1
Improving the quality of data collected in EQ-5D-5L valuation studies: a summary of the EQ-VT research methodology programme

Shah K[1], Rand-Hendriksen K[2], Ramos JM[3], Prause AJ[3], Stolk E[5]

INTRODUCTION

In the EQ-5D-5L valuation studies carried out in 2012 and 2013 (Devlin et al., 2013; Ramos-Goñi et al., 2013; Versteegh et al., 2013; Xie et al., 2013), a number of data quality issues were observed and described as warranting further investigation, particularly with respect to the composite time trade-off (TTO) data. These included:

- few observations between -0.5 and 0 and clustering of values at -1;
- spikes: clustering of valuations at certain "round numbers" (1, 0.5, 0, -0.5, -1);
- lower than expected values for mild health states;
- inconsistencies - for example, many respondents valuing health state 55555 higher than health states that dominate it;
- low number of worse than dead (WTD) valuations.

The first three issues, which relate to the distribution of observed TTO values, were considered problematic as they suggest that the discriminative capability of the valuation tasks is too low to deal adequately with the increased granularity of the EQ-5D-5L descriptive system. For example, in some of the EQ-5D-5L valuation studies, more than 20% of all TTO tasks resulted in values of 0 or 0.5. The extension of the descriptive system from three to five levels was intended to improve the instrument's sensitivity to small changes and mild health problems. The valuation tasks must be sufficiently responsive to these small changes as well.

The inconsistencies were considered problematic as they undermine the face validity of the results and add a lot of noise to the data, which impairs the modeling. While perfect consistency may not be expected because respondents are likely to be uncertain about the values that they provide, the number of inconsistencies observed

1. Office of Health Economics, London, UK
2. University of Oslo, Department of Health Management and Health Economics, Oslo, Norway
3. EuroQol Group, Rotterdam, the Netherlands
4. Erasmus University Rotterdam, iMTA, Rotterdam, The Netherlands

31th Scientific Plenary Meeting of the EuroQol Group, Stockholm, Sweden, September 25-27, 2014, Proceedings: 1-18

was considered to be unacceptably larger. For example, all health states logically dominate 55555 (the worst state in the descriptive system), yet in each of the EQ-5D-5L valuation studies, at least 20% of respondents valued one or more health states as being worse than 55555. Similarly, about 10% of respondents gave lower values to very mild health states than they did to more severe (and logically worse) health states. Some of these inconsistencies involved small differences in utilities, but one-third of the inconsistent responses seem particularly problematic because they involved large differences in utilities (>0.5). Such responses suggest that respondents were behaving inconsistently when presented with the initial BTD/WTD sorting question[1] or were making mistakes due to learning effects, ordering effects, interviewer effects, fatigue, confusion about the TTO procedure, etc.

In addition, low proportions of WTD values were observed in some of the valuation studies. It is unclear whether this reflects current public opinion about the severity of EQ-5D-5L health states or if it is an artefact of the valuation methodology.

It was considered that the observed problems may not be inherent to the nature of the composite TTO task, but could relate to the way in which the task was implemented in the valuation software, the EQ-VT. Following a meeting in June 2013 attended by members of the VMWG, PIs of recently completed valuation studies and expert advisors, a number of modifications to the protocol for valuation EQ-5D-5L using EQ-VT were proposed. Some of these were implemented immediately:

- advice to study teams to use small teams of dedicated interviewers;
- use of routine quality control checks;
- introduction of three practice TTO tasks (in addition to the "wheelchair example");
- inclusion of a prompt each time the respondent completes a task to make sure that they are happy with their answer.

A number of other proposed solutions were considered to warrant empirical investigation. These solutions are being tested in a multi-country research programme seeking "to address the issues with the EQ-VT and to further develop the Group's valuation methodology" (VMWG, 2013). The aim of this paper is to inform EQ members about the approach and emerging findings of the research programme; and to provide the basis for a discussion about lessons learned and how to incorporate the findings into an improved version of EQ-VT for use in future EQ-5D-5L valuation studies.

1. The second step in all composite TTO tasks involves choosing between 0 years in full health and 10 years in the health state under evaluation. In effect, the respondent is being asked whether they consider the health state to be better than or worse than dead.

METHODS

Study design

Four experimental studies were undertaken in Spain (Tenerife), the Netherlands, Norway and the UK. Each study included a control arm and one or more experimental arms. Data collection began in March 2014 and is expected to conclude in August 2014. In the studies conducted in Spain, the Netherlands and Norway, a fixed set of 10 EQ-5D-5L health states were used in the TTO tasks, corresponding to block 10 in the existing EQ-VT design[2]. In the UK study, the health states used in the TTO tasks were hand-picked specifically to address the aims of that study[3]. The EQ-VT also included paired comparison tasks from a discrete choice experiment (DCE; not reported in this paper). All of the valuation tasks were implemented in EQ-VT, which ensured that all aspects of the tasks were standardized across the studies except for the language in which they were presented.

The experimental studies in Spain, the Netherlands and Norway are being replicated within the context of EQ-5D-5L valuation studies currently underway in Japan, Hong Kong and Singapore, respectively. The results of those studies are not reported here since they are ongoing at the time of writing (July 2014).

