTO responses may be rounded to 1-year or 5-year units, possibly to reduce response effort (Fig. 1). Figure 5 illustrates the frequency of 5-year, 1-year, and half- year unit TTO responses. The results show that more than 40% of the Spanish TTO responses were either +10, +5, 0, or -5, regardless of sequence, and that the frequency of these 5-year unit responses increased from 30% to 40% in the Dutch data, representing a reduction in TTO response precision with sequence. Half-year unit responses potentially indicated a small gain in precision and should be half of each sample. The frequencies of half-year unit responses were clearly less than 50% and decreased from 19% to 12% and from 14% to 12% in the Dutch and Spanish samples, respectively. Furthermore, all

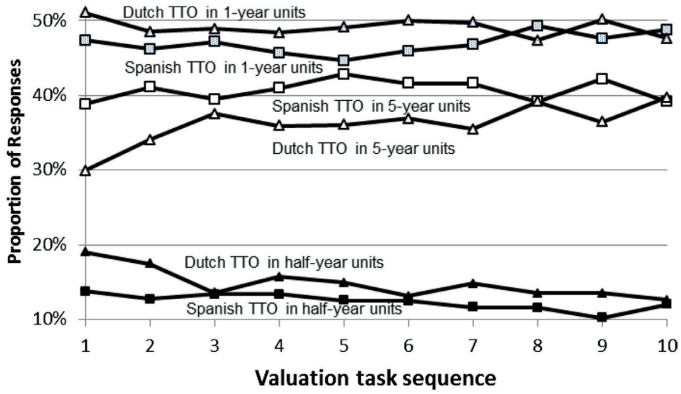

Figure 5. - TTO rounding by sequence. TTO, time trade-off.

86 modal TTO responses in the Dutch and Spanish studies were in 1-year units and most (77% and 87%, respectively) were in 5-year units. It should be noted, however, that even though the proportion of 5-year values and 1-year values is large across respondents, only a small number of respondents give only 5-year values (2% and 36% in the Dutch study and 6% and 47% in the Spanish study).

DISCUSSION

Using data from 22,225 respondents, we found that sequence had no effect on paired comparison response precision, but may induce greater rounding in TTO responses. The CR lines (Fig. 3) decrease over the initial tasks, illustrating that those respondents may be learning the task or establishing heuristics that govern their responses of all similar tasks. The first six tasks for each US study were lifespan pairs that involved the trade-off between reduced lifespan and HRQOL. This emphasis on a single attribute (i.e., lifespan) may have induced the formation of time-specific heuristics compared with latter pairs that traded two losses in HRQOL with common duration. Aside from some wavering, the MR lines (Fig. 4) appear flat (i.e., relative risks range from 1.1 to 0.9), illustrating that response precision was not associated with sequence. The greater variability seen in the TTO MR is likely attributable to its use of non-binary responses.

With TTO, it can be argued that the proportion of half-year responses, theoretically, should be similar to integer-year responses (1- or 5-year units), given the assumption that the distribution of preferences could be considered continuous. The results show that the proportion of half-year responses was less than half and decreased with sequence, although at a different rate in the Spanish data than in the Dutch data. Nevertheless, such TTO rounding had

no effect on the relative risk of a modal TTO response. This absence of effect may be attributed to the fact that most modal TTO responses are in 5-year units (i.e., rounding increases the likelihood of MR).

The low and falling proportion of half-year unit TTO responses is striking, but the correct interpretation is not straightforward. The procedure used to identify a half-year unit response requires overshooting the point of indifference and backtracking half a year. For example, a respondent who has a TTO value of 6.5 for a health scenario would be offered 10 years of perfect health, followed by 0, 5, 6, and 7 (overshoot) before stating indifference at 6.5 years. Similarly, a respondent who has a TTO value of 3.5 for a health scenario is offered 10 years of full health, followed by 0, 5, 4, 3 (overshoot) before stating indifference at 3.5 years. The reduction in elicited half-year unit response could represent satisficing, but it could also reflect a reluctance to backtrack, reluctance to overshoot, or a genuine satisfaction with the level of precision offered by sticking to whole years. Which of these explanations is at play could possibly be determined through strategic manipulation of the routing, such as removing the half-year correction, altering the step size to half a year, or giving respondents multiple alternatives (i.e., more than two scenarios in a choice set) at each step. Regardless of explanation, the results show that sequence influences the frequency of half- year responses; however, the infrequency of half-year responses suggests that the potential loss of information is limited.

The apparent and increasing frequency of 5-year unit responses is more troubling because the loss of information is large. The results suggest that most of the respondents are attracted to these 5-year unit responses, increasing the risk of bias. The extent of these primacy effects and their attraction may be caused by digit preference, satisficing, or cognitive biases, such as anchoring, and should be investigated further. Based on the paired comparison results, randomizing or reducing the number of paired comparison tasks does not appear to influence response precision; however, generalizability, practicality, and precautionary considerations remain.

These considerations are largely related to the design of DCEs: What is the optimal number of tasks that should be included in a survey? Should later tasks be down weighted? Should tasks be randomized? It has been proposed that certain variations in survey design (e.g., increases in the number of tasks, scenarios, and attributes) increase respondent burden and fatigue, thus contributing to ordering effects and response variability [10,19,28]. Despite a growing interest in identifying the optimal design for DCEs, the existing literature remains inconclusive and the results of this study failed to identify any benefits from decreasing the number of tasks, down weighting later tasks, or randomizing tasks.

Shortening a health preference survey may limit the breadth of the results (e.g., too few attributes) and collect insufficient data to calculate preferences on attributes, particularly

if sample size is small [18,29]. In their widely cited article, Hensher et al. [18] found 4 and 8 tasks to be insufficient to estimate preferences for attributes that were selected less often but concluded that this could remedied by presenting 24 to 32 tasks without overburdening respondents. Similarly, Carlsson and Martinsson [29] compared the results of 12 and 24 tasks and found no evidence of sequence effects, but they did report a significantly higher dropout rate for the longer survey. The results of these studies, however, contradict other findings. In a valuation of travel time, Hensher [28] reported that increasing the number of tasks significantly decreased participant response time and significantly affected the outcome of the study. These results were echoed by Chung et al. [30], who concluded the ideal number of tasks to be six per survey. Although it has been noted that researchers should use careful pretesting to identify the optimal number of tasks to include in a DCE [30], our results did not find any sequence effects in the DCE, possibly due to their simplicity (two alternatives with two attributes). Still, additional research is needed to rectify these discrepancies.

The primary limitation of this study is that each study included a maximum of 37 tasks because these components were designed to be completed in less than 30 minutes. Evidence, however, from health valuation studies with more than 40 tasks will be explored in future work. In fact, Craig et al. are currently in the beginning stages of a study that will allow respondents to complete hundreds of pairs. Our sensitivity analyses on the time it takes to complete a task by difficulty indicated, however, that the time needed to answer a task is shorter for easy tasks than for difficult tasks. This should be taken into account in the design of a study.

Another limitation of the present study is that trends in the relative risk of MR may underrepresent losses in TTO precision due to rounding, because most TTO MR are in 5-year units. The use of MR allowed for a uniform summary of trends in TTO and paired comparison response precision, but did not compensate rounding. The proportion of 5-year units and 1-year units is quite large across respondents. Only a few respondents, however, use only 5-year responses or 1-year responses. Future studies may investigate whether rounding is a greater concern in subgroups of respondents, particularly those with low numeracy. The conclusion from this analysis is that sequence effects are present more in TTOs than in DCEs, but both show some learning effect. In summary, the results of this study failed to identify any benefits from decreasing the number of DCE tasks, down-weighting later DCE tasks, or randomizing DCE tasks.

Acknowledgments

We thank Carol Templeton and Michelle Owens at Moffitt Cancer Center for their contributions to the research and creation of this article.

Source of financial support: Funding support for this research was provided by a National Cancer Institute R01 grant (1R01CA160104), the EuroQol Group (EQ Project 2013130), and Dr. Craig's support account at Moffitt Cancer Center.

REFERENCES

10. Barge S, Gehlbach H. Using the theory of satisficing to evaluate the quality of survey data. Res High Educ 2012;53:182–200.
11. Craig B, Reeve BB. Methods Report on the Promis Valuation Study: Year 1. Available from: http://labpages.moffitt.org/craigb/Publications/Report120928.pdf. [Accessed October 11, 2012].
12. Craig BM, Ramachandran S. Relative risk of a shuffled deck: a generalizable logical consistency criterion for sample selection in health state valuation studies. Health Econ 2006;15:835–48.
13. Augestad LA, Rand-Hendriksen K, Kristiansen IS, et al. Learning effects in time trade-off based valuation of EQ-5D health states. Value Health 2012;15:340–5.
14. Day B, Bateman IJ, Carson RT, et al. Ordering effects and choice set awareness in repeat-response stated preference studies. J Environ Econ Manage 2012;63:73–91.
15. Malhotra N. Completion time and response order effects in web surveys. Public Opin Q 2008;72:914–34.
16. Kjaer T, Bech M, Gyrd-Hansen D, et al. Ordering effect and price sensitivity in discrete choice experiments: need we worry? Health Econ 2006;15:1217–28.
17. Christian LM, Parsons NL, Dillman DA. Designing scalar questions for web surveys. Sociol Methods Res 2009;37:393–425.
18. Farrar S, Ryan M. Response-ordering effects: a methodological issue in conjoint analysis. Health Economics 1999;8:75–9.
19. Savage SJ, Waldman DM. Learning and fatigue during choice experiments: a comparison of online and mail survey modes. J Appl Econ 2008;23:351–71.
20. Krosnick JA. Response strategies for coping with the cognitive demands of attitude measures in surveys. Appl Cogn Psychol 1991;5:213–36.
21. Schwarz N, Sudman S, Schuman H. Context Effects in Social and Psychological Research. New York, NY: Springer-Verlag, 1992.
22. Blumenschein K, Johannesson M. An experimental test of question framing in health state utility assessment. Health Policy 1998;45:187–93.
23. Boyle KJ, Ozdemir S. Convergent validity of attribute-based, choice questions in stated-preference studies. Environ Resour Econ 2009;42:247–64.
24. Howard K, Salkeld G. Does attribute framing in discrete choice experiments influence willingness to pay? Results from a discrete choice experiment in screening for colorectal cancer. Value Health 2009;12:354–63.
25. Kamoen N, Holleman B, Mak P, et al. Agree or disagree? Cognitive processes in answering contrastive survey questions. Discl Process 2011;48:355–85.
26. Yan T, Tourangeau R. Fast times and easy questions: the effects of age, experience and question complexity on web survey response times. Appl Cogn Psychol 2008;22:51–68.

27. Hensher DA, Stopher PR, Louviere JJ. An exploratory analysis of the effect of numbers of choice sets in designed choice experiments: an airline choice application. J Air Trans Manag 2001;7:373–9.

28. De Palma A, Myers GM, Papageorgiou YY. Rational choice under animperfect ability to choose. Am Econ Rev 1994;84:419–40.

29. Heerwegh D. Explaining response latencies and changing answers using client-side paradata from a web survey. Soc Sci Comp Rev 2003;21:360–73.

30. Devlin NJ, Tsuchiya A, Buckingham K, et al. A uniform time trade off method for states better and worse than dead: feasibility study of the 'Lead Time' approach. Health Econ 2011;20:348–61.

31. Gudex C. Time Trade-Off User Manual: Props and Self-Completion Methods. Report of the Centre for Health Economics. York, United Kingdom: University of York, 1994.

32. Torrance GW, Thomas WH, Sackett DL. A utility maximization model for evaluation of health care programs. Health Serv Res 1972;7:118–33.

33. Craig B, Owens MA. Methods Report on the Child Health Valuation Study (CHV): Year 1. Available from: http://labpages.moffitt.org/craigb/Publications/CHVMethods_130917.pdf. [Accessed November 5, 2013].

34. Craig B, Owens MA. Methods Report on the Women's Health Valuation Study (WHV): Year 1. Available from: http://labpages.moffitt.org/craigb/ Publications/WHVMethods_140106.pdf. [Accessed January 10, 2014].

35. Craig B, Owens MA. 2013 United States Measurement and Valuation of Health Study (2013 US MVH): Methods Report. Available from: http://labpages.moffitt.org/craigb/Publications/ MVHMethods_140116. pdf. [Accessed January 27, 2014].

36. Janssen BMF, Oppe M, Versteegh MRN, et al. Introducing the composite time trade-off: a test of feasibility and face validity. Eur J Health Econ 2013;14(Suppl):S5–13.

37. Hensher DA. Revealing differences in willingness to pay due to the dimensionality of stated choice designs: an initial assessment. Environ Resour Econ 2006;34:7–44.

38. Carlsson F, Martinsson P. How much is too much? Environ Resour Econ 2008;40:165–76.

39. Chung C, Boyer T, Han S. How many choice sets and alternatives are optimal? Consistency in choice experiments. Agribusiness 2011;27:114–25.

Note: Juan M. Ramos-Goñi has contributed to this paper on the analysis, results interpretation and paper writing.

4

Chapter 4

Does the Introduction of the Ranking Task in Valuation Studies Improve Data Quality and Reduce Inconsistencies? The Case of the EQ-5D-5L

Juan M. Ramos-Goñi,
Kim Rand-Hendriksen,
Jose Luis Pinto-Prades

Value Health. 2016 Jun;19(4):478-86

ABSTRACT

Background: Time trade-off (TTO)-based valuation studies for the three-level version of the EuroQol five-dimensional questionnaire (EQ-5D) typically started off with a ranking task (ordering the health states by preference). This was not included in the protocol for the five-level EQ-5D (EQ-5D-5L) valuation study.

Objectives: To test whether reintroducing a ranking task before the composite TTO (C-TTO) could help to reduce inconsistencies in C-TTO responses and improve the data quality.

Methods: Respondents were randomly assigned to three study arms. The control arm was the present EQ-5D- 5L study protocol, without ranking. The second arm (ranking without sorting) preceded the present protocol by asking respondents to rank the target health states using physical cards. The states were then valued in random order using C-TTO. In the third arm (ranking

and sorting), the ranked states remained visible through the C-TTO tasks and the order of valuation was determined by the ranking.

The study used only 10 EQ-5D-5L health states. We compared the C-TTO–based inconsistent pairs of health states and ties.

Results: The final sample size was 196 in the control arm, 205 in the ranking without sorting arm, and 199 in the ranking and sorting arm. The percentages of ties by respondents were 15.1%, 12.5%, and 12.6% for the control arm, the ranking without sorting arm, and the ranking and sorting arm, respectively. The extra cost for adding the ranking task was about 15%.

Conclusions: The benefit does not justify the effort involved in the ranking task. For this reason, the addition of the ranking task to the present EQ-5D-5L valuation protocol is not an attractive option.

Keywords: health related quality of life, ranking, EQ-5D-5L, valuation, time trade off.

INTRODUCTION

The EuroQol five-dimensional questionnaire (EQ-5D) is the most used instrument for measuring health-related quality of life to estimate quality-adjusted life-years [1]. It is short and simple, and the EQ-5D country-specific value sets allow researchers to obtain an index value (utility) for each of the health states described by the instrument [2]. The original version of the EQ-5D is composed of five dimensions (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression) and three levels of response for each dimension (roughly corresponding to no problems, some problems, and unable/extreme problems, respectively) [3]. To improve the sensitivity of the instrument and reduce ceiling effect problems in the descriptive system, the EuroQol Group has developed a new version of the instrument, in which each of the five dimensions is described at five levels (corresponding to no problems, slight problems, moderate problems, severe problems, and unable/ extreme problems) [4]. To distinguish the two versions, the old three-level version of the instrument is now referred to as EQ-5D-3L.

Because the five-level version of the EQ-5D (EQ-5D-5L) is essentially a new instrument, new valuation studies are needed to obtain country-specific value sets. The EuroQol Group conducted a set of pilot studies to develop a valuation study protocol for EQ-5D-5L valuation studies [5]. Version 1.0 of the protocol uses composite time trade-off (C-TTO) [6], which uses the traditional time trade-off (TTO) (see detailed description elsewhere [7]) for states considered to be better than dead and the lead-time TTO (see detailed description elsewhere [8]) for states considered to be worse than dead. A discrete choice experiment (DCE) was included [9]. The new protocol is structured into three main sections: 1) general introduction and background questions, 2) an example of being in a wheelchair to explain the C-TTO task followed by 10 C-TTO tasks, and 3) seven pairs of comparison tasks on a DCE. Data collection for the EQ-5D-5L valuation study has been recently performed in countries such as Spain and the Netherlands, using the computer-based EuroQol valuation technology (EQ-VT) instead of the classical pen and paper approach used in old EQ-5D-3L valuation studies.

The use of version 1.0 of the protocol to conduct EQ-VT–based valuation studies showed several issues related to the quality of the resulting data [10, 11]. The most obvious of these was a relatively high rate of logically inconsistent responses, a formal definition of which can be found in the Methods section. Briefly, consider any pair of health states for which one state is descriptively superior (better on at least one dimension, and no worse on any other). A logical inconsistency occurs when values are assigned to the two states, indicating that the respondent would prefer the worse state to the better.

Given the complexity of TTO-based valuation studies, in previous EQ-5D-3L valuation studies, TTO tasks were preceded by a ranking task and valuation using a visual analogue scale (VAS) [7, 12]. The ranking task allows respondents to make comparisons between different health states, and not merely make a comparison with full health. It also serves to familiarize

respondents with the health states and with comparisons between health states. The tacit assumption was that these tasks would help study participants when performing the more challenging TTO tasks. Nevertheless, although several studies have compared TTO values with VAS and with rankings, the potential benefits of preceding TTO tasks by a ranking task have not been examined directly. Considering the relatively high rate of logically inconsistent responses in the EQ-5D-5L valuation studies using version 1.0 of the protocol, in which no such warm-up tasks were used, we hypothesized that having the participants perform a ranking task before the TTO tasks could improve their understanding of the task and result in less inconsistent data. The objective of this study was to evaluate whether including a ranking task before the C-TTO tasks could reduce the proportion of inconsistencies and improve the quality of the C-TTO data in EQ-5D-5L valuation studies.

METHODS

Design

Respondents were randomized into three arms:

1. Control arm: The interview in the control arm used version 1.1 of the protocol. It asked each respondent to describe their own health using the EQ-5D-5L and the EuroQol VAS: basic background questions (age, sex, and experience with severe illness); a wheelchair task as an illustrative example explaining the working of the C-TTO task; 3 C-TTO practice states; 10 C-TTO tasks in which the health states to be valued were presented in random order; a feedback module; 3 structured questions regarding the subjective difficulty of C-TTO tasks; 13 paired comparison tasks from a DCE, composed of the 7 included in version 1.0 of the protocol [9] plus 6 experimental (not reported here); and 3 structured questions regarding the difficulties of the DCE (not reported here). Notice that version 1.1 of the protocol differs from version 1.0 only in the inclusion of the three practice states and the experimental DCE pair comparisons. The EQ-VT also incorporates confirmatory pop-ups for each C-TTO task in version 1.1, whereas these were not present in version 1.0.

2. Ranking without sorting: This study arm contained all the elements of the control arm. Nevertheless, description cards were available for each of the 10 health states to be valued, plus a card for "dead." Immediately after the background questions and before the C-TTO tasks, the interviewers provided the set of cards and instructed respondents to arrange the health states in an order corresponding to how good or bad the states were perceived to be. Ties were allowed. When respondents had completed the ranking task, the cards were collected and removed from view. No further mention of the ranking task was made during the interview and the remaining tasks were performed as in the control arm.

3. Ranking and sorting: The final study arm differed from the ranking without sorting arm in two ways: 1) when the ranking was complete, the cards remained conserving the ranking order visible on the table during the C-TTO valuations, and (2) instead of the random order

for health state presentation used in the control arm or the ranking without sorting arm, the order of presentation was determined by the ranking. The order of presentation was based on a procedure used in most EQ-5D-3L valuation studies (Dolan and Shaw): 1) top state, 2) bottom state, 3) state indicated as roughly midway between top and bottom states, 4) state indicated as roughly midway between top and middle states, 5) state indicated as roughly midway between middle and bottom states, and 6) the remaining states, one by one. When a state was shown on the screen, the interviewer referred to the card representing that state on the ranking on the table, such as "this is the state you are valuing," and then the C-TTO task was performed. After the C-TTO tasks, the remaining tasks were performed as in the other arms.

Health States

Because this was a methodological study, all respondents valued the same 10 health sates selected from the 86 health states included in the EQ-5D-5L valuation protocol [9]. In this protocol, each respondent is assigned to 1 of 10 health state blocks. All 10 health state blocks contain state 55555 (the worst state), one very mild health state (11112, 11121, 11211, 12111, or 21111), and 8 other health states selected such that the different blocks should have roughly equal overall severity. For this experimental study, we selected the health state block from the standard protocol that contained the largest number of health state pairs in a logical dominance relationship. The block in question contained the following health states: 12111, 11122, 42321, 13224, 35311, 34232, 52335, 24445, 43555, and 55555. All the participants were administered the same set of health states.

Sample and Data Collection

Because of the methodological characteristics of the research study, a convenience sample from a panel plus media advertisements was recruited in Canary Islands, Spain, during March 2014. The software randomly assigned 600 respondents to the three experimental arms (200 in each arm) and all interviews were performed face to face using the EQ-VT. Interviewers guided respondents throughout the interview to explain each task. All interviewers followed a specific interviewer script developed to harmonize interviewer behaviour and reduce potential interviewer effects. The interviewer script mainly instructed interviewers about how to explain the C-TTO task and explained when they should intervene, for example, when they observed a misunderstanding of the task, or how to explain the transition between the better than dead and worse than dead health states. Participants were invited to be interviewed at a central location and a small monetary incentive (€10) was offered to increase the participation rate. An agency was contracted to conduct the interviews, handle logistics, and recruit interviewers. The interviewers were, however, trained and supervised directly by the study principal investigator (PI) to ensure strict adherence to the pre-specified quality criteria on the basis of protocol compliance.

Sample size calculations were based on the rate of state 55555 not being valued as the worst state in the C-TTO tasks ($rate_{55555}$). We assumed an anticipated $rate_{55555}$ of 13%, a 95% confidence interval, and 80% power to allow us to detect a difference of 8% or more between the control and experimental arms.

Quality Control

To ascertain consistent interview performance and protocol compliance, we used a new quality control (QC) tool developed by the EuroQol Group to monitor a number of key parameters during data collection. These parameters were as follows: 1) a full explanation of the C-TTO task in the wheelchair example; 2) a minimum of 3 minutes spent explaining the C-TTO task in the wheelchair example; 3) severe inconsistencies, that is, a respondent valuing the worst EQ-5D-5L state (55555) at least 0.5 higher than any other state; and 4) a minimum of 5 minutes spent performing the 10 C-TTO tasks. Interviews were automatically flagged if any of these criteria were met. If more than 40% of the first 10 interviews performed by any interviewer were flagged, the interviewer in question was brought in for retraining. The interviewers received QC reports daily, together with feedback from the interview manager, after discussing the QC report with the study PI.

Interviewers and Interviewer Training

Six interviewers plus an interview manager conducted all the interviews. The interviewers received 3 days of training from the PI. All seven interviewers attended the first day together, during which the interview script was explained. The second day of training was dedicated to conducting 10 interviews each to identify script misunderstandings, problems, and so on, and the PI answered any queries by telephone. The third day of training was scheduled after the 10 pilot interviews. Two sessions were administered that day: the first was a roundtable discussion to share interviewers' experiences, comments, and uncertainties, and the second was to present the QC reports to the interviewers to aid in their interpretation and understanding of the interview protocol and the importance of following it.

Statistical Analysis

Descriptive analysis was performed for sample characteristics, C-TTO values, and time required for each task. Means and SDs were reported for continuous variables, and percentages were reported for discrete variables. We used box plots to represent the similarities and dissimilarities among observed values by health state and study arm.

Analysis of variance was used to explore mean differences by health state across study arms. Comparison of variance was tested by Levene's robust test because of the non-normality of observed values. Kruskal-Wallis test was used when Levene test was significant, that is, when groups appeared to be heteroscedastic. We applied Bonferroni correction for all post hoc comparisons.

We define a logical dominance relationship between two health states as follows: state A dominates state B when state A is better than state B in at least one dimension and no worse than state B in any remaining dimension. Among the 10 health states used in this study, 25 of the 45 possible pairs of health states were in a logical dominance relationship. Elicited values were considered logically inconsistent if a state (B) was assigned a value indicating that it was better than a dominating state (A). All pairs of states for which a logical dominance relationship existed were considered and inconsistencies were counted. Ties were counted separately. The number of inconsistencies for a given health state A was calculated as the sum of inconsistencies of states dominated by A plus the sum of inconsistencies of states that dominated A. Because each inconsistent pair involves two health states, the total number of possible inconsistencies is half the sum of inconsistencies by state. In the 10 C-TTO health states included, there were 25 logical dominance relationships, allowing us to compare 25 possible inconsistent pairs of health states for each respondent. For example, the state 12111 dominated all other states except state 11122, having eight possible inconsistent pairs of health states involving 12111. The state 55555 was dominated by all other states, thereby being involved in nine logical dominance relationship pairs. The proportion of inconsistent pairs of health states was calculated by counting the number of inconsistent pairs of health states and dividing this by the total number of possible inconsistent pairs of health states. For example, for the worst EQ-5D-5L health state (55555), we counted the number of inconsistent pairs for each respondent and divided the result by 9 (possible inconsistent pairs involving 55555) arm sample size. A similar analysis was performed using the position of the states in the ranking task instead of the C-TTO values. Because of technical problems, the information about ranking position in the ranking without sorting arm was lost, and so the analysis of inconsistencies and ties using ranking positions was applied only to the ranking and sorting arm.

Crude estimates of the extra cost associated with the introduction of the ranking task were generated on the basis of the assumption of a linear relationship between the interview duration and the cost for each single interview. The extra cost was calculated on the basis of the increment in the interview duration. That is, we calculated the cost for each minute of an interview, and then on the basis of the average extra time for the ranking, we estimated the extra cost of interviews in the ranking arms relative to the cost of the control arm, considering only data collection cost. Costs not related to data collection were not estimated.

RESULTS

The final sample size was 600 respondents. Of these, 196, 205, and 199 respondents were assigned to the control arm, the ranking without sorting arm, and the ranking and sorting arm, respectively. Slight, but statistically significant, differences in terms of age and self-reported VAS were observed between the study arms (Table 1).

Table 1.- Sample characteristics by study arm

Variables	Control Arm (n =196)		Ranking no sorting arm (n = 205)		Ranking and sorting arm (n = 199)	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
Age	43.96[§]	16.12	41.3[§,t]	13.89	42.81[t]	15.32
	n	**%**	**n**	**%**	**n**	**%**
Age groups						
18-24	19	9.7%	27	13.2%	23	11.6%
25-34	45	23.0%	47	22.9%	46	23.1%
35-44	47	24.0%	51	24.9%	40	20.1%
45-54	34	17.3%	36	17.6%	43	21.6%
55-64	21	10.7%	33	16.1%	26	13.1%
65-74	19	9.7%	9	4.4%	17	8.5%
75+	11	5.6%	2	1.0%	4	2.0%
Gender						
- Male	107	54.6%	108	52.7%	99	49.7%
- Female	89	45.4%	97	47.3%	100	50.3%
Experience with illness						
- Personal (%YES)	48	24.5%	38	18.5%	41	20.6%
- Relatives (%YES)	148	75.5%	140	68.3%	148	74.4%
- Other (%YES)	90	45.9%	102	49.8%	93	46.7%
Self-reported EQ-5D-5L						
Mobility						
- No problems	152	77.6%	176	85.9%	168	84.4%
- Slight problems	31	15.8%	20	9.8%	14	7.0%
- Moderate problems	12	6.1%	6	2.9%	17	8.5%
- Severe problems	1	0.5%	3	1.5%	0	0.0%
- Unable/extreme problems	0	0.0%	0	0.0%	0	0.0%
Self-care						
- No problems	190	96.9%	198	96.6%	186	93.5%
- Slight problems	3	1.5%	7	3.4%	11	5.5%
- Moderate problems	3	1.5%	0	0.0%	1	0.5%
- Severe problems	0	0.0%	0	0.0%	1	0.5%
- Unable/extreme problems	0	0.0%	0	0.0%	0	0.0%
Usual activities						
- No problems	167	85.2%	180	87.8%	171	85.9%
- Slight problems	21	10.7%	20	9.8%	15	7.5%
- Moderate problems	8	4.1%	2	1.0%	9	4.5%
- Severe problems	0	0.0%	3	1.5%	1	0.5%

Variables	Control Arm (n =196)		Ranking no sorting arm (n = 205)		Ranking and sorting arm (n = 199)	
- Unable/extreme problems	0	0.0%	0	0.0%	3	1.5%
Pain/Discomfort						
- No problems	122	62.2%	134	65.4%	121	60.8%
- Slight problems	56	28.6%	53	25.9%	53	26.6%
- Moderate problems	14	7.1%	15	7.3%	21	10.6%
- Severe problems	4	2.0%	3	1.5%	4	2.0%
- Unable/extreme problems	0	0.0%	0	0.0%	0	0.0%
Anxiety/Depression						
- No problems	147	75.0%	152	74.1%	151	75.9%
- Slight problems	35	17.9%	42	20.5%	38	19.1%
- Moderate problems	12	6.1%	8	3.9%	8	4.0%
- Severe problems	2	1.0%	3	1.5%	2	1.0%
- Unable/extreme problems	0	0.0%	0	0.0%	0	0.0%
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
VAS (mean and SD)	77.82[§,ł]	17.27	80.22[§]	12.85	79.42[ł]	16.43

§ Differences statistically significant between control and ranking without sorting arms
[ł] Differences statistically significant between control and ranking and sorting arms
ŧ Differences statistically significant between ranking without sorting and ranking and sorting arms

The QC analysis indicated very low and statistically non-significant percentages of flagged interviews of 5%, 6%, and 9% in the control arm, the ranking without sorting arm, and the ranking and sorting arm, respectively. None of the interviewers reached the preassigned threshold of 40% flagged interviews that would indicate need for retraining.

The percentage of interviews in which the health state 55555 was not valued as the worst state in C-TTO tasks was 13.27%, 10.73%, and 13.57% in the control arm, the ranking without sorting arm, and the ranking and sorting arm, respectively. The proportion of health state pairs for which inconsistent values were assigned was 1.8%, 1.7%, and 2.2% in the control arm, the ranking without sorting arm, and the ranking and sorting arm, respectively (Table 2).

None of these differences was statistically significant. Nevertheless, when we examined the ties in the C-TTO values, we observed a statistically significant difference (P value 00.002) in total ties between the control arm (15.1%) and the ranking without sorting arm (12.5%) and the ranking and sorting arm (12.6%), as shown in Table 3.

Table 2. Inconsistencies (within respondent) by pairs of health states

	Profiles										
Control Arm		**12111**	**11122**	**42321**	**13224**	**35311**	**34232**	**52335**	**24445**	**43555**	**55555**	**Inconsistencies by state**
	12111	-										0.2 % (9)
	11122	-	-									0.1 % (6)
	42321	3	-	-								0.4 % (21)
	13224	3	2	-	-							0.4 % (21)
	35311	1	-	-	-	-						0.1 % (4)
	34232	2	2	-	-	-	-					0.1 % (5)
	52335	0	2	14	-	-	-	-				0.5 % (24)
	24445	0	0	-	8	-	-	-	-			0.4 % (21)
	43555	0	0	4	6	-	-	-	-	-		0.5 % (24)
	55555	0	0	0	2	3	1	8	13	14	-	0.8 % (41)
							Total					1.8 % (88)
Ranking no sorting Arm		**12111**	**11122**	**42321**	**13224**	**35311**	**34232**	**52335**	**24445**	**43555**	**55555**	**Inconsistencies by state**
	12111	-										0.2 % (10)
	11122	-	-									0.1 % (5)
	42321	2	-	-								0.6 % (30)
	13224	3	3	-	-							0.3 % (17)
	35311	3	-	-	-	-						0.1 % (5)
	34232	1	1	-	-	-	-					0.1 % (4)
	52335	1	0	18	-	-	-	-				0.5 % (27)
	24445	0	0	-	9	-	-	-	-			0.4 % (18)
	43555	0	0	6	1	-	-	-	-	-		0.4 % (18)
	55555	0	1	4	1	2	2	8	9	11	-	0.7 % (38)
							Total					**1.7 % (86)**
Ranking and sorting Arm		**12111**	**11122**	**42321**	**13224**	**35311**	**34232**	**52335**	**24445**	**43555**	**55555**	**Inconsistencies by state**
	12111	-										0.5 % (27)
	11122	-	-									0.3 % (15)
	42321	3	-	-								0.5 % (24)
	13224	9	5	-	-							0.4 % (22)
	35311	9	-	-	-	-						0.2 % (11)
	34232	2	6	-	-	-	-					0.2 % (10)
	52335	1	1	17	-	-	-	-				0.5 % (25)
	24445	1	1	-	4	-	-	-	-			0.3 % (13)
	43555	1	1	3	3	-	-	-	-	-		0.6 % (30)
	55555	1	1	1	1	2	2	6	7	22	-	0.9 % (43)
							Total					2.2 % (110)

*Numbers are absolute frequency by pair (row, column)

**Inconsistencies for a specific health state are calculated summing all inconsistent pairs that involve the specific health state. Proportions are calculated over the total number of possible inconsistencies given the sample size.

Table 3. Ties (within respondent) by pairs of health states

		Profiles										
Control Arm		**12111**	**11122**	**42321**	**13224**	**35311**	**34232**	**52335**	**24445**	**43555**	**55555**	**Inconsistencies by state**
	12111	-										3.8 % (188)
	11122	-	-									2.2 % (109)
	42321	45	-	-								2.6 % (126)
	13224	37	44	-	-							3.2 % (157)
	35311	50	-	-	-	-						1.5 % (73)
	34232	31	36	-	-	-	-					1.8 % (90)
	52335	11	15	39	-	-	-	-				2.5 % (122)
	24445	8	8	-	31	-	-	-	-			2.6 % (129)
	43555	4	4	24	25	-	-	-	-	-		3.3 % (160)
	55555	2	2	18	20	23	23	57	82	103	-	6.7 % (330)
							Total					15.1 % (742)
Ranking no sorting Arm		**12111**	**11122**	**42321**	**13224**	**35311**	**34232**	**52335**	**24445**	**43555**	**55555**	**Inconsistencies by state**
	12111	-										3.8 % (197)
	11122	-	-									2.1 % (110)
	42321	48	-	-								2 % (104)
	13224	41	46	-	-							2.8 % (145)
	35311	48	-	-	-	-						1.1 % (58)
	34232	30	33	-	-	-	-					1.5 % (76)
	52335	13	14	27	-	-	-	-				2.1 % (107)
	24445	8	8	-	26	-	-	-	-			1.9 % (98)
	43555	5	5	17	20	-	-	-	-	-		2.7 % (137)
	55555	4	4	12	12	10	13	53	56	90	-	5 % (254)
							Total					12.5 % (643)
Ranking and sorting Arm		**12111**	**11122**	**42321**	**13224**	**35311**	**34232**	**52335**	**24445**	**43555**	**55555**	**Inconsistencies by state**
	12111	-										3.6 % (178)
	11122	-	-									2.2 % (107)
	42321	36	-	-								1.9 % (96)
	13224	35	43	-	-							2.8 % (138)
	35311	41	-	-	-	-						1.1 % (54)
	34232	26	25	-	-	-	-					1.3 % (67)
	52335	15	15	23	-	-	-	-				1.8 % (91)
	24445	10	9	-	26	-	-	-	-			2.4 % (118)
	43555	8	8	21	20	-	-	-	-	-		2.8 % (138)
	55555	7	7	16	14	13	16	38	73	81	-	5.3 % (265)
							Total					12.6 % (626)

* Numbers are absolute frequency by pair (row, column)

** Ties for a specific health state are calculated summing all inconsistent pairs that involve the specific health state. Proportions are calculated over the total number of possible inconsistencies given the sample size.

Table 4. Inconsistencies and ties (within respondent) by pairs of health states in rank order positions of Ranking and sorting Arm

		Profiles										
Inconsis-tencies		**12111**	**11122**	**42321**	**13224**	**35311**	**34232**	**52335**	**24445**	**43555**	**55555**	**Inconsistencies by state**
	12111	-										0.6 % (29)
	11122	-	-									0.3 % (15)
	42321	4	-	-								0.9 % (44)
	13224	10	6	-	-							0.6 % (29)
	35311	5	-	-	-	-						0.2 % (9)
	34232	4	7	-	-	-	-					0.3 % (16)
	52335	3	1	17	-	-	-	-				0.6 % (30)
	24445	0	0	-	7	-	-	-	-			0.2 % (11)
	43555	1	1	18	4	-	-	-	-	-		0.7 % (33)
	55555	2	0	5	2	4	5	9	4	9	-	0.8 % (40)
							Total					2.6 % (128)
Ties		**12111**	**11122**	**42321**	**13224**	**35311**	**34232**	**52335**	**24445**	**43555**	**55555**	**Inconsistencies by state**
	12111	-										0.3 % (15)
	11122	-	-									0.2 % (9)
	42321	1	-	-								0.5 % (27)
	13224	8	4	-	-							0.4 % (22)
	35311	1	-	-	-	-						0 % (2)
	34232	5	4	-	-	-	-					0.2 % (9)
	52335	0	1	8	-	-	-	-				0.6 % (33)
	24445	0	0	-	9	-	-	-	-			0.3 % (16)
	43555	0	0	18	0	-	-	-	-	-		0.7 % (36)
	55555	0	0	0	1	1	0	24	7	18	-	1 % (51)
							Total					2.1 % (110)

* Numbers are absolute frequency by pair (row, column)

**Inconsistencies/ties for a specific health state are calculated summing all inconsistent pairs that involve the specific health state. Proportions are calculated over the total number of possible inconsistencies/ties given the sample size.

In general, the distributions of values assigned to each of the 10 health states were similar across study arms (Fig. 1). Some statistically significant differences were observed in the variability of values for single health states between study arms, but no discernible pattern was observed. In addition, the mean observed value for state 35311 was significantly lower in the control arm than in the ranking and sorting arm (0.53 vs. 0.68; P ¼ 0.006). No statistically significant

differences were observed across arms for mean number of moves in the C-TTO task or in mean time used on each C-TTO task. The proportion of states assigned values at 1, 0.5, 0, and 0.5 was slightly lower in the ranking arms than in the control arm, but the observed differences were small and not statistically significant.

The mean interview duration was shorter by about 7 minutes in the control arm than in the ranking arms (Table 5). Taking into account that the mean interview duration in the control arm was about 45 minutes, a cost increase of at least 15% for including a ranking task in the EQ-VT was estimated. As presented in Table 5, structured feedback responses from study participants after the completion of the TTO task were different across study arms. In general, the ranking arms were found to be more complex than the control arm.

