TITLE

An empirical study of two alternative comparators for use in time-trade off studies

AUTHORS

Koonal Shah, MSc, Office of Health Economics

Brendan Mulhern, MRes, University of Technology Sydney

Louise Longworth, PhD, Brunel University London

MF (Bas) Janssen, PhD, EuroQol Executive Office

CORRESPONDING AUTHOR

Koonal Kirit Shah

Office of Health Economics, Southside 7th floor, 105 Victoria Street, London SW1E 6QT, UK

kshah@ohe.org

+44 20 7747 8856

ACKNOWLEDGEMENTS

This study was funded by the EuroQol Research Foundation, and all authors are EuroQol members. However, the views expressed do not necessarily reflect the views of the EuroQol Research Foundation.

We are grateful for the contributions of Nancy Devlin, Liz Flower, Rachel Ibbotson, Arnd Jan Prause, Juan Manuel Ramos-Goñi, Kim Rand-Hendriksen, Knut Stavem, EllyStolk and Ben van Hout. We also wish to thank the interviewers and respondents who took part in the study; the anonymous reviewers; and participants at the 1st meeting of the International Academy of Health Preference Research where preliminary findings of the study were discussed.

KEYWORDS

EQ-5D; full health; time trade-off; stated preference

Abstract

Introduction: Studies to produce value sets for preference-based measures of health require a full health upper anchor to be defined if the values are to be used to calculate quality-adjusted life years. Recent value sets derived for the EQ-5D-5L instrument have described the upper anchor as "full health" whereas older valuation studies for the EQ-5D used the best health state in the descriptive system (11111). It is unclear whether this change could have led to differences in the values obtained. The objective of this study is to assess differences in time trade-off (TTO) valuations using two different comparators (full health and 11111).

Methods: Preferences for EQ-5D-5L health states were elicited from a broadly representative sample of the UK general public. TTO data were collected using computer-assisted personal interviews. Respondents were randomly allocated to one of two arms, each using a different comparator health state. Respondents completed 10 or 11 TTO valuations and a series of follow-up questions examining their interpretations of the term "full health".

Results: Interviews with 443 respondents were completed in 2014. The differences in mean values across arms are mostly small and non-significant. The two arms produced data of similar quality. There is evidence of interviewer effects. Health state 11111 was given a value of 1 by 98.2% of the respondents who valued it.

Conclusions: EQ-5D-5L values elicited using the composite TTO approach are not greatly affected by whether full health or 11111 is used as the comparator health state.

An empirical study of two alternative comparators for use in time-trade off studies

Introduction

Preference-based measures of health are commonly used to assess the impacts of health interventions and to inform calculations of quality-adjusted life years (QALYs) in economic evaluations. In order to use data from these measures in the estimation of QALYs, the instrumentsmust be accompanied by "value sets" which provide, for each health state described by the descriptive system, a valuesummarising how good or bad that health state is considered to be by a representative sample of the general population. The values lie on a scale anchored at 1 ("full health") and 0 (dead), with values of less than 0 assigned to health states considered to be "worse than dead".

Several choice-based methods are available to elicit health state values, including time trade-off (TTO) and standard gamble. In TTO tasks, survey respondents face a series of choices between two hypothetical "lives": one involving a period of time in an impaired health state; the other involving a shorter period of time in a comparator health state involving no health problems. The valuation of the impaired health state is calculated according to how much time in the comparator state the respondent is willing to give up at the point at which they are indifferent between the two lives. Within the choice task, the comparator state could be defined either as being in the best level of each dimension on the measure being valued, or in generic terms such as "healthy", "full health" or "perfect health".

Different generic preference-based measures have taken alternative approaches to this issue. In the standard gamble exercises used to value the Health Utilities Index 3 [1] and SF-6D [2] instruments, health states were valued in relation to the best states defined by therespective descriptive systems of the measures.In the TTO exercise used to value the Assessment of Quality of Life II instrument, the comparator state was "excellent health", though additional wording was included to describea lack of problems in the dimensions of health described by that measure [3].

In most of the valuation studies for the widely-used EQ-5D instrument [4], including the UK EQ-5D value set study [5], values were sought relative to the best health state in the descriptive system using TTO (denoted by the descriptor 11111; where 1 indicates "no problems" on each of the five health dimensions). Recently, however, EQ-5D valuation studies have used a comparator of "full health", including those seeking to value the new five-level version of the instrument, the EQ-5D-5L [6,7]. The EQ-5D-5L was developed to address perceived concerns about the EQ-5D's lack of sensitivity to small changes in health (and in particular its ability to capture mild health problems adequately) [8]. It

comprises the same core five dimensions (mobility; self-care; usual activities; pain/discomfort; anxiety/depression) as the EQ-5D, but increases the available response options ("levels") from three to five.

