

A comparison of three binary choice methods for health state valuation

Brendan Mulhern¹, Aki Tsuchiya^{1,2}, Nancy Devlin³, Ken Buckingham⁴, Donna Rowen¹ & John Brazier¹

1 Health Economics and Decision Science, School of Health and Related Research,
University of Sheffield, Regent Court, Sheffield, S1 4DA, UK

2 Department of Economics, University of Sheffield, Sheffield, S1 4DT

3 Office of Health Economics, 12 Whitehall, London, SW1A 2DY

4 University of Otago, New Zealand

Corresponding author:

Brendan Mulhern, School of Health and Related Research, University of Sheffield, Regent Court, 30 Regent St, Sheffield, S1 4DA, United Kingdom, e-mail: b.mulhern@sheffield.ac.uk; Tel no: +44 (0)114 222 0794

Acknowledgements

This study is part of a wider project (Preparation for the Re-valuation of the EQ-5D Tariff (PRET) project) funded by the MRC Methodology programme (ref.G0901500). We are grateful to all respondents who took part. Research ethics approval has been given by the ethics committee at the School of Health and Related Research in line with University of Sheffield research governance and ethics requirements. The usual disclaimer applies.

Wider PRET project team

Nick Bansback, John Brazier, Ken Buckingham, John Cairns, Nancy Devlin, Paul Dolan, Arne Rise Hole, Louise Longworth, Brendan Mulhern, Donna Rowen, Aki Tsuchiya

Introduction

There are a range of preference elicitation methods that can be used to generate health state utility values. These values are subsequently used in the calculation of Quality Adjusted Life Years (QALY) to inform the economic evaluation of new and existing interventions. One key method is the Time Trade Off technique (TTO, Torrance et al., 1974). TTO was used to value the generic preference based measure EQ-5D (Brooks, 1996; Dolan, 1997), which is recommended for use by the National Institute of Health and Clinical Excellence (NICE) in the UK to generate QALYs (NICE, 2008). It has also been used to generate utility scores for condition specific preference based measures of health (Rowen et al., 2011; Yang et al., 2011).

Preference elicitation methods can either be used to value states described in a health classification system (either generic or specific) or directly to a health state. Such methods enable a utility score (anchored on a 0 to 1 scale, where 1 is equivalent to full health, 0 to dead, and negative values to states worse than dead) to be generated for all possible health states described by the classification system. This is done by modelling preferences for the health states included in the valuation study using regression. For example, the EQ-5D descriptive system assesses health across five dimensions (mobility, self care, usual activities, pain/discomfort and anxiety/depression) each with three severity levels. Therefore the descriptive system produces 243 (3^5) health states, of which 42 were directly valued using TTO by 2,997 respondents in the UK (Dolan, 1997). The TTO results were modelled using multivariate regression to produce a utility scale for every health state (with a range of -0.594 to 1). Recently, a new five level version of EQ-5D has been developed (EQ-5D-5L, Herdman et al., 2011), which assesses health on the same dimensions using five response levels (none, slight, moderate, severe, extreme). Utility values for EQ-5D-5L will be derived in the near future.

Under the conventional TTO protocol, respondents trade off life years (x) in full health against a set number of years (usually 10) in a given health state until indifference is reached between the hypothetical options (Gudex et al., 1994). The value for the health state at the indifference point is calculated as $x/10$. When the health state is perceived as worse than dead, the preference elicitation procedure changes. Respondents trade off between w years in the selected health state followed by x years in full health (where $w+x=10$) and immediate death. The different procedure used to value states worse than dead and subsequent transformation required to derive the utility value has raised concerns regarding the comparability of utility values produced using the different procedures (Lamers et al., 2006).