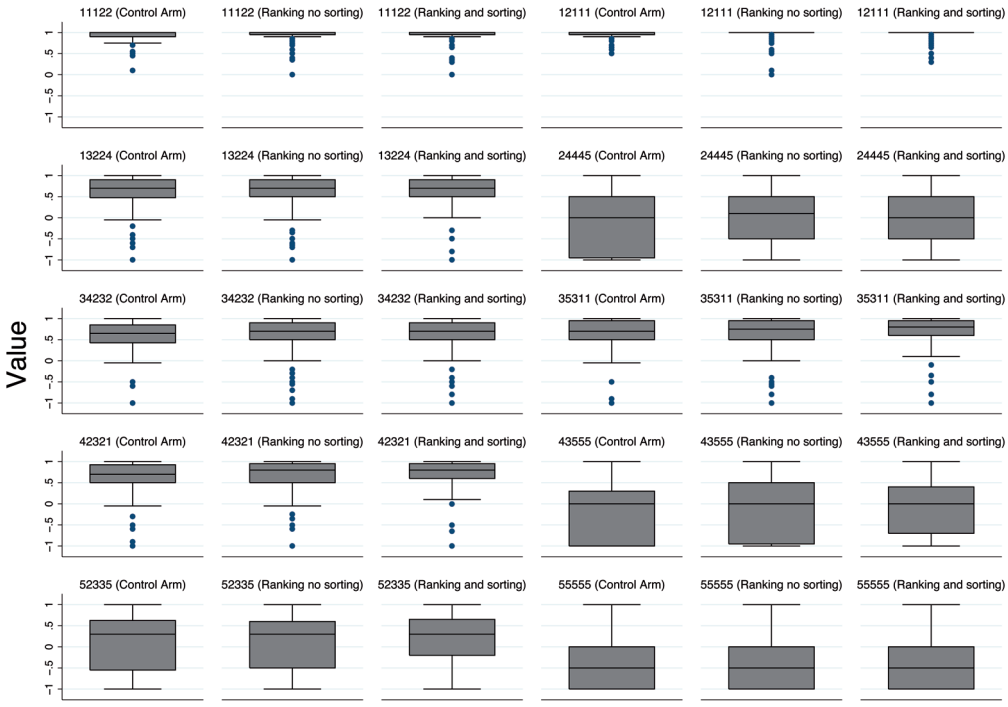

Fig. 1.- Boxplot of observed values by health state and study arm.

Table 5.- Times and feedback by study arm

Variables	Control Arm (n =196)		Ranking no sorting (n = 205)		Ranking and sorting (n = 196)	
	Mean	SD	Mean	SD	Mean	SD
Interview duration (min)	44.62	11.65	51.58	11.70	50.93	11.27
Single TTO task mean time (sec)	63.93	51.75	63.51	55.16	66.04	53.25
Single TTO task mean moves	6.69	3.20	6.61	3.63	6.72	3.58
Single TTO task mean time on WTD part (sec)	36.56	39.23	40.62	43.46	37.77	35.57
Single TTO task mean moves on WTD part	5.54	3.17	5.21	3.59	5.09	3.90
	n	%	n	%	n	%
Level of difficulties with the task						
1 = Agree	33	16.8%	52	25.4%	52	26.1%
2	48	24.5%	47	22.9%	45	22.6%
3	45	23.0%	37	18.0%	47	23.6%
4	22	11.2%	27	13.2%	16	8.0%
5 = Disagree	48	24.5%	42	20.5%	39	19.6%
Easy to understand						
1 = Agree	148	75.5%	148	72.2%	135	67.8%
2	26	13.3%	33	16.1%	26	13.1%
3	18	9.2%	16	7.8%	30	15.1%
4	3	1.5%	5	2.4%	2	1.0%
5 = Disagree	1	0.5%	3	1.5%	6	3.0%
Easy to tell						
1 = Agree	142	72.4%	137	66.8%	129	64.8%
2	23	11.7%	39	19.0%	26	13.1%
3	25	12.8%	15	7.3%	25	12.6%
4	2	1.0%	9	4.4%	14	7.0%
5 = Disagree	4	2.0%	5	2.4%	5	2.5%

DISCUSSION

This study presents results of the comparison of data from performing the EQ-5D-5L C-TTO valuation tasks and two experiments including a ranking task before the C-TTO tasks. We did not observe any indication that preceding the C-TTO valuation by a ranking task reduces the proportion of logically inconsistent responses. We observed that the overall proportion of tied values was lower in the two ranking arms than in the control arm. Nevertheless, although the difference was statistically significant, the reduction was small and tied values were less troubling than logical inconsistencies. Preceding C-TTO tasks by a ranking task did not result

in any statistically significant difference in the proportion of respondents assigning higher values to the worst state (55555) than for at least one other state.

The observed proportion of respondents assigning state 55555 a value higher than at least one other state was slightly but non-significantly higher in the ranking and sorting arm than in the other two arms. This could have been due to random variability between the arms, but it could have also been caused in part by the order of presentation of the health states. On the basis of the presentation order used in most EQ-5D-3L valuation studies when going from ranking to VAS valuation, the respondents were first presented with the top state from their ranking, then the bottom state, followed by sequential bisection. In EQ-5D- 3L valuation studies, this presentation order was used to help respondents assign values on the VAS. For this study, the same presentation order was introduced to test whether it resulted in lower rates of logical inconsistency. Incremental (top-down or bottom-up) presentation was considered but was discarded because of concerns that it could induce substantial anchoring effects. From the results, it would appear that the chosen presentation order either has no impact or is detrimental to a respondent's ability to perform C-TTO consistently. Unlike the transition from ranking to VAS in EQ-5D-3L valuation studies, in which the ordering may be useful because all previous responses remained visible on the VAS when moving from one state to the next, TTO values are assigned individually and previous responses are not visible to respondents.

In the ranking and sorting arm, there were more tied values based on C-TTO tasks than on the ranking task. This could be explained in part by clustering of values (sometimes referred to as "spikes") in the distribution of values elicited in TTO tasks; although the exact mechanisms are not fully understood, some values appear to be more attractive than others in the task, resulting in a higher prevalence of values at the end points, and "round" values (1, 0.5, 0, 0.5, and 1). The higher proportion of ties in C-TTO values compared with the ranking task could also result from health states being considered simultaneously in the ranking task, whereas they are presented sequentially in C-TTO tasks.

The fact that we found a small difference between the control (version 1.1 of the protocol) and experimental arms and the observed improvement in data quality from the Spanish valuation study (version 1.0 of the protocol) [10] may suggest that version 1.1 used in the control arm, adding three practice states and some explanatory pop-ups, combined with rigorous training of interviewers and close QC during the data collection process, may be sufficient to improve to an acceptable level the quality of the data of the EQ-5D-5L valuation protocol version 1.0 proposed by Oppe et al. [9], as implemented in Spanish and Dutch valuation studies.

In 2012, Rand-Hendriksen and Augestad [13] found that respondents in EQ-5D-3L valuation studies were more concerned with impairments on pain/discomfort and anxiety/depression dimensions when performing TTO valuation than when comparing the same states

in a ranking task. This suggests that the presentation form of the valuation task influences, which aspects of health respondents focus on, which could in theory limit the usefulness of ranking as a primer to TTO valuation. Nevertheless, differences between TTO and ranking described in the Rand- Hendriksen and Augestad study, although consistent across several valuation study data sets, were relatively small and related to ordering of health states that describe impairments on different dimensions. Sensitivity variation for different dimensions of health between valuation methods does not help explain rates of logical inconsistencies.

There are some limitations that should be considered when interpreting these results. We have included 10 health states, but the selection of these states could limit the generalizability of our results. The selection of a single block of health states from the pool of states used in EQ-5D-5L valuation studies was motivated by a need for comparability with other studies and maximizing statistical power with a limited sample size. From the 10 blocks of health states used in EQ-5D-5L valuation studies, we selected the block with the largest number of health state pairs in a logical dominance relationship. The fact that we lost the information about the ranking task in the ranking without sorting arm is unfortunate. We were, however, primarily concerned with the respondents' behavior in the C-TTO task, analyses of which were not impaired by this loss of data.

CONCLUSIONS

Reintroduction of the ranking task did not appear to result in discernible improvements in terms of reduced logical inconsistencies. Ranking before C-TTO did reduce the proportion of tied states, but only a 2% to 3% reduction was observed. The extra cost associated with its introduction because of the extra time for the ranking task was estimated to be about 15%. The observed, limited benefit of the ranking task on the quality of data and frequency of inconsistencies does not justify the cost of adding this task to valuation exercises. Thus, although we cannot conclusively declare that including a ranking task is a bad idea, we do not recommend reintroducing ranking in EQ-5D valuation on the basis of this study. Future studies should consider introducing a qualitative component to better understand the results presented here.

Acknowledgments

We are grateful to Irene Ávila Pérez for her assistance in the data collection process, to Rosana García and the whole interviewer team for their great work in organizing and conducting the interviews, and to Arnd Jan Prause and Job de Bruine for their technical assistance. We are also especially grateful to Elly Stolks and Koonal Shah for their continuous and useful contribution to the study design.

Source of financial support: This study was funded by a research grant (reference number 2013320) awarded by the EuroQol Group.

REFERENCES

1. Wisløff T, Hagen G, Hamidi V, et al. Estimating QALY gains in applied studies: a review of cost-utility analyses published in 2010. Pharmacoeconomics 2014;32:367–75.
2. Szende A, Oppe M, Devlin N. EQ-5D Value Sets: Inventory, Comparative Review and User Guide (Vol. 2). Dordrecht, The Netherlands: Springer, 2007. [3] EuroQol Group. EuroQol—a new facility for the measurement of health-related quality of life. Health Policy 1990;16:199–208.
3. Herdman M, Gudex C, Lloyd A, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res 2011;20:1727–36.
4. Devlin NJ, Krabbe PF. The development of new research methods for the valuation of EQ-5D-5L. Eur J Health Econ 2013;14(Suppl. 1):S1–3.
5. Janssen BM, Oppe M, Versteegh MM, Stolk EA. Introducing the composite time trade-off: a test of feasibility and face validity. Eur J Health Econ 2013;14(Suppl. 1):S5–13.
6. Dolan P. Modeling valuations for EuroQol health states. Med Care 1997;35:1095–108.
7. Augustovski F, Rey-Ares L, Irazola V, et al. Lead versus lag-time tradeoff variants: does it make any difference? Eur J Health Econ 2013;14(Suppl. 1):S25–31.
8. Oppe M, Devlin NJ, van Hout B, et al. A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. Value Health 2014;17:445–53.
9. Ramos-Goni JM, Pinto-Prades JL, Oppe M, et al. Valuation and modeling of EQ-5D-5L health states using a hybrid approach. Med Care 2014. [Epub ahead of print].
10. Versteegh MM, Vermeulen KM, Prenger R, et al. Dutch tariff for the 5 level version of EQ-5D. Value Health (in press).
11. Shaw JW, Johnson JA, Coons SJ. US valuation of the EQ-5D health states: development and testing of the D1 valuation model. Med Care 2005;43:203–20.

5

Chapter 5

Quality Control Process for EQ-5D-5L Valuation Studies

Juan M. Ramos-Goñi,
Mark Oppe,
Bernhard Slaap,
Jan J.V. Busschbach,
Elly Stolk

Value Health. 2017 Mar;20(3):466-473

ABSTRACT

Background: The values of the five-level EuroQol five-dimensional questionnaire (EQ-5D-5L) are elicited using composite time trade-off and discrete choice experiments. Unfortunately, data quality issues and interviewer effects were observed in the first few EQ-5D-5L valuation studies. To prevent these issues from occurring in later studies, the EuroQol Group established a cyclic quality control (QC) process.

Objectives: To describe this QC process and show its impact on data quality.

Methods: A newly developed QC tool provided information about protocol compliance, interviewer effects, and mean values by health state severity. In a cyclic process, this information is initially used to evaluate whether new interviewers meet minimal quality requirements and later to provide feedback about how their performance may be improved. To investigate the impact of this cyclic process, we compared the quality of the data in Dutch and Spanish valuation studies that did not have this QC process with that in the follow-up studies in the same countries that used the QC process. Data quality was measured using protocol violations, variability between interviewers, the proportion of inconsistent responders, and clustering of composite time trade-off values.

Results: In Spain, protocol violations were reduced from 87% in the valuation study to 5% in the follow-up study and in the Netherlands from 20% to 8%. In both countries, interviewers performed more homogeneously in the follow-up studies. The number of inconsistent respondents was reduced by 23.2% in Spain and 23.6% in the Netherlands. Values were less clustered in the follow-up studies.

Conclusions: The implementation of a strict QC process in EQ-5D-5L valuation studies increases interviewer protocol compliance and promotes data quality.

Keywords: economic, health status index, life valuation, quality control, quality of life.

INTRODUCTION

The five-level EuroQol five-dimensional questionnaire (EQ-5D-5L) is a health-related quality-of-life instrument consisting of five dimensions (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression), each with five levels of response (no problems, slight problems, moderate problems, severe problems, and extreme problems/unable) [1]. Instruments such as the EQ-5D are of great interest to clinical researchers and health economists to measure the benefit of health technologies. Two main reasons explain this interest. On one hand, the simplicity of the EQ-5D allows including it in any data collection process at a low burden for patients. On the other hand, the possibility to assign preference-based index values to the collected data makes it possible to use it in economic evaluation.

To develop a preference-based scoring algorithm, valuation studies to link EQ-5D-5L responses to index values are needed. To assess those values, the EuroQol Group developed a standardized protocol for such valuation studies [2,3]. It was implemented in a computer-assisted personal interview approach, called the EuroQol valuation technology (EQ-VT). The protocol centred around two valuation techniques: composite time trade-off (C-TTO) and discrete choice experiment (DCE). The C-TTO was developed and field-tested as part of a multinational research program [2,4]. It used the conventional TTO task for valuing health states considered better than death (BTD) [4,5], whereas it used lead-time TTO for health states considered worse than death (WTD) [4–9]. To promote comparability across EQ-5D-5L valuation studies, the interview was fully scripted and embedded in the EQ-VT. The script provided instructions about what standards and goals interviewers should achieve as well as text suggestions for what to say. The instructions section about how to explain C-TTO was notably detailed, anticipating the complexity of the C-TTO interviewer for both the respondent and the interviewer. Interviewers used an example health state ("being in a wheelchair") to explain the BTD and the WTD elements of the C-TTO task, by showing how the iterative procedure works, and interviewers are required to discuss the possibility that health states can be considered WTD. Spain and the Netherlands were among the first countries that used the EQ-VT for national valuation studies to obtain value sets for the EQ-5D-5L in 2012 and 2013, respectively. After the data were collected, preliminary analyses by both the Spanish and the Dutch research teams indicated interviewer effects: some interviewers systematically elicited higher values, lower values, or more inconsistent values than other interviewers [10,11]. This was not anticipated, because interviewer effects were not observed in a preceding pilot study that tested the application of the C-TTO technique [4]. A notable difference between the pilot and the national valuation studies was the experience of the interviewers with the C-TTO technique. The interviewers in the pilot study were researchers who participated in developing the interviewer instructions, whereas the interviewers in the Spanish and a large part of the interviewers in the Dutch valuation studies were inexperienced in conducting

TTO experiments before participating in the EQ-VT studies. This led us to suspect that the observed interviewer effects could be caused by insufficient compliance with the protocol. With post hoc data cleaning it may be possible to mitigate biases in the resulting value set because of the presence of interviewer effects and data quality issues. Nevertheless, exclusion criteria are controversial and exclusions will reduce the sample size, which affects the power to estimate the value set and jeopardizes the representativeness of the sample of respondents in terms of background variables. Acknowledging these difficulties, solutions were sought in tools to enhance protocol compliance so as to reduce interviewer effects and improve data quality. Such an approach was inspired by evidence regarding the benefit of quality control (QC) along randomized clinical trials [12–15]. Those QC processes are based on continuous data monitoring and various checks during data collection. Our specific case, however, is more challenging because we aimed to develop a standardized QC process that can be used in multiple countries, whereas preferences can vary across countries. Thus, it is not possible to distinguish between valid and invalid responses on the basis of values that might be obtained. We have developed a QC methodology and a software (EQ-VT QC tool) that does not place any previous assumptions on the values that might be obtained. We exploited the fact that the EQ-VT captures time and position of each mouse click, as well as the valuation data, which enables identification of possible patterns that emerge in valuation data in relation to key characteristics of each interviewer's approach to data collection. In this article, we describe the QC process for EQ-5D-5L valuation studies and explore the improvements in the resulting data, by comparing data from the first series of EQ-VT studies in Spain and the Netherlands, where the QC process was not available, with later data sets collected in the same countries, with the QC process in place.

METHODS

QC Reports

The EQ-VT QC tool produced standardized reports including figures, tables, and the explanation of its content. In Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2016.10.012, you will find the full QC reports combined for both the Spanish studies. The instructions and handling of DCE are simpler than those of C-TTO and therefore the QC report focused more on the C-TTO than on the DCE. The QC report can be grouped into four main sections: sample demographic characteristics, assessment of protocol compliance, assessment of interviewer effects in the data, and assessment of the consistency of the data with respect to health state severity (measured as level sum scores). The first section is self-explanatory and the content of the latter three sections of the QC report is summarized in Table 1.

The QC report started presenting the number of interviews, the demographic profile of the sample, and the number of interviews per interviewer. This was followed by figures that related to protocol compliance, which are present per interviewer. For instance, the available timings and positions of the mouse clicks provide the time interviewers used for explaining the C-TTO task and how many moves in the iterative procedure the interviewers showed to the respondents. This duration was assessed separately for the BTD part and the WTD part of the C-TTO. Using this information, it was possible to determine whether interviewers violated the protocol. For instance, short durations of the whole interview or parts of the interview suggested that the interviewer rushed through the instructions. Furthermore, the position of the mouse clicks can show whether the interviewer omitted demonstrations of parts of the EQ-VT functionality, such as not showing the WTD task to the respondent.

The next section of the QC report reviewed the values per interviewer such as the proportion of non-traders (i.e., respondents who give all health states the value of 1.00), the proportion of zero values, the proportion of negative values, and the proportion of respondents who value the state 55555 better than at least another state. This part of the report also included a comparison of the mean value over all health states and the overall SD per interviewer. For all interviewers individually there were figures with the distribution of the values from the -1.00 to 1.00 utility scale. Ideally, these figures should show, and not much, differences between interviewers. The last section of the QC report focused on the assessment of the consistency of the pooled data over all interviewers with respect to health state severity.

For several items, criteria were chosen to distinguish between compliance and noncompliance. If items were in the range of noncompliance, an interview was given a "flag." The QC tool reported a table with the number and proportion of "flagged interviews" per interviewer. An interview was flagged if one of the following criteria was met:

1. the WTD element was not shown in the wheelchair example; 2. the time spent on explaining the C-TTO task in the wheelchair example was less than 3 minutes;
2. a respondent spent less than 5 minutes to complete the 10 C-TTO tasks; or
3. the value for state 55555 was not the lowest and it was at least 0.5 higher than that of the state with the lowest value.

The judgment about protocol compliance of the DCE task was operationalized by looking for suspicious response patterns. An example is that a respondent always chooses the health state on the left. The report included a summary of the suspicious DCE responses in a table. This table was organized as follows: interviewer (column 1); the number of interviews completed (column 2); the mean amount of time taken (in minutes) to complete the seven DCE tasks

(column 3); and the number of respondents who used suspicious response patterns of choices across all seven DCE tasks (columns 4–7).

As mentioned earlier, quality issues seem less a problem for DCE because compliance with the DCE instructions seems easier. We did not determine a minimum standard at forehand to define compliance to the DCE protocol.

QC Process

Before the study, interviewers were trained: they had to conduct practice interviews before participating in the actual data collection. The practice interviews were reviewed with the QC tool and interviewers were given feedback on their performance. The QC reports from the first 10 interviews done per each interviewer were used to evaluate whether they met the minimum quality requirements to contribute to data collection. When a new interviewer had conducted 10 interviews and 4 or more were flagged, all 10 interviews conducted by that interviewer were removed from the database and he or she required retraining. After a further 10 interviews, the performance was re-evaluated. If again 4 or more interviews were flagged, these interviews were also removed and the interviewer was removed from the interviewer team. The 40% threshold was selected over more stringent cut-off points because the criteria listed earlier do not capture interviewer performance perfectly; sometimes respondents might be responsible for flags. Only when problems seemed persistent, interviewer performance was considered to be the main problem. In later stages of the data collection, the QC reports allowed the study teams to reflect on interviewer's performance and gave them continuous feedback about how to improve. This cyclic nature of the process provided a continuous stream of information that allowed the interviewers to keep improving their skills during the entire data collection period.

Table 1.- Quality control report description

Figure/Table	Aim/Expectations
	Assessment of protocol compliance
Interview total duration	The purpose of these 3 figures is to inform whether an interviewer systematically short tasks. Interviewers may want to perform fast interviews to finish their work earlier. We expect that low between and within interviewer variability. In other words, all interviewers take similar time and all respondents take similar time for each interviewer.
Amount of time taken to complete each C-TTO task	
Amount of time taken to complete each DCE task	
Amount of time spent in the wheelchair example	The purpose of these 3 figures is to inform how long the interviewers take to explain the C-TTO task, as shortcutting in the explanation could influence the respondent to do so. We expect that all interviewers make similar explanation following the interviewer script, making mean times to be similar among interviewers, but we also expect that interviewers makes always the same explanation, so we also expect low variability within interviewer. But some variability is expected within interviewer due to specific respondent's questions or doubts.
Time spent in the BTD element of the wheelchair example	
Time spent in the WTD element of the wheelchair example	
Moves performed in the wheelchair example	The purpose of these 3 figures is to inform about how well the iterative procedure (the process to move up or down the number of full health years in the C-TTO task) was explained. Few moves could not explain how to reach the preferred respondents' responses. We expect large number of moves across all interviewers on each interview. So high means, but low variability within interviewer. As in the above section respondent's questions could lead in more or less moves.
Moves performed in the BTD element of the wheelchair example	
Moves performed in the WTD element of the wheelchair example	
Percentage of interviews in which the worse-than-death element of the wheelchair example was used	The purpose of this figure is to inform about whether interviewers explain/ or at least shown the WTD element at the wheelchair example. We expect that interviewers always show the WTD when they introduce the C-TTO task. This is a key indicator for protocol compliance. If the WTD element of the C-TTO task is not explained the WTD responses will be bias producing zero censor values.
	Assessment of interviewer effects in the data
Percentage of respondents whose TTO data contain at least one 'inconsistency' in relation to health state 55555	The purpose of these two figures is either to inform about respondents misunderstanding or laziness. In one hand, lazy respondents could short the tasks by expressing their indifference point in the first step of the iterative procedure, if they do that for the 10 C-TTO tasks they are considered as non-trader. However, there could be real non-traders as very religious respondent. In the other hand, valuing the state 55555 higher than other states could be a signal of task misunderstanding as the 55555 is the worst possible health state defined by the EQ-5D-5L. We expect few inconsistent/non-traders respondents. For example, many inconsistent respondents for a specific interviewer could mean poor task explanation, even when time and moves look appropriate.
Non traders	

Figure/Table	Aim/Expectations
Percentage of health states given a value of exactly 0 in the TTO tasks	The purpose of these 2 figures is to inform about possible issues with WTD element of the C-TTO tasks. For example, many 0 values (Spike at 0) with small number of negative values could indicate either that interviewer is preventing WTD values or the interviewer is not explaining well the WTD element of the C-TTO task. We expect similar results across all interviewers.
Percentage of health states given a value of less than 0 in the C-TTO tasks	
Mean and SD of C-TTO value	The purpose of this set of figures is not only to identify whether interviewers are influencing respondents, but the side to where responses are biased and the size of the bias. We expect that interviewers have similar mean and SD value, but also similar distribution of values if no bias is present. These distributions are challenging to assess, given the fact that we do not know a priori what the "correct" mean values and distributions should be. Therefore, these figures are interpreted by comparing the data from each interviewer to the pooled data from all interviewers. In this way we can see which interviewers can be considered as outliers. This evaluation is also helpful to appraise to what extent differences in interview style that become apparent from the protocol compliance section might affect the data.
Distribution of responses for each specific interviewer	
Assessment of the consistency of the data with respect to health state severity	
Mean and SD of C-TTO values, by level sum score	The purpose of this figure is to inform about the logical basis of the results. For instance, an indication of low quality data is observing low mean values for mild states or high values for severe states, as it could be a consequence of obtaining key values in the iterative procedure (spikes). We expect health states with lower level sum scores to have higher mean value that those with higher level sum scores. But we also expect the opposite for SD, in other words, we expect more agreement in slight health states than in severe health states.
Overall C-TTO value distribution	The purpose of these figures is to inform about possible spikes and gaps in range of values. Interviewers may be similar when they are compared against each other, but they could be all producing similar influence over respondents. With these figure we can prevent this fact. Expectations very much depend of the country, i.e., cultural/religion traditions, etc.
C-TTO value distribution: by level sum score	

Examples of full reports can be found on the SM.
When the within variability of the aggregate data for one interviewer is too high compared with others, an outlier is present.

Data

We used data from the Spanish and Dutch EQ-5D-5L valuation studies [10,11] that were conducted without the QC process, and data from two follow-up studies in the same countries in which the QC process was implemented [16,17]. The Spanish and Dutch valuation studies followed the EQ-5D-5L valuation protocol as described by Oppe et al. [3], now known as version 1.0 of the protocol. The protocol had three main sections. In the first section, interviewers explained the purpose of the study and respondents were requested to value their own health using the EQ-5D-5L and were asked about their background characteristics. The second section of the interview consisted of the C-TTO tasks. This started with the interviewer explaining the C-TTO task using the example of being in a wheelchair as the health state. After this explanation, the respondents were asked to value 10 EQ-5D-5L health states using C-TTO. The last section of the interview consisted of a DCE in which respondents were requested to answer seven paired comparisons, each consisting of two EQ-5D-5L states. The two follow-up studies used an updated version of the protocol, known as version 1.1, which included the same C-TTO and DCE tasks as version 1.0. Nevertheless, several improvements were implemented: 1) three practice states (mild: 21121; severe: 35554; and moderate but difficult to imagine: 15411) were added immediately after the wheelchair example to better prepare respondents for the C-TTO task; 2) respondents were offered the possibility to confirm their response before starting the next task; and 3) the cyclic QC process was implemented as described in previous sections. The same interviewer instructions were used in both versions, except for the added instructions about the three practice states in version 1.1. All interviews for all studies were performed face-to-face using the EQ-VT platform.

The two follow-up studies were part of a methodological research program that was launched to address the data quality issues that were reported in the first wave of the EQ-5-5L valuation studies (version 1.0). These follow-up studies compared data collected using version 1.1 of the protocol with data collected using experimental versions of protocol 1.1, during which further modifications of the protocol were tested. For the present assessment of the QC process, we use data collected with only version 1.1 of the protocol in the follow-up studies to avoid confounding with other changes to the protocol.

Health States

The protocol for EQ-5D-5L valuation studies included 86 health states in the design of the C-TTO task. Only 10 of those states were also used in the follow-up studies. We restricted the comparison of the C-TTO data generated by the different versions of the protocol to those 10 states: 12111, 11122, 42321, 13224, 35311, 34232, 52335, 24445, 43555, and 55555. The DCE design was the same for all studies and included 196 health states distributed over 28 blocks of seven pairs of states.

Data Collection

C-TTO and DCE responses used for the comparison were derived from 597 participants included in the analysis, of whom 89 (Spain) and 107 (the Netherlands) participated in the valuation studies and 196 (Spain) and 205 (the Netherlands) participated in the follow-up studies. Respondents from the valuation studies were recruited from a panel using quota sampling, and those from the follow-up studies using convenient samples.

Interviews in the Dutch and Spanish valuation studies were conducted by 21 and 32 trained interviewers, respectively. The training in Spain consisted of one half-day session covering the interviewer instructions. In the Netherlands, the training took a whole day and consisted of a presentation of the components of the interview, discussion of the interviewer instructions, practice interviews in pairs, and a discussion of difficult interview elements. In addition, the Dutch principal investigator reached out at least once to each interviewer after data collection had started to discuss the interviewer's experiences.

For the follow-up studies, six trained interviewers in the Netherlands and seven in Spain, different from the valuation studies in both countries, conducted all interviews during March to April 2014. The training in Spain now consisted of a 3-day workshop, covering study background and aim, interviewer script, 10 practice interviews for each interviewer, plus a round table to share/comment interview issues and to review the QC reports for the 10 practice interviews. The training in the Netherlands involved the same 1- day training session as the valuation study. In both follow-up studies, interviewers were monitored at least weekly using the EQ-VT QC tool, which described the quality of their interviews.

Analysis

For between-study comparisons, we used proportions to present protocol compliance results. Means, standard errors, and variation coefficients (SD/mean) of duration and number of moves in the wheelchair example for both BTD and WTD values were used to explore the harmonization level of the C-TTO explanation within each study. We compared the interviewer effects by using a graphical presentation of kernel distributions of values for each interviewer. In addition, we compared the overall distribution of values and the distribution of values for state 55555 to illustrate values of the most severe health state. Finally, we considered the proportion of inconsistent respondents. In this analysis, an inconsistent respondent is defined as a respondent who values at least one pair of logically dominant health states inconsistently. We used the Paretian Classification of Health Change [18] in the definition of a logical dominance relationship between two health states; that is, state 1 dominates state 2 when state 1 is better than state 2 on at least one dimension, and no worse than state 2 on any remaining dimension. Therefore, the value of state 1 should be higher than the value of state 2, and when it is lower we considered it as an inconsistency.

RESULTS

The results showed that protocol compliance was an issue in the Spanish valuation study: 87% of interviews had a protocol violation. For example, the interviewers did not explain the WTD element of the wheelchair example in 71% of interviews. In contrast, the Spanish follow-up study had 5% of protocol violations and 0.5% of interviews omitting the WTD in the wheelchair example (Table 2). In the Dutch valuation study, protocol compliance was less of an issue, with interviewers violating the protocol in 20% of cases. Nevertheless, improvements were still made because the proportion of protocol violations dropped to 8% in their follow-up study.

Table 2. Protocol compliance

	Spain		The Netherlands	
	Valuation study	**Follow-up study**	**Valuation study**	**Follow-up study**
Sample size	89	196	107	205
Proportion of interviews flagged (N)	87% (77)	5% (10)	20% (21)	8% (17)
Proportion of interviews where the WTD element was not used in the WC example (N)	71% (63)	0.5% (1)	9% (10)	1% (3)
Proportion of interviews where interviewer did not spend at least 180 seconds (3 minutes) on the wheelchair example (N)	76% (68)	0.5% (1)	5% (5)	2% (4)
Proportion of interviews where interviewer did not spend 5 minutes on the 10 TTO tasks (N)	34%(30)	4% (7)	3% (3)	2% (5)

Various indicators were also affected. For example, the average time taken to explain the C-TTO task using the wheelchair example, the average time per TTO task, and the number of moves used in the iterative procedure all increased. In addition, the variation coefficients were smaller for durations and moves spent on the C-TTO explanation in the follow-up studies compared with valuation studies: they all showed more homogeneity (Table 3).

Table 3. Wheelchair example (duration and number of moves)

	Spain				The Netherlands			
	Valuation study		Follow-up study		Valuation study		Follow-up study	
	Mean (SE)	Variation Coeff.	Mean (SE)	Variation Coeff.	Mean (SE)	Variation Coeff.	Mean (SE)	Variation Coeff.
Total time (sec.)	132 (14)	1.21	661 (15)	0.32	433 (15)	0.35	368 (8)	0.33
Time on BTD element (sec.)	110 (13)	1.10	446 (11)	0.35	297 (11)	0.38	241 (6)	0.36
Time on WTD element (sec.)	22 (7)	3.09	215 (7)	0.44	136 (8)	0.60	126 (4)	0.49
Total moves	9.1 (1.0)	1.01	42.3 (1.0)	0.33	25.8 (1.2)	0.49	24.6 (0.8)	0.44
Moves on BTD element	7.3 (0.9)	1.10	25.5 (0.7)	0.38	17.6 (1.0)	0.56	15.3 (0.5)	0.50
Moves on WTD element	1.7 (0.4)	2.30	16.8 (0.6)	0.48	8.1 (0.7)	0.88	9.3 (0.4)	0.64

With respect to the interviewer effects, the distribution of values per interviewer was more homogeneous in the follow-up studies compared with that in the valuation studies in both countries (Fig. 1). In particular, the Spanish valuation study showed 11 of the 32 interviewers eliciting no WTD values, whereas all interviewers did in the follow-up study (Fig. 1A, 1B); further details are provided in the Supplemental Materials. The Dutch valuation study showed 1 out of 21 interviewers eliciting no WTD values, whereas all interviewers did in the follow-up study (Fig. 1C, 1D).

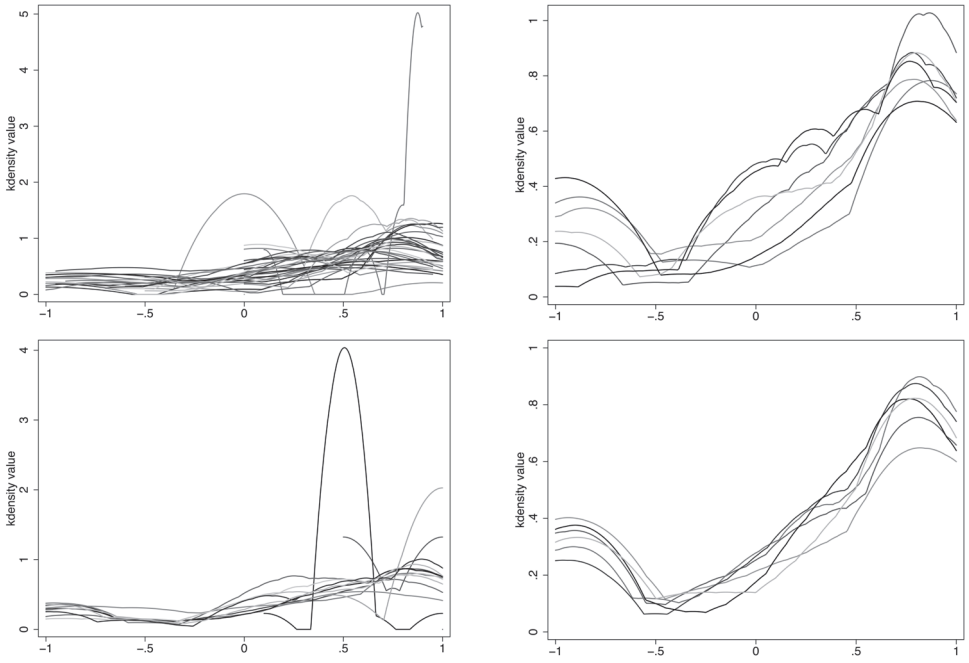

Figure 1. - Distribution of values over the 10 health states by interviewer.

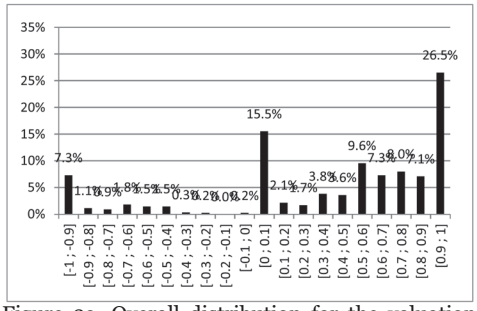

Figure 2a. Overall distribution for the valuation study (Spain)

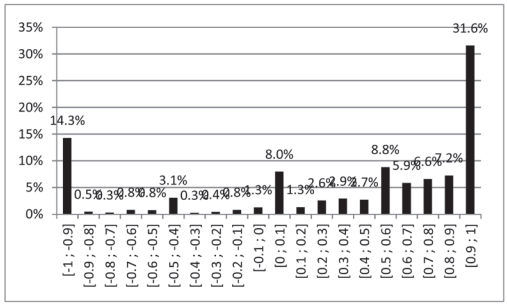

Figure 2b. Overall distribution for the follow-up study (Spain)

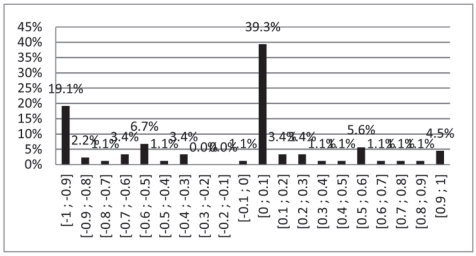

Figure 2c. 55555 distribution for the valuation study (Spain)

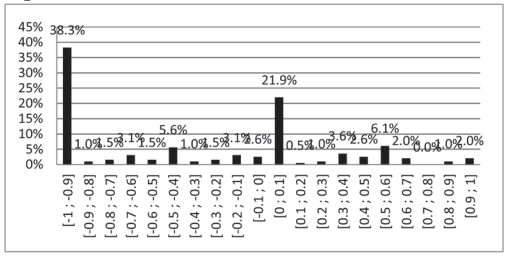

Figure 2d. 55555 distribution for the follow-up study (Spain)

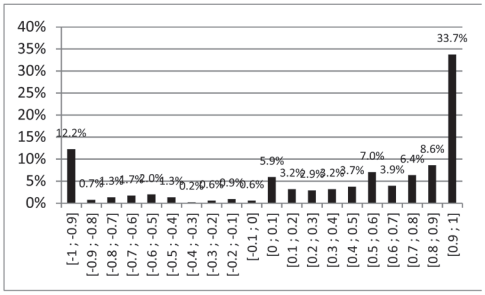

Figure 2f. Overall distribution for the valuation study (The Netherlands)

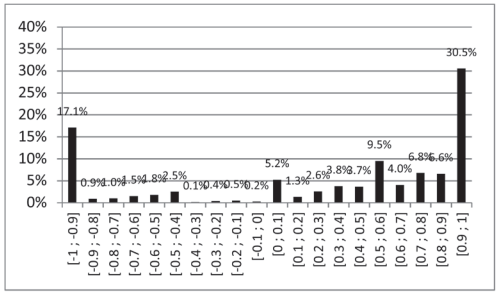

Figure 2g. Overall distribution for the follow-up study (The Netherlands)

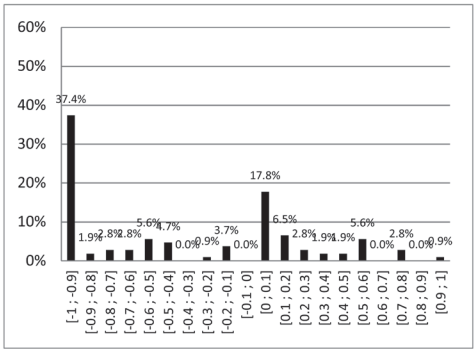

Figure 2h. 55555 distribution for the valuation study (The Netherlands)

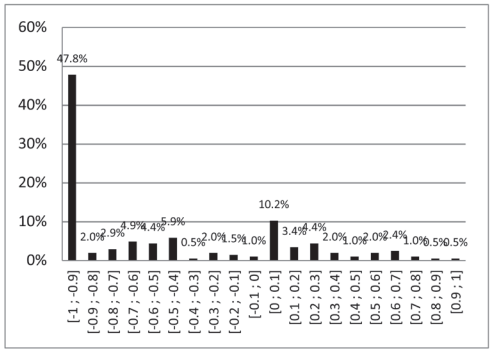

Figure 2i. 55555 distribution for the follow-up study (The Netherlands)

Figure 2.- Overall and pits state (55555) distributions

Both the overall distribution and the distribution for the state 55555 showed less clustering of values at 0 in the follow-up studies. The proportion of negative values was higher in the follow-up study, lowering the mean observed value for state 55555 in both countries. The gap of values between 0.5 and 0 shown in the valuation studies is mitigated in the follow-up studies (Fig. 2).