It is unclear how the specification of the comparator health state affects the valuations elicited. For example, recently conducted valuation studies have reported lower than expected mean values for very mild health states. In the EQ-5D-5L value set for England study [9], the mean value elicited forthehealth state comprising *slight* problems in walking about and no problems with the other four dimensions (21111)was similar to the mean value elicited for the corresponding EQ-5D(-3L) health state (*some* problems in walking about and no problems with the other four dimensions; a more severe state, but also denoted as 21111)[5]. Similar results were observed for the other very mild health states, both in the England study [9] and in parallel studies conducted elsewhere [6,7]. This is surprising as these EQ-5D-5L health states are by design milder than the corresponding EQ-5D-3L health states – level 2 denotes "slight" problems in EQ-5D-5L, compared to "some" problems in EQ-5D-3L.

This finding could represent changes in population preferences since the early EQ-5D valuation studies were conducted, but could also be explained by several methodological developments, such as the use of computer-assisted surveys and changes to the operationalisation of the valuation tasks. Some of the research supporting these developments has been reported [10], but the change in comparator health state from 11111 to full health has never,to the best of our knowledge, been investigated. It is not clear whether the two descriptors can be considered to be equivalent, empirically. To the best of our knowledge, the value of 11111, and how to interpret the gap between 11111 and full health, if any, has never previously been investigated (one study has reported an observed mean value of less than 1 for 11111 [11], but it is not clear how this value was obtained).

The primary objective of this study was to assess differences in TTO health state valuations (particularly the valuation of very mild health states) using two different comparator health states: full health and EQ-5D-5L health state 11111. Our null hypothesis was that the valuations would not depend on the choice of comparator health state. We were interested primarily in the impact of the comparator health state on the average values observed. Further, given that a possiblestudy outcome could be to recommend the use of an alternative comparator, we were also interested in the impact that such a change might have on the face validity of the data and the overall value elicitation process (commonly assessed by examining the extent of inconsistencies in the responses and respondents' reported difficulties in understanding and completing the tasks).

Further objectives were to elicit the value of 11111 itself, and to examine people's interpretations of the term "full health" and alternative labels for the comparator health state.

Methods

Administration of survey

Stated preference data were collected from a sample of members of the UK general public using a valuation questionnaire based closely on the protocol for valuing EQ-5D-5L [10]. The EuroQol Group Valuation Technology (EQ-VT), a digital aid developed specifically for EQ-5D-5L valuation studies, was used to administer the valuation tasks and to capture the response data. The EQ-VT was used as the basis for computer-assisted personal interviews, undertaken by a team of three experienced interviewers working for Sheffield Hallam University. The interviewers completed training on the specifics of the methodology and procedures for this study, and were asked to follow step-by-step instructions and a script in order to minimise interviewer bias. All interviews were carried out in a one-to-one setting in the homes of respondents.

The sample comprised adult members of the general public in South Yorkshire, UK. Ten areas were selected by identifying every 18^{th} or19^{th} area from an alphabetical list of 185 towns in the region. A central point in each area was then selected, with the relevant postcode entered into the AFD Names and Numbers software [12] to generate a database of residential properties around this point. Letters of invitation to participate in the study were sent to the selected addresses.

A "minimum quota" approach was used to ensure that the sample was broadly representative of the general population in terms of age and gender. The survey was given ethics approval by the Ethics Committee of the University of Sheffield's School of Health and Related Research.

Survey instrument

Each respondent completed the following tasks (in order): self-reported health using EQ-5D-5L and the EuroQol visual analogue scale (EQ-VAS); sociodemographic questions; introduction to the TTO exercise (including four warm-up tasks); 10 or 11 TTO valuation tasks (depending on the study arm); structured feedback questions regarding the TTO tasks; 13discrete choice experiment tasks (not reported in this paper); structured feedback questions regarding the discrete choice experiment tasks;and an opportunity to leave further feedback using an open-ended text box. Immediately after completing these tasks using EQ-VT, respondents were asked to complete a short pen-and-paper follow-up questionnaire (see below).