Table 1.1 Summary of studies included in the research programme

Study / modification	Issue(s) that this modification is seeking to address	Country
11111/full health study [a]	Lower than expected values for mild health states	UK [b]
Ranking study	Inconsistencies Spikes	Spain [b] Japan [c]
Iteration study	Spikes Low number of WTD valuations	Singapore [c] Norway [b]
Feedback module	Inconsistencies	Norway, Singapore, Spain, Hong Kong, Netherlands
BTD/WTD split study	Inconsistencies	Hong Kong [c] Netherlands [b]

a. The UK study also included an assessment of whether the valuations are affected by the order in which the dimensions are presented. This is not reported here but details can be found in the 2014 EuroQol Plenary poster presentation of that study.

b. experimental study specifically designed to address issues with EQ-VT

c. study being conducted within the context of a full EQ-5D-5L valuation study

2. 12111, 11122, 42321, 13224, 35311, 34232, 52335, 24445, 43555, 55555
3. 21111, 11121, 11112, 11223, 21232, 43331, 32442, 55233, 34155, 55555, 11111

Experimental arms

In all countries, the control arm followed the existing protocol for EQ-5D-5L valuation studies, except for the introduction of the feedback module in selected studies and the exclusion of the DCE tasks in the Norway study. Respondents allocated to the control arm completed the following (in order):

- self-reported health using EQ-5D-5L and EuroQol visual analogue scale (EQ-VAS);
- basic background questions;
- 10 TTO tasks;
- structured feedback questions regarding the TTO tasks;
- a feedback module that allowed respondents to reconsider their TTO responses (selected studies only - see Table 1.1);
- up to 13 paired comparison tasks from a DCE (not reported);
- structured feedback questions regarding the DCE tasks (not reported).

Each experimental arm applied one modification to the EQ-VT interview. The experimental arms are summarised in Table 1.2, along with the specific hypotheses being tested in each study.

Table 1.2 Summary of experimental arms and hypotheses being tested in each study

Study / modification	Hypotheses tested	Summary of experimental arm(s)
11111/ full health study	Use of 11111 rather than full health as the Life A comparator leads to lower values for mild health states	Test 1: In the TTO tasks, Life A was described in terms of time spent in health state 11111 rather than time spent in full health
Ranking study	Reintroduction of a ranking task prior to TTO valuation reduces inconsistencies and improves overall data quality	Test 1: Introduction of a rank ordering task as a warm-up exercise prior to the TTO tasks (ranking no sorting) Test 2: Introduction of a rank ordering task both as a warm-up exercise *and* to determine the order in which the health states will be presented in the TTO tasks (ranking and sorting)
Iteration study	Alternative routing procedures introduced in order to reduce spikes and routing-induced biases.	Test 1: Respondents are randomly assigned to different starting points (0, 4, and 8 years) and different sizes of incremental iterative steps (1 or 2 years). Test 2: Non-stopping TTO (NS-TTO) – incremental routing that does not stop once indifference is encountered, but continues for a fixed number of iterations. Two variants: top-down and bottom-up.
Feedback module	Presenting respondents with the rank ordering implied by their TTO valuations leads to reduced inconsistencies as they will identify and flag problematic valuations for removal from the data	No experimental arms: the feedback module was completed by all respondents in selected studies. After completing the TTO tasks, each respondent was presented with the rank ordering implied by their TTO valuations. They were asked to review their responses and to flag any that they felt should be reconsidered.
BTD/WTD split study	Separating the BTD and WTD elements of the TTO task promotes consistency of the valuations	Test 1: Explanation of and valuation of the WTD element takes place *after* all health states have been valued using the BTD element of the task

Data collection

Each study team sought to recruit a sample that was well-balanced in terms of basic background characteristics such as age group and gender. However, strict representativeness was not required for the purposes of the experimental studies.

For each study, the fieldwork was undertaken by a small group of dedicated interviewers, consisting of academics, master students trained in HTA, and persons who had prior experience with valuation exercises. All interviewers attended training sessions during which they were given intensive training on the methodology and study procedures. The interviewers were instructed to play an active role throughout the valuation interviews. They followed the latest version of the "Instructions for Interviewers" document, using this to guide respondents through the interview, one step at a time. Adapted versions of the Instructions were developed for the experimental arms.

In the Spain and UK studies, respondents were randomly allocated to the control or test arms. In the Netherlands study, a crossover design was used whereby half of the interviewers started with the control arm and then proceeded to the test arm; and the other interviewers started with the test arm and then proceeded to the control arm. The rationale for this approach was concern about the possibility of spillover effects between the methods and interviewers getting confused about having to follow two different protocols simultaneously. In the Norway study, the first 400 respondents were randomly allocated to the control arm or to the test arm involving different start points and iterative step sizes. When this study arm was complete, the NS-TTO responses from the Singapore study were considered, and a decision was made to commence interviews of a further 200 respondents in the NS-TTO setting.

Throughout the data collection period, the interim data were monitored closely by the study teams on a daily basis using the EQ-VT quality control tool. All data were monitored at the interviewer level; if a given interviewer was found to be generating unusual or poor quality data, that interviewer was contacted by the study team and received additional guidance. In accordance with the Code of Conduct for EQ-5D-5L valuation studies (VSWG, 2013), specific attention was paid to the following aspects of interviewer performance: explanation of the worse than dead task using the wheelchair example; time spent on the wheelchair example; clear inconsistencies within the TTO valuations; and total amount of time spent on the TTO tasks. Other statistics monitored on a regular basis for the purpose of quality control included: proportion of round-number TTO values such as 0 and 1; mean TTO values for health states vs. the "misery scores" (sum of the five attribute levels; a proxy for severity) of those health states; and proportion of respondents giving unusual sets of DCE responses (e.g. choosing state A in all of the tasks).

Methods of analysis

Descriptive analyses were used to examine the following data: sample background characteristics, time taken to complete the valuation interviews and individual tasks, and the TTO values. Cognitive debriefing exercises provided additional information on how the tasks and modifications were perceived by respondents and interviewers.