The QC process also had an impact on the proportion of respondents with one or more inconsistent responses. The proportion of Spanish respondents who had at least one inconsistent response was 48.3% for the valuation study, whereas this proportion dropped to 25.1% in the follow-up study. In the Netherlands, these proportions were 43.9% and 19.5% for the valuation and the follow-up study, respectively. Differences in proportions of inconsistent respondents were significant ($P < 0.0001$).

DISCUSSION

This article reported on the effect of implementing a QC process on EQ-5D-5L valuation studies, using four data sets, two of which were collected with QC and two without QC. The results provide evidence that our QC process improved the quality of the valuation data, but the effect size of the improvements varied between countries because of the marked difference in the data quality obtained in their valuation studies without the QC process. In this study, we assumed that the increased time taken for the interviews and the increased number of moves in the iterative C-TTO task reflect greater interviewer and respondent engagement. Moreover, the reduction in inconsistent responses and a lower clustering of values were also seen as improvements. One can argue that the extensive training and QC influenced interviewer responses more. This might be true, but it is difficult to see that why inconsistent and clustered responses represent better answers. All in all, the QC process seemed to improve the data in a valuation study, whereas uncertainty exists about the quality of captured data if QC is not adopted.

The QC process presented here, although custom-made for EQ- 5D-5L valuation studies, was built on the same principles as those of the traditional QC process to check units of production [19]. There are, however, obvious limitations in our case. Our QC process was more challenging, because our unit of production was a set of subjective values elicited in an interview so that neither the validity of the values nor the validity of the interview can be directly appraised. Each interview was unique making arbitrary the definition of a valid interview. It was unrealistic to expect that each interviewer will use exactly the same wording in all interviews, or as his or her colleagues; it depended on the questions from the respondents. This led us to focus on averages and variability, rather than using interviews as units, making it possible to harmonize interviewer performance and reduce potential bias in

respondents' responses. The effect of the QC process then partly came via the continuous monitoring and feedback process instigated by the principal investigator, and was not simply a result of taking out the bad units. Rather, to account for respondent interviewer interaction, we established a conservative threshold of 4 flagged interviews out of 10 as the limit to stop and retrain the interviewer. This was our analogy process of stopping and reviewing our production system. The information about the appropriate actions to take when issues are encountered can be found elsewhere [20].

Study Limitations

The QC process is probably not the only factor that caused the observed differences between valuation and follow-up studies. Small modifications beyond the QC process were also introduced from version 1.0 used in valuation studies to version 1.1 used in follow-up studies, such as the introduction of practice states and a confirmatory pop-up screen. Nevertheless, it is unlikely that these additions to the protocol can substantially improve interviewer compliance with the protocol, which was achieved by monitoring the time interviewers spent on each part of the interview, as part of the QC process.

Arguably, the improvement in training efforts in Spain may have had an important impact on the data as well, impeding conclusions about causality or about to what extent improvements can be attributed to the QC process. Nevertheless, because we also observed an improvement in the Netherlands, where the initial training efforts were comparable across the two studies, it is reasonable to assume that at least part of the effect size can be attributed to the QC. Further studies should clarify which parts of the QC process are most helpful. Another limitation of this study is that differences between respondents may affect results. It should, however, affect only the distribution of values by interviewer comparison, but it should not affect neither the overall distribution nor the overall inconsistency rate.

CONCLUSIONS

The results in this article support the decision of the EuroQol Group to extend its original EQ-5D-5L valuation protocol 1.0 with a QC tool. The impact of the QC process on the characteristics of a data set is large. We therefore recommend an uptake of similar strategies in future valuation studies, that is, transparency about interviewers' selection and training and the kind of feedback they received about their performance. Key characteristics of the raw data need to be reported as well to make possible judgment about the quality. Past valuation studies, including most of the EQ-5D-3L valuation studies, lack this kind of transparency. It is likely that efforts to prevent data quality issues across valuation studies will help to improve the determination of difference that related to cultural, methodological, analytical, or procedural choices. The implementation of the cyclic QC on EQ-5D-5L valuation studies increased

interviewer protocol compliance, reduced differences in an interviewer's elicited values, and significantly improved data quality.

Acknowledgments

Most of the research effort is funded by the EuroQol Research Foundation. Additional funds came from the Academic Centers. All authors are associate with the EuroQol Group. We express our gratitude to Koonal Shah for the texts he provided to accompany the tables and figures in the QC reports. We also express their gratitude to Oliver Rivero-Arias for the useful feedback while preparing this article.

REFERENCES

1. Herdman M, Gudex C, Lloyd A, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res 2011;20:1727–36.
2. Devlin N, Krabbe P. The development of new research methods for the valuation of EQ-5D-5L. Eur J Health Econ 2013;14(Suppl.):1–3.
3. Oppe M, Devlin NJ, van Hout B, et al. A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. Value Health 2014;17:445–53.
4. Janssen BM, Oppe M, Versteegh MM, et al. Introducing the composite time trade-off: a test of feasibility and face validity. Eur J Health Econ 2013;14(Suppl. 1):S5–13.
5. Oppe M, Rand-Hendriksen K, Shah K, et al. EuroQol protocols for time trade-off valuation of health outcomes. Pharmacoeconomics 2016;34:993–1004.
6. Robinson A, Spencer A. Exploring challenges to TTO utilities: valuing states worse than dead. Health Econ 2006;15:393–402.
7. Augustovski F, Rey-Ares L, Irazola V, et al. Lead versus lag-time trade-off variants: does it make any difference? Eur J Health Econ 2013;14 (Suppl. 1):S25–31.
8. Luo N, Li M, Stolk EA, et al. The effects of lead time and visual aids in TTO valuation: a study of the EQ-VT framework. Eur J Health Econ 2013;14(Suppl. 1):S15–24.
9. Devlin N, Buckingham K, Shah K, Tsuchiya A, Tilling C, Wilkinson G, vanHout B. A comparison of alternative variants of the lead and lag time TTO. Health Econ. 2013 May;22(5):517–32.
10. Ramos-Goñi JM, Pinto-Prades JL, Oppe M, et al. Valuation and modeling of EQ-5D-5L health states using a hybrid approach. Med Care (published online ahead of print December 17, 2014). http://dx.doi.org/ 10.1097/QAI.0000000000000492.
11. Versteegh MM, Vermeulen KM, Prenger R, et al. Dutch tariff for the 5 level version of EQ-5D. Value Health 2016;19(4):343–52.
12. Knatterud GL. Methods of quality control and of continuous audit procedures for controlled clinical trials. Control Clin Trials 1981;1:327–32.
13. Vantongelen K, Rotmensz N, van der Schueren E. Quality control of validity of data collected in clinical trials. EORTC Study Group on Data Management (SGDM). Eur J Cancer Clin Oncol 1989;25:1241–7.
14. De Pauw M. Quality control in data monitoring of clinical trials. Acta Urol Belg 1994;62:31–5.
15. Buyse M. Centralized statistical monitoring as a way to improve the quality of clinical data. March 24, 2014. Available from: http://www. appliedclinicaltrialsonline.com/centralized-statistical-onitoring-way-improve-quality-clinical-data. [Accessed June 10, 2016].
16. Ramos-Goñi JM, Rand-Hendriksen K, Pinto-Prades JL. Reintroduction of the ranking task in valuation studies: improved data quality and reduced level of inconsistencies? The case for EQ-5D-5L. Value Health. 2016;19(4):478-86.

17. Shah K, Rand-Hendriksen K, Ramos-Goñi JM, Prause AJ, Stolk E. Improving the quality of data collected in EQ-5D-5L valuation studies: a summary of the EQ-VT research methodology programme. 31ST EuroQol Group Scientific Plenary. Stockholm, Sweden. Sep 2014.

18. Devlin NJ, Parkin D, Browne J. Patient-reported outcome measures in the NHS: new methods for analysing and reporting EQ-5D data. Health Econ 2010;19:886–905.

19. Radford GS. The Control of Quality in Manufacturing. New York, NY: Ronald Press Co., 1922.

20. Purba FD, Hunfeld JA, Iskandarsyah A, Fitriana TS, Sadarjoen SS, Passchier J, Busschbach JJ. Employing quality control and feedback to the EQ-5D-5L valuation protocol to improve the quality of data collection. Qual Life Res. 2016 Oct 31. [Epub ahead of print].

6

Chapter 6

Dealing with the health state ‘dead’ when using discrete choice experiments to obtain values for EQ-5D-5L heath states

Juan M. Ramos-Goñi,
Oliver Rivero-Arias,
María Errea,
Elly A. Stolk,
Michael Herdman,
Juan M. Cabasés

Eur J Health Econ. 2013 Jul;14 Suppl 1:S33-42

ABSTRACT

Objective: To evaluate two different methods to obtain a dead (0)—full health (1) scale for EQ-5D-5L valuation studies when using discrete choice (DC) modeling.

Method: The study was carried out among 400 respondents from Barcelona who were representative of the Spanish population in terms of age, sex, and level of education. The DC design included 50 pairs of health states in five blocks. Participants were forced to choose between two EQ-5D-5L states (A and B). Two extra questions concerned whether A and B were considered worse than dead. Each participant performed ten choice exercises. In addition, values were collected using lead-time trade-off (lead-time TTO), for which 100 states in ten blocks were selected. Each participant performed five lead-time TTO exercises. These consisted of DC models offering the health state 'dead' as one of the choices—for which all participants' responses were used (DC_{dead})—and a model that included only the responses of participants who chose at least one state as worse than dead (WTD) (DC_{WTD}). The study also estimated DC models rescaled with lead-time TTO data and a lead-time TTO linear model.

Results: The DC_{dead} and DC_{WTD} models produced relatively similar results, although the coefficients in the DC_{dead} model were slightly lower. The DC model rescaled with lead-time TTO data produced higher utility decrements. Lead-time TTO produced the highest utility decrements.

Conclusions: The incorporation of the state 'dead' in the DC models produces results in concordance with DC models that do not include 'dead'.

Keywords: Discrete choice methodology Time trade-off Health state 'dead' EQ-5D-5L EuroQol Group

INTRODUCTION

The EQ-5D is one of the most widely used preference-based instruments. In 2009, the EuroQol Group released a new version (EQ-5D-5L) of the instrument that included five levels of severity in each dimension, as opposed to three in the original version [1]. For the new instrument to generate a set of societal values for the 3,125 health states, it had to distinguish five levels of severity in five dimensions.

Previous valuation studies had predominantly used time trade-off (TTO) to obtain social preferences from which value sets for EQ-5D health states could be modelled [2–5].

However, increasing the number of health states from 243 to 3,125 made it considerably more costly and complicated to conduct valuation studies based on an interview method such as TTO. Conventional TTO also has problems with health states worse than the state 'dead' [6]. These issues led the EuroQol Group to explore new approaches to obtain social values for health states, notably discrete choice (DC) methodology.

In a typical DC task, respondents compare two different options (paired comparison) and indicate which one they prefer. Discrete choice experiments (DCE) have been used extensively in areas such as marketing and transport but not so much in health economics. The use of DCE for health- state valuation is a relatively recent development. Potential advantages include the relative ease of comprehension and administration of ordinal tasks and its greater reliability. DC models may also avoid some of the biases associated with traditional valuation methods [7]. Stolk et al. [8] demonstrated that DC modeling with the classic EQ-5D (three-level) instrument produces values that are congruent with values obtained by other valuation techniques, TTO in particular. That result confirmed previously published findings [9–12].

A question that arises about the use of DC for health-state valuation concerns how to anchor the values produced by the choice model onto the dead (0)—full health (1) scale that is required to compute quality-adjusted life years. One strategy is to use DC data in combination with TTO data. This would entail deriving values from DC data and then using values from TTO to rescale those DC values. The need to collect TTO data alongside a DC study, however, might make the valuation study more complex than necessary. So, instead, the DC task could be designed in such a way that a value for 'dead' can be extracted from the DC responses and then used to anchor the values. One way to do this is by explicitly comparing the health state 'dead' to the EQ-5D-5L health states that are being judged in the DC task. An objection on theoretical grounds is that responses obtained from choices comparing heath states to dead may violate the random utility theory underlying the DC model. This happens when a subset of respondents consider all health states to be better than dead—for example, due to their religious beliefs. The size and effect of the bias are yet unknown; in practice, the bias may be small. Indeed, when this approach was adopted for the valuation of EQ-5D-3L health states [8], the results were promising. Whether or not this will also be so when it is used for EQ-5D-

5L valuation will be expanded upon in this paper.

The primary objective of the study reported here was to examine the results of two different approaches to rescale DC models incorporating 'dead' into the utility scale as an anchor point and to compare the results with those obtained anchoring on lead-time TTO. A secondary objective was to evaluate the effect of excluding DC responses elicited from those who did not consider any health state to be worse than the health state dead.

METHODS

This pilot study used both a DC and a lead-time trade-off (lead-time TTO) approach to produce values for the set of 3,125 (5^5) health states defined by the EQ-5D-5L instrument. As a detailed description of each approach in the context of health-state valuation can be found elsewhere [8, 13], only a brief summary will suffice here. The study design followed recommendations from the EuroQol Group Valuation Task Force and was part of a multi-country initiative to explore methodological uncertainties about the valuation protocol for a new EQ-5D-5L value set.

Valuation of EQ-5D-5L health states

DC method

In the DC method, the respondents were asked to state their preference between two health states, A and B. This comparison of health states produces data that were subsequently analyzed to produce values on a latent scale. The profiles did not mention either their duration or what happens after these states. The DC design was generated using a Bayesian efficient approach [14] and consisted of 50 pairs of health states allocated to five blocks. These amounts were set in order to have sufficient power to estimate health-state values based on the proportions of choices between the pairs of states. To allow anchoring of the values on the 'dead—full health' scale, we extended the DC task by asking whether state A was worse than dead (WTD) and whether state B was WTD.

Lead-time TTO

The lead-time TTO method is an extension of the traditional TTO [13]. In a classic TTO, participants complete one task for health states considered better than dead and another task for those considered WTD. Lead- time TTO consists of a single task: to choose between Life A (T years in full health) and Life B [10 years in full health (lead time) plus 5 years in a target health state (disease time)]. All respondents start with Life A versus Life B where T = 15 years in 11111; depending on whether they choose A or B, the value of T is raised or lowered until the participants feel that A and B are the same. The lead-time TTO design was constructed with a Federov algorithm that allowed model parameters to be estimated without bias and with

minimal variance [15]. The final lead-time TTO design contained 100 states in ten blocks.

Data collection

Four hundred persons, who were representative of the Spanish population in terms of age, gender, and education, took part in this study. An online survey administered via the EuroQol Valuation Technology (EQ-VT) software was used to collect DC and lead-time TTO responses. The final survey included the EQ-5D-5L questionnaire, ten DC tasks, and five lead-time TTO tasks as well as demographic questions. Participants were also queried about the difficulty of the DC and lead-time TTO tasks and how well they had understood them. The EQ-VT randomly assigned each participant to a DC block and a lead-time TTO block. In both types of block, the tasks were presented in random order. Given the number of participants, the study yielded an average of 80 observations for each DC pair (400 participants 9 10 states/50 pairs) and 20 observations for each lead-time TTO state (400 participants 9 5 states/100 states).

A survey company administered the study in Barcelona (June 2011). The researchers JMRG, ME, MH, and JC supervised the data collection with assistance from the EuroQol Group. Participants were recruited using telephone directories for the metropolitan area of Barcelona, personal contacts, a database of panellists, or 'snowballing' from contacts of participants included in this study.

Eight groups, each with an average of ten respondents, were recruited per day during 6 days, yielding the target of 400 participants. Each participant was assigned a computer and given an ID number and a password. Two computer rooms were available for each session. Interviews were conducted by two trained interviewers and four members of the Spanish Valuation Team (JMRG, ME, MH, and JC).

Statistical analysis

The sample as well as the DC and lead-time TTO responses were described with descriptive statistics. Four statistical models were used to estimate EQ-5D value sets: (1) a conditional logistic model, which produced the health-state values based only on choices between health states, thus ignoring responses to the dead questions (N = 397; henceforth DC_{TTO}; (2) a rank-ordered logistic model, which was then used on the full DC dataset and included responses to the dead questions (N = 397, henceforth DC_{dead}); (3) a rank-ordered logistic model, which used data only on those participants who chose at least one state worse than dead (N = 195, henceforth DC_{WTD}); a linear regression model, which used the lead-time TTO responses (N = 373; henceforth called lead-time TTO). The three models that were estimated with DC responses had to be rescaled to indicate that 0 stands for dead and that 1 forms the upper bound for full health. This was achieved using the additional 'dead' questions in the DC experiments in the case of DC_{dead} and DC_{WTD}. For the DC_{TTO} model, the worst health state predicted on the lead-time TTO model (profile 55555) was taken as an anchor

point to rescale the arbitrary scale of the conditional logistic model. Details on each model are given below.

DC_{TTO} model

In the case of DC, the values are not directly observable and have to be calculated from the responses to the choice exercise. We assume that the participants choose the health state that gives them higher utility, so this can be modelled as a conditional logistic model. As such, the independent variable Y_I represents the choice of participant I between A or B. The model assumes a value decomposition in two parts, explainable by V_{iA} plus an error e_i. If errors are assumed to be random and to show a type 1 extreme value distribution, a conditional logistic model emerges [8, 16, 17]. Let us assume that component V of the value can be explained with an additive model:

$$(1)\ V_{iA} = \sum_{j=1}^{J} x_{iAj} \cdot \beta_j$$

where X_{iAj} are 20 dummies {0, 1}, per participant i, representing the severity levels for each dimension of EQ-5D- 5L for state A. Then β_j will represent the coefficient for each independent variable j.

Accordingly, it is possible to estimate the coefficients of the model and thus to extrapolate values that have not been observed within the population by using the linear part of the DC_{TTO} model. The values obtained from the linear part of the model shown above are on an arbitrary scale. In order to rescale the values from the DC_{TTO} model, the extreme negative value estimated in the lead-time TTO model (55555) was used to anchor the DC_{TTO} 55555 health state to that value. Therefore, both models produce the same index value for the 55555 health state. To obtain a full set of utility decrements, every coefficient of the DC model is divided by the scalar ($55555_{\text{lead-time TTO}}$ - 1)/ (55555_{DCTTO} - 1). The outcome of this transformation for each coefficient yields the utility decrements for the DC_{TTO} model.

DC_{dead} model

A rank-order logistic analysis was performed for the DC_{dead} model [8]. In the same way as for a conditional logistic model, a two-part decomposition is assumed for the value. Where V_{iA}, this model can be written as follows:

$$(2)\ V_{iA} = \sum_{j=1}^{20} X_{iAj} \cdot \beta_j + X_{i\,dead} \cdot \beta_{dead}$$

Values are therefore obtained from the linear part (above) of the model on an arbitrary scale, as they are in the DC_{TTO} model. For this DC_{dead} model, the anchor point is the health state

dead. Since the value for dead has to be 0, each coefficient is divided by β_{death}; ensuring $\beta'_{death} = -1$. The final function to estimate index values is given by:

$$(3)\ V_{iA} = 1 - \sum_{j=1}^{20} X_{iAj} \cdot \beta'_j + X_{i\,dead} \cdot \beta'_{dead}$$

Where $\beta'_j = \beta_j / abs(\beta_{dead})$ $\beta'_j = \beta_j / abs(\beta_{dead})$.

DC_{WTD} model

The DC_{WTD} model was estimated as a rank-order logistic model similar to the DC_{dead} model. For this case, the data were restricted to responses from participants who chose at least one state worse than dead. This model was used to evaluate whether including participants who did not choose any state worse than dead would bias the coefficient estimates.

Lead-time TTO model

For lead-time TTO responses, a linear model was estimated. The specification of the model in its general form is:

$$(4)\ Y_i = \sum_{j=1}^{n} x_{ij} \cdot \beta_j + \varepsilon_i$$

where Y_i represents the observed values from lead-time TTO data for participant i. A continuous variable, which takes values between -2 and 1, was created. The lead-time TTO values T from the survey were transformed into a -2 and 1 scale using the formula (T - T_lead)/(T_total - T_lead). In our design, T_lead = 10 indicates that the additional years in full health occur at the beginning of the exercise, and T_total = 15 indicates the sum of T_lead and disease time (5 years). The independent variables X_{ij} are 20 dummies {0, 1} for each participant i, representing the severity levels for each dimension of EQ-5D-5L. β_j represents the coefficients for each independent variable

j; e_i represents the errors for each participant i. Different specifications used in previously published examples were explored in order to fit the best model [2–5]. However, none of the models led to improved goodness of fit measured with log-likelihood, nor did they correct any inconsistencies in the models' coefficients. Therefore, the lead-time TTO model presented in this study was estimated using a simple ordinary least squares model. Finally, a function to estimate values for each health state was created using the regression model specified in the following equation:

$$(5)\ Y_i = 1 - (\beta_0 + \beta_1 \cdot mo2_i + \beta_2 \cdot mo3_i + \beta_3 \cdot mo4_i + \beta_4 \cdot mo5_i + \cdots + \beta_{20} \cdot ad5_i + \varepsilon_i)$$

with mo2, mo3,mo4, mo5, sc2, sc3..., ad4 and ad5 indicating the corresponding dummy for the EQ-5D-5L severity level.

To compare the four models, we used descriptive statistics and quantile–quantile plots (Q-Q plots) of the value sets obtained from the different models. A Q-Q plot sets off estimates of the quantiles of two distributions against each other, and the pattern of points it displays is used to compare the two distributions of value sets. In addition, the value sets produced for each model are compared using the mean square difference (MSD) and concordance correlation coefficient (CCC) [18]. All values for the 3,125 health states are estimated by each of the estimated models. For each one: one comparison (model 1 vs. model 2), the MSD is calculated as follows:

$$(6)\ MSD_{model1\ vs\ model2} = \frac{\sum_{i=1}^{3,125}(indexvalue_{model1_i} - indexvalue_{model2_i})^2}{3,125}$$

All statistical analyses were performed on STATA 11 MP (StataCorp LP, College Station, TX).

RESULTS

Sample characteristics

The study cohort comprised 400 persons with a mean age (standard deviation, SD) of 44.1 (16.9) years; and 59.7 % (239) were male (Table 1). More than half were employed or freelance and 15 % were retired.

Table 1: Descriptive statistics of sample (n=400)

Age (mean, SD)	**44.1 (16.9)**	
Gender	N	%
Male	239	59.7
Female	161	41.3
Employment status		
- Domestic tasks	13	3.25
- Employed or freelance	201	50.25
- Student	39	9.75
- Retired	59	14.75
- Unemployed	60	15
- Missing	28	7
Education		
- Higher education	110	27.5
- High school	175	43.75
- Primary school	86	21.5
- Missing	29	7.25
Experience severe illness		
- Self	63	15.75
- Relatives	278	69.5
- Other	136	34

SD standard deviation

[a] Data are presented as the number (N) of subjects with the percentage of total subject cohort given in parenthesis, unless stated otherwise

Less than half (43.75 %; 175) were in full health (11111). Few reported extreme or severe problems in any dimension of the EQ- 5D-5L (three was the maximum number of respondents reporting extreme problems in the 'usual activities' dimension; see Table 2).

Table 2: Distribution of EQ-5D-5L responses (N (%)) across participants

	Mobility	Self care	Usual activities	Pain/Discomfort	Anxiety/Depression
Level					
No problems	337 (84.9)	383 (96.5)	352 (88.7)	239 (60.2)	271 (68.3)
Slight problems	35 (8.8)	8 (2)	31 (7.8)	119 (30)	95 (23.9)
Moderate problems	21 (5.3)	5 (1.3)	10 (2.5)	30 (7.6)	22 (5.5)
Severe problems	3 (0.8)	0 (0)	1 (0.3)	8 (2)	9 (2.3)
Unable/Extreme	1 (0.3)	1 (0.3)	3 (0.8)	1 (0.3)	0 (0.0)

Descriptive statistics

The DC responses were 61.7 % for state A and 38.3 % for state B. Reflecting differences in the impact of dimensions and levels on health status, not all choices followed the misery index (sum of the levels across domains) order. For example, the observed probability for choosing state 55534 over state 33355 was 0.852. Only 2.4 % of all respondents thought that state 55534 was WTD and 14.81 % thought that 33355 was WTD (Table 3).

Some inconsistencies were observed in the estimated lead-time TTO valuations. For example, health state 55253 had a lower mean value (-0.4) than health state 55255 (-0.147) (P = 0.0004), even though the latter clearly dominates in term of severity of the five health domains (Table 4). A total of 195 (48.75 %) participants using DC and 216 (54 %) using lead-time TTO rated at least one state as WTD.

Table 3: Discrete choice responses for the 50 paired scenarios included in the valuation exercise

Profile A (Misery index)	Profile B (Misery index)	A (%)	WTD (%) A	WTD (%) B	Profile A (Misery index)	Profile B (Misery index)	A (%)	WTD (%) A	WTD (%) B
11445 (15)	32115 (12)	58.02	2.47	8.64	33223 (13)	21232 (10)	85.54	2.41	7.23
13334 (14)	45441 (18)	19.75	3.70	13.58	33432 (15)	15551 (17)	37.04	2.47	6.17
14122 (10)	54231 (15)	55.42	6.02	25.30	34134 (15)	45325 (19)	93.83	2.47	7.41
14533 (16)	21542 (14)	24.69	3.70	13.58	34255 (19)	35221 (13)	44.74	2.63	9.21
14552 (17)	55325 (20)	93.83	7.41	40.74	35235 (18)	42325 (16)	10.53	0.00	15.79
15351 (15)	14312 (11)	51.32	2.63	14.47	35252 (17)	32254 (16)	33.33	7.41	18.52
15555 (21)	53455 (22)	78.31	6.02	24.10	35312 (14)	14422 (13)	74.36	2.56	20.51
21235 (13)	12243 (12)	24.69	2.47	8.64	41114 (11)	24142 (13)	98.72	3.85	37.18
21445 (16)	55141 (16)	24.36	2.56	24.36	41312 (11)	24253 (16)	37.04	2.47	16.05
21522 (12)	25324 (16)	62.96	9.88	24.69	42122 (11)	31325 (14)	88.46	1.28	10.26
22341 (12)	45145 (19)	74.36	2.56	20.51	42153 (15)	53151 (15)	96.15	1.28	17.95
22544 (17)	35452 (19)	85.19	4.94	16.05	42255 (18)	55524 (21)	48.68	3.95	13.16
23122 (10)	12415 (13)	18.42	1.32	5.26	42441 (15)	21415 (13)	71.08	4.82	12.05
23134 (13)	14314 (13)	85.53	6.58	17.11	43245 (18)	34324 (16)	61.73	2.47	6.17
23231 (11)	25323 (15)	70.37	3.70	27.16	43412 (14)	13342 (13)	51.81	8.43	15.66
23442 (15)	25414 (16)	83.95	3.70	19.75	43514 (17)	23321 (11)	83.33	0.00	6.41
23451 (15)	34354 (19)	79.01	6.17	30.86	44115 (15)	21455 (17)	32.53	9.64	39.76
24453 (18)	41331 (12)	87.65	2.47	30.86	44151 (15)	53242 (16)	75.00	6.58	17.11
25235 (17)	13413 (12)	83.95	2.47	13.58	44234 (17)	33441 (15)	60.24	3.61	21.69
31451 (14)	45431 (17)	80.72	4.82	10.84	45515 (20)	34433 (17)	14.10	5.13	24.36
31452 (15)	13141 (10)	37.04	12.35	32.10	51331 (13)	22421 (11)	85.90	7.69	23.08
31521 (12)	43152 (15)	84.21	0.00	18.42	51552 (18)	35513 (17)	13.25	0.00	7.23
32211 (9)	14211 (9)	88.89	1.23	12.35	54121 (13)	44322 (15)	80.77	1.28	12.82
32241 (12)	51525 (18)	40.79	3.95	17.11	54424 (19)	15321 (12)	67.11	1.32	9.21
33111 (9)	32545 (19)	61.45	10.84	19.28	55534 (22)	33355 (19)	85.19	2.47	14.81

Table 4: Mean lead-time TTO values and percentage of values WTD for the health states included in the valuation exercise

Profile	Value	Std error	WTD (%)	Profile	Value	Std error	WTD (%)	Profile	Value	Std error	WTD (%)	Profile	Value	Std error	WTD (%)
11112	0.786	0.323	4.76	14335	0.041	0.852	18.18	25555	-0.184	0.978	31.82	44415	-0.068	0.700	36.84
11114	0.363	0.614	10.53	14411	-0.006	0.887	33.33	33133	0.483	0.746	9.52	52221	0.503	0.813	11.76
11115	0.075	0.667	27.78	14413	0.081	0.913	33.33	33331	0.263	0.809	10.53	52225	0.379	0.567	19.05
11121	0.629	0.630	10.53	14415	0.264	0.703	11.11	33333	0.470	0.641	10.00	52251	-0.061	0.933	22.73
11122	0.456	0.739	16.67	14441	-0.277	0.920	40.91	33334	0.471	0.365	10.53	52255	-0.038	0.920	33.33
11141	0.335	0.887	17.65	21111	0.664	0.439	0.00	33345	0.008	0.651	25.00	52324	0.161	0.603	31.58
11144	-0.087	0.719	21.05	21112	0.505	0.647	14.29	35251	-0.129	0.790	38.10	52521	-0.216	0.920	47.37
11145	0.274	0.686	33.33	21115	0.326	0.656	23.81	35525	-0.035	0.929	35.00	52525	0.081	0.901	28.57
11211	0.562	0.781	9.52	22251	-0.050	0.998	37.50	41111	0.635	0.492	5.00	52551	-0.608	1.010	65.00
11212	0.422	0.623	8.70	22521	0.224	0.838	26.09	41115	-0.009	0.906	36.36	52555	-0.406	0.826	50.00
11221	0.534	0.572	9.09	22525	0.183	0.815	17.39	41141	0.161	0.566	26.32	53251	0.150	0.630	33.33
11245	-0.053	0.799	38.89	22551	0.036	0.728	16.67	41143	0.266	0.695	21.05	53521	0.093	0.923	22.73
11411	0.571	0.561	14.29	22553	0.253	0.654	16.67	41145	-0.075	0.733	33.33	53555	-0.337	0.964	47.37
11413	0.447	0.746	5.88	22555	-0.463	0.887	56.25	41343	-0.100	0.823	30.00	55221	0.329	0.605	10.53
11415	0.119	0.860	33.33	23255	-0.187	0.623	31.58	41411	0.421	0.365	5.26	55225	-0.197	0.838	44.44
11441	0.075	0.905	35.00	25221	-0.053	0.972	31.25	41413	0.032	0.863	31.82	55235	0.003	0.942	35.00
11445	-0.134	0.778	42.11	25225	-0.113	0.898	40.00	41415	-0.175	0.921	31.25	55251	-0.287	0.950	52.17
12111	0.681	0.536	11.11	25251	-0.110	0.785	42.86	41441	0.184	0.505	26.32	55253	-0.400	1.062	44.44
12112	0.624	0.525	5.26	25255	0.105	0.595	38.10	41445	0.286	0.736	11.11	55255	-0.147	0.888	41.18
14111	0.266	0.536	15.79	25455	-0.053	0.763	31.58	44111	0.059	0.870	31.25	55521	0.167	0.651	23.81
14113	0.328	0.808	15.00	25521	0.389	0.374	5.26	44113	0.256	0.683	11.11	55523	-0.114	0.772	47.62
14115	0.308	0.650	15.79	25525	0.097	0.937	23.53	44115	-0.289	0.987	50.00	55525	-0.337	0.810	42.11
14141	-0.130	0.907	36.36	25531	0.189	0.694	26.09	44141	0.233	0.582	22.22	55551	-0.289	0.857	44.44
14143	0.002	0.903	40.00	25551	0.074	0.676	23.81	44145	-0.215	1.027	39.13	55553	-0.329	0.909	52.38
14145	0.050	0.703	31.58	25553	0.026	0.795	36.84	44411	0.125	0.645	25.00	55555	-0.545	0.935	52.63

* Std error: Standard error

Models

For the estimation of the three DC models, we omitted two respondents from the analysis because their DC choices were always A or always B; the 328 responses without a logical order among state A, state B, and dead were also omitted. For the lead-time TTO model, it was necessary to clean the dataset for inconsistencies. In this case 24 respondents with the same value for all TTO tasks were excluded from the analysis, as were two respondents for whom data were missing due to technical problems.

Several model specifications were explored. However, only main effects models are presented here. The others did not perform better in terms of having fewer inconsistencies or maximizing the likelihood function. In order to allow comparison among the models' coefficients, we present here the rescaled coefficients for the three final DC models. The DC_{WTD} model has the highest likelihood value (-1,401.549), but DC_{TTO} performs better than DC_{dead} (-1,791.37 vs. -2,700.25 respectively) (Table 5).

Regarding the rescaling method for DC models, the value for 55555 was estimated with a lead-time TTO model to be -0.535. This value was used to anchor the DC_{TTO} model, which previously had a value of -5.491 for state 55555. The ratio to rescale the coefficients was abs [(-5.491 - 1)/(-0.535 - 1)] = 4.228. The final rescaled coefficients for DC_{TTO} are $\beta'_j = \beta_j/4.228$. In DC_{dead} models, the dead state has a value of 0. The coefficient for the dead state b_{dead} in the DC_{dead} model is -6.494, since this coefficient must be -1 (meaning that the dead state has a value of 0). The rescaled coefficients are then $\beta'_j = \beta_j/ 6.494$. If the coefficient for the dead state b_{dead} in the DC_{WTD} model is -5.346, then the rescaled coefficients are $\beta'_j = \beta_j/5.346$.

In general, values in the lead-time TTO model were lower than in any of the DC rescaled models due to the estimated intercept value of 0.452. However, there are several inconsistencies for some estimated coefficients. In all of the estimated models, for example, the coefficient for moderate problems (level 3) in the pain/discomfort domain is positive, although not statistically significant. Other inconsistencies are statistically significant: the lower coefficients for slight (level 2) compared to moderate problems (level 3) in the self-care domain for the three DC models and in the mobility and usual-activities domain for DC. The value of the 55555 state in the DC_{dead} model (0.100) was higher than the corresponding value for the DC_{WTD} model (-0.004); however, for both DC_{dead} models, these values were much higher than that in the lead-time TTO model (-0.535).

The two DC dead models are in concordance, with DC_{dead} versus DC_{WTD} having CCC = 0.848, and DC_{TTO} versus lead-time TTO having CCC = 0.725 as well. However, the concordance among the remaining models is lower: (1) DC_{WTD} vs. DC_{TTO} : CCC: 0.677; (2) DC_{dead} versus DC_{TTO}: CCC = 0.478; (3) DC_{dead} versus lead-time TTO: CCC = 0.239; (4) DC_{WTD} vs. lead-time TTO: CCC = 0.349. Compared to DC models, lead-time TTO produced

Table 5: Parameter estimates for the models based on data derived by DC and lead-time TTO

	DC_{dead} Model N = 397 Observation = 21,852			DC_{WTD} Model N=195 Observation = 9,726			DC_{TTO} Model N = 397 Observation = 7,940			lead-time TTO Model N = 373 Observation = 1,864			DC_{dead} rescaled	DC_{WTD} rescaled	DC rescaled
	Coef.	Std. Err.	P>z	Coef.	Std. Err.	P>z	Coef.	Std. Err.	P>z	Coef.	Std. Err.	P>t	Coef.	Coef.	Coef
MO2	-0.365	0.098	0.00	-0.418	0.141	0.00	-0.449	0.108	0.00	-0.042	0.092	0.652	-0.056	-0.109	-0.106
MO3	-0.370	0.093	0.00	-0.446	0.134	0.00	-0.408	0.105	0.00	-0.091	0.131	0.489	-0.057	-0.111	-0.096
MO4	-1.021	0.106	0.00	-1.150	0.154	0.00	-1.115	0.12	0.00	-0.128	0.056	0.022	-0.157	-0.266	-0.264
MO5	-1.445	0.108	0.00	-1.470	0.154	0.00	-1.596	0.127	0.00	-0.251	0.102	0.014	-0.223	-0.331	-0.378
SC2	-0.292	0.091	0.00	-0.298	0.134	0.03	-0.239	0.102	0.02	0.098	0.098	0.319	-0.045	-0.061	-0.057
SC3	-0.273	0.088	0.00	-0.288	0.128	0.02	-0.224	0.098	0.02	0.133	0.123	0.280	-0.042	-0.054	-0.053
SC4	-1.018	0.108	0.00	-1.016	0.153	0.00	-1.118	0.124	0.00	-0.256	0.056	0.000	-0.157	-0.26	-0.265
SC5	-1.041	0.081	0.00	-0.922	0.117	0.00	-1.132	0.089	0.00	-0.042	0.106	0.693	-0.160	-0.226	-0.268
UA2	-0.428	0.085	0.00	-0.488	0.125	0.00	-0.51	0.095	0.00	-0.111	0.08	0.165	-0.066	-0.103	-0.121
UA3	-0.475	0.088	0.00	-0.562	0.126	0.00	-0.479	0.099	0.00	-0.175	0.099	0.076	-0.073	-0.124	-0.113
UA4	-0.781	0.084	0.00	-0.812	0.128	0.00	-0.839	0.092	0.00	-0.124	0.057	0.030	-0.120	-0.145	-0.198
UA5	-0.872	0.094	0.00	-0.757	0.134	0.00	-0.953	0.11	0.00	-0.267	0.087	0.002	-0.134	-0.176	-0.225
PD2	-0.098	0.093	0.29	-0.245	0.137	0.07	-0.034	0.104	0.74	-0.021	0.086	0.811	-0.015	-0.028	-0.008
PD3	0.004	0.097	0.97	-0.115	0.145	0.43	0.091	0.109	0.40	0.036	0.105	0.730	0.001	0.001	0.022
PD4	-0.922	0.107	0.00	-1.003	0.156	0.00	-0.893	0.121	0.00	-0.238	0.056	0.000	-0.142	-0.217	-0.211
PD5	-1.213	0.112	0.00	-1.057	0.157	0.00	-1.441	0.133	0.00	-0.348	0.093	0.000	-0.187	-0.275	-0.341
AD2	-0.398	0.095	0.00	-0.332	0.139	0.02	-0.412	0.104	0.00	0.055	0.092	0.553	-0.061	-0.082	-0.098
AD3	-0.760	0.103	0.00	-0.621	0.143	0.00	-0.819	0.12	0.00	-0.015	0.056	0.790	-0.117	-0.164	-0.194
AD4	-1.079	0.108	0.00	-0.961	0.152	0.00	-1.146	0.126	0.00	-0.161	0.105	0.125	-0.166	-0.246	-0.271
AD5	-1.271	0.103	0.00	-1.164	0.142	0.00	-1.369	0.12	0.00	-0.176	0.041	0.000	-0.196	-0.278	-0.324
Intercept	NA	NA	NA	NA	NA	NA	NA	NA	NA	-0.452	0.066	0.000	NA	NA	NA
DEAD	-6.494	0.187	0.00	-5.346	0.262	0.00	NA	NA	NA	NA	NA	NA	NA	NA	NA
	Log L = -2,700.2528			Log L = -1,401.5487			Log L = -1,791.3742			R-squared = 0.1066					
							Pseudo R2 = 0.2202								

lower values for practically every health state (Fig. 1c, e, f). Both DC_{dead} and DC_{WTD} models estimated very similar values (Fig. 1a).