TTO tasks

In the TTO tasks, a "composite" approach was used. This involved beginning with "conventional" TTO for all health states, and shifting to "lead-time" TTO [13] if the respondent indicated that they considered the health state to be worse than dead. Details of the composite TTO approach, as well as the results of an empirical study supporting its use in the valuation of EQ-5D-5L health states, are described by Janssen et al. [14].

The two hypothetical lives were labelled as Life A (time in the comparator health state – either full health or 11111) and Life B (time in the EQ-5D-5L health state under evaluation). At the respondent's point of indifference, the value for the health state can be calculated (in the simplest case, with zero temporal discounting) as follows:

$U = t/10$ for conventional TTO (better than dead health states)

or $U = (t - 10)/10$ for lead-time TTO (worse than dead health states)

where U is the value (utility) and t is the number of years in Life A at the respondent's point of indifference.For more information, see Oppe et al. [10].

Follow-up questionnaire

The follow-up questionnaire comprised the following tasks: (1) paired comparison task in which respondents were asked to indicate whether they considered 11111 and full health to be "the same as each other", and if not, to explain what makes them different from each other; (2) EQ-VAS rating of four health states, including 11111 and full health; (3) ranking task in which respondents were asked to rank six health state descriptions (full health, perfect health, no health problems, 11111, healthy) in order of how much they would want to live in them; (4) open-ended question inviting respondents to indicate any aspects of health and quality of life that they consider to be important but are not captured by the five EQ-5D dimensions. These follow-up tasks were developed by the study team to elicit additional information about the comparability between 11111 and full health, and to inform the ongoing research agenda around the measurement of quality of life beyond the existing EQ-5D descriptive system.

Study design

Respondents were randomly allocated to one of two study arms. In the control arm, Life A was described in terms of time in "full health". In the test arm, Life A was described in terms of time in 11111. SeeFigure 1.

<Figure 1 about here>

The EQ-5D-5L health states were hand-picked by the study team so as to cover a variety of mild, moderate and severe health problems, with three very mild health states (21111, 11121, 11112) included because of their particular relevance to the study objectives. These health states can each be described as having alevel "sum score" (a proxy for severity; calculated by summing the five dimension levels – i.e. 2+1+1+1+1=6) of six. The majority of health states included were taken from the experimental design used to select health states for the EQ-5D-5L valuation studies. The remainder were commonly occurring health states that were relatively well-represented in the self-reported health data in another large-scale public preference survey [15].

Respondents in the test arm each evaluated 10 health states using TTO. Respondents in the control arm evaluated the same 10 health states and additionally valued health state 11111. The order in which the health states were presented was randomised with the exception of 11111 which was always presented last. This is because of concerns that if some respondents consider being asked to value 11111 trivial or frivolous, they might pay less attention to the remaining valuation tasks.

The EQ-5D-5L dimensions were presented in three different orderings to test for order effects. Each respondent was randomly allocated to one of the three orderings, and saw the same ordering through all of the valuation tasks. The results of this part of the study are reported elsewhere [16].

Methods of analysis

Descriptive statistics (mean, median, standard deviation) of the health state values were compared across arms and against their sum scores. The distributions of values were also compared across arms.

The study design included a number of pairs of health states whereby one state can be considered to logically "dominate" the other (e.g. 21232 dominates 32442 because it is better on the first four dimensions and no worse on the fifth). A respondent can be described as having a logical inconsistency if, for any given pair of dominant-dominated health states, they give a higher value to the dominated state [17]. The propensity to give inconsistent valuations was compared across arms.

Interviewer effects were assessed by comparing the distributions of values given by respondents interviewed by each of the three interviewers. We also estimated a linear regression model of the form:

$$y = X\beta + \epsilon$$

where y is the health state value, X represents the explanatory variables, and ϵ represents the error term capturing other factors. Study arm, interviewer (one dummy

variable for each interviewer) and sequence (a proxy for interviewer learning effects: a dummy variable taking a value of 1 if the interview was one of the first 20 undertaken by the interviewer, and 0 otherwise) were included as the explanatory variables.

Descriptive statistics of respondents' self-reported health were examined, with particular focus on respondents who self-reported as being in health state 11111. Responses to the feedback questions were analysed by comparing the proportions of respondents who agreed or strongly agreed with each feedback statement across arms.

Comparisons across arms were assessed using chi-squared and two-sample t-tests. Analyses were undertaken using Microsoft Excel and STATA 11.2 software.