To evaluate the consistency of the observed responses, we identified all pairs of health states in which one health state can be said to dominate the other. The definition used was as follows: A dominates B when A is better than B on at least one dimension, and no worse than B on any of the dimensions. We then defined an inconsistency as an observation of B (the dominated health state) being given a higher value by a respondent than A (the dominant health state)[4]. In this report, the primary statistic reported is the percentage of respondents whose TTO values contained at least one inconsistency was calculated (out of a maximum of 25 in the Spain, the Netherlands and Norway; and of 28 in the UK), along with the percentage of respondents who valued health state 55555 higher than at least one other health state (recall that all health states dominate 55555, so nine of the possible inconsistencies involve 55555).

To evaluate clustering effects, we examined the proportion of valuations clustered at round numbers (-1, -0.5, 0, 0.5, 1).

Since the primary aim of the UK study was to assess whether the choice of the Life A comparator affects the values of the very mild health states (those with a misery score of 6, namely 21111, 11121 and 11112), we examined the mean values given to these three states:

(i) The statistical significance of any observed differences was assessed using appropriate test statistics;
(ii) Results (as of mid-July 2014);
(iii) Samples and completion of the tasks.

Table 1.3 summarises the background characteristics of the samples. We did not observe any substantial differences between the control and test arms of any of the studies in terms of sample composition.

Table 1.4 shows the average time taken to complete the valuation interview and individual TTO tasks in each study and study arm. In the Spain study, the interviews that included ranking tasks lasted for about eight minutes longer than those that did not. In the Norway study, time spent on each TTO task and on the feedback module was

4. An alternative approach, as reported in Ramos-Goñi et al. et al. (2014), involves counting as inconsistencies any observations of the dominant and dominated health states being given the same value.

distinctly shorter in the NS-TTO arm than in the other arms. However, the total interview time was not substantially reduced, indicating that more time has been spent on instructions and in other parts of the interview. In the Netherlands study, the separation of the BTD and WTD tasks increased overall interview duration by about three minutes. In the UK study, respondents spent slightly longer completing the tasks when Life A was described in terms of time spent in 11111 rather than time spent in full health.

Table 1.3 Background characteristics of the samples

	Spain	Netherlands	Norway	UK
N (completed interviews)	600	405	598	231
Age (years)				
- < 35	34.5%	19.8%	32.6%	16.9%
- 35 to 54	41.8%	30.9%	35.5%	26.0%
- ≥ 55	23.7%	49.4%	31.9%	57.1%
Gender				
- Male	52.3%	34.1%	45.5%	44.2%
- Female	47.7%	65.9%	54.5%	55.8%
Experience of illness				
- in self	21.2%	22.7%	28.1%	36.4%
- in family	72.7%	75.3%	77.6%	63.6%
- in caring for others	47.5%	50.1%	32.4%	39.4%

Table 1.4 Time taken to complete the valuation interview and individual TTO tasks

Study	Arm	Mean (interview) (minutes)	Median (interview) (minutes)	Mean (TTO task) (seconds)	Median (TTO task) (seconds)	Mean (FB module) (seconds)	Median (FB module) (seconds)
Spain	Control	42.6	41.7	63.9	47.3	126.9	112.8
	Test 1 (rank no sorting)	51.6	51.3	63.5	46.9	133.0	116.5
	Test 2 (rank + sorting)	50.9	50.7	66.0	52.0	129.7	114.4
Netherlands	Control	35.2	32.7	59.4	45.6	115.0	98.0
	Test (BTD/WTD split)	38.2	36.4	65.4	50.4	121.1	112.5
Norway	Control	32.9	31.2	69.1	63.5	143.8	127.4
	Test 1 (iteration)	32.2	30.9	67.2	60.8	154.9	132.2
	Test 2 (NS-TTO)	28.7	25.4	48.5	44.1	97.7	70.2
UK	Control	24.0	22.3	52.9	34.2	N/A	N/A
	Test (11111)	26.4	24.5	61.7	42.4	N/A	N/A

Valuation data

Figures 1.1a to 1.1d compare, for each study, the mean TTO value for health states with different misery scores for the control and test arms.

In the Spain study, the observed TTO values are similar across arms for the very mild (misery 6 and 7) and very worst (misery 25) health states. The inclusion of the ranking task leads to higher mean values for all other health states (misery 12 to 22). In the Netherlands study a similar effect was observed, whereby the test arm seemed to provide higher values for moderately severe health states. However, these differences were not found to be statistically significant in either country. In the Norway study, mean scores were virtually identical for the control arm and the pooled sample of varied starting point and iteration size groups. NS-TTO resulted in somewhat lower values, while the bottom-up variant resulted in substantially lowered values across all health states. In the UK study, the difference in mean values for the very mild (misery 6) health states across the arms was small and not statistically significant (p=0.19), though we observe higher mean values for all of the other health states in the full health (control) arm. Health state 11111, which was valued by respondents for whom full health was the Life A comparator, was almost always given a value of 1.