To attempt to standardise the preference elicitation procedure for states both better and worse than dead, the Lead Time -Time Trade Off (LT-TTO) was developed (Robinson & Spencer, 2006; Devlin et al., 2010; 2011). Using this method, a lead time in full health is added to the overall profile before the health state occurs, regardless of whether it is better or worse than dead. For example, if lead time was 10 years, respondents trade off between living in full health for 10 years followed by a selected health state for 10 years against living in full health for a duration (x) between 0 and 20 until indifference is reached. This allows the values for the health state to be negative (where $x<10$), and the TTO values for all health states to be calculated using the same method (i.e. $x/10$). Initial testing has demonstrated that LT-TTO is a feasible method to elicit

preferences, but methodological concerns remain including the extent to which respondents may 'exhaust' or use up all of the lead time in full health if health states are particularly severe (Devlin et al., 2011).

While TTO and LT-TTO elicit a value for each state from each individual respondent, there are methods that model aggregate preferences by only eliciting ordinal preferences from individuals. These include Discrete Choice Experiments (DCE). The use of DCE methods in health economics has increased in recent years (DeBekker-Grob et al., 2010). In a standard DCE task, respondents are asked to provide ordinal preferences over pairwise choices, where the choices consist of multidimensional health states (Ryan & Gerard, 2003). Where DCE is used to value health states, respondents would be asked to choose between health state profiles that include a number of attributes and associated levels. Results are modelled using regression to produce a coefficient for each level of each attribute. DCE studies can be used to derive utility values for health state descriptive systems, but preferences must be anchored on the full health (1) – dead (0) utility scale. This can be done both analytically, for example by anchoring on mean preferences derived for a particular state from TTO, and also methodologically. A methodological approach for deriving utility values using DCE was developed by Bansback et al. (2012), who incorporated an attribute for duration into the DCE scenarios in addition to EQ-5D health state profiles. Responses were modelled to be equivalent to the indifference point in TTO by estimating coefficients for the interaction between EQ-5D dimension and duration, and using the coefficients to calculate the value of health states by solving the equivalence relationship for a binary choice situation between living in a given health state for a specific duration of time and living in full health for a shorter duration. This approach is known as DCE_{TTO} .

Health state valuation tasks may be difficult for respondents: TTO, because it involves an iterative process where respondents are asked to reach the point of indifference by choosing between options where one option changes based on their previous answer; and DCE, because it involves a choice between two options where all of the attributes included in the options should inform responses, and may differ from task to task. It is therefore important to understand factors that may impact on the validity of responses, including the acceptability of the techniques to respondents. Research comparing DCE and TTO found that both techniques have acceptable predictive validity (Hakim and Pathak, 1999) and equivalent respondent comprehension and overall level of completion (Ratcliffe et al., 2011). However these studies did not test the DCE_{TTO} methodology which may be more difficult than standard DCE due to addition of duration.

The strategies and processes used by respondents to complete TTO and DCE tasks is also an important factor to understand as they may influence the validity of responses, or inform the design of valuation studies. Robinson et al., (1997) found that respondents in a TTO study may use a "threshold of tolerability" to establish whether a state is severe enough for them to trade any time at all. In qualitative work, San Miguel et al., (2005) found that respondents in a DCE study introduced additional information and assumptions to help them answer the questions. It has also been found that respondents may focus on key attributes, and may not attend to all attributes included, both because the attribute is not relevant to the individual, and also to simplify the task (Ryan et al., 2009).

The subjective importance to respondents of the actual health dimensions included in the hypothetical scenarios is an important factor to assess. This relates to the dimensions that respondents attend to first when decision making. Values for both generic and condition specific preference based measures are mostly derived from the general population, and different descriptions of health dimensions differ in their level of importance. A severe level of a key health dimension may carry more weight, and it is important to understand the qualitative hierarchy of dimension importance that influences respondent's answers when completing valuation tasks. Quantitative information about the importance of dimensions is available from the regression coefficients for each level of each dimension of a health state descriptive system, and the dimensions with the most subjective importance may or may not be the same as those dimensions with the largest weight. Quantitative information about the importance of EQ-5D-5L dimensions is not currently available. Furthermore, little is known about the qualitative importance of the five EQ-5D-5L dimensions and associated response levels.