Results

The interviews were conducted between May and October 2014. In accordance with the ethical approval for this study, respondents who did not complete the interview in full were excluded from the analysis (n=6). We also excluded the data for 13 respondents who gave the same value in all of the TTO tasks. This is consistent with the exclusion criteria used in previous valuation studies [18]. Excluding these individuals results in a sample of 443 respondents, of whom 227 (51.2%) were allocated to the control arm and 216 (48.8%) to the test arm.

Older individuals (35.9% of the sample are aged 60 and over) and males (58.2%)were overrepresentedin comparison to the general population [19].The sample was also relatively well-educated, with 44.5% of respondents educated to university degree level or equivalent. We observed no statistically significant associations between study arm and these or other background characteristics (chi-squared test; in all cases $p>0.05$).

The difference in mean completion times across the arms was not statistically significant (t-test; $p=0.61$).

Valuation data

Table 1 summarises the descriptive statistics for the health states valued. The mean values are higher in the control arm for all health states except 21111, although for most states the differences are small and non-significant.

<Table 1 about here>

Figure 2 groups the health states by their sum scores and shows that the mean TTO values decrease as the sum score increases, an indicator of the face validity of the data. Higher mean values were observed in the control arm for the very mild health states (i.e. those with asum score of six), but the difference was not statistically significant at the 5% level (t-test; $p=0.10$).

<Figure 2 about here>

Figure 3 shows, for each arm, the overall distribution of TTO values for all health states combined except for 11111 which was valued only by respondents in the control arm. The relatively large number of observations at 0.95 and 1 reflect the inclusion of the three very mild health states in the design of this study (in the experimental design used to select health states for the EQ-5D-5L valuation studies, most blocks contain only one health state with a sum score of six). We observe some clustering at other "round-number" values, specifically at -1, 0 0.5 and 1, but overall the distributions are smoother than those observed in previous EQ-5D-5L studies [6,7,9].

<Figure 3 about here>

Overall, 112 respondents (25.3%) gave valuations that included at least one inconsistency (as defined above). The mean number of inconsistencies per respondent was 0.4. We do not observe a statistically significant association between study arm and the propensity to give inconsistent valuations (chi-squared test; p=0.16).

Of the 227 respondents who valued health state 11111, 223 (98.2%) gave it a value of 1. The lowest value given to 11111 was 0.9.

Interviewer effects

A team of three interviewers was used (INV1 – 115 interviews, INV2 – 169 interviews, INV3 – 159 interviews). The random allocation of respondents to study arms resulted in an uneven distribution at the interviewer-level, with 57.1% of respondents interviewed by INV3 but only 45.8% of respondents interviewed by INV1 being allocated to the control arm.Figure 4 highlights the differences in the data collected by the three interviewers. For example, respondents interviewed by INV1 were more (less) likely to give health states a value of -1 (0).

<Figure 4 about here>

The regression analysis suggests that interviewer effects are present. The coefficients for interviewer (INV1: $p<0.01$; INV3; $p=0.02$) and sequence ($p<0.01$) were statistically significant; the coefficient for study arm was not ($p=0.18$).

Self-reported health

When asked about their own level of health today (i.e. on the day of the interview), 224 respondents (50.6%) self-reported as being in health state 11111. Of these 224 respondents, 187 (83.5%) self-reported an EQ-VAS score of less than 100, indicating that despite having no problems with the five dimensions covered by EQ-5D, they considered their level of health to fall short of the EQ-VAS upper anchor of "best

imaginable health". The mean (median) EQ-VAS score for respondents self-reporting as being in 11111 was 89.1 (90).

Feedback

In the structured feedback questions, all respondents were asked to indicate their level of agreement (via five-point Likert items ranging from strongly agree to strongly disagree) with the following three statements: (1) "It was easy to understand the questions I was asked"; (2) "I found it easy to tell the difference between the lives I was asked to think about"; and (3) "I found it difficult to decide on the exact points where Life A and Life B were about the same". The vast majority of respondents agreed or strongly agreed with statements 1 and 2 (91.4% and 90.3%, respectively). Opinion regarding statement 3 was more divided, with 50.7% of the respondents agreeing or strongly agreeing with it. We do not observe a statistically significant association between study arm and the propensity to agree or strongly agree with any of the three statements.

Follow-up questions

Responses to the follow-up questions are available for 436 respondents. These data are unavailable for the remaining seven respondents due to a recording error. It is not expected that the missing data will differ systematically from those of rest of the sample.