Figures 1.1a- 1.1d Mean TTO values by misery score for each country and test arm

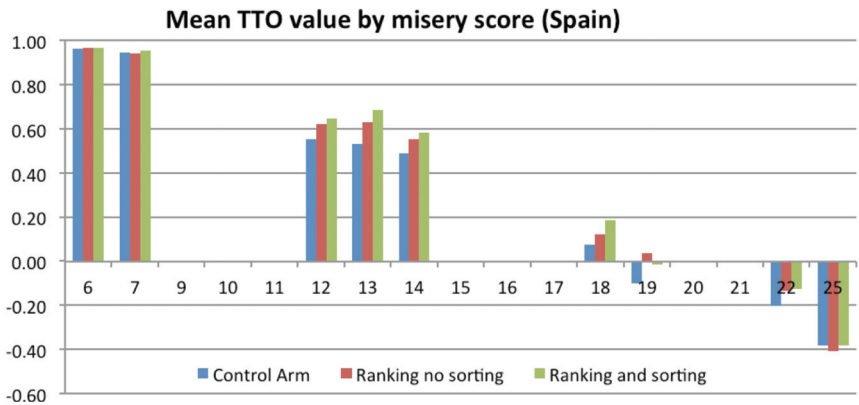

Figure 1.1a Spain

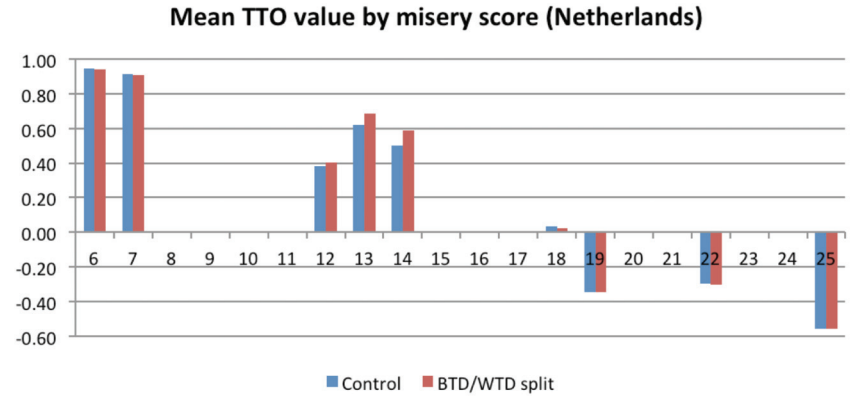

Figure 1.1b Netherlands

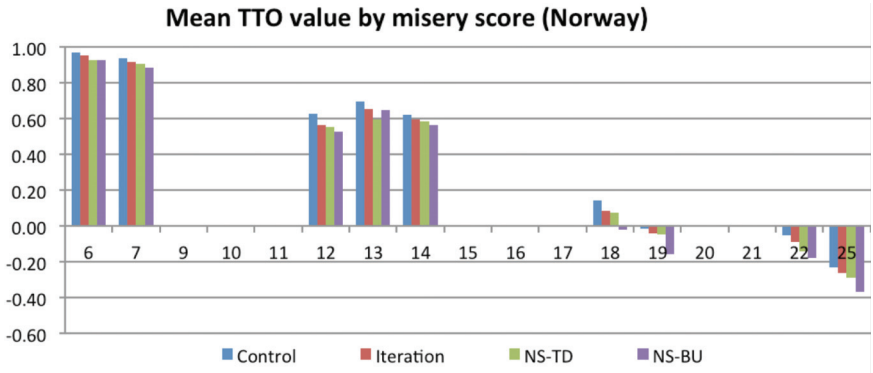

Figure 1.1c Norway

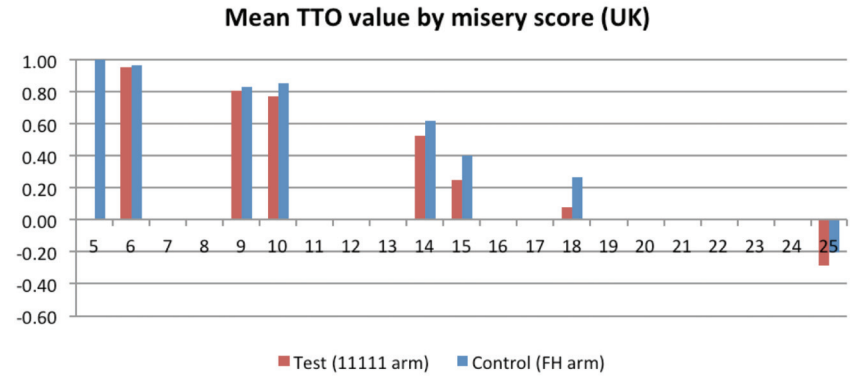

Figure 1.1d UK

Figures 1.2a to 1.2e compare, for each study, the TTO valuation distributions (for all health states) for the control and test arms. Considering all health states and respondents, the modifications under evaluation did not seem to affect the value distribution in the studies in Spain or the Netherlands. In the Norway study we observe that NS-TTO reduced the proportion of values clusters at round numbers. The varied starting point and iteration size groups displayed significantly less clustering at -0.5, 0 and 0.5. In the UK study we observe larger spikes at 0.95 and 1 compared to the other studies, which is consistent with the inclusion of three very mild health states in the UK study design (21111, 11121 and 11112). Respondents in the 11111 (test) arm gave fewer round number values than those in the full health (control) arm. The difference in propensity to give round number values across the arms is statistically significant ($p<0.01$).