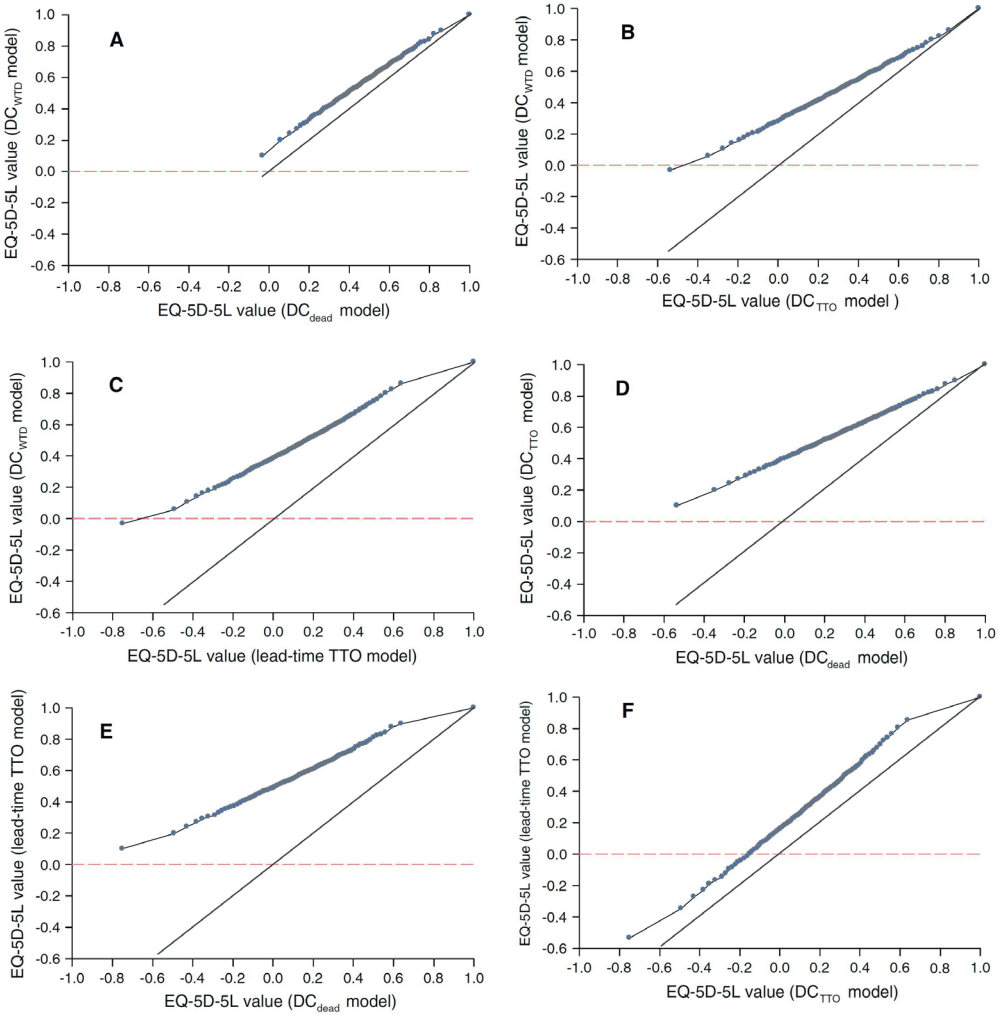

Figure 1: Quantile-Quantile plots for comparison of values obtained from DC_{dead}, DC_{WTD}, DC_{TTO}, and lead-time TTO models

The MSD for differences between the 3,125 states in both DC_{dead} models is 0.009. However, the MSD for the differences with the lead-time TTO model are 0.217, 0.142, and 0.045 for the DC_{dead}, DC_{WTD}, and DC_{TTO} models, respectively. The MSD for the differences with DC_{TTO} are 0.091 and 0.044 for DC_{dead} and DC_{WTD}, respectively.

DISCUSSION AND CONCLUSIONS

In the study reported here we compared two approaches for rescaling DC values on the dead (0)—full health (0) scale to obtain an EQ-5D-5L value set that can be used in economic evaluation. The two approaches were: (1) DC incorporating an additional judgmental task in which the health state 'dead' is assessed against other health states; and (2) a DC model anchoring on lead-time TTO values.

None of the estimated models were completely consistent in terms of regression coefficients. All models had some positive coefficients. Also, to be consistent, a model must meet the condition that each dimension should satisfy an increasing order in the absolute value of the coefficients for each level of severity. According to the results, each of the models did satisfy the condition for some dimensions but not for all. The DC_{TTO} model did not satisfy the condition more often than the DC_{dead} models, and its rescaled results produced higher utility decrements than both rescaled DC_{dead} models. The rescaled DC_{WTD} model differs less from rescaled DC_{TTO} than from rescaled DC_{dead}.

However, we have to take into account that the intercept for the lead-time TTO model was extremely high, which leads to health state values that lack face validity. For example, a person with slight mobility problems has a value of \0.55, which is ridiculous when compared to the previous EQ-5D value set [2–5].

The reason for the inconsistencies in the logistic regression results is not clear. On the one hand, these inconsistencies could be explained by the fact that the DC design included only 50 pairs of health states, which may be inadequate to yield sufficient information (and thus power) to estimate the logistic models (some coefficients were not statistically significant). On the other hand, more power (thus, a larger sample size) may be needed for each pair of health states when the number of pairs is fixed. When the data were applied to the Spanish arm of the multi-country study, the inconsistencies in the DC model disappeared [19]; however that study had both more pairs (200) and more observations per pair. The questions touching upon dead, which are necessary for the DC_{dead} models, were only conducted in the Spanish pilot study. Therefore, the analysis of DC_{dead} models could not be extended to all countries for the sake of comparison. In that light, it would make sense to increase the number of pairs in the DC design that touch upon dead and also to increase the power per pair as this approach would ensure that future studies conducted by using a DC model incorporating dead will be consistent for the whole multi-country dataset.

On comparing the results of the modeling exercise for all participants versus those who rated at least one state as WTD, we found that the DC_{dead} and DC_{WTD} models produced similar results, with the only difference being the position of 'dead'. In particular, we found higher utility decrements and thus lower health state values for EQ-5D- 5L states when the participants who did not rate any state as WTD were removed from the analysis. However,

this may not amount to bias and may simply reflect the preferences of the population. Whatever the reason, the impact on actual results was not large. It should be kept in mind that this was not a direct comparison, as the participants it covered were not identical. From a mathematical point of view and based on the RUT theory, estimation may fail respondents did not answer the TTO part of the exercises appropriately. Some individuals reported that they could not decide when they were indifferent between both lives because they always preferred Life B. This indecisiveness could explain the illogical results obtained with the lead- time TTO model. In general, the respondents needed less assistance on the DC part of the survey, but many did comment on the difficulty of making choices between health states. The difficulties they encountered in the survey tasks emphasize the important role of the face-to-face interviews that are also part of the study design. DC and lead-time TTO elicitation techniques require the respondents to compare health states with 'dead'; this question was posed directly in each of the DC exercises and indirectly in each of the lead-time TTO exercises. From the results we can deduce that a state was more frequently considered WTD in indirect (lead-time TTO) than direct questions (DC ? dead), possibly due to the fact that in lead-time TTO the distinction between negative and positive values was not explicitly made. This fact could explain the lower values observed for the lead-time TTO method and hence the DC_{TTO}.

Previous studies have investigated the incorporation of the health state dead in the DC task [8, 16, 17]. However, none of these used the EQ-5D-5L to allow a direct comparison. Stolk et al. [8] used the classic three-level version of EQ-5D. Our results do not confirm those obtained by Stolk et al., probably because their comparison was made with classic instead of lead-time TTO. Also, the five-level version makes the DC task more complicated for the respondents, and this complexity might have led some participants to make random choices when they could not decide between health states A and B.

DC_{dead} models produce correlated results with slight differences (no bias). Incorporating the health state dead into the general DC technique produces results in concordance with the DC_{TTO}. DC modeling warrants further research to optimize the design if it is to be used to estimate EQ-5D-5L value sets. The lead-time TTO produces very high utility decrements, and its consistency among responses is lower than that of DC models.

Acknowledgments

The authors wish to express their gratitude to the reviewers of the manuscript. They are especially grateful to Paul Krabbe for his helpful comments.

REFERENCES

1. 1. Herdman, M., Gudex, C., Lloyd, A., Janssen, M., Kind, P., Parkin, D., Bonsel, G., Badia, X.: Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual. Life Res. 20(10), 1727–1736 (2011)
2. 2. Dolan, P.: Modeling valuations for EuroQol health states. Med. Care 35(11), 1095–1108 (1997)
3. 3. Shaw, J.W., Johnson, J.A., Coons, S.J.: US valuation of the EQ- 5D health states: development and testing of the D1 valuation model. Med. Care 43(3), 203–223 (2005)
4. 4. Badia, X., Roset, M., Herdman, M., Kind, P.: A comparison of United Kingdom and Spanish general population time trade-off values for EQ-5D health states. Med. Decis. Making 21(1), 7–16 (2001)
5. 5. Lamers, L.M., McDonnell, J., Stalmeier, P.F., Krabbe, P.F., Busschbach, J.J.: The Dutch tariff: results and arguments for an effective design for national EQ-5D valuation studies. Health Econ. 15, 1121–1153 (2006)
6. 6. Craig, B.M., Busschbach, J.J.: Toward a more universal approach in health valuation. Health Econ. 20(7), 864–875 (2011)
7. 7. Brazier, J., McCabe, C.: Is there a case for using visual analogue scale valuations in CUA' by Parkin and Devlin. A response: 'yes there is a case, but what does it add to ordinal data? Health Econ. 16(6), 645–647 (2007). discussion 649-51
8. 8. Stolk, E.A., Oppe, M., Scalone, L., Krabbe, P.F.: Discrete choice modeling for the quantification of health states: the case of the EQ-5D. Value Health 13(8), 1005–1013 (2010)
9. 9. Hakim, Z., Dev, S.P.: Modelling the EuroQol data: a comparison of discrete choice conjoint and conditional preference modelling. Health Econ. 8(2), 103–116 (1999)
10. 10. Salomon, J.A.: Reconsidering the use of rankings in the valuation of health states: a model for estimating cardinal values from ordinal data. Popul Health Metr. 19, 1–12 (2003)
11. 11. McCabe, C., Brazier, J., Gilks, P., Tsuchiya, A., Roberts, J., O'Hagan, A., Stevens, K.: Using rank data to estimate health state utility models. J Health Econ. 25, 418–431 (2006)
12. 12. Ratcliffe, J., Brazier, J., Tsuchiya, A., Symonds, T., Brown, M.: Using DCE and ranking data to estimate cardinal values for health states for deriving a preference-based single index from the sexual quality of life questionnaire. Health Econ. 18, 1261–1347 (2009)
13. 13. Devlin, N., Buckingham, K., Shah, K., Tsuchiya, A., Tilling, C., Wilkinson, G., van Hout, B.: A comparison of alternative variants of the lead and lag time TTO. Health Econ. 22(5),517–532 (2013). doi:10.1002/hec.2819
14. 14. Bliemer, M., Rose, J.M., Hess, S.: Approximation of Bayesian efficiency in experimental choice designs. J. Choice Model. 1(1), 98–127 (2008)
15. 15. Fedorov V. Theory of Optimal Experiments. New York (1972) 16. Coast, J., Flynn, T.N., Natarajan, L., Sproston, K., Lewis, J.,Louviere, J.J., Peters, T.J.: Valuing the ICECAP capability index for older people. Soc. Sci. Med. 67(5), 874–882 (2008)

16. 17. Flynn, T.N., Louviere, J.J., Marley, A.A., Coast, J., Peters, T.J.: Rescaling quality of life values from discrete choice experiments for use as QALYs: a cautionary tale. Popul Health Metr. 22, 6–12 (2008)

17. 18. Lin, L.: A concordance correlation coefficient to evaluate reproducibility. Biometrics 45, 255–268 (1989)

18. 19. Krabbe PFM, Devlin NJ, Stolk EA, Shah KK, Oppe M, Van Hout B, Quik EH, Pickart S, Xie F. Multinational evidence on the feasibility and consistency of the discrete choice model in quantifying health states for the EQ-5D-5L (submitted)

Chapter 7

An EQ-5D-5L value set based on Uruguayan population preferences

Federico Augustovski,
Lucila Rey-Ares,
Vilma Irazola,
Osvaldo Ulises Garay,
Oscar Gianneo,
Graciela Fernández,
Marcelo Morales,
Luz Gibbons,
Juan M. Ramos-Goñi

Qual Life Res. 2016 Feb;25(2):323-33

ABSTRACT

Purpose: To derive a value set from Uruguayan general population using the EQ-5D-5L questionnaire and report population norms.

Methods: General population individuals were randomly assigned to value 10 health states using composite time trade off and 7 pairs of health states through discrete choice experiments. A stratified sampling with quotas by location, gender, age and socio-economic status was used to respect the Uruguayan population structure. Trained interviewers conducted face-to-face interviews. The EuroQol valuation technology was used to administer the protocol as well as to collect the data. OLS and maximum likelihood robust regression models with or without interactions were tested.

Results: We included 794 respondents between 20 and 83 years. Their characteristics were broadly similar to the Uruguayan population. The main effects robust model was chosen to derive social values. Values ranged from -0.264 to 1. States with a misery index = 6 had a mean predicted value of 0.965. When comparing the Uruguayan population with the Argentinian EQ-5D-5L crosswalk value set, the prediction for states which differed from full health only in having one of the dimensions at level 2 were about 0.05 higher in Uruguay. The mean index value, using the selected Uruguayan EQ-5D-5L value set, for the general population in Uruguay was 0.954. In general, older people had worse values and males had slightly better values than females.

Conclusion: We derived the EQ-5D-5L Uruguayan value set, the first in Latin America. These results will help inform decision-making using economic evaluations for resource allocation decisions.

Keywords: Quality of life, EuroQol, Preferences, Value set, Uruguay

INTRODUCTION

Worldwide, healthcare funders are increasingly using economic evaluations to inform their decisions related to health care and the adoption of new technologies. There are many multi-attribute utility-based instruments (MAUI) for the assessment of Quality of Life (QoL), such as the Health Utilities Index (HUI) 3, the Finnish 15D, the SF6D and the EQ-5D, the most widely used MAUI in published cost- utility analyses [1, 2]. The National Institute for Health and Care Excellence (NICE) from England and Wales states that health effects in cost-effectiveness analyses should be expressed in quality-adjusted life years (QALYs), utilities should be based on public preferences and use of EQ-5D is recommend [3].

The EQ-5D is a generic instrument commonly used to measure patient-reported QoL. In order to help inform decision-making in economic evaluations, it is used to assign a preference value to the amount of time living on the reported health status. Many countries, some from Latin America (LA), have derived population value sets for the EQ-5D [4–7].

The classic version of the EQ-5D comprises five dimensions with three severity levels and a visual analogue scale (EQ–VAS) [8]. Dimensions are mobility, self-care, usual activities, pain/discomfort and anxiety/depression. Recently, the EuroQol Group developed a new version of the instrument with the same dimensions but five levels of response instead of the original three in the classic version, recently recalled EQ-5D-3L. The new labels for the response levels range from no problems to unable/extreme problems with three intermediate levels (slight, moderate and severe problems). This new version, called EQ-5D-5L, describes 3125 possible health states, and it was designed to improve the measurement properties of the EQ-5D-3L, reducing the ceiling effect and increasing the discriminatory power [9]. Recent studies confirm the higher discriminatory power and lower ceiling effect for the 5L version as compared to the 3L [10–13]. A five-digit number (one per dimension, on the same order than in the instrument) usually codifies the EQ-5 D health states; this code is usually called state profile. Each digit varies from 1 to 5 representing no problems level to unable/extreme problems level, respectively. For example, the worst state (“pits”) is represented by 55555 while being unable to walk but with no problems in the rest of dimensions is represented by 51111. The misery index is a proxy measure of the severity of the state. It is defined as the addition of the five digits of its profile; for example, the state 23221 has a misery index equal to 10. The EuroQol have recently developed an EQ- 5D-5L valuation study protocol aiming to harmonize 5L valuation studies around the world [14]. To implement and facilitate this standard protocol, the EuroQol group developed specific software, the EuroQol Valuation Technology (EQ-VT) [15].

Uruguay is located in the southern cone of South America and according to the last census has 3,286,314 inhabitants [16]. By the end of 2007, Uruguay started a healthcare reform, encompassing healthcare delivery, financing and management, establishing an Integrated

National Health System with the objective of improving the quality, access and efficiency of healthcare services. Since the reform, the Ministry of Health regularly uses Health Technology Assessment (HTA) and economic evaluations in decision-making process involving the financial coverage and provision of high-cost technologies for the whole population, through the National Resource Fund (Fondo Nacional de Recursos) [17, 18]. Uruguay had no previous value set, and it is the first country in LA to undertake a general population valuation study for the EQ- 5D-5L instrument. The objectives of our study were to obtain social preferences and derive the value set from Uruguayan general population using the EQ-5D-5L questionnaire, as well as to report the population norms.

METHODS

Protocol

We used a standardized interview protocol developed by the EuroQol group based on the obtained evidence from a set of conducted pilot studies [19]. The protocol consists in five different sections: (1) a general welcome, (2) introduction to the research and completion of background information, (3) the set of composite time trade-off (C- TTO) tasks, (4) the set of discrete choice (DC) experiment tasks and (5) general thank you and goodbye. We also collected socio-demographic information and health literacy measurement using the Short Assessment of Health Literacy-Spanish questionnaire (SAHL-S).

Eliciting preferences methods

Traditionally, EQ-5D-3L valuation studies were mainly based on time trade-off (TTO) methods [4]. However, the TTO version used in the "Measurement and Value of Health" protocol [20] had some problems, especially with the transformation of values from states considered to be worse than dead [21, 22]. Lead-Time TTO, Lag-Time TTO [23] and C-TTO [24], more recent TTO versions, were tested on several pilot studies. Based on the results of those studies, the EuroQol Group concluded that the EQ-5D-5L protocol should include the C-TTO version [14].

The conventional TTO approach has two different tasks, one for states considered to be better than death (BTD) and another one for states considered to be worse than death (WTD). For the valuation of states BTD, respondents are required to choose between living 10 years in a specific health state (life B) or X years in full health (life A). The amount of time X in life A is varied between 0 and 10 years. For states WTD, the procedure is conceptually and operationally different, and participants have to choose between dying immediately or live X years in a specific state followed by 10–X years in full health (life B).

Lead-time TTO and lag-time TTO methods add extra time in full health to the health

state to be valued (Life B). The lead-time approach used in this protocol places the extra time before, being the options living 10 years in full health and later 10 years in the given health state or X amount of time (between 0 and 20 years) in full health in life A. The main characteristic of these variants is that the iterative trading process allows the participant to move between negative and positive values without explicitly thinking about whether the state is worse or better than being dead.

C-TTO involves the use of traditional TTO approach for states considered to be BTD, and a lead-time TTO for states considered to be WTD, combined in a unique task. So, when the participant exhausts the 10 years in full health in the traditional TTO task and does not want to spend any time in full health as the evaluated state is very bad (WTD) he is switched to the lead-time TTO component.

The interview protocol included 86 EQ-5D-5L health states (selected using Monte Carlo simulation [14]) to be evaluated using C-TTO divided in ten blocks with similar representation of all severity levels. All the blocks included one very mild state with only one health state with mild problems (i.e. 21111) and the pits state (55555). Respondents were randomly assigned to one of the ten C-TTO blocks. The presentation order of the states within each block was also randomly generated by the EQ-VT.

The protocol included a DC experiment as a secondary valuation technique [14]. A DC experiment is an ordinal elicitation technique that has received recent attention for eliciting EQ-5D-5L values [25, 26]. DC experiments require individuals to make a pairwise comparison between two different scenarios, being in our case two EQ-5D-5L health states. The protocol included 196 DC pairs of EQ- 5D-5L health states divided in 28 blocks of seven pairs with similar misery index. Respondents were randomly assigned to one DC block, and the order of each pair and its position on the screen (i.e. left or right) were also randomly generated by the EQ-VT.

Quality control

The initial EQ-5D-5L valuation studies found some interviewer effects that could affect data quality, as it is reported by Ramos-Goñi et al. [27]. For further valuation studies, the EuroQol Group decided to create a quality control tool to monitor interviewer performance. This tool mainly evaluates interviewer protocol compliance through four key parameters of the C-TTO task: the full explanation and the time spent on the wheel chair example, the time used to complete the 10 C-TTO tasks and the presence of large inconsistencies. The DC experiments section was monitored to detect unusual response patterns (i.e. AAAAAAA, ABABABA). Interview quality was checked weekly. Based on weekly results, we decided whether we had to retrain or drop interviewers from the team. After quality control analysis, researchers (FA, UG and LRA) gave feedback and retrained interviewers when necessary.

Sampling and data collection

Uruguay was geographically stratified. The study took place in the following locations: Uruguay capital city, Montevideo; and the departments of Maldonado and Paysandú. Quotas by location, gender, age and socio-economic status replicated the Uruguayan population structure [16].

The EuroQol Group recommends including 10,000 C-TTO responses in the valuation studies. Since each participant values 10 health states, the initial sample size was 1000 individuals. The power calculation was based on precision requirements for the estimation of the C-TTO mean [27].

Twenty-one trained interviewers administered the questionnaire using the EQ-VT. During the face-to-face interviews, respondents had the control of the computer most of the time and the interviewers were available to assist and monitor the process.

The valuation exercise started explaining the objectives of the research, then the respondents' filled out the EQ-5D-5L and rated their current health state using the EQ–VAS. Additionally, they gave background information (age, sex, educational level and their experience with illness). Prior to completion of the C-TTO tasks, participants received an explanation (using as an example a life living in wheel chair) and completed three mock states of different severity in order to verify their understanding. Later on, they completed 10 C-TTO and 7 DC experiments. Upon completion of the tasks, participants answered follow-up questions related to the difficulty and comprehension.

Once both tasks and the follow-up questions were completed, the interviewers asked the respondents to complete the Short Assessment of Health Literacy–Spanish (SAHL-S) instrument [28]. This questionnaire evaluates health literacy through 18 multiple-choice questions combining word recognition and comprehension. Low health literacy is defined by identifying 14 or fewer correct items. Respondents were asked to read aloud 18 medical terms (word recognition), and the interviewer assessed comprehension through the multiple-choice question. Health literacy is a construct that reflects the capacity to obtain, process and understand health information and services needed to make appropriate health decisions [29].

Statistical analysis

We describe the sample characteristics using means and standard deviations for continuous variables and percentages for discrete variables. The EQ-5D-5L descriptive system from the recruited sample is presented by age group.

Valuation data from C-TTO tasks and data from DC experiments were available from the collected data. We initially tried to follow the hybrid approach reported by Ramos-Goñi et al. [27]. However, in our case, the DC models had several logical inconsistencies, leading us to

base our analysis on the C-TTO data only. The DC and hybrid analyses are available from the authors upon request.

We started the C-TTO analysis using the classic ordinary least square (OLS) model. However, our data had problems that prevented us from using this approach: heteroscedasticity and significant outliers. Thus, we opted for using robust regression [30]. This regression method basically applies a different weight to each observation based on how far away it is located from the median of the population sample. In this way, the impact of the outliers is reduced and the heteroscedasticity problem is addressed. We used a tuning constant based on the bi-weight function to calculate each weight in the model. A tuning constant of 7 is usually used, in order to confer similar model efficiency as the OLS model, and assuming no heteroscedasticity or outliers [31]. We set this value to 8.5 to include a broader range of values for each respondent, without losing the logical consistency of the model. We analysed the distribution of weights according to the misery index value in order to explore the contribution of different groups of observations to the final model estimations.

For the model specification, we started with a 20-parameter main effects model, using the response values as dependent variables and health states as explanatory variables. We created a dummy variable Dij indicating whether the dimension i is at level j. For example, we created variables MO2, MO3, MO4 and MO5 for mobility dimension, indicating whether the mobility dimension is at level 2 or 3 or 4 or 5, respectively. Similar sets of variables were created for each dimension. In order to explore alternative model specifications and performance, we added interaction terms to the main effects model. We evaluated traditional Nj terms (1 if at least one dimension is at level j) and Nij terms (1 if at least one dimension is at level i or j). We also tested the following interactions terms: (1) D1, number of movements away from full health beyond the first; (2) I2, number of dimensions at level 2 or 3 beyond the first; (3) C3, number of dimensions at level 3, 4 or 5 beyond the first; (4) K45, number of dimensions at level 4 or 5; (5) I45, number of dimensions at level 4 or 5 beyond the first; (6) O2, 1 if all dimensions are at level 1 or 2; (7) Z2, 1 if at least one dimension is at level 2 or 3 and one is at level 4 or 5; and (8) Z3, number of dimensions at level 2 or 3 given that at least one dimension was at level 4 or 5. We used a stepwise approach to decide whether to keep the interaction terms in the model or leave them out.

In this manuscript, we present three models: (1) the OLS model for comparison purpose, (2) the main effects robust model and (3) the robust model including best interaction terms. All statistical analyses were performed on STATA 11 MP [26], using the "regress" command for OLS and "rreg" for the robust regression.

Expanding the modelling exercise, we have performed additional analyses to check differences on preference values by educational level. We have performed an ANOVA test for crude preferences and we have also added dummies by levels of education on the final estimation of the model. None of these results were statistically significant (data not shown).

Exclusion criteria

Interviewers with low-quality performance, fulfilling prespecified criteria regarding interview quality (i.e. too little time spent explaining the task, no explanation of the lead time section, C-TTO responses with clear inconsistencies or too little time to perform all TTO tasks) were excluded during the data collection process. We also excluded respondents meeting two additional criteria: (1) having a positive slope on the relationship between their values and the misery index of the health states. This means, that the respondent poorly understood the task, as he/she provided higher utility values for worse health states; and (2) respondents who valued all states at the same value, except non-traders (i.e. subjects who value all states as 1).

Model performance

We used four criteria to evaluate the performance of the model: (1) logical consistency of parameters, (2) goodness of fit, (3) prediction accuracy and (4) parsimony. A set of model parameters is said to be logically consistent if predictions for logically better health states (ex: 12111 is logically better than 13111) are higher than the predictions for logically worse health states. In our models, it means that MO2 \ MO3 \ MO4 \ MO5, and so on for SC, UA, AD and PD dimensions. We used Akaike information criterion (AIC) and Bayesian information criterion (BIC) to evaluate goodness of fit, adjusted by the number of model parameters. We calculated the mean square error (MSE) and the mean absolute error (MAE) to evaluate prediction accuracy. The principle of parsimony stated that when two competing models are similar in terms of performance parameters, the simplest model should be selected. These four criteria were used to compare different model specifications using different interaction terms.

Comparison between predicted values from different models and to the Argentinian crosswalk 5L value set

In order to compare predictions form different models, we calculated the estimated values from robust models and compared them with the weighted means of the 86 TTO health states included. Those predictions were also compared to the ones from the crosswalk 5L value set from Argentina, (the 3L value set that uses a mapping function between the 3L and the 5L versions). Because of the EQ- 5D-5L crosswalk value set for Argentina has not been previously estimated [32], it was calculated specifically for this study following the methodology proposed by van Hout et al. [33].

RESULTS

Study recruitment took place between October 2013 and June 2014. We started the field work with 21 interviewers, who were trained, and evaluated on a weekly basis at the beginning of the data collection phase. Based on the quality control analysis, we decided to keep 11 interviewers with good performance, which resulted in the exclusion of 220 interviews conducted at this stage by 10 poor-quality interviewers. Excluded participants had similar age, gender and educational level than the remaining sample. We periodically analysed the remaining interviews during data collection and decided to stop data collection when our analysis showed the robustness of the results. The study sample had 805 respondents between 20 and 83 years old. Eleven subjects met exclusion criteria, leaving 794 subjects in the final sample. Sample characteristics were similar to the Uruguayan population in terms of gender. However, younger as well as higher educated categories were slightly over-represented in our sample, though utility values did not significantly differ by educational level. Nearly 44 % of the population had low health literacy despite the fact that more than 80 % had educational attainment of at least some secondary level (see Table 1).

Table 1: Study sample and Uruguayan general population characteristics

	Study sample (794)		Uruguayan population (Census 2011**)
	(n)	%	%
Socio-demographic characteristics			
Age group (years)			
20-39	386	48.6	42.4
40-59	271	34.1	35.2
60+	137	17.3	22.4
Female	439	55.3	52.0
Educational attainment *			
Primary level	137	17.3	36.2
Secondary level	406	51.3	44.5
Tertiary level	249	31.4	18.5
Low Health literacy *	346	43.7	
Experience with serious Illness			
in itself	180	22.7	
in family	450	56.7	
in caring for others	425	53.5	
SALHS score – Mean (SD) *	14.3 (2.5)		

*Not all responses for these questions complete. We estimate the denominator with the non-missing values. ** Percentages for the age categories were calculated using the population between 20 and 79 years for the ease of comparison with our sample, and the census estimations for educational level comprise inhabitants older than 25 years of age. SALHS: Short Assessment of Health Literacy-Spanish questionnaire; SD: Standard Deviation

Forty-four per cent of the sample reported no problems on any dimension of self-reported EQ-5D-5L. Older respondents reported more problems in all dimensions, and the mean self-reported VAS also decreased with increasing age, and was smaller in women (Table 2).

Table 2: Self-reported health using EQ-5D-5L descriptive system and EQ VAS

Age		20-39		40-59		60+		Total	
		n	%	n	%	n	%	n	%
Mobility	No problems	367	95.10%	214	79.00%	84	61.31%	665	83.80%
	Slight problems	16	4.10%	39	14.40%	31	22.63%	86	10.80%
	Moderate problems	2	0.50%	13	4.80%	16	11.68%	31	3.90%
	Severe problems	1	0.30%	5	1.80%	6	4.38%	12	1.50%
	Unable to walk	0	0.00%	0	0.00%	0	0.00%	0	0.00%
Self-Care	No problems	383	99.20%	257	94.80%	124	90.51%	764	96.20%
	Slight problems	2	0.50%	9	3.30%	8	5.84%	19	2.40%
	Moderate problems	0	0.00%	2	0.70%	4	2.92%	6	0.80%
	Severe problems	1	0.30%	3	1.10%	1	0.73%	5	0.60%
	Unable to	0	0.00%	0	0.00%	0	0.00%	0	0.00%
Usual activities	No problems	365	94.60%	234	86.40%	104	75.91%	703	88.50%
	Slight problems	15	3.90%	18	6.60%	25	18.25%	58	7.30%
	Moderate problems	5	1.30%	14	5.20%	6	4.38%	25	3.10%
	Severe problems	0	0.00%	2	0.70%	1	0.73%	3	0.40%
	Unable to	1	0.30%	3	1.10%	1	0.73%	5	0.60%
Pain/ Discomfort	No problems	275	71.20%	166	61.30%	73	53.28%	514	64.70%
	Slight problems	87	22.50%	67	24.70%	41	29.93%	195	24.60%
	Moderate problems	19	4.90%	20	7.40%	14	10.22%	53	6.70%
	Severe problems	4	1.00%	16	5.90%	9	6.57%	29	3.70%
	Extreme problems	1	0.30%	2	0.70%	0	0.00%	3	0.40%
Anxiety/ Depression	No problems	252	65.30%	170	62.70%	94	68.61%	516	65.00%
	Slight problems	91	23.60%	63	23.30%	27	19.71%	181	22.80%
	Moderate problems	35	9.10%	24	8.90%	10	7.30%	69	8.70%
	Severe problems	7	1.80%	13	4.80%	5	3.65%	25	3.10%
	Extreme problems	1	0.30%	1	0.40%	1	0.73%	3	0.40%
Visual Analogue Scale	Mean	83.19		78.49		71.81		79.63	
	Standard error	0.68		1.04		1.66		0.58	
	25th Percentile	80		70		55		70	
	50th Percentile (median)	85		80		75		80	
	75th Percentile	90		90		90		90	

The results from the OLS model showed a logical inconsistency between the coefficients associated with slight and moderate problem in usual activities dimension. This inconsistency was not observed when using the same model specification with a robust regression approach. Both robust model estimations reported that the main effects and the main effects with interactions were logically consistent. The goodness of fit of both robust models was similar, gaining only 0.4 % of relative improvement in AIC or BIC with the interaction terms. Similarly, the prediction accuracy of the main effect robust model was similar to the main effect model with interaction (MSE = 0.002) (Table 3).

Table 3. Model estimations: comparison of OLS, main effect robust model, and main effect with interaction terms.

	OLS Model N=794			*(Final Value set)* Robust Estimation Model N=794 (Tune = 8.5)			Robust estimation model with interactions N=794 (Tune = 8.5)		
	Coefficient	Std. error	P-value	Coefficient	Std. error	P-value	Coefficient	Std. error	P-value
MO2	0.0767	0.019	0.00	0.0140	0.016	0.37	0.0514	0.017	0.00
MO3	0.1019	0.020	0.00	0.0322	0.016	0.05	0.0857	0.020	0.00
MO4	0.1906	0.022	0.00	0.1077	0.018	0.00	0.0930	0.020	0.00
MO5	0.3435	0.020	0.00	0.2987	0.016	0.00	0.2688	0.021	0.00
SC2	0.0249	0.019	0.18	0.0256	0.015	0.09	0.0529	0.015	0.00
SC3	0.0820	0.021	0.00	0.0609	0.017	0.00	0.0875	0.021	0.00
SC4	0.1282	0.021	0.00	0.1169	0.017	0.00	0.0801	0.021	0.00
SC5	0.2616	0.019	0.00	0.2734	0.016	0.00	0.2120	0.023	0.00
UA2	0.0710	0.020	0.00	0.0424	0.016	0.01	0.0691	0.016	0.00
UA3	0.0512	0.021	0.01	0.0455	0.017	0.01	0.0714	0.018	0.00
UA4	0.1303	0.021	0.00	0.1183	0.017	0.00	0.0905	0.019	0.00
UA5	0.2101	0.019	0.00	0.2315	0.016	0.00	0.1775	0.022	0.00
PD2	0.0260	0.018	0.14	0.0171	0.014	0.23	0.0450	0.015	0.00
PD3	0.0820	0.021	0.00	0.0607	0.017	0.00	0.0974	0.022	0.00
PD4	0.2160	0.019	0.00	0.1870	0.015	0.00	0.1511	0.022	0.00
PD5	0.2833	0.021	0.00	0.2705	0.017	0.00	0.2184	0.022	0.00
AD2	0.0320	0.020	0.11	0.0095	0.016	0.55	0.0329	0.014	0.00
AD3	0.0884	0.022	0.00	0.0435	0.018	0.01	0.0885	0.019	0.02
AD4	0.1509	0.021	0.00	0.1043	0.017	0.00	0.0832	0.018	0.00
AD5	0.1809	0.019	0.00	0.1771	0.016	0.00	0.1381	0.019	0.00
Const.	0.0104	0.020	0.61	0.0126	0.016	0.44	-	-	-
D1							-0.0192	0.015	0.00

	OLS Model N=794			*(Final Value set)* Robust Estimation Model N=794 (Tune = 8.5)			Robust estimation model with interactions N=794 (Tune = 8.5)		
	Coefficient	Std. error	P-value	Coefficient	Std. error	P-value	Coefficient	Std. error	P-value
I45^2							0.0140	0.002	0.19
LogL	-5707.448			-3290.65			-3274.3		
AIC	11456.9			6623.3			6592.6		
BIC	11603.5			6769.9			6746.1		
MSE	0.003			0.002			0.002		
MAE	0.04			0.03			0.03		
U(55555)	-0.28			-0.26			-0.28		

Bold values indicate logical inconsistencies. Parameter abbreviations are described in the text. LogL: Log likelihood; AIC: Akaike information criterion; BIC: Bayesian information criterion; MSE: mean standard error; MAE: mean absolute error; U(55555): utility value of the pits state.

Based on the similarities of these parameters and taking into account the parsimony criterion, we chose the robust estimation of the main effects model as the most appropriate model for the Uruguayan value set. The technical appendix shows how to estimate an individual state value, as well as the code to do it in Stata. In the electronic supplementary material, we report the utility values for the 3125 different states.

The weights applied to the responses in the robust model range from 0 to 1. The distribution of weights, according to different values of the misery index, showed how the respondents' opinions have more discrepancies as the misery index increases (Table 4). For example, for a misery index = 6, the 10th percentile is 0.992, meaning that 90 % of the responses have almost the same impact in the model estimations (i.e. weight close to 1). However, the 10th percentile for those states with a misery index = 21 is 0.449, meaning that at least 10 % of the responses for these states will be considered as half important in the model estimations (i.e. weight close to 0.5). The selected model predicts index values that range from -0.264 to 1. States with a misery index = 6 had a mean predicted value of 0.965 (Table 4). The predictions from the model with interactions are slightly lower at the top and bottom of the scale than predictions form the main effects model, i.e. it has a lowest prediction of -0.288 for the pits state and has a mean prediction for misery index 6 states of 0.95. When comparing the prediction for a misery index = 6 in the Uruguayan population with the Argentinian EQ-5D-5L crosswalk value set, the Uruguay values are about 0.05 higher, and they are also higher than the Argentinian values across the misery index spectrum.