When asked to compare 11111 and full health, 305 respondents (70.1%) stated that they considered the two descriptions to be the same as each other. The respondents who did not consider them to be the same offered explanations such as:

- "Full health is a collection of factors - physical, psychological and social wellbeing. Someone can have everything on the left-hand side of the list and still not be in full health because of loneliness"
- "[11111] seems to stress physical capabilities. Full health must include emotional and mental condition."
- "Being in full health means nothing is wrong, whereas overweight unhealthy people may not have any problems walking."

374 respondents (86.0%) gave full health an EQ-VAS score of 100 (mean score: 98.6). By contrast, 253 respondents (58.2%) gave 11111 an EQ-VAS score of 100 (mean score: 95.1).

Of the six health state descriptions included in the ranking task, "perfect health" was most often ranked as the state that respondents most wanted to live in (ranked best or joint-best by 60.5% of respondents), followed by (in order) "full health", "best

imaginable health", "no health problems", 11111 and "healthy". Full health and 11111 were ranked best/joint-best by 42.7% and 20.9% of respondents, respectively.

177 respondents (40.6%) stated that there were aspects of health that they considered to be important but were not covered by the five EQ-5D dimensions. These included vision, hearing, energy and sociability. A similar number of respondents (202; 46.3%) stated that there were important aspects of quality of life that were not covered by EQ-5D.

Discussion

The results of this study suggest that the EQ-5D-5L health state values elicited using the composite TTO approach are not greatly affected by whether full health or 11111 is used as the comparator health state. We examined a number of standard measures of the quality and face validity of the TTO data. The propensity for respondents to give inconsistent valuations was not found to be associated with the choice of comparator health state. In both study arms, higher average values were observed for health states that can be considered to be relatively milder, and lower average values were observed for more severe health states. This suggests that the data have acceptable face validity, and we did not observe differences across the arms in this respect.

Overall, the average values were higher in the control arm. However, this was not the case for all of the health states, and for the very mild health states the observed difference was not statistically significant. Further, the difference appears to be driven by interviewer effects (respondents interviewed by INV1 tended to give lower values overall andwere allocated disproportionately to the test arm) rather than by inherent differences between the arms. Interviewer-led administration of complex stated preference surveys is usually preferred because of the need for interviewers to explain instructions and to guide the respondents [20,21]. However, this can result in interviewer bias because different individuals have different interviewing styles.

The results presented here demonstrate that the differences in values between the recent EQ-5D-5L valuation studies and the earlier EQ-5D value sets are not explained by the change in the description of the comparator health state. However, it should also be noted that the low mean values for the very mild health states in the EQ-5D-5L value set for England study [9]were not observed in either arm of the current study. This suggests that other factors (such as the use of different groups of interviewers or other changes to the valuation protocol) were more influential than the choice of comparator state in determining the health state values.

Almost all of the respondents who valued 11111 via TTO gave it a value of 1. Yet many respondents indicated through their responses to the follow-up questions that they did

not consider full health and 11111 to be the same, with full health rated and ranked higher than 11111 overall. In the EQ-5D-5L valuation protocol it is not possible to make trades of less than six months, which means that the highest value (other than 1) that can be given to a health state is 0.95. It is conceivable that many respondents consider living in full health for 10 years to be better than living in 11111 for 10 years (in which case 11111 should have a value of less than 1) but would not be willing to trade as much as six months of life in order to live in full health rather than 11111. Nevertheless, this study provides no empirical evidence to suggest that assigning a value of 1 to health state 11111 is problematic.

The finding that many respondents self-reported as being in 11111 whilst also self-reporting an EQ-VAS score of less than 100 is not unexpected. It is consistent with the results of the ranking task, in which the EQ-VAS upper anchor – "best imaginable health" – was ranked higher than 11111 overall. A detailed analysis of the responses to the follow-up questions is reported elsewhere [22]. We would encourage qualitative research that seeks to further develop our understanding of what the different candidate upper anchor represent and how the various descriptors are interpreted by respondents. Although the evidence suggests that the TTO values elicited are largely unaffected by whether full health or 11111 is used as the comparator health state, there may be other reasons for preferring one comparator or the other. The term "full health" does not translate exactly into some languages, such as Arabic [23], which makes comparisons across countries challenging. On the other hand, the literature makes it clear that QALYs are anchored at full health rather than at an instrument-specific best health state [24]. Indeed, TTO is used to value not only EQ-5D-5L health states but also health states defined by other health-related quality of life instruments [25], so a consistent upper anchor is needed in order for the values elicited to be fully comparable.