External respondent related factors and background characteristics may also impact the results of health state valuation studies. Dolan & Roberts (2002) found that age, gender and marital status influenced responses to TTO tasks, and respondents' own experiences have been found to impact on choices made in both TTO and in non health state valuation DCE studies (Jansen et al., 2000; San Miguel et al., 2005). It has also been established that respondents who find valuation tasks complex are less likely to be educated to college level (Wittenburg & Prosser, 2011).

Iterative TTO and LT-TTO procedures can be conceptualised as multiple binary choice tasks following a similar format used to represent DCE_{TTO} scenarios (Tsuchiya et al., 2011). This means that the iterative task process can be simplified (Bosch et al. 1998), and direct comparisons with DCE_{TTO} can be carried out. Furthermore the binary choice tasks are amenable to completion using a variety of media including Computer Assisted Personal Interview (CAPI) and on-line, which produce comparable results for binary choice questions (Mulhern et al., 2011). However, note that since individuals do not report their point of indifference, there is a fundamental shift in the focus of the analysis, from determining a mean over individual preference, to modelling group preferences in other ways that do not rely on group means.

This study aims to investigate the validity and acceptability of binary choice versions of TTO, LT-TTO and DCE_{TTO} using Computer Assisted Personal Interview (CAPI) methods with EQ-5D-5L health states, and to our knowledge is the first to compare binary choice conceptualisations of TTO and DCE. The processes respondents use to complete health state valuation tasks and the influence of a range of external factors and demographics on responses are also assessed. This includes an investigation of the importance of the EQ-5D-5L dimensions in the decision making process. Research investigating these issues will add to the literature about how health state valuation tasks are completed, and why particular preferences are given. This study is part of the wider Preparation for the Re-valuation of the EQ-5D Tariff (PRET) project (see Tsuchiya & Mulhern, 2011 for a description of the earlier stages of the project).

Methods

Valuation question format

The format of the three types of binary choice questions used in this study is displayed in Figure 1. In binary choice TTO (question type I), scenario A includes an EQ-5D-5L health state with an associated duration level and scenario B presents full health for a shorter duration (therefore eight pieces of health state and duration information that are included in each question). In binary choice LT-TTO (type II), scenario A presents full health for a certain duration followed by an EQ-5D-5L health state for a certain duration, and scenario B presents full health for a specified duration (meaning 10 pieces of information in each question). DCE_{TTO} (type III) presents an EQ-5D health state with an associated level for duration for both scenarios A and B (therefore 12 pieces of information in each question).

Three tasks of each of the question types were set (see Table 1) For types I and II, three EQ-5D-5L states defined as mild, moderate and severe were selected, and the same states were used across both question types. A duration level was selected to go with each state, and the full health duration was varied in accordance with the selected health state duration level. For type III, the same three states were presented as scenario A, with a state of similar severity presented as scenario B. For two of the tasks, time spent in the health state varied across the pairs, and duration was constant across the third example scenarios. Members of the research team selected the states to use across the scenarios with the aim of setting the duration spent in the health state and full health (for types I and II), or the health state pairs (in type III) to provide a difficult choice for respondents that would enable us to investigate the strategies and processes used to answer the questions in more depth than if the decision about which scenario to choose was easier to make.

Follow up question format

After completing three tasks of a given question type, follow up probing questions were used to investigate the issues related to question acceptability and task completion (see appendix 1). The majority of the probing questions took the format of tick boxes, with a free text question available to allow for further issues to be raised by respondents. The questions were developed through a series of pilot studies with a convenience sample of academic and non-academic university employees. The questions were conceptualised across four categories: task completion process and acceptability; potential difficulties answering the questions; importance of EQ-5D-5L dimensions; and external influences on response. One set of follow up questions was devised to investigate issues specific to each question type, and these appeared after the three binary choice question examples (five questions were completed for each). A further set of general follow up questions were included to assess issues specific to both types of valuation task.