Table 4: Model predictions comparison, and with the crosswalk 3L Argentinean value set

Misery index	Percentile 10% for Weights	Weighted observed TTO values	Robust Estimation Model	Robust estimation model with interactions	3L Argentinean value set by cross-walk mapping
6	0.992	0.953	0.965	0.950	0.906
7	0.987	0.927	0.947	0.925	0.845
9	0.966	0.855	0.878	0.871	0.750
10	0.942	0.796	0.827	0.837	0.732
11	0.938	0.761	0.801	0.790	0.689
12	0.943	0.744	0.732	0.741	0.517
13	0.895	0.666	0.650	0.667	0.454
14	0.819	0.600	0.593	0.608	0.407
15	0.860	0.557	0.544	0.558	0.403
16	0.843	0.532	0.514	0.520	0.376
17	0.848	0.449	0.462	0.457	0.421
18	0.730	0.358	0.340	0.351	0.229
19	0.758	0.354	0.285	0.296	0.164
20	0.735	0.215	0.252	0.224	0.175
21	0.449	0.095	0.161	0.143	0.082
22	0.613	0.150	0.140	0.138	0.069
25	0.707	-0.300	-0.264	-0.288	-0.376

The mean index values, using the selected Uruguayan EQ-5D-5L value set, for the general population in Uruguay is 0.954. Older people have worse health-related quality of life for all paired comparisons (highest P value \ 0.001). Males had slightly higher values than females (Table 5), but this difference was not significant.

Table 5: Population norms for EQ-5D-5L in Uruguay

EQ-5D-5L (Index values)	Age	20-39	40-59	60+	Total
Total sample	**Mean**	0.972	0.942	0.93	0.954
	Standard error	0.003	0.006	0.009	0.003
	25th Percentile	0.961	0.927	0.904	0.945
	50th Percentile (median)	1	0.973	0.961	0.978
	75th Percentile	1	1	1	1
Males	**Mean**	0.974	0.951	0.957	0.963
	Standard error	0.003	0.01	0.009	0.004
	25th Percentile	0.967	0.963	0.94	0.961
	50th Percentile (median)	1	0.978	0.978	0.978
	75th Percentile	1	1	1	1
Females	**Mean**	0.97	0.935	0.907	0.947
	Standard error	0.004	0.007	0.014	0.004
	25th Percentile	0.961	0.916	0.867	0.935
	50th Percentile (median)	1	0.97	0.947	0.978
	75th Percentile	1	1	0.989	1

DISCUSSION

The EQ-5D-5L is a recently developed instrument. Only a few countries conducted valuation studies, and only the Spanish study was recently published, though authors did not recommend the use of the reported value set [27]. This is to our knowledge the first study that provides a population-based EQ-5D-5L value set in Latin America.

The choice of a robust estimation for modelling the C-TTO data is also a novel approach in EQ-5D valuation exercises. While other studies present models mainly based on OLS or random effects estimations [4], we selected a robust estimation based on the observed between respondent variability. Some extreme differences in opinions in our sample made some of the OLS or random effects coefficients to be logically inconsistent, due to the significant heteroscedasticity and the presence of outliers. Robust regression tries to solve these issues by weighting opinions less strongly if they are extreme. Extreme opinions have less impact depending on how extreme they are, and the values close to the majority (median) have the greatest weight. In our estimation, we relaxed the robust condition of the estimation as much as possible, stopping when inconsistencies in coefficients were found. We included

all responses from the sample, as there is no 0 weight for any response. This was a balance point we chose in order to have logical results but also to incorporate everybody's opinions.

Based on the parsimony criteria, we think that a more complex model should be preferred only when the improvement compared to a simpler one is large enough to overcome the complexity. In our estimation, the improvement of the best model with interaction terms tested was marginal compared to the main effects model. That led us to prefer the main effects model instead of the model with the interaction terms.

Given the fact that there is not any EQ-5D-5L value set currently recommended in the literature, and no previous EQ-5D-3L value set was available for Uruguay, we decided to compare our results with the EQ-5D-5L crosswalk value set derived from the original 3L set for Argentina [5], being a close country with similar socio-economic characteristics. In our selected model, we had slightly higher values in Uruguay in the entire severity spectrum. Taking into account the changes on the descriptive system of the 3L and 5L versions of EQ-5D, [9] it was something expected. For example, the levels for the misery index 6 states on the 5L version are by definition less severe ("slight") compared to the same levels in the 3L version ("some"). Also, the higher observed index value for the pits state can be explained, as the level for the mobility dimension has changed from "confined to bed" in the 3L version, to "unable to walk" in the 5L version, making the description of the pits state (55555) in the 5L version better than the corresponding state (33333) in the 3L version. As both anchor values have been moving up in our estimations, it is expected that the whole scale move up according these anchors. Population norms derived for Uruguay showed to be consistent and similar to international population norms previously published [34].

One limitation of this study is the use of a quota (i.e. non-probabilistic) sample. Though our sample was broadly representative of the socio-demographic characteristics of the Uruguayan population, younger and higher educated individuals were slightly over-represented. However, the age difference was small. Additionally, although the proportion of participants with tertiary education was slightly higher in the sample compared with national data, utility values did not significantly differ according to educational level. Another limitation is the fact that we have not used the information from the DC experiment. The recently published study from Spain reported the feasibility of obtaining an EQ-5D-5L value set using a hybrid approach, combining the C-TTO and DC data [27]. However, in our initial analysis, when we tested our DC models, they showed several inconsistencies, and these could not be solved through a hybrid modelling approach. In addition, we have seen in our C-TTO responses some extreme differences in opinions, mainly regarding severe health states. This fact could also explain the inconsistencies found in the DC models. As far as we are aware, there is no available "robust" estimation method for analysing DC data, limiting our capacity to include this information in our estimations. Another limitation of this study is the fact

that we have not performed an internal or an external validation of our predictions. Given the final sample size of our study, which was somewhat smaller than originally intended, and that the requirements for internal validation reduce statistical power (i.e. randomly splitting the sample and evaluate how the model derived in the estimation set applies to the validation set), we included all responses in the model estimation.

We obtained the EQ-5D-5L value set that will be implemented in Uruguay, which is the first country in Latin America to undertake such a study. The use of these values will help researchers, in Uruguay and eventually in other similar socio-economic countries, in conducting cost-utility studies based on the specific preferences of the general population to inform decision makers' resource allocation decisions.

Acknowledgments

Special thanks to Arnd Jan Prause for the support and Elka Pérez and Gastón Díaz from "Equipos Mori" for their great work and commitment.

Technical Appendix

In this manuscript, the value set for Uruguay has been presented (see Table 3). This appendix describes how to obtain the utility value for a specific health state. Notice that the model coefficients should be interpreted as the disutility of moving from having no problems in that particular domain (level 1) to the specific level of response of each domain.

Given the profile of a specific health state, $L_{MO}L_{SC}\ L_{UA}L_{PD}L_{AD}$, and given the final model and coefficients to derive them, the formula to obtain the utility value for each health state is as follow:

$U(L_{MO}L_{SC}L_{UA}L_{PD}L_{AD}) = 1 - MO(L_{MO}) - SC(L_{SC}) - UA(L_{UA}) - PD(L_{PD}) - AD(L_{AD})$—Deviation from full health.

Where $U(L_{MO}L_{SC}L_{UA}L_{PD}L_{AD})$ denotes the utility for the state $L_{MO}L_{SC}L_{UA}L_{PD}L_{AD}$, L_{MO} denotes the response level on mobility domain, $MO(L_{MO})$ denotes the coefficient of the level L_{MO} on mobility domain (and the same for rest of domains), and Deviation from full health is the model constant. When the level of a given domain is no problems (1), the coefficient of that domain is 0. As there is no movement from no problems, no disutility is associated.

Example 1

U(25413) = 1 - MO2 – SC5 - UA4 - PD1 (=0) - AD3-Deviation from full health = 1 - 0.0140 - 0.2734 - 0.1183 - 0 - 0.0435 - 0.0126 = 0.5382

Example 2

U(31412) = 1 - MO3 - SC1 (=0) - UA4 - PD1 (=0) - AD2-Deviation from full health = 1 - 0.0322 - 0 - 0.1183 - 0 - 0.0095 - 0.0126 = 0.8274

Example 3

U(11111) = 1 - MO1 (=0) - 0 (SC1) - UA1(=0) - PD1 (=0) - AD1 (=0) = 1
(Notice that 11111 represents full health, so the deviation from full health is not applicable here).

Stata code

```
//This code calculates the utility values for a given data set
//The variable representing mobility domain has to be named MO, SC for self-care, UA for usual activities, PD for pain/discomfort and AD for anxiety/depression
gen Utility = 1.
recast double Utility
//MO
replace Utility = Utility - 0.0140 if MO == 2 replace Utility = Utility - 0.0322 if MO == 3
replace Utility = Utility - 0.1077 if MO == 4 replace Utility = Utility - 0.2987 if MO == 5
//SC
replace Utility = Utility - 0.0256 if SC == 2 replace Utility = Utility - 0.0609 if SC == 3 replace Utility = Utility - 0.1169 if SC == 4 replace Utility = Utility - 0.2734 if SC == 5
//UA
replace Utility = Utility - 0.0424 if UA == 2 replace Utility = Utility - 0.0455 if UA == 3
replace Utility = Utility - 0.1183 if UA == 4 replace Utility = Utility - 0.2315 if UA == 5
//PD
replace Utility = Utility - 0.0171 if PD == 2 replace Utility = Utility - 0.0607 if PD == 3 replace Utility = Utility - 0.1870 if PD == 4
replace Utility = Utility - 0.2705 if PD == 5
//AD
replace Utility = Utility - 0.0095 if AD == 2 replace Utility = Utility - 0.0435 if AD == 3
replace Utility = Utility - 0.1043 if AD == 4 replace Utility = Utility - 0.1771 if AD == 5
//Deviation from full health
replace Utility = Utility - 0.0126 if (MO ! = 1 | SC ! = 1 | UA ! = 1 | PD ! = 1 | AD ! = 1)
```

REFERENCES

1. 1Hawthorne, G., Richardson, J., & Day, N. A. (2001). A comparison of the Assessment of Quality of Life (AQoL) with four other generic utility instruments. Annals of Medicine, 33(5), 358–370.
2. Brauer, C. A., Rosen, A. B., Greenberg, D., & Neumann, P. J. (2006). Trends in the measurement of health utilities in published cost-utility analyses. Value Health, 9(4), 213–218.
3. Guide to the methods of technology appraisal 2013. (2013). Process and methods guides. National Institute for Heatlh and Care Excellence (NICE): UK.
4. Szende, A., Oppe, M., & Devlin, N. (2007). EQ-5D value sets: Inventory, comparative review and user guide (Vol. 2). Dordrecht: Springer.
5. Augustovski, F. A., Irazola, V. E., Velazquez, A. P., Gibbons, L., & Craig, B. M. (2009). Argentine valuation of the EQ-5D health states. Value Health, 12(4), 587–596.
6. Zarate, V., Kind, P., Valenzuela, P., Vignau, A., Olivares-Tirado, P., & Munoz, A. (2011). Social valuation of EQ-5D health states: The Chilean case. Value Health, 14(8), 1135–1141.
7. Viegas Andrade, M., Noronha, K., Kind, P., Maia, A. C., Miranda de Menezes, R., De Barros Reis, C., et al. (2013). Societal preferences for EQ-5D health states from a Brazilian population survey. Value in Health Regional Issues, 2(3), 405–412.
8. EuroQol, G. (1990). EuroQol–a new facility for the measurement of health-related quality of life. Health Policy, 16(3), 199–208.
9. Herdman, M., Gudex, C., Lloyd, A., Janssen, M., Kind, P., Parkin, D., et al. (2011). Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Quality of Life Research, 20(10), 1727–1736.
10. Pattanaphesaj, J., & Thavorncharoensap, M. (2015). Measurement properties of the EQ-5D-5L compared to EQ-5D-3L in the Thai diabetes patients. Health Qual Life Outcomes, 13(1), 14.
11. Golicki, D., Niewada, M., Buczek, J., Karlinska, A., Kobayashi, A., Janssen, M. F., & Pickard, A. S. (2015). Validity of EQ-5D-5L in stroke. Quality of Life Research, 24(4), 845–850.
12. Conner-Spady, B. L., Marshall, D. A., Bohm, E., Dunbar, M. J., Loucks, L., Khudairy, A. A., & Noseworthy, T. W. (2015). Reliability and validity of the EQ-5D-5L compared to the EQ-5D-3L in patients with osteoarthritis referred for hip and knee replacement. Quality of Life Research, 24(7), 1775–1784.
13. Greene, M. E., Rader, K. A., Garellick, G., Malchau, H., Freiberg, A. A., & Rolfson, O. (2014). The EQ-5D-5L Improves on the EQ-5D-3L for health-related quality-of-life assessment in patients undergoing total hip arthroplasty. Clinical Orthopaedics and Related Research. doi:10.1007/s11999-014-4091-y.
14. Oppe, M., Devlin, N. J., van Hout, B., Krabbe, P. F., & de Charro, F. (2014). A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. Value Health, 17(4), 445–453.

15. EuroQol Valuation Technology. (2011). Retrieved January, 2015, from https://www.valuationstudy.org.

16. Censo. (2011). Retrieved September 2014, from http://www.ine. gub.uy/censos2011/index.html.

17. Modificación de la integración de la comisión asesora del Formulario Terapéutico Nacional. (2010). Decreto N 04/010 (2010th ed.). Montevideo: Presidencia Repu´blica Oriental del Uruguay.

18. Fondo Nacional de Recursos. Retrieved January 2015, from http://www.fnr.gub.uy.

19. Devlin, N. J., & Krabbe, P. F. (2013). The development of new research methods for the valuation of EQ-5D-5L. The European Journal of Health Economics, 14(Suppl 1), S1–S3.

20. Dolan, P. (1997). Modeling valuations for EuroQol health states. Medical Care, 35(11), 1095–1108.

21. Tilling, C., Devlin, N., Tsuchiya, A., & Buckingham, K. (2010). Protocols for time tradeoff valuations of health states worse than dead: A literature review. Medical Decision Making, 30(5), 610–619.

22. Lamers, L. M. (2007). The transformation of utilities for health states worse than death: Consequences for the estimation of EQ-5D value sets. Medical Care, 45(3), 238–244.

23. Augustovski, F., Rey-Ares, L., Irazola, V., Oppe, M., & Devlin, N. J. (2013). Lead versus lag-time trade-off variants: Does it make any difference? The European Journal of Health Economics, 14(Suppl 1), S25–S31.

24. Janssen, B. M., Oppe, M., Versteegh, M. M., & Stolk, E. A. (2013). Introducing the composite time trade-off: A test of feasibility and face validity. European Journal of Health Economics, 14(Suppl 1), S5–S13.

25. Krabbe, P. F., Devlin, N. J., Stolk, E. A., Shah, K. K., Oppe, M., van Hout, B., et al. (2014). Multinational evidence of the applicability and robustness of discrete choice modeling for deriving EQ-5D-5L health-state values. Medical Care, 52(11), 935–943.

26. Ramos-Goni, J. M., Rivero-Arias, O., Errea, M., Stolk, E. A., Herdman, M., & Cabases, J. M. (2013). Dealing with the health state 'dead' when using discrete choice experiments to obtain values for EQ-5D-5L heath states. European Journal of Health Economics, 14(Suppl 1), S33–S42.

27. Ramos-Goni, J. M., Pinto-Prades, J. L., Oppe, M., Cabases, J. M., Serrano-Aguilar, P., & Rivero-Arias, O. (2014). Valuation and modeling of EQ-5D-5L health states using a hybrid approach. Medical Care. doi:10.1097/MLR.0000000000000283.

28. Lee, S. Y., Stucky, B. D., Lee, J. Y., Rozier, R. G., & Bender, D. E. (2010). Short assessment of health literacy-Spanish and English: A comparable test of health literacy for Spanish and English speakers. Health Services Research, 45(4), 1105–1120.

29. Selden, C. R., Zorn, M., Ratzan, S. C., & Parker, R. M. (2000). Health literacy, current bibliographies in medicine. Bethesda, MD: National Institutes of Health.

30. Huber, P. J. (1964). Robust estimation of a location parameter. The Annals of Mathematical Statistics, 35(1), 73–101.

31. Hamilton, L. (1991). srd1: How robust is robust regression? Stata Technical Bulletin Reprints, 1, 169–175.

32. EQ-5D-5L Crosswalk value sets. Retrieved January 2015, from http://www.euroqol.org/about-eq-5d/valuation-of-eq-5d/eq-5d-5l- value-sets.html.

33. van Hout, B., Janssen, M. F., Feng, Y. S., Kohlmann, T., Busschbach, J., Golicki, D., et al. (2012). Interim scoring for the EQ-5D-5L: Mapping the EQ-5D-5L to EQ-5D-3L value sets. Value Health, 15(5), 708–715.

34. 34. Szende, A., Janssen, B. M., & Cabase´s, J. M. (2014). Self-reported population health: An international perspective based on EQ-5D. Dordrecht, Heidelberg, New York, London: Springer.

Chapter 8

Combining continuous and dichotomous responses in a hybrid model

Juan M. Ramos-Goñi,
Benjamin Craig,
Mark Oppe,
Ben van Hout

EuroQol Working paper series 16002

ABSTRACT

In survey research, continuous responses may represent a value, a lower or upper bound of a value, or a range of values (e.g., the value of my car is $10,000, is greater than or equal to $9,000 or is between $9,000 and $10,999). Dichotomous responses may represent an inequality in value (e.g., the value of my car is higher than the value of that car). Given a dataset with continuous and dichotomous responses, the hyreg command estimates the parameters of a hybrid regression model by maximizing a single likelihood function, namely the product of the likelihoods of continuous and dichotomous responses. Analogous to combining, for example, intreg and logit commands, this paper demonstrates the hyreg command using simulated data and includes an example of an econometric specification from health preference research.

Keywords: TTO, DCE, Hybrid regression

INTRODUCTION

In survey research, respondents are commonly asked to consider the location of objects along a scale using multiple tasks. For example, they may be asked to choose between 2 objects to express an inequality in value (e.g., car A is preferred to car B [A>B]). Such dichotomous responses facilitate the consideration of multiple differences in attributes and may mimic real world behaviour (i.e., action or inaction). Alternatively, respondents may be asked to place objects at points along a scale (e.g., I am willing to pay $10,000 for car A) or within intervals on a scale (e.g., I am willing to pay between $ 9,000 and $ 12,000 for car A). Respondents may be asked whether an object is located above or below a threshold (e.g., car A>$9,000; open interval). Unlike dichotomous responses, continuous responses (i.e., points or intervals along a scale) are more precise, but can be more cognitively burdensome for respondents as well as they require greater numeracy or understanding of labels. Whether a survey instrument captures the location of objects relative to other objects or at a point, within an interval, or above/below a threshold on a scale, survey researchers require an analytical approach that takes into account all available evidence. Notice that it is assumed high correlation between the different types of responses as all survey questions are related to the same objects, however, this assumption should be tested first.

This paper introduces the hyreg command, which allows the estimation of a regression model with both continuous and dichotomous responses by maximizing a single likelihood function, namely the product of the likelihoods of dichotomous and continuous responses. Like many innovations, this hybrid approach was borne out of necessity: specifically, Oppe and van Hout created an econometric approach, for modelling EQ-5D-5L valuation data that integrates dichotomous responses from discrete choice experiments (DCE; i.e., health A prefer to health B) with continuous responses from an iterative choice-based task, known as the time-trade off (TTO) [1,2]. Using an iterative process, the TTO task identifies the respondent's value of a health description in terms of years in full health (equivalent statement; i.e., health A for 10 years then die = full health for 8 years then die). Oppe and van Hout proposed to combine TTO and DCE responses in a single model, calling it the hybrid approach [2]. They suggest a maximum likelihood estimation of the product of the likelihood functions of a normal distribution for point observations (TTO responses) and a logistic model for dichotomous observations (DCE responses) based on the difference between the alternatives [3-5]. However, further review of the TTO responses revealed that their distribution was largely uniform with clustering on specific numbers of the iteration process, which complicated their interpretation [6].

During a scientific meeting in August 2014, Ramos-Goñi and Craig considered ignoring the equivalence statements of the TTO task and focusing on the iterative procedure that led up to the statement. The TTO choice-based process iteratively creates open and closed

intervals (e.g., full health for 5 years < health state A for 10 years < full health for 8 years) as a means of narrowing in on the equivalence statement. As an exploratory analysis, these intervals were included as the dependent variable in the intreg command, which produced results with greater face validity than regular linear regression on the equivalence statements alone.

Built from previous work on the hybrid approach [2-5], Ramos-Goñi and Craig decided to integrate the interval responses from the TTO with the dichotomous responses from the DCE under a common likelihood specification, which led to the development of the hyreg command. The hyreg command further extends the hybrid approach to include 2 distributions (logistic and normal) and a multiplicative function of scaling (i.e., as hetprobit or intreg using het(#) option). The hyreg command also allows the dichotomous and continuous responses to have different distributions (logistic and normal) and have different independent variables to model scaling terms. Although originally developed for health preference research, the hyreg command can be used by anyone interested in combining continuous and dichotomous responses in a single maximum likelihood function to estimate the parameters of a regression model on a scale (e.g., sweetness, pain, wealth, value).

Description

Hyreg fits a hybrid model with both continuous and dichotomous responses by maximizing a single likelihood function.

Syntax

hyreg *depvar1* [*depvar2*] [*indepvars*] [if] [in] , datatype(*varname*) [interval contdist(*normal* | *logistic*) dichdist(*normal* | *logistic*) ll(#) ul(#) hetcont(*varlist*) hetdich(*varlist*) noconstant vce(oim | opg | robust | cluster *varname*) *maximize options*]

hyreg works in a similar way to most other Stata regression commands. Each observation includes one response described using one or two dependent variables (depvar1, depvar2) and one binary variable specified by datatype() (1 indicating that the response is continuous and 0 indicating that the response is dichotomous). A continuous response can be either a point or an interval (i.e., as for intreg). A dichotomous response is binary (0 or 1; i.e., as for probit). If the observations include only points and dichotomous responses, only one dependent variable is required (depvar1). If the observations also include interval responses, the hyreg command requires both the "interval" option and two dependent variables (depvar1, depvar2) indicating the boundaries of the interval. With the "interval" option, a point response is indicated when the two dependent variables have the same value (i.e., depvar1=depvar2). For open intervals (i.e., where either the left or right bounds are censored), the open end of the interval is represented by a missing value. In summary, the specification of depvar1 and depvar2 depend on the type of observation:

Type of observation		depvar1	depvar2
point observation	a = [a,a]	a	a
interval observation	[a,b]	a	b
left-censored observation	(-inf,b]	.	b
right-censored observation	[a,inf)	a	.
dichotomous observation	c	c	.

a, b ϵ]-∞,+ ∞[and c ϵ {0,1}

The contdist() and dichdist() options indicate the distributions for the continuous and dichotomous responses to be used in the maximum likelihood estimator. Point responses can have a lower limit (ll) and upper limit (ul; i.e., as tobit) and the scaling of each distribution may be associated with independent variables (e.g., heteroskedasticity). Therefore, the hyreg command includes distributional modifiers, namely ll(), ul(), hetcont(*varlist*), and hetdich(*varlist*). The default distributions are normal distribution for continuous responses and logistic distribution for dichotomous responses and do not include any modifiers.

OPTIONS

Model

datatype(*varname*) specifies the variable name containing the indicators of response type. An observation is 0 when a dichotomous response is present and 1 when a continuous response is present. datatype() is required.

interval is specified in the presence of a second dependent variable (depvar2). This second dependent variable allows the inclusion of intervals among the continuous responses (i.e., depvar1 is the lower bound, depvar2 is the upper bound) The open end of an interval is indicated by a missing value. With this option, a point response is indicated when the two dependent variables have the same value (i.e., depvar1=depvar2).

contdist(*normal* | *logistic*) specifies the distribution that the model fits over the continuous responses.

normal fits a normal distribution for continuous responses.

logistic fits a logistic distribution for continuous responses

dichdist(normal | logistic) specifies the distribution that the model fit over the dichotomous responses.

normal fits a normal distribution for dichotomous responses, as a probit model does.

logistic fits a logistic distribution for dichotomous responses, as a logistic model does.

ul(#) right-censoring limit such that all point responses greater than or equal to this limit are treated as censored.

ll(#) left-censoring limit such than all point responses less than or equal to this limit are treated as censored.

hetcont(*varlist***)** specifies the independent variables in the scaling function for the continuous distribution (i.e., lnsigma).

hetdich(*varlist***)** specifies the independent variables in the scaling function for the dichotomous distribution (i.e., lntheta).

noconstant suppresses the constant term (intercept) in the model of the scaled variable.

SE/Robust

vce(*vcetype*) specifies the type of standard error reported, which includes types that are derived from asymptotic theory (**oim, opg**), that are robust to some kinds of misspecification (robust), that allow for intragroup correlation (***cluster*** *clustvar*).

Maximization

maximize options: difficult, technique(algorithm spec), iterate(#) , nolog, trace, gradient, showstep, hessian, showtolerance, tolerance(#), ltolerance(#), nrtolerance(#), nonrtolerance, and init(*init specs*).

EXAMPLE

To illustrate how hyreg works we created a dataset of 1,000 respondents with 17 responses for each respondent. The 17 responses for each respondent include: 1 point response (Value of car A), 4 open interval responses (value of car A is higher or lower than a threshold), 5 closed interval responses (Value of car A is between X and Y) and 7 dichotomous responses representing inequalities (car A preferred to car B). This leads to 1,000 point responses. 2,000 left-censored intervals. 2,000 right-censored intervals. 5,000 closed intervals and 7,000 dichotomous responses. The values for continuous responses are scaled between 5-15 meaning thousands of dollars and the values for dichotomous responses are 0-1 (1 if car B is chosen). The responses (N=17,000) has been stored in the file hyreg_data.dta.

. use hyreg_data.dta

. describe

```
  obs:        17,000
 vars:            37                          4 Mar 2016 20:14
 size:     2,312,000
```

variable name	storage type	display format	value label	variable label
id	long	%8.0g		
task	int	%8.0g		Task ID on study design
order_id	byte	%8.0g		Order of task presentation
colour	byte	%8.0g		Colour of the car (5 different colours)
wheels	byte	%8.0g		Type of the wheels (5 typess)
sport_4x4	byte	%8.0g		Type of car from sport to 4x4 type (5 types)
audio	byte	%8.0g		Type of audio system (5 types)
doors	byte	%8.0g		Type and number of doors (5 options)
B_colour	float	%9.0g		Colour of the car (5 different colours)
B_wheels	float	%9.0g		Type of the wheels (5 typess)
B_sport_4x4	float	%9.0g		Type of car from sport to 4x4 type (5 types)
B_audio	float	%9.0g		Type of audio system (5 types)
B_doors	float	%9.0g		Type and number of doors (5 options)
value	float	%8.0g		Responses (0-1 for inequalities and missing for continious values)
method	float	%9.0g		Binary variable indicating the type of method used to obtain the response
lower	double	%10.0g		Lower limit for the intervlas and inequality responses
upper	double	%10.0g		Upper limit for the intervlas and missing for inequality responses
colour2	float	%9.0g		Dummy (independent variable for the model)
colour3	float	%9.0g		Dummy (independent variable for the model)
colour4	float	%9.0g		Dummy (independent variable for the model)
colour5	float	%9.0g		Dummy (independent variable for the model)
wheels2	float	%9.0g		Dummy (independent variable for the model)
wheels3	float	%9.0g		Dummy (independent variable for the model)
wheels4	float	%9.0g		Dummy (independent variable for the model)
wheels5	float	%9.0g		Dummy (independent variable for the model)
sport_4x42	float	%9.0g		Dummy (independent variable for the model)
sport_4x43	float	%9.0g		Dummy (independent variable for the model)
sport_4x44	float	%9.0g		Dummy (independent variable for the model)
sport_4x45	float	%9.0g		Dummy (independent variable for the model)
audio2	float	%9.0g		Dummy (independent variable for the model)
audio3	float	%9.0g		Dummy (independent variable for the model)
audio4	float	%9.0g		Dummy (independent variable for the model)
audio5	float	%9.0g		Dummy (independent variable for the model)
doors2	float	%9.0g		Dummy (independent variable for the model)
doors3	float	%9.0g		Dummy (independent variable for the model)
doors4	float	%9.0g		Dummy (independent variable for the model)
doors5	float	%9.0g		Dummy (independent variable for the model)

```
Sorted by: id
```

The variable “method” indicates the type of response (0 for dichotomous responses; 1 for continuous responses).

. tab method

```
. tab method
```

Binary variable indicating the type of method used to obtain the response	Freq.	Percent	Cum.
0	7,000	41.18	41.18
1	10,000	58.82	100.00
Total	17,000	100.00	

The data for each respondent is as follows: for the continuous responses, the independent variables represent the description of car A (colour to doors). For dichotomous responses, we also include variables describing the alternative (car B; B_colour to B_doors).

```
. list id-upper if id ==2590
```

	id	task	order_id	colour	wheels	sport~x4	audio	doors	B_colour	B_wheels	B_spor~4	B_audio	B_doors	value	method	lower	upper
1.	2590	53	1	2	2	4	3	4	.	.	.	.	.	.	1	10	11
2.	2590	50	2	2	4	5	5	3	.	.	.	.	.	.	1	12	12.5
3.	2590	83	3	1	1	2	1	1	.	.	.	.	.	.	1	5	5.25
4.	2590	52	4	1	1	4	2	5	.	.	.	.	.	.	1	6.5	7
5.	2590	54	5	4	2	1	1	5	.	.	.	.	.	.	1	7	7.5
6.	2590	49	6	1	3	1	2	2	.	.	.	.	.	.	1	.	7.5
7.	2590	56	7	4	5	4	1	3	.	.	.	.	.	.	1	.	10
8.	2590	55	8	3	5	3	3	2	.	.	.	.	.	.	1	10	.
9.	2590	51	9	5	1	1	5	2	.	.	.	.	.	.	1	10	.
10.	2590	86	10	5	5	5	5	5	.	.	.	.	.	.	1	13.5	13.5
11.	2590	113	11	5	2	1	1	1	1	1	4	3	1	1	0	1	.
12.	2590	135	12	5	4	5	5	5	3	5	5	3	5	1	0	1	.
13.	2590	122	13	2	1	3	5	4	4	1	3	2	1	1	0	1	.
14.	2590	27	14	3	4	1	3	2	2	4	4	4	5	0	0	0	.
15.	2590	105	15	2	4	5	2	3	4	5	1	2	5	0	0	0	.
16.	2590	169	16	1	1	4	4	5	3	2	1	1	5	0	0	0	.
17.	2590	11	17	3	5	2	1	1	4	2	5	5	1	0	0	0	.

Prior to analysis, all variables representing the attributes of cars A are recoded as dummy variables for continuous responses. In case of dichotomous responses, the dummy variables of car B are subtracted from the variables of car A so that the recoded dummy variables represent the differences between car A and B.

The default specification includes a normal distribution for continuous responses and the logistic distribution for dichotomous responses. For purposes of simulation, the constant term was dropped. To incorporate the open and closed intervals, the hybrid command must include the “interval” option and a second dependent variable (i.e., depvar2). In this case, the hybrid model estimates are as follows:

```
. hyreg lower upper colour2-doors5 , datatype(method) interval nocons nolog
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
model						
colour2	1.344763	.0884481	15.20	0.000	1.171408	1.518119
colour3	1.024789	.0984525	10.41	0.000	.8318259	1.217753
colour4	2.312743	.0966544	23.93	0.000	2.123303	2.502182
colour5	2.488422	.096671	25.74	0.000	2.298951	2.677894
wheels2	1.611344	.084098	19.16	0.000	1.446515	1.776173
wheels3	1.267405	.0979788	12.94	0.000	1.07537	1.45944
wheels4	1.952684	.0962394	20.29	0.000	1.764058	2.141309
wheels5	2.224468	.0888176	25.05	0.000	2.050388	2.398547
sport_4x42	1.692613	.08951	18.91	0.000	1.517176	1.868049
sport_4x43	1.25203	.0942327	13.29	0.000	1.067337	1.436723
sport_4x44	2.083887	.0956769	21.78	0.000	1.896363	2.27141
sport_4x45	2.069083	.0882375	23.45	0.000	1.896141	2.242025
audio2	1.596118	.0838017	19.05	0.000	1.431869	1.760366
audio3	1.338447	.0986938	13.56	0.000	1.145011	1.531883
audio4	2.303966	.0946902	24.33	0.000	2.118377	2.489555
audio5	2.929835	.0952562	30.76	0.000	2.743136	3.116533
doors2	2.094111	.0885908	23.64	0.000	1.920476	2.267746
doors3	1.96096	.0985397	19.90	0.000	1.767826	2.154094
doors4	3.071502	.0900714	34.10	0.000	2.894965	3.248039
doors5	3.217783	.088312	36.44	0.000	3.044694	3.390871
lnsigma						
_cons	1.253981	.0095759	130.95	0.000	1.235213	1.27275
lntheta						
_cons	-.8315329	.0344434	-24.14	0.000	-.8990408	-.764025

```
        1000  continuous uncensored observations
        2000  continuous left-censored observations
        2000  continuous right-censored observations
        5000  continuous interval observations
        7000  dichotomous observations
```

With a normal distribution for the dichotomous responses, the hybrid model estimates are as follows:

```
. hyreg lower upper colour2-doors5 , datatype(method) contdist(normal) dichdist(normal) interval nocons nolog
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
model						
colour2	1.361203	.0896599	15.18	0.000	1.185472	1.536933
colour3	1.022867	.0993353	10.30	0.000	.8281736	1.217561
colour4	2.339319	.0979142	23.89	0.000	2.147411	2.531227
colour5	2.504011	.0976231	25.65	0.000	2.312673	2.695349
wheels2	1.623739	.0849729	19.11	0.000	1.457195	1.790283
wheels3	1.249896	.0986605	12.67	0.000	1.056525	1.443267
wheels4	1.933073	.0974855	19.83	0.000	1.742005	2.124141
wheels5	2.208389	.0898828	24.57	0.000	2.032222	2.384556
sport_4x42	1.700799	.0901446	18.87	0.000	1.524119	1.877479
sport_4x43	1.246483	.0949004	13.13	0.000	1.060482	1.432484
sport_4x44	2.083483	.0964581	21.60	0.000	1.894428	2.272537
sport_4x45	2.059563	.0890736	23.12	0.000	1.884982	2.234144
audio2	1.616608	.084741	19.08	0.000	1.450518	1.782697
audio3	1.32147	.0995103	13.28	0.000	1.126434	1.516507
audio4	2.314139	.0954957	24.23	0.000	2.126971	2.501307
audio5	2.933666	.0959325	30.58	0.000	2.745641	3.12169
doors2	2.103319	.0894827	23.51	0.000	1.927936	2.278702
doors3	1.968084	.0994225	19.80	0.000	1.773219	2.162948
doors4	3.069516	.0906838	33.85	0.000	2.891779	3.247253
doors5	3.203245	.0896107	35.75	0.000	3.027611	3.378879
lnsigma						
_cons	1.253046	.0095754	130.86	0.000	1.234279	1.271814
lntheta						
_cons	1.363679	.0322494	42.29	0.000	1.300471	1.426886

```
        1000  continuous uncensored observations
        2000  continuous left-censored observations
        2000  continuous right-censored observations
        5000  continuous interval observations
        7000  dichotomous observations
```

With a logistic distribution for the continuous responses, the hybrid model estimates are as follows:

```
. hyreg lower upper colour2-doors5 , datatype(method) contdist(logistic) dichdist(logistic) interval nocons nolog
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
model						
colour2	1.263469	.088078	14.34	0.000	1.090839	1.436099
colour3	.901853	.0974724	9.25	0.000	.7108105	1.092896
colour4	2.206804	.0954919	23.11	0.000	2.019643	2.393965
colour5	2.416781	.0961424	25.14	0.000	2.228345	2.605216
wheels2	1.612092	.0836299	19.28	0.000	1.448181	1.776004
wheels3	1.202755	.096541	12.46	0.000	1.013538	1.391972
wheels4	1.885458	.0961765	19.60	0.000	1.696956	2.073961
wheels5	2.228747	.0881642	25.28	0.000	2.055949	2.401546
sport_4x42	1.694819	.0899525	18.84	0.000	1.518515	1.871123
sport_4x43	1.2378	.0927731	13.34	0.000	1.055968	1.419631
sport_4x44	2.049346	.0947613	21.63	0.000	1.863618	2.235075
sport_4x45	2.083945	.0872743	23.88	0.000	1.912891	2.254999
audio2	1.59081	.0831156	19.14	0.000	1.427906	1.753713
audio3	1.240593	.0976139	12.71	0.000	1.049273	1.431913
audio4	2.22149	.0940198	23.63	0.000	2.037215	2.405766
audio5	2.921267	.0947456	30.83	0.000	2.735569	3.106965
doors2	2.166209	.087646	24.72	0.000	1.994426	2.337992
doors3	1.950307	.0972055	20.06	0.000	1.759787	2.140826
doors4	3.025481	.0884988	34.19	0.000	2.852026	3.198935
doors5	3.259518	.087192	37.38	0.000	3.088624	3.430411
lnsigma						
_cons	.6971123	.0109073	63.91	0.000	.6757343	.7184902
lntheta						
_cons	-.8290934	.0349098	-23.75	0.000	-.8975153	-.7606715

```
        1000  continuous uncensored observations
        2000  continuous left-censored observations
        2000  continuous right-censored observations
        5000  continuous interval observations
        7000  dichotomous observations
```

With a logistic distribution for the continuous responses and a normal distribution for the dichotomous responses, the hybrid model estimates are as follows:

```
. hyreg lower upper colour2-doors5 , datatype(method) contdist(logistic) dichdist(normal) interval nocons nolog
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
model						
colour2	1.280671	.0894655	14.31	0.000	1.105322	1.45602
colour3	.8992594	.0982985	9.15	0.000	.7065979	1.091921
colour4	2.23168	.0967706	23.06	0.000	2.042013	2.421347
colour5	2.431784	.0971365	25.03	0.000	2.2414	2.622168
wheels2	1.62347	.0845193	19.21	0.000	1.457815	1.789124
wheels3	1.18437	.0972124	12.18	0.000	.9938369	1.374903
wheels4	1.864428	.0972932	19.16	0.000	1.673736	2.055119
wheels5	2.212532	.0891569	24.82	0.000	2.037788	2.387276
sport_4x42	1.702845	.0903023	18.86	0.000	1.525855	1.879834
sport_4x43	1.232745	.0933047	13.21	0.000	1.049871	1.415619
sport_4x44	2.048329	.0954312	21.46	0.000	1.861287	2.235371
sport_4x45	2.075412	.0880446	23.57	0.000	1.902848	2.247976
audio2	1.611453	.0838582	19.22	0.000	1.447094	1.775813
audio3	1.224691	.0982763	12.46	0.000	1.032073	1.417309
audio4	2.232013	.0948568	23.53	0.000	2.046097	2.417929
audio5	2.926141	.0953487	30.69	0.000	2.739261	3.113022
doors2	2.175888	.0884085	24.61	0.000	2.002611	2.349165
doors3	1.956895	.0979885	19.97	0.000	1.764841	2.148949
doors4	3.021881	.0890481	33.94	0.000	2.847349	3.196412
doors5	3.245708	.0884435	36.70	0.000	3.072362	3.419054
lnsigma						
_cons	.6961352	.0109066	63.83	0.000	.6747586	.7175119
lntheta						
_cons	1.361181	.0327107	41.61	0.000	1.297069	1.425293

```
      1000  continuous uncensored observations
      2000  continuous left-censored observations
      2000  continuous right-censored observations
      5000  continuous interval observations
      7000  dichotomous observations
```