Some limitations of the study design should be mentioned.The study arms were less well matched than desirable. Following the computerised randomisation procedure, respondents interviewed by INV3 were more likely to be allocated to the control arm than those interviewed by INV1. Future studies should consider making provisions to ensure that each interviewer is allocated an equal number of respondents in each arm. Using a larger interviewer team might also have lessened the impactof any one interviewer on the overall study results, though this potential benefit should be considered alongside the fact that larger interviewer teams are generally more difficult to train and monitor than smaller teams.

The follow-up questions offer useful insights and supporting evidence. However, it should be noted that these questions had not been piloted or used in previous research, so we

cannot be certain that they were well understood by respondents (though the interviewers did not report any perceived difficulties in understanding).

Notwithstanding these limitations, we have demonstrated in this paper that the choice of comparator health state does not greatly affect EQ-5D-5L values elicited using the composite TTO approach.

REFERENCES

[1]Furlong W, Feeny D, Torrance GW, et al. Multiplicative Multi-attribute utility function for the HUI Mark 3 (HUI3) system: a technical report. McMaster University Centre for Health Economics and Policy Analysis Working Paper 98-11. Hamilton: McMaster University; 1998.

[2] Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. Journal of Health Economics 2002;21:271-292.

[3]Richardson J, Peacock S, Day N, Iezzi A. The Assessment of Quality of Life (AQoL) II instrument. Derivation of the scaling weights using a multiplicative model and econometric second stage correction.Melbourne: Centre for Health Economics, Monash University; 2004.

[4]Kind P, Brooks R, Rabin R (eds). EQ-5D concepts and methods: a developmental history. Dordrecht: Springer; 2005.

[5] Dolan P. Modeling valuations for EuroQol health states. Medical Care 1997;35(11):1095-1108.

[6]Xie F, Pullenayegum E, Bansback N, et al. The Canadian EQ-5D-5L valuation study: an exploratory analysis. 30th Scientific Plenary Meeting of the EuroQol Group, Montreal, Canada, September 12-14 2013, Proceedings: 1-17. Rotterdam: EuroQol Group; 2014.

[7] Ramos-Goñi JM, Pinto-Prades JL, Cabasés JM, Rivero-Arias O. Valuation and modeling of EQ-5D-5L health states using a hybrid approach. Medical Care; In Press.

[8]Herdman M, Gudex A, Lloyd MF, Janssen, Kind P, Parkin D, Bonsel G, Badia X. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Quality of Life Research 2011;20(10):1727-1736

[9] Van Hout B, Devlin NJ, Shah KK, et al. An EQ-5D-5L value set for England. Final report for the Department of Health. London: Office of Health Economics; 2014.

[10]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 in Health 2014;17:445-453.

[11]Jelsma J, Hansen K, de Weerdt W, et al. How do Zimbabweans value health states? Population Health Metrics 2003;1:11-20.

[12] AFD Names and Numbers. http://www.afd.co.uk/product_namesandnumbers.asp

[13]Robinson A, Spencer A. Exploring challenges to TTO utilities: valuing states worse than dead. Health Economics 2006;15:393-402.

[14] Janssen BMF, Oppe M, Versteegh MM, Stolk EA. Introducing the composite time trade-off: a test of feasibility and face validity. European Journal of Health Economics;14(S1):S5-S13.

[15]Mulhern B, Bansback N, Brazier J, et al. Preparatory study for the re-valuation of the EQ-5D tariff (PRET): Methodology report. Health Technology Assessment 2014;18(12).

[16] [Details removed to protect the blind review process]

[17] Devlin NJ, Hansen P, Kind P, Williams A. Logical inconsistencies in survey respondents' health state valuations - a methodological challenge for estimating social tariffs. Health Economics 2003;12:529-544.

[18]Szende A, Oppe M, Devlin N. EQ-5D valuation sets: an inventory, comparative review and users' guide. Dordrecht: Springer; 2007.

[19] Office for National Statistics. 2011 Census, Population Estimates by single year of age and sex for Local Authorities in the United Kingdom. Available from http://www.ons.gov.uk/ [Accessed October 23, 2014]

[20] Bridges, JFP, Hauber AB, Marshall D, et al. Conjoint analysis applications in health – a checklist: a report of the ISPOR Good Research Practices for Conjoint Analysis Task Force. Value in Health 2011;14:403-413.