Study design

CAPI interviews were used to administer the health state valuation and follow up questions. Each respondent completed two types of binary choice questions and associated type-specific follow up questions. This was followed by the general feedback questions relating to both valuation methods. Each valuation task was presented as both the first and second of the two completed by respondents, and therefore there were six versions of the survey overall. Respondents also completed demographic questions, self reported general

health (on five point scale from “excellent” to “poor”), health and life satisfaction questions (on a scale of 0 (not satisfied at all) to 10 (completely satisfied)), and EQ-5D-5L at the start of the interview. Following completion of the interview, interviewers completed three questions about the environment in which the interview was conducted.

Recruitment

Interviews were carried out with a representative sample of the UK general population by trained and experienced interviewers. Respondents were recruited by knocking on one in every 10 doors in selected postcodes in five areas of the UK. The interviewer explained the project and gained consent from an eligible member of the household (i.e. aged over 18 and of an age and gender quota where interviews were still required for the sample to be representative). The questions were presented to respondents on a laptop, with the interviewer reading out all of the content displayed on the screen, and recording the response. Interviews were conducted in a one-to-one setting, and participants were able to stop the interview at any time. The recruitment procedure and overall study design gained ethical approval from the University of Sheffield School of Health and Related Research ethics committee.

Analysis

Descriptive statistics including frequency and cross tab analyses were used to assess the results to the follow up questions. Significance testing between demographic groups was carried out using Chi square tests.

Results

Sample demographics, response and interview information

In total, interviewers knocked on 1,783 houses to achieve 306 interviews (a response rate of 17.2%). Of those who did not participate, 789 (44.3%) were not at home or unavailable, 333 (18.6%) refused, and 355 (19.8%) were out of scope (i.e. if no one in the house fitted the age and gender quota groups that were still to be completed). The response rate for eligible contacts was 48%. Table 2 presents demographic information and data relating to the interview environment. Overall the sample was generally representative of the UK general population and the majority self reported good health and high levels of health and life satisfaction. The majority of respondents displayed a good understanding of the task, and concentrated on the questions. The majority of the interviews were conducted in a quiet environment with no distraction from other activities in the household.

Task comparison and acceptability

Overall, 52.3% of respondents reported that both of the question types that they completed were of equal difficulty (table 3). Of those who indicated a different level of difficulty across the questions, type II was perceived as the easiest followed by type I and type III. Furthermore, 36% of those that completed both type I and type III and 47% of those who completed both types II and III stated that the TTO (i.e. types I and II) binary choice questions were easier to complete. This indicates that the binary choice conceptualisation of both TTO tasks may be more acceptable to respondents than DCE_{TTO} questions.

The majority of the sample (71.2%) reported that the layout of the questions meant that they could be answered easily. However, across all three question types, over half the sample reported that they sometimes or always found it difficult to complete the task, with the most difficulties being reported by those who completed question type I (TTO) first (Figure 2). The difference in reported levels of difficulty between the groups is significant ($p < 0.01$). Of the overall group, 17% of respondents reported that DCE_{TTO} questions encouraged them to think about external influences the most when responding, and this is higher than the TTO (9%) and LT-TTO (10%) questions. However the majority (64%) reported that the questions were equivalent in this regard.

Attention to attributes

Overall, 43% of those completing question type I, 33% of those completing type II and 24% of those completing type III indicated that they always completed the task by only considering the most important attribute, and the difference in response between the tasks across the questions is not significant ($p = 0.07$). The majority of the sample agree that they only consider the attributes that are subjectively important to them when completing the tasks, and this is generally consistent irrespective of which question the respondent found the easiest (Figure 3). However, 35% of those who complete question type III (DCE_{TTO}) indicated that they did not only consider the most important attribute, indicating that they are assessing a number of attributes when choosing between the options.

Importance of individual task attributes

Importance of health state and duration attributes

Respondents were asked to indicate which single attribute included in the valuation task (i.e. EQ-5D dimensions and duration) was most important in the decision making process (see Table 4). In types I and II, the duration spent in full health is consistently indicated as the most important attribute, and this is followed by the duration in the health state. When all task attributes are included, the EQ-5D dimension with the highest number of respondents indicating that it is the most important in the decision making process is mobility, and this is consistent across the question types.