Figures 1.2a - 1.2e TTO value distributions by country and test arm

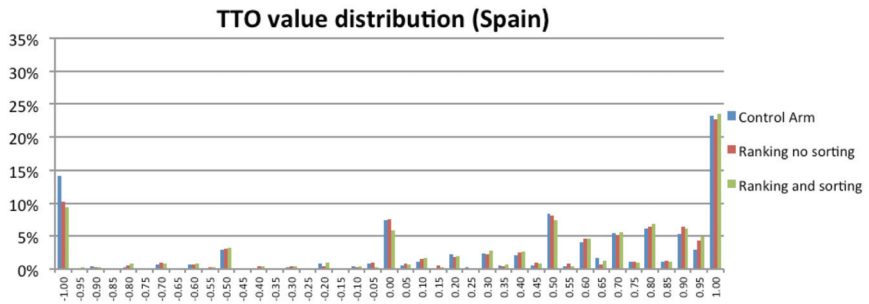

Figure 1.2a Spain

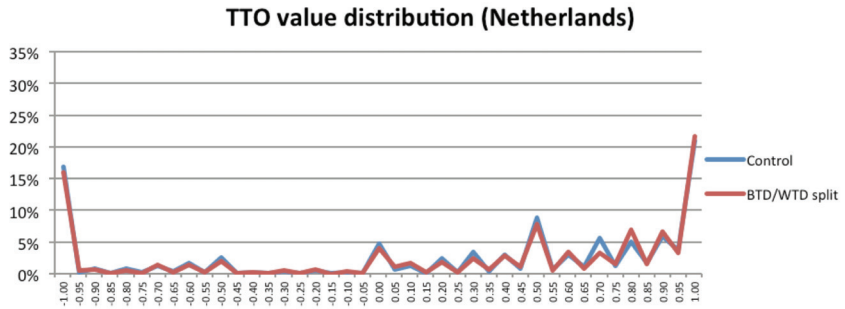

Figure 1.2b Netherlands

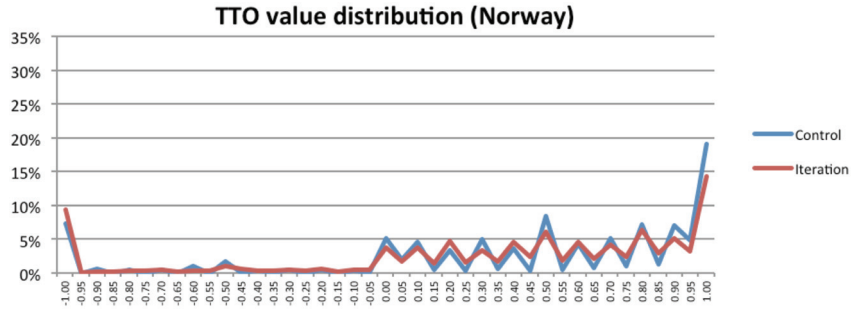

Figure 1.2c Norway (control vs. iteration test arm)

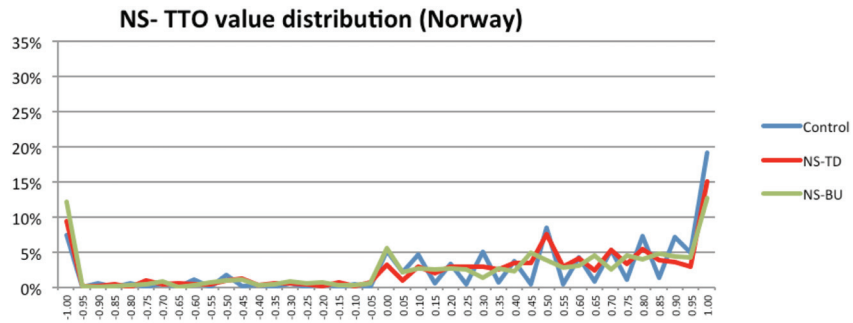

Figure 1.2d Norway (control vs. top-down NS-TTO vs. bottom-up NS-TTO)

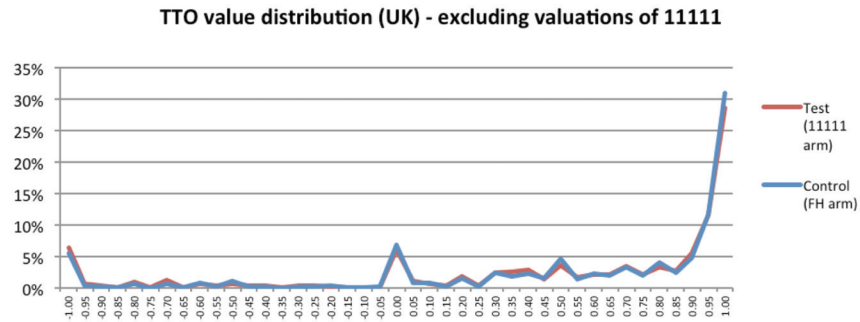

Figure 1.2e UK

Figures 1.3a to 1.3d compare, for each of the studies, the percentage of respondents with one or more inconsistencies in the control and test arms. For the studies that included the feedback module, the figures also show the impact of removing valuations that had been flagged by respondents.

The modifications tested in the Netherlands (separating the BTD and WTD task) and Spain (inclusion of a ranking task) sought specifically to promote the consistency of the data. The results from the Spain study suggest a potential but modest effect of the ranking task on consistency, restricted to test arm whereby the rank ordering task was strictly used for warm-up purposes and not to guide the order of health states in the TTO task. Specifically, we observe a lower inconsistency rate involving health state 55555, but not the overall inconsistency rates. In the other test arm, ranking and sorting, consistency rates were higher than in the control arm. The feedback module reduced the level for inconsistencies about approximately 1% in each arm. This improvement was relatively modest in comparison to the other countries. However, by adopting a weak dominance criterion as the basis for the inconsistency count rather than the strict dominance criterion used in this paper, Ramos-Goñi et al. (2014) found a statistically significant reduction in the number of inconsistent pairs of health states, due to a reduction of ties.

In the Netherlands study, both the separation of the BTD and WTD tasks and the use of the feedback module appeared to have promoted consistency of the data, lowering the overall proportion of respondents with inconsistencies. In the control arm, 80.5% of the respondents were fully consistent (after the removal of valuations flagged in the feedback module, this increased to 87%). In the BTD/WTD split arm, 83.6% of the respondents were fully consistent. After the removal of responses flagged in the feedback module, this proportion further increased to 91.8 %, suggesting that the effects were independent of each other. Considering the pooled data across both arms, the adoption of the feedback module improved the level of consistency from 82% to 89.4% (p=0.003). Similarly, the proportion of respondents who did not value 55555 the worst health state was reduced using the feedback module from 5.7% to 2.7% (p=0.036). Again, inconsistency rates were higher in the control arm than in the BTD/WTD split arm, both before the feedback module (6.3% vs. 5.0%; p=0.57) and after it (3.4% vs. 2.0%; p=0.39).