[21] Shah KK, Lloyd A, Oppe M, Devlin NJ. One-to-one versus group setting for conducting computer-assisted TTO studies: findings from pilot studies in England and the Netherlands. European Journal of Health Economics 2013;14(S1):S65-S73.

[22][Details removed to protect the blind review process]

[23]Papadimitropoulos M, Elbarazi I, Blair I, et al.An investigation of the feasibility and cultural appropriateness of stated preference methods to generate EQ-5D-5L values in the United Arab Emirates. OHE Research Paper. London: Office of Health Economics; 2015.

[24] Brazier J, Ratcliffe J, Salomon JA, Tsuchiya A. Measuring and valuing health benefits for economic evaluation. Oxford: Oxford University Press; 2007.

[25] Brazier J, Rowen D. NICE DSU Technical Support Document 11: Alternatives to EQ-5D for generating health state utility values. Sheffield: Decision Support Unit; 2011. Available from http://www.nicedsu.org.uk [Accessed November 27,2014]

Figure 1. Screenshots depicting TTO task in the control arm (upper) and test arm (lower)

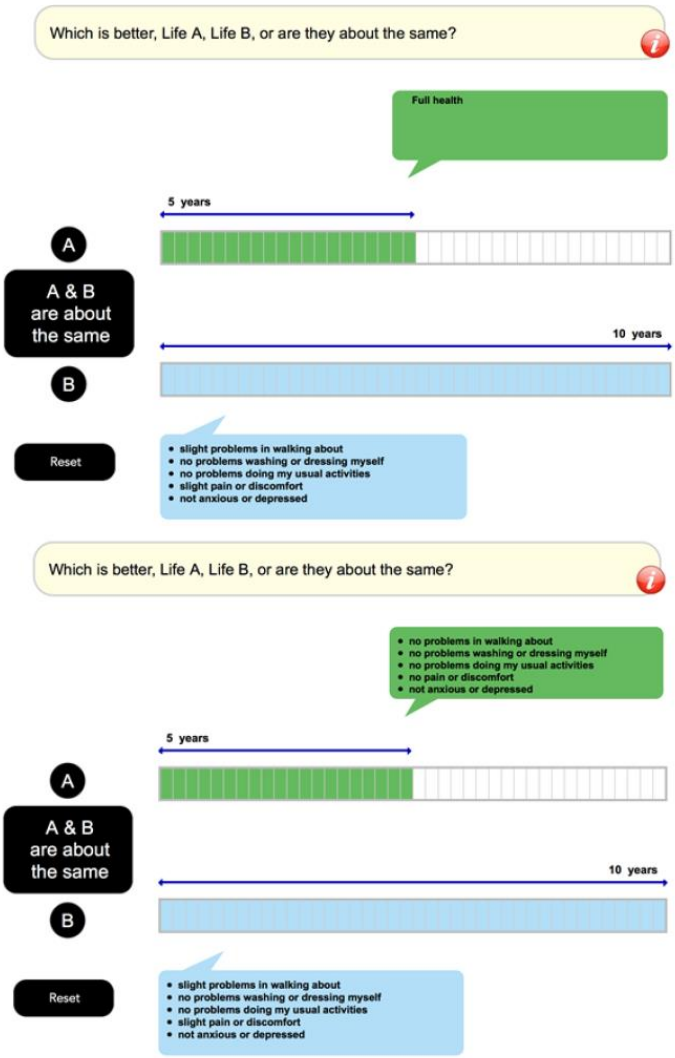


Figure 2. Mean TTO value, by sum score

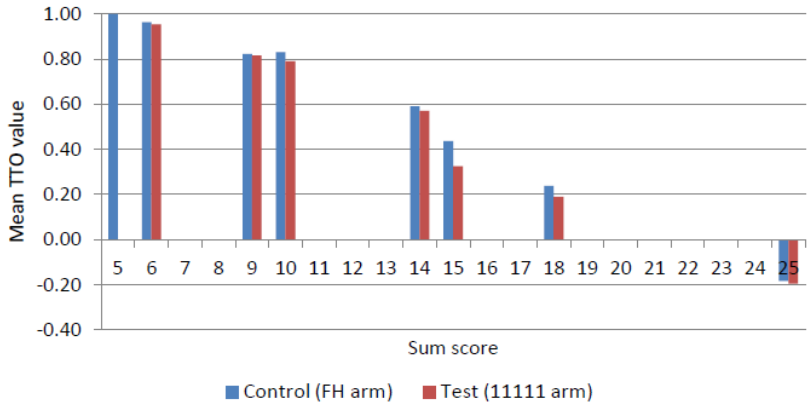


Figure 3. TTO valuation distribution across all health states (except 11111), by arm