Importance of EQ-5D-5L dimensions

Respondents were asked to rank all five EQ-5D-5L dimensions (excluding the duration attributes) in order of importance in the decision making process, and the results overall are displayed in Figure 4. Mobility was ranked as most important by the largest number of respondents (29.4%), with Pain/Discomfort ranked as the most important dimension by 24.5%. Anxiety/Depression was ranked as the least important dimension in the decision making process by the highest frequency of respondents (40.5%). Furthermore, the highest frequency across each of the rankings corresponds with the order the dimensions appear in the classification system. When the results are assessed by question type, a similar pattern is established as Mobility is ranked as the most important dimension across all question types, with Anxiety/Depression ranked as the least important (figure 5). A large proportion of the sample reported that they were able to tell the difference between the EQ-5D-5L dimension response levels slight/moderate, moderate/severe and severe/extreme, but

9.2% reported that they could never tell the difference between severe and extreme, and 20.5% reported that they could not tell the difference between severe and extreme in some situations (Figure 6).

Influence of external factors and background characteristics on response

Overall, 269 (87.9%) of respondents reported that they imagined themselves living in the health state. However, 30.7% reported that their own health experiences influenced their response, 26.1% reported that other people's experiences influenced their response, 31.4% reported that both groups influenced response and 11.8% reported that neither group influenced response. Of the 269 respondents reporting that they imagined themselves in the health states, 90 (33.4%) reported that their own health influenced their response, 66 (24.5%) reported that their response was influenced by other people with poor health, 79 (29.4) said both and 34 (12.6%) said neither of these groups (see figure 7). Table 5 reports the influence of a range of other external factors.

Overall, 78% of the sample indicate that they always or often consider how the health state would impact on their feeling about their health and life, 48% report that they would always or often consider the impact of the health state on their life and financial situation. Furthermore, 36% of the sample indicate that they would always or often choose a longer duration to spend time with others, but 40% report that they rarely or never did this.

Figure 8 displays the impact of considerations about other people on response across a selection of background characteristics. Overall, 85.3% of the sample report that their responses are influenced by considerations about how the health state would affect other people close to them either 'sometimes' or more often. Respondents who are married or with partner are significantly more likely to indicate that their answers are influenced by how the health state would affect those around them ($p < 0.01$). There is no overall difference in response for those with or without children aged under 18 ($p = 0.18$) or those with dependents aged over 18 ($p = 0.58$).

Figure 9 displays the impact of age and level of responsibility to others on response across a number of key demographic variables. Overall, 79.1% of the sample report that their age and level of responsibility impacts on their responses at least sometimes. There are no significant differences regarding how age and responsibilities impact on response by marital status ($p = 0.11$), having children ($p = 0.54$), having dependents aged over 18 ($p = 0.16$), being employed ($p = 0.51$) or by age group ($p = 0.51$).

Discussion:

This study used CAPI methods to investigate issues related to the completion of health state valuation tasks using binary choice presentations of the methods, and found that the tasks were acceptable, but both the TTO and LT-TTO tasks may be easier for respondents to complete than the DCE_{TTO} task. It was also found that the majority of respondents do not attend to all of the health attributes when completing the tasks, and there are a range of external factors that influence the response. Secondly, we investigated the EQ-5D-5L descriptive system. The importance of each dimension to general population respondents was assessed and we also investigated whether respondents can differentiate between the five response levels. When respondents rank the order of importance of the EQ-5D-5L dimensions, there is some evidence of an ordering effect where they are matched with the ordering in the descriptive system (and, indeed, the ranking question).