With a normal distribution for the continuous responses and a logistic distribution for the dichotomous responses, but using heteroscedasticity in both types of responses, the hybrid model estimates are as follows:

```
. hyreg lower upper colour2-doors5 , datatype(method) interval nocons nolog hetcont(colour2-doors5) hetdich(colour2-doors5)
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
model						
colour2	1.215438	.097968	12.41	0.000	1.023424	1.407452
colour3	.9361072	.104181	8.99	0.000	.7319161	1.140298
colour4	2.180632	.1056835	20.63	0.000	1.973496	2.387767
colour5	2.362725	.103163	22.90	0.000	2.160529	2.564921
wheels2	1.445028	.0919531	15.71	0.000	1.264803	1.625252
wheels3	1.303646	.1042452	12.51	0.000	1.099329	1.507963
wheels4	1.793598	.1098829	16.32	0.000	1.578231	2.008965
wheels5	2.087382	.1000534	20.86	0.000	1.891281	2.283483
sport_4x42	1.564673	.1012316	15.46	0.000	1.366263	1.763083
sport_4x43	1.206338	.106827	11.29	0.000	.9969612	1.415715
sport_4x44	1.985284	.1095693	18.12	0.000	1.770532	2.200036
sport_4x45	2.032859	.0948608	21.43	0.000	1.846935	2.218782
audio2	1.320055	.0992176	13.30	0.000	1.125592	1.514518
audio3	1.260709	.1089929	11.57	0.000	1.047086	1.474331
audio4	2.092974	.1011976	20.68	0.000	1.89463	2.291318
audio5	2.684734	.1051208	25.54	0.000	2.478701	2.890767
doors2	2.435524	.123007	19.80	0.000	2.194434	2.676613
doors3	2.448527	.1164185	21.03	0.000	2.220351	2.676703
doors4	3.355396	.1015937	33.03	0.000	3.156276	3.554516
doors5	3.590764	.1046951	34.30	0.000	3.385565	3.795962
lnsigma						
colour2	-.0832428	.0352389	-2.36	0.018	-.1523097	-.0141758
colour3	-.0246814	.035188	-0.70	0.483	-.0936487	.0442858
colour4	-.0234911	.0401845	-0.58	0.559	-.1022513	.0552692
colour5	.0251924	.0349982	0.72	0.472	-.0434028	.0937876
wheels2	-.1091662	.0340023	-3.21	0.001	-.1758095	-.0425229
wheels3	-.0830598	.0404313	-2.05	0.040	-.1623037	-.0038158
wheels4	.0169463	.0365636	0.46	0.643	-.054717	.0886095
wheels5	-.0086818	.0349379	-0.25	0.804	-.0771587	.0597952
sport_4x42	-.1038585	.0357219	-2.91	0.004	-.1738721	-.0338448
sport_4x43	-.1569476	.0378076	-4.15	0.000	-.2310492	-.0828461
sport_4x44	-.0657407	.0371338	-1.77	0.077	-.1385217	.0070403
sport_4x45	-.0460983	.0335391	-1.37	0.169	-.1118338	.0196372
audio2	-.1309788	.0335284	-3.91	0.000	-.1966931	-.0652644
audio3	-.0523165	.0380126	-1.38	0.169	-.1268198	.0221868
audio4	.0679624	.0328714	2.07	0.039	.0035356	.1323891
audio5	-.0473537	.0349757	-1.35	0.176	-.1159047	.0211974
doors2	-.3984228	.0414845	-9.60	0.000	-.4797309	-.3171147
doors3	-.3664519	.0416811	-8.79	0.000	-.4481453	-.2847585
doors4	-.3148747	.0382326	-8.24	0.000	-.3898091	-.2399402
doors5	-.2867316	.0378698	-7.57	0.000	-.360955	-.2125082
_cons	1.642797	.0339145	48.44	0.000	1.576325	1.709268

```
lntheta
     colour2    .023478   .0899211     0.26   0.794    -.1527642    .1997202
     colour3   -.0629817   .0820764    -0.77   0.443    -.2238484     .097885
     colour4   -.0534122   .0902596    -0.59   0.554    -.2303178    .1234933
     colour5    .1208891   .0844548     1.43   0.152    -.0446393    .2864175
     wheels2   -.0317514   .0774907    -0.41   0.682    -.1836304    .1201276
     wheels3    .0524146   .0735013     0.71   0.476    -.0916453    .1964746
     wheels4    .0642982   .0870393     0.74   0.460    -.1062956     .234892
     wheels5    .1947116   .0747065     2.61   0.009     .0482895    .3411337
  sport_4x42   -.0981628   .0623218    -1.58   0.115    -.2203112    .0239856
  sport_4x43    .0271457   .0672912     0.40   0.687    -.1047426     .159034
  sport_4x44    .0568761    .070923     0.80   0.423    -.0821306    .1958827
  sport_4x45    .1017443    .073639     1.38   0.167    -.0425855    .2460741
     audio2   -.2327513   .0786215    -2.96   0.003    -.3868465   -.078656
     audio3   -.0611286   .0756355    -0.81   0.419    -.2093715    .0871143
     audio4    .2391177    .069731     3.43   0.001     .1024474    .3757881
     audio5    .1768521   .0680425     2.60   0.009     .0434911     .310213
     doors2    .1420926   .0798455     1.78   0.075    -.0144017    .2985869
     doors3   -.3092774   .0716293    -4.32   0.000    -.4496682   -.1688867
     doors4   -.2833589   .0629347    -4.50   0.000    -.4067087   -.1600091
     doors5   -.0915457   .0704827    -1.30   0.194    -.2296893    .0465979
      _cons   -.8620898   .0375803   -22.94   0.000    -.9357458   -.7884338
------------------------------------------------------------------------------