One-hundred and thirty-four of the Netherland study respondents (35.9%) flagged one or more of their TTO responses using the feedback module. The most frequently flagged health states were 13224 and 52335. In total, 5.3% of all TTO values were flagged. Removing the flagged valuations from the data set had no significant effect on the mean TTO values. The majority of respondents who used the feedback module to flag responses had no inconsistencies in their TTO data (N=96) and used the module for other reasons. Twenty-eight respondents who did have inconsistencies did not use the feedback module. The most frequently observed inconsistencies involved health states 55555/24445 (N=15) and 52335/42321 (N=16). The former was more often resolved than the latter using the feedback module.

In the Norway study, we observed a non-significant increase in the proportion of respondents with one or more inconsistent responses in all but one of the test arms, while the proportion of respondents who were inconsistent in their ordering of health state 55555 seemed to be slightly reduced. The exception to this was the test arm offering 8 years in full health as starting point, in which respondents were less than half as likely to produce inconsistent responses. After excluding responses flagged in the feedback module, only 4% of inconsistent respondents remained, nearly all of which involved health state 55555.

In the UK study, the largest proportion of inconsistent respondents was observed (22.3% in the control arm; 33.7% in the test arm). However, these statistics cannot be compared directly to the other studies because a different block of health states was used with a greater number of potentially inconsistent responses. The difference between arms is not statistically significant at the 5% level.

Figures 1.3a - 1.3d Proportions of respondents with inconsistent responses by country and test arm