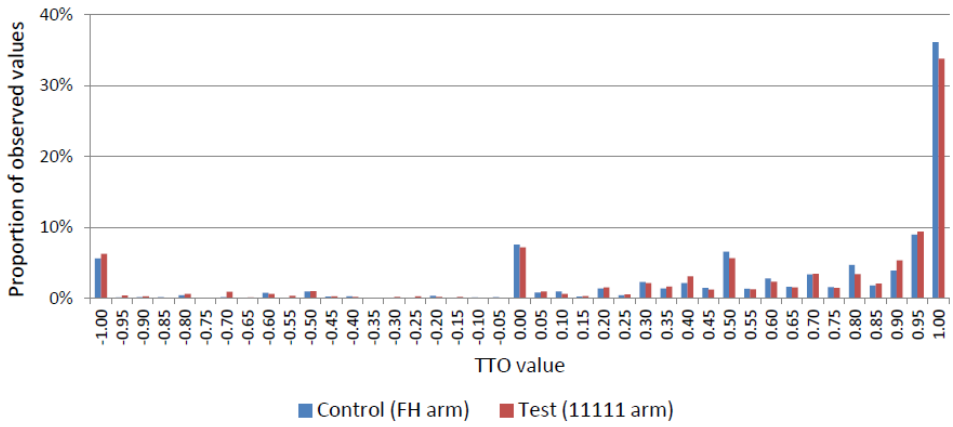
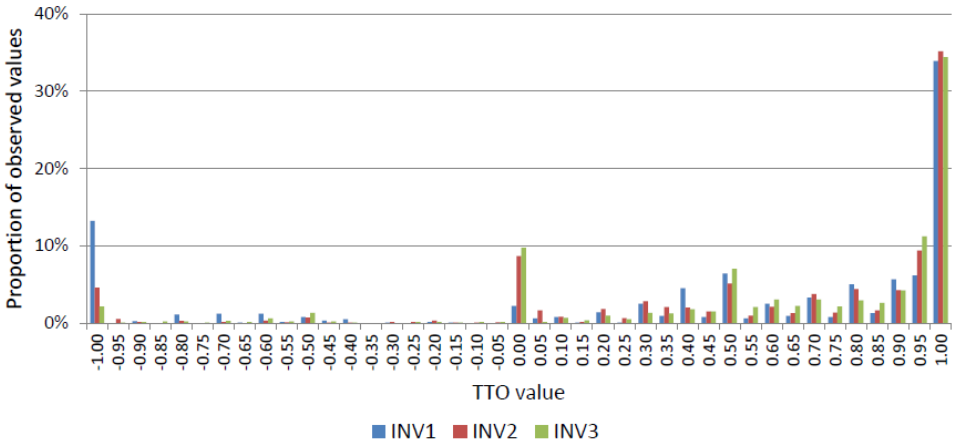


Figure 4. TTO valuation distribution across all health states (except 1111), by interviewer



TABLES

Table 1. Descriptive statistics for the health state valued, by arm

		Control (FH arm)				Test (11111 arm)					
State	Sum score	n	Mean	Median	SD	n	Mean	Median	SD	One-way t-test p-value	Two-way t-test p-value
11111	5	227	0.999	1.00	0.009	0				N/A	N/A
21111	6	227	0.957	1.00	0.117	216	0.962	1.00	0.119	0.672	0.657
11121	6	227	0.970	1.00	0.079	216	0.962	1.00	0.116	0.218	0.437
11112	6	227	0.961	1.00	0.091	216	0.936	1.00	0.195	0.045*	0.090
11223	9	227	0.823	0.95	0.291	216	0.814	0.95	0.335	0.390	0.780
21232	10	227	0.830	0.95	0.264	216	0.789	0.90	0.306	0.068	0.137
43331	14	227	0.591	0.75	0.485	216	0.571	0.70	0.482	0.331	0.662
32442	15	227	0.437	0.50	0.507	216	0.325	0.45	0.598	0.017*	0.033*
55233	18	227	0.292	0.50	0.603	216	0.264	0.40	0.621	0.311	0.623
34155	18	227	0.184	0.35	0.593	216	0.116	0.30	0.647	0.123	0.246
55555	25	227	-0.182	0.00	0.554	216	-0.194	0.00	0.576	0.410	0.819

* statistically significant at the 5% level