        1000  continuous uncensored observations
        2000  continuous left-censored observations
        2000  continuous right-censored observations
        5000  continuous interval observations
        7000  dichotomous observations
```

Saved results

hyreg stores the following in e():

Scalars

e(rank)	rank of e(V)
e(N)	number of observations
e(ic)	number of iterations
e(k)	number of parameters
e(k_eq)	number of equations in e(b)
e(k_dv)	number of dependent variables
e(converged)	1 if converged, 0 otherwise
e(rc)	return code
e(N_clust)	number of clusters
e(ll)	log likelihood
e(k_eq_model)	number of equations in overall model test
e(df_m)	model degrees of freedom
e(chi2)	chi-squared
e(p)	p-value for model chi-squared test

Macros

e(cmd)	used command
e(chi2type)	Wald type of model chi-squared test
e(opt)	type of optimization
e(predict)	program used to implement predict
e(vcetype)	title used to label Std. Err.
e(clustvar)	name of cluster variable
e(vce)	vcetype specified in vce()
e(user)	name of likelihood-evaluator program
e(ml_method)	type of ml method
e(technique)	maximization technique
e(which)	max or min; whether optimizer is to perform maximization or minimization
e(depvar)	names of dependent variable
e(properties)	b V

Matrices

e(b)	coefficient vector
e(V)	variance-covariance matrix of the estimators
e(ilog)	iteration log (up to 20 iterations)
e(gradient)	gradient vector
e(V_modelbased)	model-based variance

Functions

e(sample)	marks estimation sample

METHODS AND FORMULAS

hyreg fits, by maximum likelihood, a hybrid regression model, $x\beta + \varepsilon$, where β denotes the vector of model coefficients, x denotes the independent variables of the model and ε represents the error term. The dependent variable of the model, yj, depends on the type of observation: yj is a continuous point response for observations j ∈ C and yj is a dichotomous response for observations j ∈D. Caution is warranted when merging responses from different techniques into a single estimator [5]. The variance of the continuous responses may not be equal to the variance found in dichotomous responses. The continuous and dichotomous error terms may even have entirely different distributions. Oppe and van Hout used a normal distribution for continuous responses and a logistic distribution for dichotomous responses [2], obtaining the following log-likelihood formula:

(Formula 1)

$$lnL = -\frac{1}{2}*\sum_{j\in C}\left\{\ln(2\pi\sigma^2) + \left(\frac{y_j\text{-}x\beta}{\sigma}\right)^2\right\} + \sum_{j\in D}\left\{ln\left(\frac{1}{1+e^{(-x\beta')}}\right)*y_j + ln\left(\frac{e^{(-x\beta')}}{1+e^{(-x\beta')}}\right)*(1\text{-}y_j)\right\}$$

This formula includes only point and dichotomous responses and serves as the default specification for the hyreg command. Noting that the logit coefficients of the dichotomous model, β' may not be on the same scale as the coefficients of the continuous model, β, due to distributional differences, they introduced a proportional rescaling parameter θ, such that β' = β /θ:

(Formula 2)

$$lnL = -\frac{1}{2}*\sum_{j\in C}\left\{\ln(2\pi\sigma^2) + \left(\frac{y_j\text{-}x\beta}{\sigma}\right)^2\right\} + \sum_{j\in D}\left\{ln\left(\frac{1}{1+e^{(-x\beta/\theta)}}\right)*y_j + ln\left(\frac{e^{(-x\beta/\theta)}}{1+e^{(-x\beta/\theta)}}\right)*(1\text{-}y_j)\right\}$$

For the hyreg command, this log-likelihood was extended to allow for left-censored (L), right-censored (R), and closed intervals (I) obtaining the formula (i.e., as intreg):

(Formula 3)

$$lnL = -\frac{1}{2}*\sum_{j\in C}\left\{\ln(2\pi\sigma^2) + \left(\frac{y_j\text{-}x\beta}{\sigma}\right)^2\right\} + \sum_{j\in L} ln\left(\Phi\left(\frac{y_{Lj}\text{-}x\beta}{\sigma}\right)\right) + \sum_{j\in R} ln\left(\Phi\left(\frac{-(y_{Rj}\text{-}x\beta)}{\sigma}\right)\right) + \sum_{j\in I} ln\left(\Phi\left(\frac{y_{2j}\text{-}x\beta}{\sigma}\right) - \Phi\left(\frac{y_{1j}\text{-}x\beta}{\sigma}\right)\right) + \sum_{j\in D}\left\{ln\left(\frac{1}{1+e^{(-x\beta/\theta)}}\right)*y_j + ln\left(\frac{e^{(-x\beta/\theta)}}{1+e^{(-x\beta/\theta)}}\right)*(1\text{-}y_j)\right\}$$

Alternatively, the distribution of the error terms may be the same for the continuous and dichotomous responses (i.e., normal-probit or logistic-logit), obtaining the following 2 log-likelihood formulae respectively:

(Formula 4)

$$lnL = -\frac{1}{2}*\sum_{j\in C}\left\{\ln(2\pi\sigma^2) + \left(\frac{y_j\text{-}x\beta}{\sigma}\right)^2\right\} + \sum_{j\in L} ln\left(\Phi\left(\frac{y_{Lj}\text{-}x\beta}{\sigma}\right)\right)$$

$$+\sum_{j\in R} ln\left(\Phi\left(\frac{-(y_{Rj}-x\beta)}{\sigma}\right)\right)$$
$$+\sum_{j\in I} ln\left(\Phi\left(\frac{y_{2j}-x\beta}{\sigma}\right)-\Phi\left(\frac{y_{1j}-x\beta}{\sigma}\right)\right)$$
$$+\sum_{j\in D}\left\{\ln\left(\Phi\left(\frac{-x\beta}{\theta}\right)\right)*(1-y_j)+ln\left(\Phi\left(\frac{x\beta}{\theta}\right)\right)*y_j\right\}$$

(Formula 5)

$$lnL=\sum_{j\in C} ln\left(\frac{e^{\frac{-(y_j-x\beta)}{\sigma}}}{\sigma*\left(1+e^{\frac{-(y_j-x\beta)}{\sigma}}\right)^2}\right)$$
$$+\sum_{j\in L} ln\left(\frac{1}{1+e^{\frac{-(y_{Lj}-x\beta)}{\sigma}}}\right)$$
$$+\sum_{j\in R} ln\left(1-\frac{1}{1+e^{\frac{-(y_{Rj}-x\beta)}{\sigma}}}\right)$$
$$+\sum_{j\in I} ln\left(\left(\frac{1}{1+e^{\frac{-(y_{2j}-x\beta)}{\sigma}}}\right)-\left(\frac{1}{1+e^{\frac{-(y_{1j}-x\beta)}{\sigma}}}\right)\right)$$
$$+\sum_{j\in D}\left\{ln\left(\frac{1}{1+e^{(-x\beta/\theta)}}\right)*y_j+ln\left(\frac{e^{(-x\beta/\theta)}}{1+e^{(-x\beta/\theta)}}\right)*(1-y_j)\right\}$$

Technical note:

For implementation purpose, the hyreg command estimates ln(σ) and ln(θ), instead of σ and θ directly. These parameters, ln(σ) and ln(θ), may be modelled using separate regressions to allow for heteroskedasticity (i.e., as hetprobit or using the het option of the intreg command).

The specific case of TTO and DCE data

The EQ-5D-5L valuation datasets include point responses from a TTO task and paired comparison responses from a DCE task. These tasks have 3 potential implications, which serve to demonstrate the capabilities of the *hyreg* command. First, the rescaling parameter for the TTO responses may be proportionally associated with the EQ-5D-5L attributes (i.e., heteroskedasticity). In other words, worse health implies greater potential variability in value. Second, the point responses equal to -1 were recorded as -1, not allowing for responses less than -1 (left-censoring; i.e., as *tobit*). Third, independent variables in the paired comparison

responses represent additive differences between the attributes of the alternatives, x_A-x_B (i.e., as *logit*) [6].

The three implications are demonstrated as modifications to formula 3:

(Formula 6)

$$lnL = -\frac{1}{2} * \sum_{j \in C'} \left\{ \ln\left(2\pi(e^{z\gamma})^2\right) + \left(\frac{y_j - x\beta}{e^{z\gamma}}\right)^2 \right\} + \sum_{j \in L'} ln\left(\Phi\left(\frac{-1 - x\beta}{e^{z\gamma}}\right)\right) + \sum_{j \in D} \left\{ ln\left(\frac{1}{1 + e^{(-(x_A - x_B)\beta/\theta)}}\right) * y_j + ln\left(\frac{e^{(-(x_A - x_B)\beta/\theta)}}{1 + e^{(-(x_A - x_B)\beta/\theta)}}\right) * (1 - y_j) \right\}$$

Where $\sigma = e^{z\gamma}$ (i.e., ln(σ)=zγ), C' represents TTO responses greater than -1, L' represents TTO responses of -1, and x_A and x_B represent the attributes of alternatives A and B in the paired comparisons.

Alternatively, some analysts may be accustomed to maximizing conditional log-likelihood functions to fit models of dichotomous responses (i.e., as *clogit*). Unlike Formula 6, these functions include separate observations for each alternative (no differences) and assemble the observations in groups [7]. However, this approach is not directly amenable to the integration with continuous responses, particularly normal distributions. For the modelling of a scaled variable using continuous and dichotomous responses, *hyreg* provides a common framework for normal and logistic distributional specifications, separates the distributional specifications by response type (e.g., normal-logit), allows censoring of points and lower and upper bounds, and can relax homoscedasticity assumptions.

Acknowledgements

We are grateful to the EuroQol Research Foundation for covering the fees of the authors in preparing this manuscript.

REFERENCES

1. Oppe M, Devlin NJ, van Hout B, Krabbe PF, de Charro F. A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. Value Health. 2014;17(4):445-53.
2. Oppe M, van Hout B. The optimal hybrid: experimental design and modeling of a combination of TTO and DCE. EuroQol Group Proceedings. 2013. Available at: http://www.euroqol.org/uploads/media/EQ2010_-_CH03_-_Oppe_-_The_optimal_hybrid_-_Experimental_design_and_modeling_of_a_combination_of_TTO_and_DCE.pdf. Accessed October 11, 2014.
3. Rowen D, Brazier J, Van Hout B. A comparison of methods for converting DCE values onto the full health-dead QALY scale. Med Decis Making. 2015 Apr;35(3):328-40. doi: 10.1177/0272989X14559542. Epub 2014 Nov 14.
4. Ramos-Goñi JM, Pinto-Prades JL, Oppe M, Cabasés JM, Serrano-Aguilar P, Rivero-Arias O. Valuation and Modeling of EQ-5D-5L Health States Using a Hybrid Approach. Med Care. 2014 Dec 17. [Epub ahead of print]
5. Craig BM, Busschbach JJ. The episodic random utility model unifies time trade-off and discrete choice approaches in health state valuation. Popul Health Metr. 2009 Jan 13;7:3. doi: 10.1186/1478-7954-7-3. Available at: http://ww.pophealthmetrics.com/content/7/1/3
6. Craig BM, Runge SK, Rand-Hendriksen K, Ramos-Goñi JM, Oppe M. Learning and satisficing: an analysis of sequence effects in health valuation. Value Health. 2015 Mar;18(2):217-23. doi: 10.1016/j.jval.2014.11.005. Epub 2015 Feb 2.
7. Train, K. Discrete Choice Methods with Simulation. Second edition. Cambridge University Press, 2009.

About the authors

Juan M. Ramos-Goñi is Senior researcher at the EuroQol Research Foundation, The Netherlands.

Benjamin M. Craig is Associate Professor of Economics at the University of South Florida, USA.

Mark Oppe is Senior researcher at the EuroQol Research Foundation, The Netherlands.

Ben van Hout is Professor at the University of Sheffield, UK.

9

Chapter 9

Handling data quality issues to estimate the Spanish EQ-5D-5L Value Set using a hybrid interval regression approach

Juan M. Ramos-Goñi,
Benjamin M. Craig,
Jose Luis Pinto-Prades,
Mark Oppe,
Yolanda Ramallo-Fariña,
Nan Luo,
Oliver Rivero-Arias

Value in Health 2017. In press.

ABSTRACT

Background: The Spanish five-level EuroQol five-dimensional questionnaire (EQ-5D-5L) valuation study was the first to use the EuroQol Valuation Technology protocol, including composite time trade-off (C-TTO) and discrete choice experiments (DCE). In this study, its investigators noticed that some interviewers did not fully explain the C-TTO task to respondents. Evidence from a follow-up study in 2014 confirmed that when interviewers followed the protocol, the distribution of C-TTO responses widened.

Objectives: To handle the data quality issues in the C-TTO responses by estimating a hybrid interval regression model to produce a Spanish EQ-5D-5L value set.

Methods: Four different models were tested. Model 0 integrated C-TTO and DCE responses in a hybrid model and models 1 to 3 altered the interpretation of the C-TTO responses: model 1 allowed for censoring of the C-TTO responses, whereas model 2 incorporated interval responses and model 3 included the interviewer-specific protocol violations. For external validation, the predictions of the four models were compared with those of the follow-up study using the Lin's concordance correlation coefficient.

Results: This stepwise approach to modeling C-TTO and DCE responses improved the concordance between the valuation and follow-up studies (concordance correlation coefficient: 0.948 [model 0], 0.958 [model 1], 0.952 [model 2], and 0.989 [model 3]). We recommend the estimates from model 3, because its hybrid interval regression model addresses the data quality issues found in the valuation study.

Conclusions: Proto-col violations may occur in any valuation study; handling them in the analysis can improve external validity. The resulting EQ-5D-5L value set (model 3) can be applied to inform Spanish health technology assessments.

Keywords: economic, health status index, life valuation, quality of life.

BACKGROUND

In 2012, the EuroQol Group developed a new standardized protocol (version 1.0) to perform country-specific valuation studies for the five-level EuroQol five-dimensional questionnaire (EQ-5D-5L) using EuroQol Valuation Technology (EQ-VT) [1]. The EQ-VT protocol was developed to elicit health preferences through face-to-face interviews using two valuation techniques, the composite time trade-off (C-TTO) [2,3] and a discrete choice experiment (DCE) [4]. Each respondent completed C-TTO tasks for 10 EQ-5D-5L health states and forced-choice pair comparisons for seven pairs of EQ-5D-5L health states without duration. The C-TTO was a modified version of the traditional TTO technique [5,6], which used the traditional TTO technique for health states considered to be better than immediate death (BTD) and a lead-time TTO technique [7–9] for states considered to be worse than immediate death (WTD).

The C-TTO task entailed a series of consecutive and adapted choices terminating when respondents stated indifference. Because of the complexity of the task, the EQ-VT protocol included an example of this task (being in a wheelchair), which was designed to facilitate and standardize interviewers' explanations. In a previous publication, we described the Spanish EQ-5D-5L valuation study [10]. During this initial analysis, interviewer effects were identified, which were attributed to protocol violations by specific interviewers. Some interviewers did not explain the WTD sections of the C-TTO task and respondents may not have been aware of these sections, leading to fewer WTD values. In fact, evidence from a follow-up study performed in Spain [11], which used an updated protocol version, showed that when interviewers properly explained the WTD sections of the C-TTO task, a higher proportion of negative numbers were observed [12], altering the distribution of the C-TTO responses. In addition, some interviewers did not properly explain the wheelchair example or showed only a few steps from the iterative procedure to respondents [12]. These participants may have responded imprecisely either because they were not aware of the full iterative procedure or to avoid the time and effort needed to reach their accurate indifference points (i.e., satisficing) [13]. We hypothesized in this study that the C-TTO responses in the Spanish EQ-5D-5L valuation study, although not being as precise as we had expected, still contain valuable information about health preferences from the Spanish population. We demonstrate that such information can be retrieved by assessing individual's paths during the iterative procedure when completing each C-TTO task. At this time, we have no reason to believe that the DCE responses in the valuation study were affected by the protocol violations in the C-TTO tasks.

The primary objectives of this article were to introduce an analytical approach based on hybrid interval regression models (jointly incorporating C-TTO and DCE responses), which updates our previous work [10] to handle the data quality issues commented earlier, and to

produce an EQ-5D-5L value set for health technology assessments in Spain. Furthermore, we assessed the external validity of the resulting value set by comparing its estimates with those of a follow-up study.

METHOD

Data

The Spanish EQ-5D-5L valuation study has been previously reported in the literature [10,12], and therefore we describe it only briefly here. The valuation study included 1000 face-to-face interviews conducted in 2012 following the EQ-VT protocol version 1 [1]. After applying exclusions, the analytical sample included 9730 C-TTO responses on 86 health states and 7000 DCE responses on 196 pairs of health states. The sample was representative of the Spanish general population with respect to age and sex.

We used C-TTO and DCE responses from a follow-up study conducted also in Spain in 2014 to assess the external validity of the models described later [11]. This follow-up study was performed in only one Spanish region (Canary Islands), and therefore it was not representative of the Spanish population. Nevertheless, it included the quality control process currently recommended by the EuroQol Group to improve data quality. The original aim of the follow-up study was to test the effect of adding a ranking task to the protocol and its results showed that this addition had no significant effect. Therefore, the data from all study arms of the follow-up study were used for external validation.

The C-TTO Iterative Procedure

The C-TTO task used an iterative procedure (Fig. 1) composed of a series of consecutive and adapted choices terminating when respondents stated indifference. Across its four sections, boxes indicate the possible C-TTO responses (i.e., values) and the arrows represent steps from one value to another. Each C-TTO task started (Start box) by asking whether the respondent preferred 10 years in full health or 10 years in the EQ-5D-5L state. If the respondent preferred 10 years in the EQ-5D-5L state (double arrow up from 1 to 1), the same question was asked again to confirm the extreme value. If the respondent preferred 10 years in full health over 10 years in the EQ-5D-5L state (i.e., double arrow down from 1 to 0), the next question was whether the respondent preferred 0 years in full health (i.e., die immediately) or 10 years in the EQ-5D-5L state.

In the iterative procedure (Fig. 1), the “immediate death” question separated the BTD and WTD scenarios (0 at centre left). If the respondent preferred 10 years in the EQ-5D-5L state (i.e., BTD state; double arrow up from 0 to 0.5), the next question was whether the respondent preferred 5 years in full health or 10 years in the EQ-5D-5L state. If the respondent preferred

to die immediately over 10 years in the EQ-5D-5L state (i.e., WTD state; double-dash arrow from 0 to 0 on the left), the next question was a confirmation of the response but in a lead-time TTO scenario, that is, 10 years in full health versus 10 years in full health followed by 10 years in the EQ-5D-5L state. If the respondent preferred 10 years in full health (double arrow down from 0 to −0.5), the next question was whether the respondent preferred 5 years in full health or 10 years in full health followed by 10 years in the EQ-5D-5L state. If the respondent preferred 10 years in full health followed by 10 years in the EQ-5D-5L state (double arrow horizontal from 0 to 0.05), the iterative procedure changed back to the BTD scenario and the next question asked whether the respondent preferred 0.05 years in full health or 10 years in the EQ-5D-5L state. After these initial steps (double arrows to −0.5, 0.05, 0.5, and 1), the iterative procedure imposed 1-year increments/decrements (i.e., single arrows) followed by half-year corrections (i.e., single-dash arrows) depending on the respondent's preferences. Respondents who visited the BTD scenario after the three initial steps and switched later to the WTD scenario, that is, preferred to die immediately over 10 years in the EQ-5D-5L state (double-dash arrow from 0 to 0 on the right), also had to complete the WTD confirmatory question. This was, however, only once per state.

Although respondents were allowed to go from −0.05 to 0 (immediate death), they were not allowed to go from 0 to −0.05 because of a survey programming error (elbow arrows from 0 to −0.5).

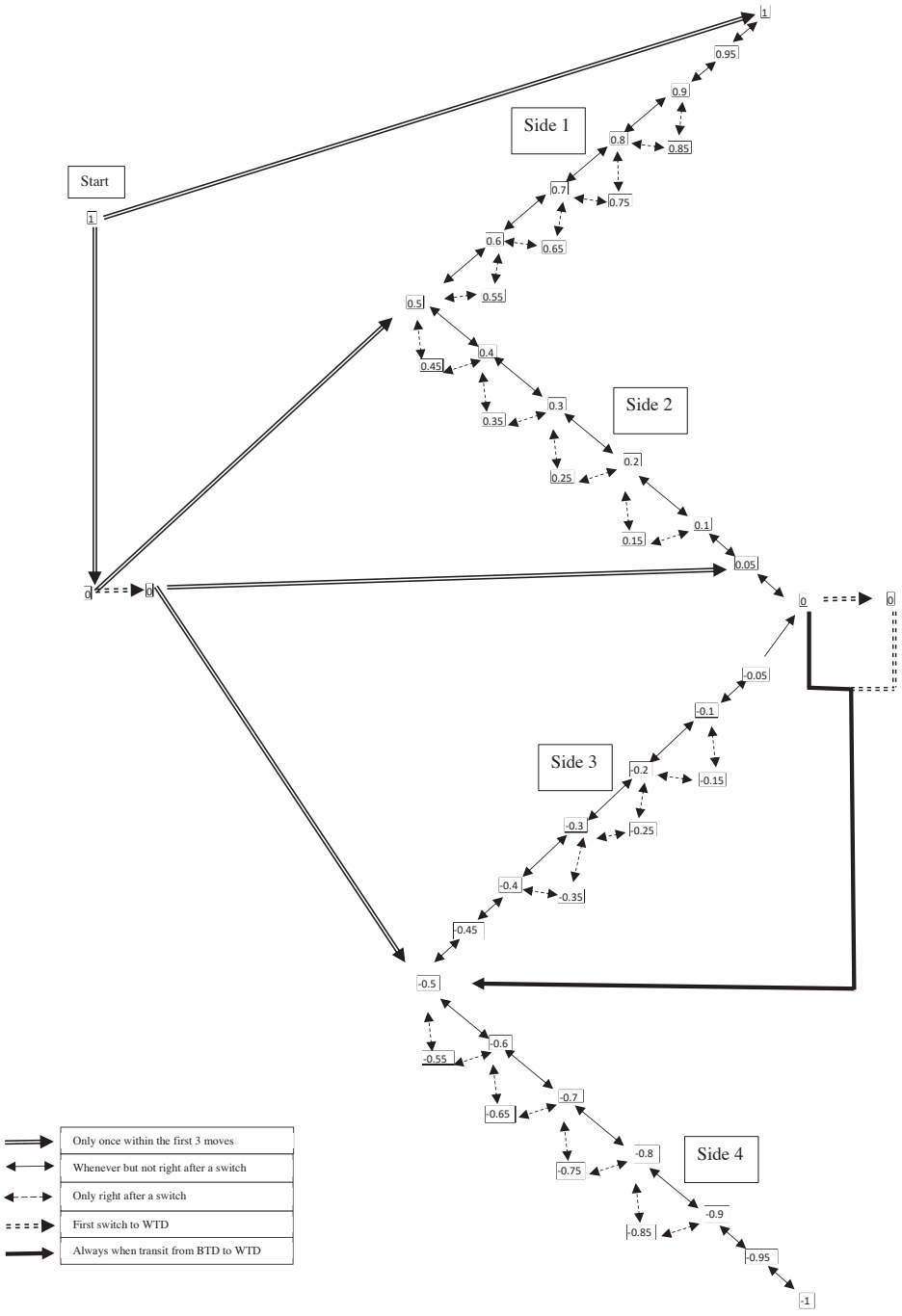

Fig. 1 – C-TTO Iterative procedure. C-TTO, composite time trade-off.

ANALYSIS

Modelling

In a previous publication, we developed and estimated a hybrid model using C-TTO and DCE responses [10]. This initial hybrid model (model 0) assumed normality, homoscedasticity, and that respondents completed the C-TTO tasks accurately. In this study, we followed an analytical approach that relaxed the initial assumption about the accuracy of the C-TTO responses. Specifically, we reconsidered censoring, respondent uncertainty, and protocol violations on the C-TTO tasks [12] as follows.

Censoring of C-TTO responses at –1. The C-TTO task had a minimum TTO value bounded at –1 by design and produced responses in the range [–1, 1]. Nevertheless, feedback from interviewers suggested that some respondents would have responded beyond –1 if allowed, which corroborates the findings of Attema et al. [14]. Because values may be in the range (–∞,1], we relaxed this lower bound assumption and considered responses at the lower bound (–1) to be censored, similar to the open intervals produced by DCE responses (A > B) [15].

Inaccuracy of C-TTO responses. The EQ-VT recorded the full path in the C-TTO iterative procedure for each state presented. Using these paths, we built intervals for each state for each respondent. Instead of considering only the final indifference point, this interval assessment used all path information in a conservative manner. Specifically, we observed four response patterns (see examples of each in Supplemental Materials 1 found at http://dx.doi.org/10.1016/j.jval.2017.10.023):

1. Straight-lining: This refers to an uninterrupted path, only up or only down, that leads to extreme values of a section, namely, 1, 0.95, 0.05, 0, –0.05, –0.95, and –1, using the minimum number of steps. We refer to this response behavior as straight-lining because it represents repeated choices of the same alternative until the end of a C-TTO section.
2. Satisficing: This refers to an uninterrupted path, only up or only down, that leads to non-extreme values of a section using the minimum number of steps. We observed that many interviewers used few number of steps to explain the C-TTO tasks using the wheelchair example, leading us to suspect that some respondents were not trained to perform sufficient steps in the C-TTO iterative procedure to express their values accurately [12]. The literature refers to this lack of engagement as satisficing [13] and its prevalence varies by interviewer [12].

For straight-lining and satisficing, we constructed the intervals by section of the iterative procedure (Fig. 1). For example, in section 1, the path 1 → 0 → 0.5 → 0.6 may imply that the

respondent had an indifference point of 0.6 or that the respondent was insufficiently engaged to express his or her indifference point accurately within the interval [0.5, 1]. If the path was 1 → 0 → 0.5 → 0.6 → 0.7, then the interval becomes [0.6, 1] and so on. In section 2, the path 1 → 0 → 0.5 → 0.4 → 0.3 may imply an indifference point of 0.3 or an interval of [0, 0.4]. Similar intervals were derived for paths included in sections 3 and 4.

1. Circling: This is a path that circles around a value, by going up and down the iterative procedure. For example, the path 1 → 0 → 0.5 → 0.6 → 0.7 → 0.65 → 0.6 → 0.65 → 0.7 may imply a respondent's indifference point of 0.7 or an interval as [0.6, 0.7].
2. Wandering: This is a path that wanders up and down in the iterative procedure without any clear pattern, implying that the respondent is having difficulty with the task. For example, the path 1 → 0 → 0.5 → 0.6 → 0.7 → 0.8 → 0.9 → 0.85 → 0.8 → 0.7 → 0.6 → 0.5 → 0.4 → 0.3 → 0.35 → 0.4 → 0.5 may imply an indifference point of 0.5 or an interval using the lowest and highest visited values [0.3, 0.9]. Wandering was like circling in that the respondent takes an inefficient path to the indifference point, but differs from circling because the respondent did not hone in or circle around a value.

In summary, we created an interval for each indifference point. In case of no switches beyond the first three steps (straight-lining and satisficing), the bounds of the interval were defined as the previous visited value and the corner of the iterative procedure section (1 for section 1, 0 for sections 2 and 3, and −1 for section 4). In the case of switches after the first three steps (circling and wandering), the bounds were defined by the minimum and maximum visited values beyond the first three steps. Nevertheless, for section 3, as the iterative procedure forced respondents to go directly from 0 to −0.5 (discussed earlier), the value of −0.5 was not considered the minimum when it was reached from 0.

In addition, we tested several interval definitions for cases with limited information (i.e., fewer than three steps). On the basis of comparing follow-up and other EQ-VT-based valuation studies [12], we decided to define intervals for such paths as follows: 1) path "1 →0," interval [−0.05, 0.05]; 2) path "1 → 0 → 0.5," interval [0.45, 1]; 3) path "1 → 0 → 0," interval [−0.05, 0.05]; and 4) path "1," interval [−0.995, 1].

To illustrate the intervals, we used a scatterplot showing each health state included in the C-TTO design with their observed mean values together with the mean upper and lower bounds of the intervals. Further details about each path that we found in the data and its corresponding interval can be found in Supplemental Materials 1.

Protocol violations. For most respondents in the valuation study (76.1%), interviewers did not show and explain the iterative procedure allowing for WTD responses, largely shifting

the lower bound up from −1 to 0 [12]. To relax the assumption of no protocol violations, all C-TTO responses that were equal to 0 and were collected without WTD explanation were considered to be censored.

We used hybrid models to sequentially apply these three assumptions to our reference case (model 0), which assumed a normal distribution of the errors for the C-TTO responses (as an ordinary least squares model) and a logistic distribution of the error differences for the DCE responses (as a conditional logit model). Coefficients for model 0 were presented previously, but in this study the constant was removed because it was not statistically significant [10]. Model 1 relaxed the lower bound assumption of the value range (i.e., C-TTO values are censored at −1). Model 2 extended model 1 by replacing C-TTO point responses with intervals (explained earlier), relaxing the interpretation of the C-TTO responses and integrating behavioral data on the path to the indifference point. Finally, model 3 extended model 2 by incorporating protocol violations (i.e., C-TTO values were censored at 0 if the respondent did not receive a WTD explanation). All four models were estimated using a cluster estimation of the standard errors on the basis of the respondent to account for multiple C-TTO and DCE responses from each respondent.

Dependent and Independent Variables

The dependent variable represented the DCE and C-TTO responses. The DCE responses were codified as a binary variable for all models. The C-TTO were codified either as indifference points (model 0) or as intervals (models 1–3) [15]. Each model included an identical set of 20 dummy variables that represented the incremental differences between the five consecutive levels (1–2, 2–3, 3–4, 4–5) within each of the five dimensions of the EQ-5D-5L (i.e., main effects). To facilitate health technology assessments, the Appendix in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2017.10.023 shows the preferred model with 20 cumulative dummy variables (1–2, 1–3, 1–4, 1–5).

Heteroscedasticity

The observed variability in C-TTO responses was not uniform across health states and the variance of C-TTO responses depended on the severity of the health state being valued [10]. We tested for homoscedasticity of the error term using a separate Tobit model for the C-TTO data with the 20 incremental dummy variables (i.e., main effects) [16]. If the homoscedasticity assumption was rejected, the statistical inference may not have been accurate.

External validation

The follow-up study had fewer protocol violations (5.2%) than did the Spanish valuation study (86.5%); therefore, we considered its data more accurate. Using the follow-up study

data, we estimated a heteroscedastic hybrid model in which C-TTO responses were censored at −1. Predictions for the 3125 EQ-5D-5L states (55) were compared with the predictions of the four aforementioned models using the Lin concordance correlation coefficient (CCC) as a measure of agreement. We evaluated model performance using 1) logical consistency of parameters and 2) CCC with the external validation model.

We also compared the predictions for the 86 health states included in the C-TTO design using scatterplots. Scatterplots were also used to compare the four models. We plotted the kernel distribution for model 0, the selected value set, and the external validation model. We plotted the kernel distribution of the 3125 values of the final selected value set, the 243 values of the previous three-level EQ-5D (EQ-5D-3L) value set [17], and the crosswalk value set derived from the EQ-5D-3L value set in Spain [18].

We performed all analyses using Stata 14 MP (StataCorp LP, College Station, TX) [19] and estimated hybrid models using the user-written hyreg command [15].

RESULTS

Straight-lining and satisficing paths were found in 22.6% and 61.3% of the C-TTO responses, respectively, whereas circling and wandering responses were observed in 10.5% and 5.6% of the responses, respectively. In general, C-TTO intervals were not symmetric around the mean: the distance from the mean C-TTO to the mean of the upper limit was greater than the distance to the mean of the lower limit. Nevertheless, the two distances were more similar for mean C-TTO near 1 than near −1 (Fig. 2).

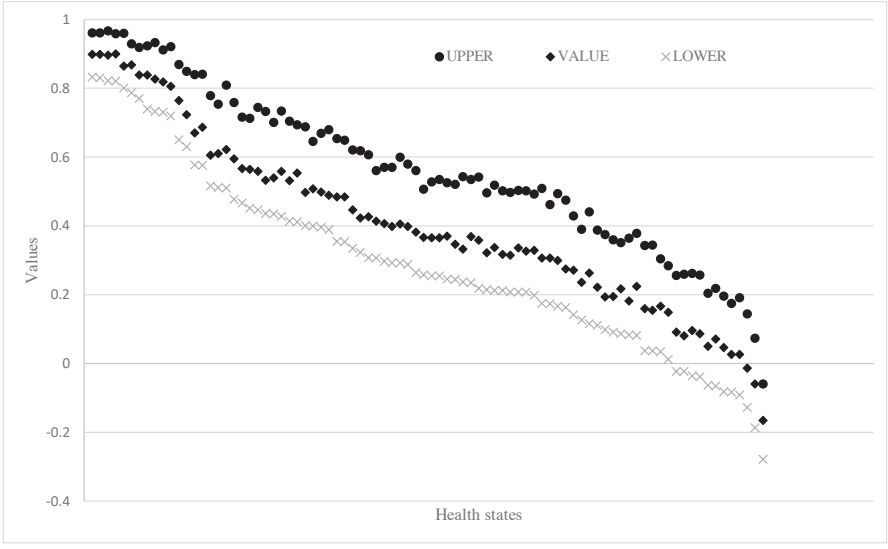

Fig. 2 – Mean of observed values and lower and upper bounds of the intervals based of the iterative procedure

Estimates of the 20 parameters' coefficients (main effects) were logically consistent for all models (Table 1). As expected, the estimated value of the pits state (55555) in model 0 was higher than in model 1 (in which C-TTO values are censored at −1). Although model 2 (intervals) had no effect on the value of the pits state (55555), its values for the mild health states were slightly higher than in model 0 or model 1. On the contrary, model 3, which addressed censoring due to protocol violations, had a lower value for the pits state (55555) as well as higher values for the mild state, widening the predicted range. In addition, we tested for homoscedasticity and re-estimated the heteroscedastic version of each model accordingly; some estimates were, however, inconsistent (see Supplemental Materials 2 found at http://dx.doi.org/10.1016/j.jval.2017.10.023).

Table 1 – Hybrid models (C-TTO & DCE) estimations

Independent variables of the model		Model 0		Model 1		Model 2		(*Value set*) Model 3*		External validation (Follow-up study)	
		Coeff. (SE)	P	Coeff. (SE)	P	Coeff. (SE)	P	Coeff. (SE)	P	Coeff. (SE)	P
Mobility	No to slight problems	0.086 (0.008)	0.000	0.086 (0.008)	0.000	0.084 (0.009)	0.000	0.084 (0.010)	0.000	0.100 (0.015)	0.000
	Slight to moderate problems	0.014 (0.009)	0.120	0.014 (0.009)	0.109	0.012 (0.010)	0.222	0.015 (0.011)	0.183	0.000 (0.016)	0.993
	Moderate to severe problems	0.131 (0.010)	0.000	0.133 (0.011)	0.000	0.134 (0.010)	0.000	0.15 (0.012)	0.000	0.128 (0.016)	0.000
	Severe problems to Unable	0.059 (0.010)	0.000	0.062 (0.010)	0.000	0.066 (0.010)	0.000	0.088 (0.011)	0.000	0.106 (0.016)	0.000
Self-Care	No to slight problems	0.058 (0.008)	0.000	0.057 (0.008)	0.000	0.052 (0.009)	0.000	0.050 (0.010)	0.000	0.037 (0.013)	0.003
	Slight to moderate problems	0.000 (0.009)	0.975	0.001 (0.010)	0.920	0.003 (0.010)	0.776	0.003 (0.012)	0.780	0.007 (0.017)	0.687
	Moderate to severe problems	0.097 (0.011)	0.000	0.099 (0.011)	0.000	0.098 (0.011)	0.000	0.111 (0.012)	0.000	0.139 (0.017)	0.000
	Severe problems to Unable	0.015 (0.009)	0.107	0.017 (0.010)	0.076	0.019 (0.010)	0.048	0.032 (0.011)	0.004	0.030 (0.016)	0.065
Usual Activities	No to slight problems	0.055 (0.008)	0.000	0.055 (0.008)	0.000	0.047 (0.009)	0.000	0.044 (0.010)	0.000	0.062 (0.014)	0.000
	Slight to moderate problems	0.005 (0.010)	0.638	0.004 (0.010)	0.670	0.004 (0.010)	0.691	0.005 (0.011)	0.663	0.001 (0.015)	0.943
	Moderate to severe problems	0.072 (0.010)	0.000	0.074 (0.010)	0.000	0.075 (0.010)	0.000	0.086 (0.012)	0.000	0.109 (0.017)	0.000
	Severe problems to Unable	0.004 (0.010)	0.685	0.006 (0.010)	0.554	0.009 (0.010)	0.374	0.018 (0.012)	0.122	0.026 (0.020)	0.191
Pain/Discomfort	No to slight problems	0.080 (0.008)	0.000	0.080 (0.008)	0.000	0.076 (0.009)	0.000	0.078 (0.010)	0.000	0.069 (0.013)	0.000
	Slight to moderate problems	0.024 (0.009)	0.008	0.024 (0.009)	0.009	0.024 (0.010)	0.022	0.023 (0.012)	0.045	0.019 (0.015)	0.225
	Moderate to severe problems	0.114 (0.011)	0.000	0.118 (0.011)	0.000	0.121 (0.010)	0.000	0.144 (0.012)	0.000	0.136 (0.018)	0.000
	Severe to extreme problems	0.106 (0.010)	0.000	0.109 (0.011)	0.000	0.114 (0.011)	0.000	0.136 (0.012)	0.000	0.164 (0.019)	0.000

Independent variables of the model		Model 0		Model 1		Model 2		(*Value set*) Model 3*		External validation (Follow-up study)	
		Coeff. (SE)	P	Coeff. (SE)	P	Coeff. (SE)	P	Coeff. (SE)	P	Coeff. (SE)	P
Anxiety/Depression	No to slight problems	0.088 (0.008)	0.000	0.087 (0.008)	0.000	0.082 (0.009)	0.000	0.081 (0.010)	0.000	0.035 (0.014)	0.010
	Slight to moderate problems	0.043 (0.010)	0.000	0.044 (0.010)	0.000	0.043 (0.010)	0.000	0.047 (0.012)	0.000	0.058 (0.018)	0.001
	Moderate to severe problems	0.123 (0.010)	0.000	0.126 (0.011)	0.000	0.126 (0.010)	0.000	0.143 (0.012)	0.000	0.144 (0.018)	0.000
	Severe to extreme problems	0.052 (0.010)	0.000	0.055 (0.010)	0.000	0.059 (0.010)	0.000	0.077 (0.011)	0.000	0.132 (0.017)	0.000
Obs. included in the model											
Cont. uncensored (C-TTO)		9,730		9,287		-		-		5,192	
Cont. left-censored (C-TTO)		-		443		443		1,467		658	
Cont. interval (C-TTO)		-		-		9,287		8,263		-	
Dichotomous observations (DCE)		7,000		7,000		7,000		7,000		4,095	
Estimated values by health state											
	U(21111)	0.914		0.914		0.916		0.916		0.9	
	U(12111)	0.942		0.943		0.948		0.95		0.963	
	U(11211)	0.945		0.945		0.953		0.956		0.938	
	U(11121)	0.92		0.92		0.924		0.922		0.931	
	U(11112)	0.912		0.913		0.918		0.919		0.965	
	U(55555)	-0.225		-0.25		-0.249		-0.416		-0.501	
Lin's CCC of model i vs External validation data		0.949		0.958		0.951		0.989		NA	

* We have censored all 0 values for the respondents who did not receive the explanation of the WTD element on wheelchair example
Note: All constant we dropped due to lack of significance

For the 86 states in the C-TTO task, the scatterplots illustrate the relationship between the predictions of model 0 (the reference case) and models 1, 2, and 3. Although censoring increased some coefficients significantly, there appears to be no discernible difference between model 0 and model 1 predictions (Fig. 3A), which may be due to the small proportion of −1 C-TTO responses (4.55%). The differences between model 0 and model 2 predictions (Fig. 3B) appeared across the range of values, and model 2 appeared to have higher values for the mild states. The differences between model 0 and model 3 predictions (Fig. 3C) appeared to increase when less than 0.5, which implied that failure to account for protocol violations increases the values of severe health states and reduces the range of values.

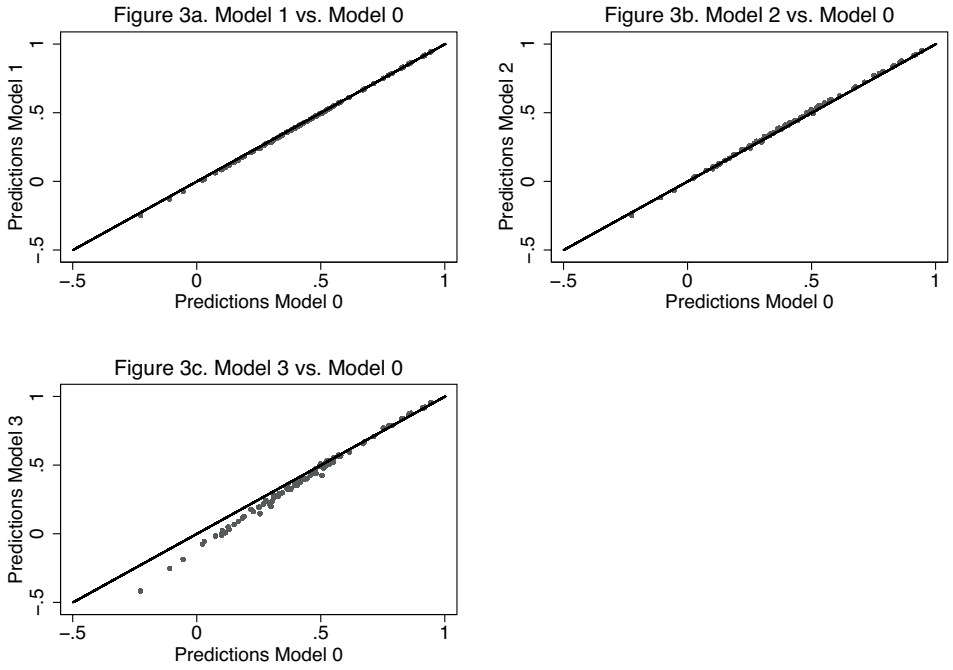

Fig. 3 – Comparison of hybrid model predictions (86 states included in the C-TTO design). C-TTO, composite time trade-off.

In terms of external validity, model 0 (the reference case) and the follow-up model had a CCC of 0.948 across the 3125 state predictions. Nevertheless, the CCC between model 3 and the follow-up model was 0.989. This result and the fact that 16 of its 20 coefficients are statistically significant led us to recommend model 3 estimates for use in health technology assessments as the Spanish EQ-5D-5L value set (Table 1).

For the 86 states in the C-TTO task, the scatterplots illustrate the relationship between the predictions of external validation model and models 0, 1, 2, and 3 (Figs. 4 A, B, C, D, respectively). The comparison with the external validation model showed that model 3 predictions agreed with the predictions of the external validation model (Fig. 4D; CCC 0.989), whereas the other models had worse agreement.

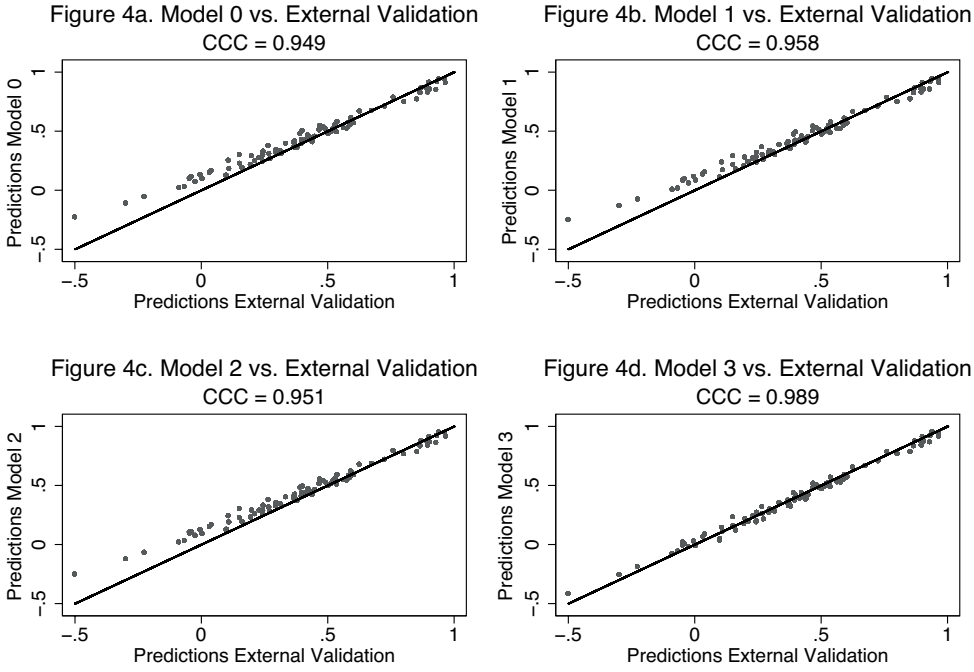

Fig. 4 – Comparison of hybrid model predictions with external validation data from the follow-up study (86 states included in the C-TTO design). CCC, concordance correlation coefficient; C-TTO, composite time trade-off.

Figure 5A further shows that the prediction distribution of model 3 overlapped more closely with the prediction distribution of the external validation model than with the predictions of the reference case (model 0). Figure 5B shows that the prediction distribution of model 3 overlapped with the distribution of crosswalk predictions [18], but not with the distribution of the EQ-5D-3L predictions [17], which is skewed and has more values less than 0 (i.e., die immediately).

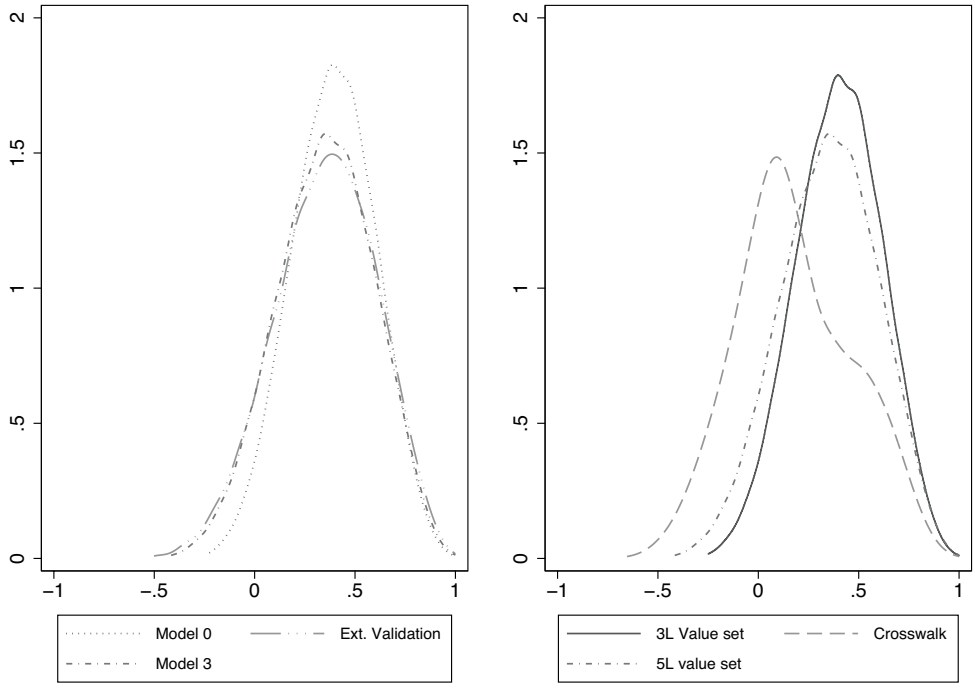

Fig. 5 – Comparison of the distribution of the selected value set (model 3).

DISCUSSION

In this article, we have presented two main findings. The first finding is that our approach to handling data quality issues in the C-TTO responses can be incorporated into the modeling of the EQ-VT data to improve the estimation of EQ-5D-5L value sets. The second finding is the reporting of an EQ-5D-5L value set based on version 1 of the EQ-VT protocol to inform health technology assessment in the Spanish setting.

The approach introduced here was developed from previous work introducing the hybrid model to estimate a value set using C-TTO and DCE responses [10,15,20,21]. The estimation of a hybrid model, although initially feasible, did not address the data quality issues encountered during the valuation study [12]. New evidence from a follow-up study in Spain suggested that the values for severe health states calculated with the reference case (hybrid model 0) were upward biased because of the data quality issues. At that time, we decided to further develop the hybrid model to allow the use of intervals and censored responses. After these post hoc adjustments, the final model (model 3) produced predictions that were closer to follow-up predictions than the reference case (model 0).

In the process of creating our approach, we developed a path analysis of C-TTO responses to produce intervals that may better represent individual preferences than indifference points (i.e., indifference point). We recognized that a C-TTO response is not only affected by interviewer behavior but also limited to half-year units by the design. The EQ-VT task generates a total of 41 unique values ranging from −1 to 1 (Fig. 1) and disallows responses less than −1 (censoring), which may not be sufficient to accurately reflect a respondent's value of a health state; more importantly, independently of interviewer's behavior, it is possible that only some respondents are capable of accurately reporting a range of values for a health state. By modeling ranges (i.e., open and closed intervals) as described by C-TTO paths, interval regression analyses benefit from both an improvement in precision (beyond 41 points) and the mitigation of behavioral imprecision (i.e., straight-lining, satisficing, circling, wandering, and censoring) in the iterative procedure. Hence, we encourage further investigation of the interval regression on the basis of pathway analysis for EQ-5D-5L valuation data or similar health preference data (e.g., standard gamble).

The final results (model 3) further suggested some differences between the EQ-5D-3L and the EQ-5D-5L value sets in Spain, which could be due to the instrument or study design. The sample of the valuation study comprised a representative sample of the Spanish population, whereas the EQ-5D-3L value set was estimated using only a representative sample of Catalonians [17]. In the original EQ- 5D-3L value set, mobility had the largest value decrement from all EQ-5D dimensions (i.e., confined to bed), but in the EQ-5D-5L, this label was replaced with "unable to walk about" and anxiety and depression had the largest decrement in the EQ-5D-5L value set. In addition to the amendments in labelling, a possible explanation is that preferences of the Spanish population have changed over time because of changes in the socioeconomic environment. For instance, the EQ-5D-3L value set was estimated more than 15 years ago, when the socioeconomic situation in Spain was in a different state than at the time of the EQ-5D-5L valuation study. The current economic situation in Spain has been associated with an increase in the number of people with mental health problems in the country [22–25], and hence the EQ-5D-5L value set reported in this article provides a more realistic representation of the current health preferences in Spain than the original EQ-5D-3L value set.

Study Limitations

This study is subject to some limitations. Definitions of intervals on the basis of limited information such as fewer than three steps could have impacted on modeling results; we have, however, shown that the final model basically replicated results of the follow-up study. The fact that we were not able to estimate a consistent heteroscedastic model made us carefully interpret the resulting P values for the model's coefficients. Nevertheless, we tried to limit the impact of this limitation by making a cluster estimation of the standard errors. In addition,

the data used for external validation are not representative of the Spanish population, but only from one province. Narrow intervals were more informative than wide ones; therefore, the interval analyses naturally emphasized the C-TTO responses of persons who knew their health preferences and understood the C-TTO task, which may affect the generalizability of the results. Finally, since the completion of this study, the EuroQol international valuation protocol has been updated to version 2, which incorporates new features to improve data quality including a quality control process and a feedback module [26] for C-TTO responses. Evidence from a new wave of studies using this updated protocol suggests that some of the problems encountered in this study during the original data collection are no longer present [27]. Future research should assess the robustness of the value set presented in this article with models using data from the new version of the protocol. The authors encourage such research to understand the implications of using the recommended EQ-5D- 5L value set in this article in health technology assessment decisions in Spain.

Conclusions

We explored statistical methods for handling data quality issues in the C-TTO task of the EQ-VT. On the basis of our findings, these analytical adjustments improved external validity and led to the development of a novel interval approach for the analysis of C-TTO responses. Given that the impact of data quality issues is predictable and not unique to the Spanish valuation study, we think that the lessons we learned can be useful to other health preference researchers. Furthermore, we recommend that, in future analysis of EQ-5D-5L valuation data, researchers consider including a similar approach to modeling C-TTO and DCE responses, particularly the examination of intervals on the basis of respondent behaviors. This article also provides a Spanish EQ-5D-5L value set that is recommended for use in health technology assessment in Spain.

Source of financial support: This study was financially supported by the Instituto de Salud Carlos III, Subdirección General de Evaluación y Fomento de la Investigación, Plan Estatal de Investigación Científica y Técnica y de Innovación 2013–2016, and the Fondo Europeo de Desarrollo Regional (grant no. PI12/02103). The EuroQol Research Foundation partially funded this work.

Supplemental Materials

Supplemental material accompanying this article can be found in the online version as a hyperlink at http://dx.doi.org/10.1016/j. jval.2017.10.023 or, if a hard copy of article, at www.valueinhealth journal.com/issues (select volume, issue, and article).

REFERENCES

1. Oppe M, Devlin NJ, van Hout B, et al. A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. Value Health 2014;17:445–53.
2. Janssen BM, Oppe M, Versteegh MM, Stolk EA. Introducing the composite time trade-off: a test of feasibility and face validity. Eur J Health Econ 2013;14(Suppl. 1):S5–13.
3. Oppe M, Rand-Hendriksen K, Shah K, et al. EuroQol protocols for time trade-off valuation of health outcomes. Pharmacoeconomics 2016;34:993–1004.
4. Lancsar E, Louviere J. Conducting discrete choice experiments to inform healthcare decision making: a user's guide. Pharmacoeconomics 2008;26:661–77.
5. Torrance GW, Thomas WH, Sackett DL. A utility maximization model for evaluation of health care programs. Health Serv Res 1972;7:118–33.
6. Dolan P. Modeling valuations for EuroQol health states. Med Care 1997;35:1095–108.
7. Robinson A, Spencer A. Exploring challenges to TTO utilities: valuing states worse than dead. Health Econ 2006;15:393–402.
8. Augustovski F, Rey-Ares L, Irazola V, et al. Lead versus lag-time trade-off variants: Does it make any difference? Eur J Health Econ 2013;14 (Suppl. 1):S25–31.
9. Luo N, Li M, Stolk EA, Devlin NJ. The effects of lead time and visual aids in TTO valuation: a study of the EQ-VT framework. Eur J Health Econ 2013;14(Suppl. 1):S15–24.
10. Ramos-Goñi JM, Pinto-Prades JL, Oppe M, et al. Valuation and modeling of EQ-5D-5L health states using a hybrid approach. Med Care 2017;55: e51–8.
11. Ramos-Goñi JM, Rand-Hendriksen K, Pinto-Prades JL. Reintroduction of the ranking task in valuation studies: Improved data quality and reduced level of inconsistencies? The case for EQ-5D-5L. Value Health 2016;19:478–86.
12. Ramos-Goñi JM, Oppe M, Slaap B, et al. Quality control process on EQ- 5D-5L valuation studies. Value Health 2017;20:466–73.
13. Craig BM, Runge SK, Rand-Hendriksen K, et al. Learning and satisficing: an analysis of sequence effects in health valuation. Value Health 2015;18:217–23.
14. Attema AE, Versteegh MM, Oppe M, et al. Lead time TTO: Leading to better health state valuations? Health Econ 2013;22:376–92.
15. Ramos-Goñi J, Craig AM, Oppe M, Van Hout B. Combining continuous and dichotomous responses in a hybrid model. Available from: https:// euroqol.org/wp-content/uploads/working_paper_ series/EuroQol_ Working_Paper_Series_Manuscript_16002_-_Juan_Ramos-Goni.pdf. [Accessed September 16, 2017].
16. Cameron AC, Trivedi PK. Microeconometrics Using Stata. College Station, TX: Stata Press, 2009.
17. Badia X, Roset M, Herdman M, Kind P. A comparison of United Kingdom and Spanish general population time trade-off values for EQ-5D health states. Med Decis Making 2001;21:7–16.

18. van Hout B, Janssen MF, Feng YS, et al. Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. Value Health 2012;15: 708–15.

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

20. Oppe M, van Hout B. The optimal hybrid: experimental design and modeling of a combination of TTO and DCE. EuroQol Group Proceedings. 2013. Available from: http://www.euroqol.org/uploads/ media/EQ2010_-_CH03_-_Oppe_-_The_optimal_hybrid_-_Experimental_ design_ and_modeling_of_a_combination_of_TTO_and_DCE.pdf. [Accessed October 11, 2014].

21. Rowen D, Brazier J, Van Hout B. A comparison of methods for converting DCE values onto the full health-dead QALY scale. Med Decis Making 2015;35:328–40.

22. Córdoba-Doña JA, Escolar-Pujolar A, San Sebastián M, Gustafsson PE. How are the employed and unemployed affected by the economic crisis in Spain? Educational inequalities, life conditions and mental health in a context of high unemployment. BMC Public Health 2016;16:267.

23. Bartoll X, Palència L, Malmusi D, et al. The evolution of mental health in Spain during the economic crisis. Eur J Public Health 2014;24:415–8.

24. Roca M, Gili M, Garcia-Campayo J, García-Toro M. Economic crisis and mental health in Spain. Lancet 2013;382:1977–8.

25. Gili M, Roca M, Basu S, et al. The mental health risks of economic crisis in Spain: evidence from primary care centres, 2006 and 2010. Eur J Public Health 2013;23:103–8.

26. Wong E, Ramos-Goñi JM, Cheung A, et al. Assessing the use of a feedback module to model EQ-5D-5L health states values in Hong Kong [published online ahead of print October 10, 2017]. Patient. http://dx. doi.org/%2010.1007/s40271-017-0278-0.

27. Purba FD, Hunfeld JAM, Iskandarsyah A, et al. The Indonesian EQ-5D-5L value set. Pharmacoeconomics 2017;35:1153–65.

10

Chapter 10

General discussion

Juan M. Ramos-Goñi

Note: This chapter uses text sections from a paper under the review process in Value in Health. Reference: *Stolk E, Ludwig E, Rand K, van Hout B, Ramos-Goñi JM. Overview, update and lessons learned from the international EQ-5D-5L valuation work: version 2 of the EQ-5D-5L valuation protocol*

This thesis has investigated strategies for conducting valuation studies of EQ-5D-5L. Specifically, it has focused on testing and improving the initial version of the EQ-5D-5L valuation protocol (version 1.0), in addition to investigating how to improve the modelling of the preference evidence.

A detailed description of version 1.0 of the EQ-5D-5L valuation protocol and its software (EQ-VT) is provided by Oppe et al. [1]. Briefly, the protocol requires that preferences are elicited in face-to-face interviews using C-TTO and DCE. C-TTO is a modified version of the conventional TTO variant employed in the influential Measurement and Valuation of Health (MVH) study [2] and most subsequent EQ-5D-3L valuation studies. For the valuation of health states considered to be better than dead, C-TTO offers respondents the conventional TTO task comprising a series of adaptive choices between x years in full health and 10 years in-a disease state. In an iterative procedure, x is varied to identify the respondent's point of indifference where the health state value is given by x/10. When a respondent considers a health state to be worse than dead, lead-time TTO is used. Respondents receive a series of choices between x years in full health and a fixed life of 10 years in full health followed by 10 years in the target state. As before, x is varied until indifference is reached and the health state is given by (x-10)/10 [3, 4]. Interviewers are instructed to use the 'wheelchair example' as a means of explaining the C-TTO task and showing the range of possible answers (i.e., both better than dead and worse than dead). After the example and 10 real C-TTO tasks, each respondent completes DCE tasks, comparing two health states and indicating which of the two is best. The protocol enables analysts to estimate EQ-5D-5L values using C-TTO responses, DCE responses, or both types of responses, as illustrated, for instance, by the hybrid models [5-7].

The EQ-5D-5L protocol was a major advance in health valuation, not just because it included an updated approach to TTO valuation, but also because it was incorporated into software for computer-assisted personal interviews: the EQ-VT. This was designed to support the process of standardizing the valuation task across studies and to manage data collection efficiently. However, EQ-VT offers the further benefit that the subjects' responses are immediately stored and available, and in addition allow meta data to be stored such as the path taken in C-TTO to a response and the time between mouse clicks. This thesis has exploited this technological advance, making it possible to contribute to the health preference research field in two ways. First, the automatic data storage facility allowed the PIs to retrieve the data on a regular basis, making it possible to monitor incoming data continuously and to implement a quality control procedure. Hence one of the main contributions of this thesis has been the introduction of a metric/ tool that facilitates quality control, and enables an exploration of how its use might affect the quality of the elicited preference data. Second, but no less important, this thesis has exploited the richness of the stored information by examining the path of values visited during the C-TTO tasks, thus allowing the definition of