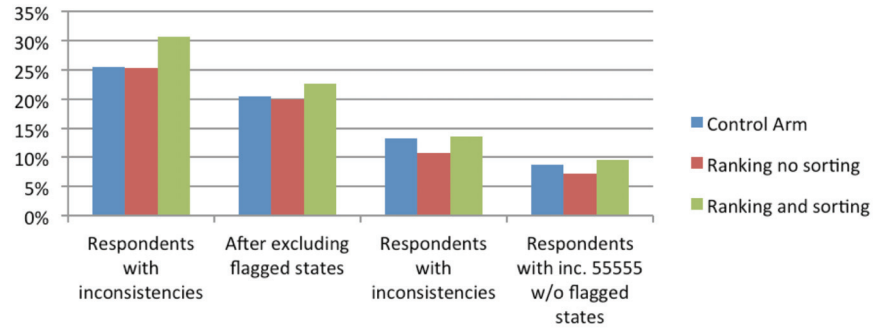

Figure 1.3a Spain

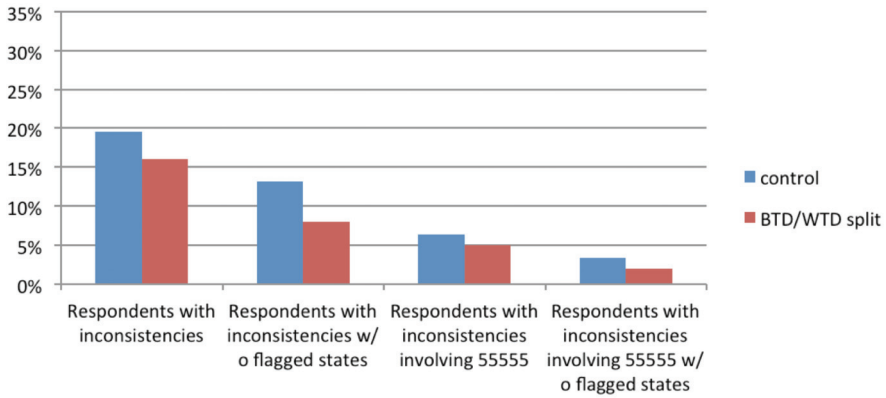

Figure 1.3b Netherlands

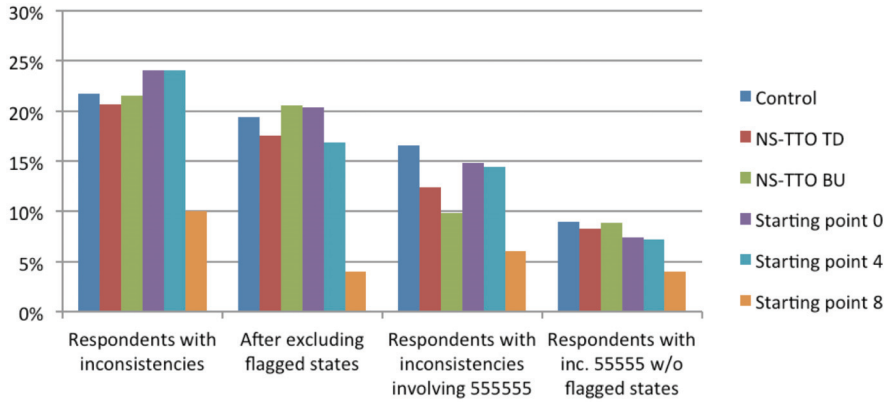

Figure 1.3c Norway

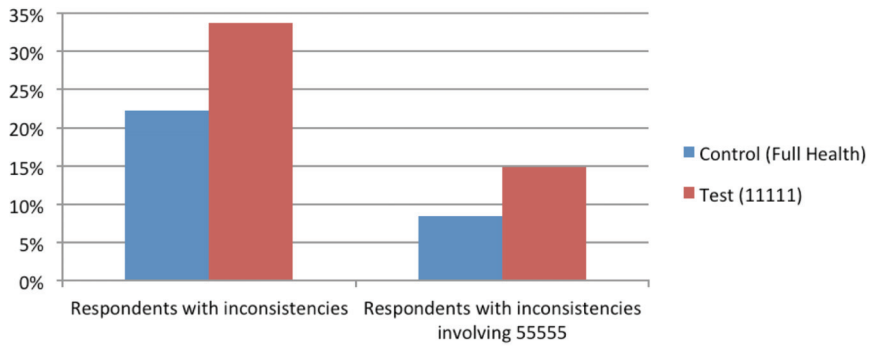

Figure 1.3d UK

Cognitive debriefing

In the Spain study, the responses to the structured feedback questions indicate that respondents considered the inclusion of the ranking tasks to make the overall valuation exercise more difficult in comparison to the current protocol.

In the Netherlands study, respondents and interviewers reported that they found the BTD/WTD split version of the task slightly more difficult to complete or explain, respectively.

In the UK study, the interviewers reported no difficulties in switching between the two Life A comparators (full health and 11111) and have indicated no preference for one comparator or the other.

DISCUSSION

In this section we first briefly summarise the emerging findings of each experiment.

Ranking study

When using a strict dominance criterion, the results from the Spain study suggest a modest effect of the ranking task on consistency, restricted to the test arm whereby the rank ordering task was strictly used for warm up-purposes and not to guide the order of health states in the TTO task. Two possible explanations are suggested for the lack of effect in the other arm (ranking and sorting). The first explanation is random variability between the samples of the arms. The second explanation is that the observed behaviour results from the algorithm used to determine the order of presentation for TTO health states in the ranking and sorting arms. However, by adopting a weak dominance criterion as the basis for the inconsistency count rather than the strict dominance criterion used in this paper, Ramos-Goñi et al. (2014) found a statistically significant reduction on the number of inconsistent pairs of health states due to a reduction of ties, thus supporting the hypothesis that the introduction of the ranking task promotes the discriminative power of the TTO task.

BTD/WTD split study

Separation of the BTD and WTD tasks was found to promote consistency in the data, despite being perceived as being more burdensome by interviewers and less easy by respondents (than the current composite approach). We need to wait for the data from the Hong Kong study to see if the effects of this modification are statistically significant.

Iteration study

Mean values were virtually identical for the control arm and the pooled sample of varied starting point and iteration size groups. The varied starting points had very limited impact on mean values for the 10 included health states, but the distribution of values was altered, significantly with fewer elicitations on round numbers (-0.5, 0, and 0.5). The proportion of WTD responses was generally low in the Norway study, making analyses of the WTD distribution tricky.

One group distinguished itself in terms of the number of inconsistent responses, or, rather the lack thereof. The proportion of respondents making at least one inconsistent response was similar for the control group and for respondents jumping to 0 and 4 years after the initial comparison of 10 years of full health to 10 years in the target state - around 22% to 24%. However, only 10% of the group that jumped to 8 years after the initial question made any inconsistent responses. The rate of inconsistent responses in the 8-year group was less than one-third of the rate of the control group. This could suggest that an incremental routing starting at the top and working down (downward titration) is easier for respondents to perform well.

NS-TTO resulted in slightly lower values across the board. The bottom-up variant of NS-TTO was disliked by both interviewers and respondents, while the top-down variant was generally met with acceptance.

11111 vs. full health study

The preliminary results indicate that the choice of Life A comparator does not affect the valuations of the mild health states: we found no statistically significant difference in the mean values given to these health states across the study arms based on the current sample size (N=231). Nevertheless, we note that the values of the mild health states are higher than was observed for similarly mild health states in the EQ-5D-5L valuation studies, which suggests that other factors (such as the use of a practice task involving a mild health state, something that was introduced after the first few EQ-5D-5L valuation studies had been completed) may be relevant. Responses to follow-up questions indicate that a sizeable minority of respondents do not agree that 11111 and full health are equivalent.

Please note that data collection in the UK is still underway at the time of writing. We are continuing to examine the significance of any other differences between the study arms.

Feedback module

The feedback module consistently promoted the consistency of the data in all countries where it was adopted. The fewest responses were flagged in the Norway study (4.3%) and the large proportion in the Netherlands (5.3%). Removing the flagged valuations resulted in all three countries in a statistically significant increase in the proportion of consistent respondents, increasing this statistic by between 2% and 7% depending on country and study arm. Measured on the pooled data, the percentage of respondents who were fully consistent increased from 72.8% to 79% in Spain; and from 82% to 89.4% in the Netherlands. However, the mean health state values were not statistically significantly affected. Interviewers reported that the feedback module was not straightforward to use and many respondents expressed dissatisfaction regarding the feedback module. Pending the data collection, the feedback module was integrated in the quality control tool, which from that point forward facilitated discussion about the module between the PIs and the interviewers. Interviewers in the Spain study reported that the respondents refused to use the module more often in the ranking test arms, which suggests that fatigue must be considered.

CONCLUSIONS

A face-to-face meeting was held in Amsterdam on 10-11 July. This was attended by members of the MAT, study PIs, and representatives from the Valuation Methodology Working Group and Value Set Working Group. The meeting attendees agreed that the research programme has been successful. The following was agreed by the group:

- The quality of the data collected in the control arms (i.e. without the modifications) of the studies appears to be much better than the quality of the data from the initial EQ-5D-5L valuation studies. This suggests that the modifications that were implemented without research (e.g. more practice TTO tasks, routine quality control checks) have had a positive impact.
- Most of the modifications appear to further improve the quality of the data. The extent of improvement varies across modifications and needs to be weighed up against any costs or risks associated with the changes.
- The quality of the data observed in this research programme is sufficient to justify sticking to the composite TTO-based valuation protocol.