intervals representing a respondent's uncertainty. In other words, this thesis describes a way to analyse respondents' behaviour during each C-TTO task.

These two contributions are related because it was the richness of the data, comprising both values and meta data, that led to important new insights with respect to the C-TTO tasks. The main lessons learned are discussed below.

1 To what extent was the proposed valuation protocol feasible and hybrid estimations possible in practice?

Chapter 2 reports on the feasibility of the EQ-5D-5L valuation protocol as it was tested in the context of a national valuation study conducted in Spain, the first in which this test was conducted. Feasibility was demonstrated in two respects: (i) collection of data as described in the protocol, (ii) estimation of a hybrid model as described by Oppe and van Hout [6]. In the case of the Spanish EQ-5D-5L valuation study, it was important / advantageous that both C-TTO and DCE data were collected with the idea to potentially combine them (hybrid estimation). In Spain, some **issues** were present in the C-TTO data. Being able also to include DCE responses in the modelling of the Spanish valuation data increased the power of the model and resulted in more precise parameter estimates than could have been derived from C-TTO data alone.

In addition, chapter 2 elaborated arguments with respect to the use of both sources of preference data in a single value set estimation (hybrid estimation). The main argument in support of this novel analytical approach is that both C-TTO and DCE are intended to measure the same concept, namely health preferences. Hence the two sources of preference data would complement each other. However, the question remains: if both techniques measured the same concept, why were their results not identical? The chapter discussed the point that any valuation technique has related limitations that could introduce bias into its estimates. For example, there is an extensive literature concerning the limitations associated with C-TTO, such as scale compatibility or loss aversion [8]. However, it appears that these limitations have not been present in the DCE data. On the other hand, there have been limitations associated with the DCE technique such as prominence effects [9] that do not appear to be present in the C-TTO data. Hence, when combining both sources of data, the results appeared to be less problematic than when using only one source of data, as the limitations related to one technique may have compensated for the limitations related to the other technique. This argument seemed to be valid in the first test. Some recently published hybrid value sets in other countries have provided extended evidence for this argument [10-13]. However, there are more reasons to justify combining C-TTO and DCE data.

Chapter 8 introduced the "interval hybrid regression model". TTO data is traditionally modelled as point estimates that do not fit within a framework of interval regression.

However, the metadata that the EQ-VT software routinely collected with the value data, revealed a significant influence of the C-TTO design task on values (moderated by task engagement levels) that changed the perception of the C-TTO data. The accuracy of some of the observations was very low, and thus an interval regression framework could also account for low engagement on the part of the respondent. This regression model combines interval responses from C-TTO with DCE responses. The intervals used in interval regression may be closed intervals such as [a, b] (values in the range a-b including both a and b, note that this includes the case of a=b, i.e. points), or the intervals may be open intervals, e.g.]-∞, a] or [a, ∞[. The latter is especially interesting; it implies that the value of a given health state "S" is lower or higher than "a" for]-∞, a] or [a, ∞[respectively. When looking at a single DCE response, for example State1 is preferred to State2, this means that the value for State1 is higher than that for State2, therefore this is an open interval. However, there is a small difference with the C-TTO open intervals; instead of having a numeric value "a" as limit of the interval, like in C-TTO, in DCE the limit of the interval is the value of the other health state presented in the same DCE task. Nevertheless, from a mathematical point of view, both valuation techniques (C-TTO and DCE) can be seen as sources of interval information. Given that interval regression was developed as an extension of the Tobit model [14], it seems natural to combine the intervals arising from C-TTO valuations with those from DCE valuations.

Taking into account the two arguments: 1) both C-TTO and DCE try to measure the same concept with no shared limitations, and 2) both C-TTO and DCE techniques produced valuation data in interval form; the use of the hybrid approach can be justified. Hence the proposed protocol is feasible and sensible. However, as stated in chapter 2, the feasibility test showed room for improvement, which motivated the next research question.

2 *Is there an explanation for the interviewer effects found in chapter 2? If so, how to modify the existing protocol to collect better data?*

The first tranche of EQ-5D-5L valuation studies were conducted in Spain, England, the Netherlands, China, and Canada using version 1.0 of the EQ-5D-5L valuation protocol [1]. Following the results of these valuation studies, concerns were raised over observations of high rates of inconsistent responses, clustering of values, low values for mild states, few worse than dead responses and interviewer effects [7,10,15-16]. To clearly define the problem, chapter 5 of this thesis analysed the valuation data in depth. This was possible because the EQ-VT software captured the entire path followed to reach a value, and the time-stamps between mouse clicks.

Exploiting the richness of this meta-data, chapter 5 reported that some interviewers systematically omitted explanation of the lead-time section of the C-TTO task and elicited no worse than dead values. Furthermore, in some interviews, interviewers used very little time

in explaining the C-TTO task; and only obtained a single C-TTO value, which could indicate that respondents minimized effort and expedited the C-TTO tasks by reducing the number of iterations. Since the iterative procedure requires a different number of steps to reach specific values, lack of effort in the task may have partially accounted for the relatively low values for mild states (it takes more steps to reach the ends of the scale), and clustering of values (a limited number of values can be attained when a C-TTO task is completed with little iteration). Of importance is that the occurrence of these issues was found to vary across interviewers, suggesting that interviewer behaviour had an effect on the tendency for respondents to use such short-cuts.

When respondents make choices that result in quick task completion; this is indicative of respondent behaviour that technically complies with the requirements of the task, but may still be detrimental to the precision of the answers that are obtained: this was presented in chapter 5 as a general phenomenon called "*satisficing*"; as described in chapter 3. Indirectly, this may have accounted for the large number of inconsistent valuations. While the worst state described by the EQ-5D-5L descriptive system is dominated by all other EQ-5D-5L states, roughly 20% of respondents valued at least one health state higher and the observed utility difference was large (>0.5 on the utility scale) in almost half of those cases [17]. An inconsistency could itself be assigned to different causes, such as task complexity, random error, or learning effects, but it could also reflect inadequate efforts from respondents who did not feel compelled to expend resources on providing optimal answers. In a broader sense, the findings seem to have indicated an, at times, low level of task engagement of respondents or interviewers, with detrimental effects on the quality of the data.

In order to collect better data using the same protocol, version 1.1 of the EQ-5D-5L valuation protocol was developed. In this version 1.1, several suggestions were implemented. Specifically, three practice states were added after the wheelchair example, to better familiarize respondents with the C-TTO task and with the severity range of health state descriptions. In addition, confirmatory pop-ups were implemented to validate answers before storing them. Furthermore, the key modification was to commence the monitoring of interviewer performance during data collection to enable timely intervention if problems were detected. Accordingly, we introduced the quality control (QC) procedure described in chapter 5, that involved the production of standardized reports periodically while the study was ongoing in order to review protocol compliance and interviewer effects.

While all new requirements increase the cost per interview and may lead to the early exclusion of some interviewers, as well as all their interviewees, the effectiveness of the QC process is by now undisputed. Reported problems such as interviewer effects [7], clustering effects [10], and severe inconsistencies [18], were dramatically reduced in studies that adhered to the QC procedure described in chapter 5 [11-12].

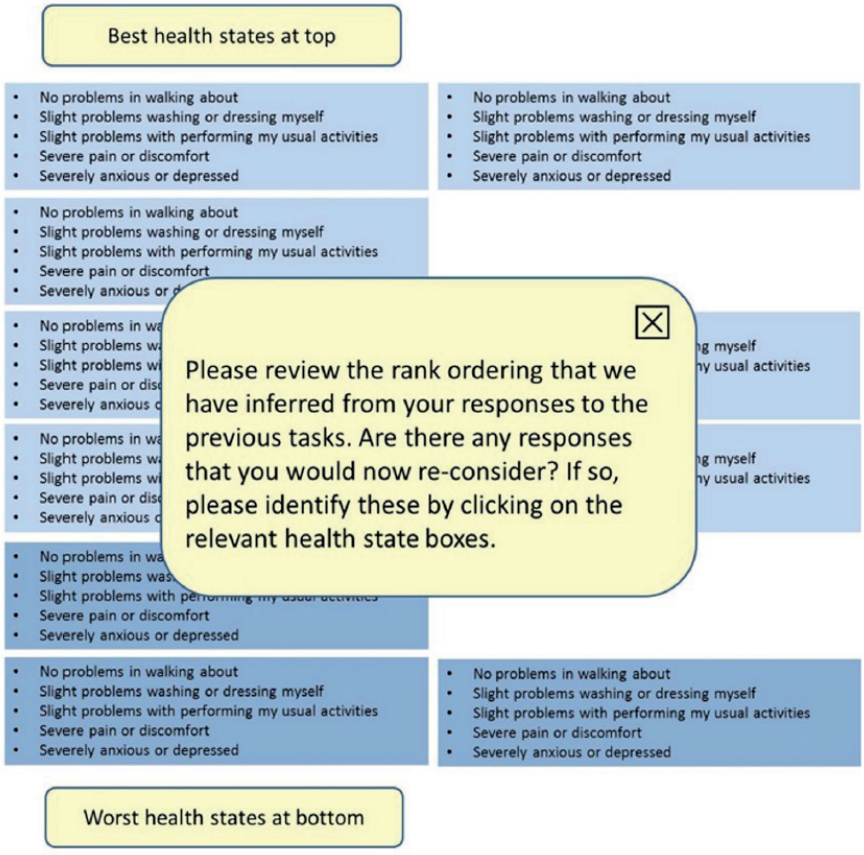

Figure 1: Example of feedback module in EQ-VT

There appears to be little room for further improvement in interviewer administered C-TTO. When version 1.1 of the EQ-5 D-5L valuation protocol was introduced, a research programme started to test several further C-TTO task modifications that were proposed as improvements. Teams from Spain, Japan, the Netherlands, Hong Kong, Norway, Singapore, Germany, and England were involved in this research programme, each allocating participants to a control group that received version 1.1 of the protocol, or an experimental group that received a modified protocol [17]. Given the great improvement produced by the inclusion of the QC procedure, only one modification deserved more attention, which was the inclusion of a feedback model after the C-TTO task completion. In this feedback module (Figure 1), the respondent was able to look at the valued health states in the order specified by him/her, and had the opportunity to flag one or multiple health states not located in his/her preferred order. No attempt was made to derive new C-TTO values but flagged responses could be removed before analysing the data. Further details about the feedback module can be found

elsewhere [12]. The other tested modifications, among which was the reintroduction of a ranking task prior to the C-TTO task, described in chapter 4, provided no substantial benefit or resulted in mixed results across countries.

All of the other modifications tested produced only marginal improvements. The reintroduction of the ranking task described in chapter 4 produced few improvements, with just a small reduction in the proportion of inconsistencies. The fact that the improvements were marginal, together with the considerable associated cost of including the ranking task, made the decision not to include this task an easy one.

3 What types of modelling techniques are more suitable for modelling the C-TTO and DCE valuation data?

While key features of raw C-TTO data have been analysed in recent years, researchers have also investigated what the findings imply for modelling the data. EQ-5D-3L TTO data were often modelled using simple linear regression or random effects models. Modelling approaches to EQ-5D-5L C-TTO data are new in that they account for censoring, heteroscedasticity, truncation, preference heterogeneity and response uncertainty. The advances made in modelling have all been driven by considerations obtained from carefully investigating aspects of the C-TTO task and the data thus generated, and by matching these to the assumptions underlying the regression models. These considerations can be broadly categorized into three groups related to: 1) the mechanics of the C-TTO task itself; 2) individual respondent behaviour; and 3) characteristics of the complete C-TTO dataset; as described in table 1.

An obvious reason to adopt a framework for censored data is that C-TTO data is left-censored at -1, but it is novel that we also consider the presence of other types of censoring. Table 1 identifies several factors contributing to the view that C-TTO responses can have a low level of accuracy and therefore may be better construed as indicating ranges of values within which the point of indifference is likely to reside, rather than discrete indifference points. For instance, left-censoring may not be limited to the bottom value of the scale but could also occur at zero for people who were not properly introduced to the worse than dead task. The latter is described when generating the Spanish value set in chapter 9.

Table 1: Overview of phenomena that characterize C-TTO data and how they can be modelled using the hyreg command developed in Chapter 8.

	Phenomenon observed in the data	Implemented solutions
1) C-TTO task mechanics	Censoring: The bottom value in the C-TTO task is -1, lower values are censored. When a health state is valued at -1, the true value could be -1 or a lower value. If interviewers do not explain the worse than dead C-TTO task, respondents may not be aware of the possibility to give values <0 and censoring at 0 may occur.	tobit or interval regression
	Truncation: all health state values have an upper bound of 1, so that the distribution of values close to 1 is not normal but left-skewed. Consequentially, people can also only err in one direction: erroneous values can be too low but not too high. Models that assume equal and normal distribution of the error term will produce biased values for mild states.	tobit or interval regression
	Smallest tradeable unit: in the EQ-VT, the smallest tradeable unit is 6 months, defining 41 discrete values rather than a continuous scale of values. Greater precision would be obtained if smaller units were offered (days, weeks, months).	interval regression
2) Individual respondent behaviour	Satisficing/Straight-lining: When the iteration procedure homes in on the indifference point, respondents may not complete the task up to intended standards and state indifference early in the sequence to escape follow-up questions.	interval regression
	Circling/Wandering: It may be the case that respondents only know the range in which the indifference point falls but are unable themselves to locate an exact value. The first stimulus within that range would trigger an indifference statement.	interval regression
	Use of the numeraire. Using time as a basis for valuing health, time preferences and extrinsic goals requiring a certain lifespan can impact on values. Consequentially, people may apply their personal maximum trade-off in life years to a group of health states. Similarly, some respondents may not accept any reduction in lifespan until a severity threshold is passed. Such phenomena may exist even when respondents have a genuine preference for one health state over another, thus obscuring differences.	tobit or interval regression
3) Characteristics of the complete C-TTO dataset	Heteroscedasticity refers to the phenomenon where the error around observed values is not constant. Medium or poor states have different variability than mild states. In general, the worse the health state, the wider the standard deviation. To prevent biased parameter estimates and/or standard errors, the error terms can be modelled based on the severity of the health state.	heteroscedastic models

In summary, this thesis also contributes to the field from the modelling point of view, not only by making novel applications of existing regression methods to the data collected (specifically, chapter 7 applies robust regression to model C-TTO data), but also by developing new methods and their related software implementation when required (see chapter 8).

The fact that the EQ-VT protocol stores each mouse click is not only useful for checking times and performing the quality control procedure described in chapter 5, but also the stored paths to reach indifference points for each task facilitates to the use of all the information given by the respondent on each task for each health state. Looking at the paths that respondents followed we were able to identify several types of behaviour: 1) Circling; 2) Wandering, 3) Satisficing and 4) Straight-lining. "Circling" is an expected behaviour as this can occur when the respondent displays uncertainty around nearby values. "Straight-lining" is also expected and reasonable behaviour as there are cases of exhausting all time or not trading any. However, both the "wandering" and "satisficing" behaviours are problematic. On the one hand, wandering behaviour occurs when there is a poor understanding of the task; hence it appears that a respondent goes through the iterative procedure without a clear idea of what s/he is doing. This reflects a high level of uncertainty about where the true value is. On the other hand, the satisficing effect occurs when the respondent is not fully engaged with the task and s/he only goes through a minimum number of steps in the iterative procedure. This also produces a high level of uncertainty about the true value, as it is unknown whether using more steps in the iterative procedure could provide a more accurate response. To the best of my knowledge, there have been no previous valuation studies that have investigated this type of behaviour, meaning that earlier value set estimations could suffer from bias. In chapter 9, a way was developed to take into account this type of uncertainty when developing a value set, and given the level of concordance obtained with external validation data, it appeared that the estimations were reasonable. Hence any valuation study that uses valuation techniques based on iterative procedures should explore this interval-based approach.

A further innovation was the introduction of models that accommodate heteroscedasticity and non-normality, as described mathematically and implemented in chapter 8. In health valuation, variability increases with severity; there is little disagreement that mild health states are good, but opinions diverge concerning how bad moderate and severe states really are. A cause for non-normality is that values for health have a maximum of 1, which is often referred to as 'truncation'. When relatively mild health states are valued, many values at 1 or close to 1 will be obtained, resulting in a skewed distribution, in which outliers can be identified only at one side of the peak. These outliers can cause bias and result in estimates that are too low, especially for mild states. Accommodating these concerns has practical value in achieving values close to 1 that are modelled well. Of note is that models for censored data can also be used to accommodate for these factors, which contribute to their popularity.

Future research

Comparison of previous EQ-5D-3L value sets with the more recently developed EQ-5D-5L value set within countries has commenced. However, the fact that population preferences for EQ-5D-3L health states were collected, in many cases, more than 20 years ago, has not been considered [19]. In the specific case of Spain, the time difference between studies was around 12 years and, as shown in chapters 2 and 9, differences with respect to the relative importance of dimensions occurred. One could argue that these differences were only due to differences in the instruments or in the valuation methods used. However, as discussed in chapter 9, it seems that these were not the only issues, at least in Spain where the economic crisis appears to have had an important impact in how the population views health problems. As the economic crisis has been worldwide, and has not only occurred in Spain, it is worth considering whether the Spanish case can be safely extrapolated to any other country. However, another reason can partially explain the observed differences. It may have been that the policies implemented to make the life of physically disabled people easier have had an impact on how the population views mobility-related problems, which could explain the reduction in the relative importance of the mobility dimension. Having stated this, and given the uncertainty concerning what really influences the population to change its health preferences, then the development of a value set necessitates a longitudinal approach rather than a transversal one. In other words, valuation studies for a specific instrument should be performed at least each 5-10 years, the exact time being related to the number of policy changes made in the country.

The remaining question is whether exactly the same method should be used again. In principle the response should be: "no", as there is always room for improvement. For example, even when explained well, the C-TTO task remains complex for respondents to carry out. One of the most difficult aspects of the task for the respondent is to give a precise value "t" for her/his point of indifference when choosing between 10 years in an impaired state and t years in full health. This can lead respondents to exhibit behavioural imprecision (satisficing, circling, and wandering) which will reduce the accuracy of their C-TTO responses. Satisficing is arguably the most problematic: once respondents start to feel that the choices are becoming difficult to make, they press the "A&B are about the same" button to terminate the task, instead of continuing until they reach their "true" point of indifference. Satisficing can thus lead to inaccurate preferences being recorded by respondents, which may bias modelling results. While some respondents may be struggling to determine their indifference point, they might be capable of reporting their indifference range of values for a health state more accurately. In addition to the behavioural imprecision of the respondent's data, the C-TTO task itself also inhibits a degree of imprecision (i.e., task imprecision) as only year and half-year values of t are included in the iteration sequence, leading to 41 discrete values, rather than a continuous range of values. Changing the termination rule in C-TTO (i.e., removing the obligation to

provide an indifference point) may reduce behavioural and TTO task imprecision. I have proposed to change the current termination rule of the TTO by utilizing the following two conditions: 1) A fixed number of steps in the iterative sequence, and 2) "A specific" interval of the range of values where the true preference is, e.g,. when a respondent is circling between values of a width of 1 year. Hence, the future of C-TTO tasks should be interval-based instead of based on indifference points.

Conclusions

The studies presented in this thesis, together with similar work accomplished elsewhere, have resulted in a detailed valuation protocol for the EQ-5D-5L instrument, paired with a quality assurance procedure and novel analytical approaches. The updated protocol has enabled teams from all over the world to successfully establish EQ-5D-5L value sets.

Despite TTO being a preferred method for health state valuation for two decades, important new insights have been achieved concerning how respondent behaviour and specific features of the valuation task work together to define the level of precision of C-TTO responses. These insights have emphasized the importance of the interviewer's role in C-TTO valuation, and made it evident that it is unlikely that any interview protocol or software for performing interviews is sufficient to guarantee proper interviewer competence, compliance, or engagement. This motivated the i**n**troduction of a QC process and of new modelling approaches. Looking back, one may wonder how widespread similar issues were in other valuation studies, and why they have not been identified (and resolved) previously. Most likely, similar issues have always been there, but were simply not noticed.

Given the changes in thinking concerning what works in valuation exercises, it would be appropriate to recognize that – at least from a valuation perspective – the rigorous approach to EQ-5D-5L valuation studies inspires trust. However, it can also be noted that the kind of insights that guide our current valuation work did not exist when EQ-5D-3L was originally valued. It is wise to remain modest with respect to claims about the qualities of older value sets derived from instruments such as EQ-5D-3L, SF-6D, AQoL and HUI, until we have returned to scrutinize them. Such research is warranted to provide users, stakeholders and society with recommendations regarding the use of the instruments in analysis and health care decision-making, to the benefit of all patients.

REFERENCES

1. Oppe M, Devlin NJ, van Hout B, Krabbe PF, de Charro F. A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. Value Health. 2014 Jun;17(4):445-53.
2. Dolan P. Modeling valuations for EuroQol health states. Med Care 1997; 35(11):1095-1108.
3. Janssen BMF, Oppe M, Versteegh MM, Stolk EA. Introducing the composite time trade-off: a test of feasibility and face validity. Eur J Health Econ 2013;14 Suppl 1:S5-13. doi:10.1007/s10198-013-0503-2.
4. Devlin NJ, Tsuchiya A, Buckingham K, Tilling C. A uniform time trade off method for states better and worse than dead: feasibility study of the 'lead time' approach. Health Econ 2011;20:348–61.
5. Craig BM, Busschbach JJ. The episodic random utility model unifies time trade-off and discrete choice approaches in health state valuation. Popul Health Metr. 2009 Jan 13;7:3.
6. Oppe M, van Hout B. The optimal hybrid: experimental design and modeling of a combination of TTO and DCE. EuroQol Group Proceedings. 2010. [available at: http://eq-5dpublications.euroqol.org/download?id=0_53738&fileId=54152, accessed June 6, 2017]
7. Ramos-Goñi JM, Pinto-Prades JL, Oppe M, Cabasés JM, Serrano-Aguilar P, Rivero-Arias O. Valuation and Modeling of EQ-5D-5L Health States Using a Hybrid Approach. Med Care. 2014 Dec 17.
8. Bleichrodt H, Pinto JL. Loss aversion and scale compatibility in two- attribute trade-offs. *J Math Psychol*. 2002;46:315–337.
9. Hawkins SA. Information processing strategies in riskless preference reversals: the prominence effect. *Organ Behav Hum Decis Proces*. 1994;59:1–26.
10. Devlin NJ, Shah KK, Feng Y, Mulhern B, van Hout B. Valuing health-related quality of life: An EQ-5D-5L value set for England. Health Econ. 2017 Aug 22. doi: 10.1002/hec.3564. [Epub ahead of print]
11. Purba FD, Hunfeld JAM, Iskandarsyah A, Fitriana TS, Sadarjoen SS, Ramos-Goñi JM, Passchier J, Busschbach JJV. The Indonesian EQ-5D-5L Value Set. Pharmacoeconomics. 2017 Jul 10.
12. Wong ELY, Ramos-Goñi JM, Cheung AWL, Wong AYK, Rivero-Arias O. Assessing the Use of a Feedback Module to Model EQ-5D-5L Health States Values in Hong Kong. Patient. 2017 Oct 10. doi: 10.1007/s40271-017-0278-0. [Epub ahead of print]
13. Ludwig K, Graf von der Schulenburg J-M, Greiner W. German value set for the EQ-5D-5L. Phamacoeconomics 2017. In press.
14. Amemiya T. Advanced Econometrics. Harvard University Press. Cambridge Massachusetts. ISBN: 0-674-00560-0.
15. Versteegh MM, Vermeulen KM, Evers, Silvia M. A. A., Wit GA de, Prenger R, Stolk EA. Dutch Tariff for the Five-Level Version of EQ-5D. Value Health 2016;19:343–52.
16. Luo N, Liu G, Li M, Guan H, Jin X, Rand-Hendriksen K. Estimating an EQ-5D-5L Value Set for China. Value Health. 2017 Apr;20(4):662-669. doi: 10.1016/j.jval.2016.11.016. Epub 2017 Feb 9.

17. Shah K, Rand-Hendriksen K, Ramos JM, Prause AJ, Stolk E. Improving the quality of data collected in EQ-5D-5L valuation studies: a summary of the EQ-VT research methodology program. 2014 EuroQol Proceedings. [available at: http://eq-5dpublications.euroqol.org/download?id=0_53918&fileId=54332, accessed June 6, 2017]

18. Al Sayah F, Johnson JA, Ohinmaa A, Xie F, Bansback N; Canadian EQ-5D-5L Valuation Study Group.. Health literacy and logical inconsistencies in valuations of hypothetical health states: results from the Canadian EQ-5D-5L valuation study. Qual Life Res. 2017 Jan 25.

19. Hernandez-Alava M, Wailoo A, Grimm S, Pudney S, Gomes M, Sadique Z, Meads D, O'Dwyer J, Barton G, Irvine L. EQ-5D-5L versusEQ-5D-3L:The Impact on Cost-Effectiveness in the United Kingdom. Value in Helath 2017. In press.

11

Chapter 11

Samenvatting

Recent is een nieuwe versie van de EQ-5D standaard vragenlijst geïntroduceerd: de EQ-5D-5L. Om de EQ-5D-5L bruikbaar te maken voor economische evaluaties, moeten nationale waarderingen voor de gezondheidstoestanden van de EQ-5D-5L worden bepaald. Er is voorgesteld dat alle landen waar dit gebeurt gebruik maken van hetzelfde, gestandaardiseerde protocol voor waarderingsstudies, zodat het mogelijk is gezondheidswaarderingen tussen landen te vergelijken. Het voorgestelde protocol voor waarderingsstudies omvat twee technieken voor het waarderen van gezondheid, genaamd: “Composite Trade-Off(C-TTO)” en ”Discrete Choice Experiments" (DCE). Dit proefschrift beschrijft de eerste ervaringen met dit gestandaardiseerde protocol en de ontwikkelingen die sindsdien hebben plaatsgevonden, met een focus op de volgende drie vragen:

1. Wat voor problemen kan men verwachten bij het gebruik van het EQ-5D-5L evaluatie protocol bij het genereren van nationale waardebepalingen?
2. Hoe kan dit protocol verbeterd worden?
3. Hoe kan de verkregen data het best verwerkt worden om een waardebepaling te ontwikkelen?

HOOFDSTUK 2 presenteert één van de eerste toepassingen van dit gestandaardiseerde protocol voor de EQ-5D-5L waardering. Het bleek goed mogelijk om gezondheidswaarderingen te bepalen op basis van de verzamelde data, maar de resultaten werden sterk beïnvloed door interviewereffecten. De conclusie van dit hoofdstuk is dat een betere implementatie van het protocol noodzakelijk is om interviewereffecten te verminderen. Het beantwoorden van de tweede vraag was een grotere uitdaging en er is daarom aanvullend onderzoek gedaan onder de paraplu van een internationaal onderzoeksprogramma. Door gebruik te maken van verschillende DCE en C-TTO datasets werd eerst onderzocht wat het effect is van: i) het aantal vragen, en ii) de volgorde waarin ze gesteld zijn op de nauwkeurigheid van de antwoorden en op het gedrag van respondenten. HOOFDSTUK 3 toont aan dat de C-TTO data gevoeliger is voor dit soort effecten dan de DCE data. Deelnemers gaven vaak hun C-TTO respons al na het beantwoorden van slechts drie vragen in de iteratieprocedure. Dat is wel een efficiënte strategie, maar het betekent ook dat de nauwkeurigheid van de waarden beperkt is tot dat wat haalbaar is in drie stappen. In HOOFDSTUK 4 werd getest of het invoeren van een rangschikkingstaak vóór de C-TTO kan helpen om inconsistenties in de C-TTO antwoorden te verminderen en de datakwaliteit te verbeteren. In HOOFDSTUK 5 werd een kwaliteitscontroleproces (QC) ontwikkeld om interviewereffecten te verminderen. Een vergelijking van de resultaten van C-TTO studies uitgevoerd met en zonder QC-proces laat zien dat implementatie van het QC-proces in EQ-5D-5L waarderingsstudies de naleving van het protocol verhoogt en de datakwaliteit bevordert. De meerwaarde van de rangschikkingstaak was gering en weegt niet op tegen de extra inspanning die het vraagt.

Om de laatste onderzoeksvraag te beantwoorden, werden in dit proefschrift meerdere manieren voor het modelleren van C-TTO en DCE data onderzocht. In HOOFDSTUK 6 werden twee methoden geëvalueerd om bij het gebruik van DCE waarderingen te verkrijgen die op een schaal liggen waar het ankerpunt 'gezond' een waarde 1.00 heeft en ankerpunt 'dood' een waarde 0.00. In de eerste methode werden ankerpunten verkregen door in de DCE taak ook de gezondheidstoestand 'dood' aan te bieden. In de tweede methode werden de DCE resultaten op basis van C-TTO-observaties herschaald. De resultaten van beide methoden waren vergelijkbaar. HOOFDSTUK 7 rapporteert over 'robuuste regressie modellen', die gebruikt zijn voor het creëren van een C-TTO 'waarderingsset' voor Uruguay. Het bleek haalbaar om op basis van uitsluitend C-TTO data waarderingen voor EQ-5D-5L gezondheidstoestanden te bepalen, dus zonder de combinatie met DCE zoals geprobeerd was in hoofdstuk 2. Er waren wel robuuste regressie modellen voor nodig vanwege de sterke heterogeniteit. In HOOFDSTUK 8 werd het hybride model uit hoofdstuk 2 dat C-TTO en DCE combineert, verder ontwikkeld. Dit hybride model werd verfijnd door rekening te houden met intervalkenmerken van de data, het afkappen van de data op 1.00 en door rekening te houden met heteroskedasticiteit. Tot slot beschrijft HOOFDSTUK 9 hoe hybride interval regressie werd toegepast om een EQ-5D-5L waarderingsset voor Spanje te produceren, ondanks dat de Spaanse C-TTO data veel problemen kende. Dit hoofdstuk laat zien dat het de validiteit van de resultaten ten goede komt, wanneer men in de verwerking van data expliciet rekening houdt met de problematische kenmerken ervan.

Conclusies

De studies die in dit proefschrift worden gepresenteerd hebben, samen met soortgelijk werk elders, geresulteerd in een gedetailleerd waarderingsprotocol voor de EQ-5D-5L, gekoppeld aan een kwaliteitscontroleprocedure (QC-proces) en nieuwe analytische benaderingen. Het bijgewerkte protocol heeft teams van over de hele wereld in staat gesteld om met succes EQ-5D-5L waarderingssets te ontwikkelen.

Het QC proces werd geïntroduceerd vanuit de gedachte dat in grootschalige CTTO studies het optreden van interviewereffecten onvermijdelijk is en dat dit soort effecten een direct gevolg is van de complexiteit van de taak voor de interviewer. Door de data direct bij ontvangst te controleren, kan een onderzoeker vroegtijdig interviewereffecten ontdekken en ingrijpen, bijvoorbeeld door de interviewers individueel feedback te geven op hun functioneren. Het is onwaarschijnlijk dat interviewereffecten op deze manier volledig weggewerkt kunnen worden, maar de impact ervan wordt sterk verminderd. Verbeterde technieken voor het analyseren en moduleren van de data zorgen voor verdere reductie van de impact van problemen in de data.

Omdat de nieuwe modeleertechnieken en het QC proces problemen adresseren die inherent zijn aan data die nodig zijn voor het waarderen van vragenlijsten zoals de EQ-5D, en verder niet voortkomen uit een specifieke toepassing, lijken de nieuwe modeleertechnieken en het

QC proces universeel toepasbaar. De nieuwe modeleertechnieken en de QC proces technieken kunnen daarom worden ingezet voor betere waarderingssets voor de EQ-5D-5L of voor andere instrumenten, zoals de SF-6D, AQoL of HUI.

12

Chapter 12

Summary

Recently, a new version of the standard EQ-5D questionnaire called EQ-5D-5L with 5 levels on each dimension was developed. To make the EQ-5D-5L suitable for use in economic evaluations, national value sets need to be developed. A standardized valuation protocol has been suggested for that purpose, to enable comparison of values across countries. This protocol included two elicitation techniques, the 'composite time trade-off' (C-TTO), and 'discrete choice experiments' (DCE). This thesis describes experiences with first use of that standardized protocol and major evolvements happening to it since, with a focus on the following three questions:

1) What problems may be encountered in the use of the EQ-5D-5L valuation protocol to generate national value sets?
2) How can the protocol be improved?
3) How can the produced data best be modelled to develop a value set?

This thesis covered those questions in detail by conducting different research experiments. **Chapter 2** reports the results obtained from one of the first valuation studies that had made use of the standardized protocol for EQ-5D-5L valuation. While it was possible to generate a value set on the basis of data collected using the protocol,, major issues were encountered in the C-TTO data that was strongly affected by interviewer effects. The conclusion of this chapter is that improvements were needed, especially in controlling the variation in the way the interviewers administer C-TTO.

The first step in answering the second research question consisted in estimating the effect of sequence on response precision and response behaviour in different health valuation studies using C-TTO or DCE, or both. **Chapter 3** shows that sequence effects were present more in TTOs than in DCEs, but both showed some learning effect, i.e., participants learned to respond efficiently within the first three tasks and may have rounded their TTO responses. In parallel, **chapter 4** tested whether reintroducing a ranking task before the composite TTO (C-TTO) could help to reduce inconsistencies in C-TTO responses and improve the data quality; and **chapter 5** developed a cyclic Quality Control (QC) process to prevent interviewer effects. Results from those research experiments, showed that the implementation of a strict QC process in EQ-5D-5L valuation studies increased interviewer protocol compliance and promoted data quality. However, the benefit does not justify the effort involved in the ranking task.

The last research question of this thesis explored multiple approaches of modelling the produced data. **Chapter 6** evaluated two different methods to anchor the DCE scale in such a way that the 'health state' full health get the value of 1.00 and dead the value 0.00. The first model was a rank ordered logistic model which used the data from a DCE offering the health state 'dead' as one of the choices. The second model was a conditional logistic model using

pair comparison data from a DCE which was rescaled with lead-time TTO data; specifically using the value of the pits health state. Results showed that both models produced concordant results. **Chapter 7** tested maximum likelihood robust regression models with and without interactions to derive a value set from Uruguayan general population using the EQ-5D-5L. Results showed that it was feasible to obtain a value set by using only C-TTO data. However, robust regression estimation was needed due to the presence of strong heterogeneity. **Chapter 8** further developed the hybrid model described in chapter 2 by introducing interval responses or censored responses and by accounting for heteroscedasticity. Finally, **chapter 9** presents an effort to handle the data quality issues in C-TTO responses by estimating a hybrid interval regression model to produce the Spanish EQ-5D-5L value set. Results indicated that this approach improves validity.

Conclusions

The studies presented in this thesis, together with similar work accomplished elsewhere, have resulted in an evolved valuation protocol for the EQ-5D-5L instrument, paired with a quality assurance procedure and novel analytical approaches. The updated protocol has enabled teams from all over the world to successfully establish EQ-5D-5L value sets.

The quality assurance procedure has been put in place with the view that in large-scale, interviewer administered CTTO studies, inevitably interviewer effects will be encountered and frequently this happens because of the complexity of the task from the interviewer perspective. By monitoring the data when it comes in, the analysts can early detect such problems and intervene to improve and harmonize interviewer performance. While these measures are unlikely to eradicate issues in the data, their presence will be strongly reduced.

Improved modelling techniques ensure that the impact of issues present in the data can be reduced as well. Given that data issues will be present anyway, suggested methods as hybrid interval regression or heteroscedastic regression may help future value set developments, not only for EQ-5D instruments but also for others like SF-6D, AQoL or HUI.

.

13

Chapter 13

Acknowledgements

As I have not followed the usual academic life of a PhD student, there has been a number of people who has played a role in my career. In order to acknowledge them in chronological order, I have to start with my parents for being always there supporting me in every sense.

When I started my professional carrier, the one who first put trust in my research skills and motivated me to enter further employ the research field, was Antonio Sedeño-Noda, my teacher in Utility Theory when I was an undergraduate student.

In third place, I would like to express my gratitude to Pedro Serrano-Aguilar, Yolanda Ramallo-Fariña and especially to Julio López-Bastida for introducing me to the economic evaluation filed and for giving me the opportunity to be fellowship visitor at HERC, where all this work started.

In fourth place, to express my gratitude to all authors of the chapters included in this book, but especially to my friends Mark Oppe and Benjamin Craig for their constant support and counsel. In this place, I also would like to express my gratitude to my colleagues at the Office of the EuroQol Research Foundation for the constant motivation during the development of this thesis.

In fifth place to express my gratitude to my promotor Jan van Busschbach for taking the responsibility of making this thesis happen /come into existence (to occur)and specially for my supervisor Elly Stolk, not only for investing a lot of time in guiding me, but also for making it a smooth process.

I would like to specially mention Oliver Rivero-Arias, to whom I am eternally grateful. There are no words to describe his important influence and impact on me, not only in this thesis development, but also in my whole career.

Finally, my deepest gratitude is for Irene Ávila-Pérez, my love, who has supported me while I spend a long time on my research projects like the ones included in the chapters of this thesis. Indeed, I also want to express my gratitude for her help in conducting the study included in chapter 4.

14

Chapter 14

List of publications

- **Ramos-Goñi JM**, Craig B, Oppe M, Ramallo-Fariña Y, Pinto-Prades JL, Luo N, Rivero-Arias O. Handling data quality issues to estimate the Spanish EQ-5D-5L Value Set using a hybrid interval regression approach. Value in Health 2017. In Press
- Jakubczyk M, Craig B, Barra M, Groothuis-Oudshoorn C, Hartman JD, Huynh E, **Ramos-Goñi JM**, Stolk EA, Rand K. Choice Defines Value: A Predictive Modeling Competition in Health Preference Research. Value in Health 2017. In press.
- Wong ELY, **Ramos-Goñi JM**, Cheung AWL, Wong AYK, Rivero-Arias O. Assessing the Use of a Feedback Module to Model EQ-5D-5L Health States Values in Hong Kong. Patient. 2017 Oct 10. doi: 10.1007/s40271-017-0278-0. [Epub ahead of print]
- Rand-Hendriksen K, **Ramos-Goñi JM**, Augestad LA, Luo N. Less Is More: Cross-Validation Testing of Simplified Nonlinear Regression Model Specifications for EQ-5D-5L Health State Values. Value Health. 2017 Jul - Aug;20(7):945-952. doi: 10.1016/j.jval.2017.03.013
- Purba FD, Hunfeld JAM, Iskandarsyah A, Fitriana TS, Sadarjoen SS, **Ramos-Goñi JM**, Passchier J, Busschbach JJV. The Indonesian EQ-5D-5L Value Set. Pharmacoeconomics. 2017 Jul 10. doi: 10.1007/s40273-017-0538-9. [Epub ahead of print]
- **Juan M Ramos-Goñi**, Mark Oppe, Bernhard Slaap, Jan J V Busschbach, Elly Stolk: *Quality Control Process for EQ-5D-5L Valuation Studies*. Value in Health 12/2016;, DOI:10.1016/j.jval.2016.10.012
- **Juan M Ramos-Goñi**, Benjamin Craig, Mark Oppe, Ben Van Hout: *Combining continuous and dichotomous responses in a hybrid model*. EuroQol WPS 2016 (16002).
- **J. M. Ramos-Goñi**, Y. Ramallo-Fariña: *eq5dds: A command to analyze the descriptive system of EQ-5D quality-of-life instrument*. Stata Journal 10/2016; 16(3).
- Feng Xie, Eleanor Pullenayegum, A. Simon Pickard, **Juan Manuel Ramos Goñi**, Minwoo Jo, Ataru Igarashi: *Transforming Latent Utilities to Health Utilities: East Does Not Meet West: Transforming Latent Utilities*. Health Economics 10/2016;, DOI:10.1002/hec.3444
- Mark Oppe, Kim Rand-Hendriksen, Koonal Shah, **Juan M. Ramos Goñi**, Nan Luo: *EuroQol Protocols for Time Trade-Off Valuation of Health Outcomes*. PharmacoEconomics 04/2016; 34(10)., DOI:10.1007/s40273-016-0404-1
- **Juan M. Ramos-Goñi**, Kim Rand-Hendriksen, Jose Luis Pinto-Prades: *Does the Introduction of the Ranking Task in Valuation Studies Improve Data Quality and Reduce Inconsistencies? The Case of the EQ-5D-5L*. Value in Health 03/2016; 19(4)., DOI:10.1016/j.jval.2016.02.002

- Federico Augustovski, Lucila Rey-Ares, Vilma Irazola, Osvaldo Ulises Garay, Oscar Gianneo, Graciela Fernández, Marcelo Morales, Luz Gibbons, **Juan Manuel Ramos-Goñi**: *An EQ-5D-5L value set based on Uruguayan population preferences*. Quality of Life Research 08/2015; 25(2)., DOI:10.1007/s11136-015-1086-4
- Xavier Bonfill, María José Martinez-Zapata, Robin Wm Vernooij, María José Sánchez, María Morales Suárez-Varela, Javier de la Cruz, José Ignacio Emparanza, Montserrat Ferrer, José Ignacio Pijoán, **Juan M Ramos-Goñi**, Joan Palou, Stefanie Schmidt, Víctor Abraira, Javier Zamora: *Clinical intervals and diagnostic characteristics in a cohort of prostate cancer patients in Spain: A multicentre observational study*. BMC Urology 07/2015; 15(1)., DOI:10.1186/s12894-015-0058-x
- **Juan Manuel Ramos Goñi**, Iván Castilla, Cristina Valcarcel Nazco, Carlos de las Cuevas Castresana, Javier Mar, Pedro Serrano Aguilar: *Coste-utilidad de asenapina frente a olanzapina para el tratamiento de los episodios maniacos de pacientes con trastorno bipolar tipo I*. Pharmacoeconomics - Spanish Research Articles 07/2015; 12(4)., DOI:10.1007/s40277-015-0042-6
- Benjamin M. Craig, Shannon K. Runge, Kim Rand-Hendriksen, **Juan Manuel Ramos-Goñi**, Mark Oppe: *Learning and Satisficing: An Analysis of Sequence Effects in Health Valuation*. Value in Health 02/2015; 18(2)., DOI:10.1016/j.jval.2014.11.005
- **Juan M Ramos-Goñi**, Jose L Pinto-Prades, Mark Oppe, Juan M Cabasés, Pedro Serrano-Aguilar, Oliver Rivero-Arias: *Valuation and Modeling of EQ-5D-5L Health States Using a Hybrid Approach*. Medical Care 12/2014; Publish Ahead of Print., DOI:10.1097/MLR.0000000000000283
- Iván Castilla, Javier Mar, Cristina Valcárcel-Nazco, Arantzazu Arrospide, **Juan M Ramos-Goñi**: *Cost-Utility Analysis of Gastric Bypass for Severely Obese Patients in Spain*. Obesity Surgery 06/2014; 24(12)., DOI:10.1007/s11695-014-1304-0
- Stefanie Schmidt, Ricard Riel, Albert Frances, José Antonio Lorente Garin, Xavier Bonfill, María José Martinez-Zapata, Maria Morales Suarez-Varela, Javier Dela Cruz, José Ignacio Emparanza, María-José Sánchez, Javier Zamora, **Juan Manuel Ramos Goñi**, Jordi Alonso, Montse Ferrer: *Bladder cancer index: cross-cultural adaptation into Spanish and psychometric evaluation*. Health and Quality of Life Outcomes 02/2014; 12(1)., DOI:10.1186/1477-7525-12-20
- Pere Castellvi, Montse Ferrer, Jordi Alonso, Arrarás JI, Escobar A, Herdman M, Martínez-Martín P, Ochoa S, Quintana JM, Rajmil L, Ramada JM, **Goñi JM**, Rebollo P, Ribera A, Ribero A, Valderas JM: *[The patient-reported outcomes in research: definition, impact, classification, measurement and assessment]*. Medicina Clínica 10/2013; 141(8)., DOI:10.1016/j.medcli.2013.07.013

- Alvaro Hidalgo-Vega, **Juan Manuel Ramos-Goñi**, Renata Villoro: *Cost Utility of Ranolazine in the Symptomatic Treatment of Patients with Chronic Angina Pectoris in Spain*. The European Journal of Health Economics 10/2013; 15(9)., DOI:10.1007/s10198-013-0534-8
- **Ramos-Goñi JM**, Oliver Rivero-Arias, Helen Dakin: *Response mapping to translate health outcomes into the generic health-related quality of life instrument EQ-5D: Introducing the mrs2eq and oks2eq commands*. Stata Journal 10/2013; 13(3).
- **Juan Manuel Ramos-Goñi**, Oliver Rivero-Arias, María Errea, Elly A. Stolk, Michael Herdman, Juan Manuel Cabasés: *Dealing with the health state 'dead' when using discrete choice experiments to obtain values for EQ-5D-5L heath states*. The European Journal of Health Economics 07/2013; 14(1)., DOI:10.1007/s10198-013-0511-2
- Maria Trujillo-Martin, P. Serrano-Aguilar, F. Monton-Álvarez, **J. M. Ramos-Goñi**: *Appropriateness of Treatments for Patients with Degenerative Ataxias: Recommendations by a Panel of Experts*.
- Pedro Serrano-Aguilar, Francisco M Kovacs, Jose M Cabrera-Hernández, **Juan M Ramos-Goñi**, Lidia García-Pérez: *Avoidable costs of physical treatments for chronic back, neck and shoulder pain within the Spanish National Health Service: A cross-sectional study*. BMC Musculoskeletal Disorders 12/2011; 12(1)., DOI:10.1186/1471-2474-12-287
- **Juan Manuel Ramos-Goni**, Oliver Rivero-Arias: *Eq5d: A command to calculate index values for the EQ-5D quality-of-life instrument*. Stata Journal 01/2011; 11(1).
- P Serrano-Aguilar, M M Trujillo-Martín, **J M Ramos-Goñi**, V Mahtani-Chugani, L Perestelo-Pérez, M Posada-de la Paz: *Patient involvement in health research: A contribution to a systematic review on the effectiveness of treatments for degenerative ataxias*. Social Science [?] Medicine 08/2009; 69(6)., DOI:10.1016/j.socscimed.2009.07.005

15

Chapter 15

Curriculum Vitae

Juan Manuel Ramos Goñi (1977, Tenerife, Spain) obtained a B.Sc. in Sciences and Statistical techniques (2000-2003) in University of La Laguna (ULL) in Tenerife, Spain. He also obtained his Master degree in Socio-Sanitary Research at the University of Castilla La Mancha (UCLM), Spain in 2014.

After his degree in Statistic Juan Manuel worked for the Canary Island Institute of Statistic (ISTAC) developing input-output tables for the Canary Island (2004-2006). Then, he moved to work for the Health Technology Assessment (HTA) Unit of the Canary Island Health Care Service (SESCS), where he started to conduct HTA reports, including economic evaluations, for the Spanish National Government (2006-2012). During this period Juan Manuel received extensive training in health economics in different Universities as Birmingham, York, Oxford or Cambridge. In 2009, he started a fellowship visitor experience at Health Economics Research Centre (HERC) of the Oxford University. The collaboration with HERC was focused on improving mapping algorithm to translate different health related quality of life measures, as SF-12, OKS or MRs, into EQ-5D utilities, initiating his EuroQol line of research. This visiting period ended on December 2010. In 2011, Juan Manuel obtained, as PI, the grant to conduct the EQ-5D-5L national value set study in Spain, which was a project in collaboration with HERC researchers and other EuroQol members. In 2012, Juan Manuel was accepted as EuroQol Group member and he joined the EuroQol Research Foundation as senior researcher in 2013.

Juan Manuel lives in Tenerife, but keeps working for EuroQol Research Foundation at the same time that he has co-created a growing consultancy firm called “Axentiva Solutions”.

16

Chapter 16

PhD Portfolio

Name PhD candidate:	Juan Manuel Ramos-Goñi
Eramus University of Rotterdam department:	Erasmus MC
PhD Period:	2016-2017
Promotor:	Prof. dr. Jan van Busschbach
Supervisors:	Dr. Oliver Rivero-Arias and Dr. Elly Stolk

1.- PhD training	2016 Big data University of La Laguna (ULL)
2.- Seminar and workshops	2016 EuroQol Academy meeting. Workshop 2016 ISPOR conference, Washington. Workshop 2016 ISPOR conference, Washington. Workshop 2016 ISPOR conference, Singapore. Workshop 2016 ISPOR conference, Singapore. Issue panel 2016 ISPOR conference, Vienna. Workshop 2016 ISPOR conference, Vienna. Workshop 2017 EuroQol Academy meeting. Workshop 2017 University of Galway. Workshop.
3.- International Conferences	2016 EuroQol Plenary meeting. Discussant 2016 EuroQol Plenary meeting. Poster presentation 2016 ISPOR Vienna. Poster presentation
4.- Teaching	September 2016 Barcelona. Master module at Pompeu Fabra University (co-tutor) May 2017 Mérida. MCDA Course (tutor) May 2017 Murcia. MCDA Course (tutor) October 2017 Zaragoza. MCDA Course (tutor) November 2017 Granada. MCDA Course (tutor)