Table 1.5 summarises the views of the group on: the face validity of the observed data; the feasibility and acceptability of each modification to different stakeholders; and the likely costs involved in implementing each modification. We have made a tentative recommendation for each modification, subject to the preliminary findings being validated once data collection is complete in all of the studies (including those being conducted within the context of full valuation studies). In addition, we recommend updating and improving the interviewer training materials and interviewer instructions document.

Table 1.5 Summary of views on the modifications (based on discussions at the face-to-face meeting on 10-11 July)

Modification	Face validity	Feasibility and acceptability to respondents, interviewers and the scientific community	Costs	Recommendation
Ranking without sort	Do not observe a significant impact on (strictly defined) inconsistencies Reduces number of states with tied values	Not liked by interviewers or respondents	Adds 8 min on average Increases complexity of the study	Do not recommend
Ranking with sort	Reduces face validity Increases inconsistencies Ranking data contains more inconsistencies than TTO data Reduces number of states with tied values	Not liked by interviewers or respondents	Adds 8 min on average Increases complexity of the study	Do not recommend
BTD/WTD split	Reduces inconsistencies	Small number of feasibility concerns reported Unclear whether the benefits would persist when implemented alongside the 8yr start point Interviewers reported that the split felt "unnatural"	Adds 3 min on average (may be shorter in countries with fewer WTD responses)	Recommend examination of BTD/WTD split alongside 8yr start point with view to recommending BTD/WTD modification if benefits persist
Iteration – starting point	Reduces inconsistencies Reduces spikes Do not observe a significant impact on mean values Need to see if results are repeated in Singapore	Is choice of starting point arbitrary? Question marks remain around impact on means Need to examine the impact on %WTD values Interviewers and respondents seem to find the 8yr start point easy, on balance	Increased number of steps/clicks required -> potential fatigue (but no major impact on time taken) Interviewer instructions would require re-working	More analysis required on 1yr vs. 2yr steps Recommend partial roll-out subject to findings being repeated in Singapore (would want to see all of the face validity improvements repeated) Seek review of results by Peep Stalmeier
Iteration – non-stopping TTO	Bottom-up reduces face validity Top-down leads to smoother distributions but other effects on face validity are unclear Values for mild states appear to be reduced (TBC)	Bottom-up unacceptable Top-down appears feasible	Reduces time taken Risks are unknown at this stage	More research and potentially re-visit when we know more
11111 (vs. Full Health) as the Life A comparator	Mean values for mild health states no different from the full health arm	May help with translation issues Seems to be acceptable to respondents and interviewers	Increases visual complexity Zero development costs Little/no increase in interview length	Recommend replacing FH with 11111 (subject to results being confirmed with full sample)
Dimension ordering	Results are different for the different orderings but we are not sure why	Unacceptable to stick to fixed ordering if we are aware of biases Our data suggests that between-respondent randomisation is necessary	No known costs Within-respondent randomisation would likely impose costs	Recommend full randomisation of dimensions between respondents (subject to results being confirmed with full sample)
Feedback module	Reduces inconsistencies (that respondents and interviewers would otherwise flag)	Could potentially be made easier for respondents to comprehend Instructions and interviewer training need work	2 min to complete on average – unsure about distribution Loss of data – unclear how much of the excluded data are perfectly good Additional instructional and monitoring requirements	Recommend current version for use Suggest experimental studies to improve functionality

We welcome comments from Plenary participants on any aspect of this paper or the EQ-VT research methodology programme. For further details, see the paper on the Spain study (Ramos-Goñi et al., 2014) and the posters on the studies in the Netherlands, Norway and the UK.

REFERENCES

Devlin, N.J., Shah, K.K., Mulhern, B., Feng, Y. and van Hout, B. (2013) An EQ-5D-5L value set for England. Paper presented at the 2013 EuroQol Scientific Plenary Meeting. Montreal, Canada.

Ramos-Goñi, J.M., Pinto-Prades, J.L., Cabasés, J.M. and Rivero-Arias, O. (2013) Spanish EQ-5D-5L Valuation project: dealing with inconsistencies of C-TTO responses. Paper presented at the 2013 EuroQol Scientific Plenary Meeting. Montreal, Canada.

Ramos-Goñi, J.M. et al. (2014) Reintroduction of the ranking task in EQ-5D valuation. Improved data quality and reduced level of inconsistencies? Paper presented at the 2014 EuroQol Scientific Plenary Meeting. Stockholm, Sweden.

Versteegh, M., Krabbe, P., Evers, S., de Wit, A., Prenger, R. and Stolk, E. (2013) A Dutch tariff for the EQ-5D-5L. Paper presented at the 2013 EuroQol Scientific Plenary Meeting. Montreal, Canada.

VMWG (Valuation Methodology Working Group). (2013) Request for proposals - Valuation Methodology Working Group. Rotterdam: EuroQol Group.

VSWG (Value Set Working Group). (2013) Code of conduct EQ-5D-5L value set study. Version 16DEC2013. Rotterdam: EuroQol Group.

Xie, F., Pullenayegum, E., Bansback, N., Bryan, S., Ohinmaa, A., Poissant, L. and Johnson, J.A. (2013) The Canadian EQ-5D-5L valuation study: an exploratory analysis. Paper presented at the 2013 EuroQol Scientific Plenary Meeting. Montreal, Canada.