In this study we have represented iterative TTO and LT-TTO as binary choice questions, and this has enabled us to compare the tasks directly with a version of DCE incorporating duration (DCE_{TTO}). To our knowledge a direct comparison has not been carried out before. Past work has suggested that both techniques in their standard form are acceptable and valid for respondents, although many respondents find the tasks difficult (Hakim & Pathak, 1999, Ratcliffe et al., 2011). Respondents in this study also found the tasks difficult to complete. However we have demonstrated that in their binary choice form, TTO and LT-TTO are easier to complete by respondents than DCE_{TTO} . This may be because the iterative nature of the task is removed, and the options incorporate less attributes than a DCE_{TTO} task.

In their conventional form, TTO and LT-TTO iterate until the point of indifference between the health state and full health is achieved, and this point is used to calculate the TTO value. This process is not followed when deriving utility values for health states using DCE_{TTO} , as only ordinal preferences are achieved for each task. However, by designing studies that incorporate many health state pairs administered to a sufficient sample size, it is possible to model the ordinal results to derive a utility scale, and the feasibility of this has been demonstrated by Bansback et al. (2012) using EQ-5D. It would also be possible to use the binary choice conceptualisations of TTO and LT-TTO to derive utility values as both include a duration attribute so can be anchored on the full health – dead scale as required. However, further work would be needed to produce a valid study design with a sufficient number of states, and also to establish the exact form that the regression model to estimate utility values would require.

It is interesting to note that discrete choice methods side step one of the more intractable difficulties that arise in aggregating individual level utility data, i.e. the problem of infinite negative utilities. Respondents refusing to accept any trade that involve durations in poor health, however short, are implying that that state of health is infinitely bad. The arithmetic mean of any set in which one member is infinite is also infinite, implying that the use of mean values with the conventional TTO might be untenable. Discrete choices methods do not derive utilities for individuals and so cannot determine the mean utility over individuals. However the problem does not go away. Some people may still have extreme values. Discrete choice questions simply do not address them. The utilities of individual respondents are not estimated, so they cannot be averaged. Discrete choice methodologies are more akin to complex voting procedures in which societal preferences are inferred by consensus. There are other methods, in addition to discrete choice, of avoiding the problem of aggregation while preserving individual level utility estimation. These include: censoring, the use of alternative measures of central tendency (the median or mode), and the use of mathematical transformations.

Past work investigating the processes used to help respondents complete DCE tasks has found that respondents introduce further assumptions and also do not attend to all attributes (San Miguel et al., 2005; Ryan et al., 2009). To some extent this is supported here as a group of respondents reported that they answer by only considering the subjectively most important attribute, and this was found consistently across question types. However there is also a subset of those completing the DCE_{TTO} questions who indicate that they attend to more than one attribute, but it is unclear how many. This is an area that warrants further investigation to establish how many attributes it is reasonable to present in binary choice health state

valuation tasks. It may be possible to improve attribute attention by improving the study design and presentation of tasks. For example, participants could be asked to consider all of the attributes, or advances in computer technology could be used to develop innovative methods for presenting the health states. The visual presentation of individuals experiencing health states (for example using video) may also encourage respondents to spend longer and think more deeply about the states that they are appraising. Further to this, it is essential in studies using binary choice methods that the number of tasks completed by respondents is kept to the minimum required to achieve valid results (which may be 10-15, Tsuchiya et al., 2011). This helps minimise respondent fatigue and the use of strategies to complete the study as quickly as possible.

This study also assessed the importance of EQ-5D-5L dimensions to general population respondents when presented in health state valuation tasks. This relates to which dimensions respondents pay attention to which may or may not be the same as the dimensions with the highest disutility. When assessing the overall ranking of EQ-5D dimensions, the results indicate that Mobility is the most important dimension followed by Pain/Discomfort, Self Care, Usual Activities, with Anxiety/Depression the least important. This order differs by question type when duration is also included as an attribute and respondents are asked to indicate which one attribute was most important when choosing between scenarios. This suggests that when respondents are asked to rank just the EQ-5D-5L dimensions there is some evidence of an ordering effect in line with the EQ-5D classification system (or the order in the ranking question). However the results differ when looking at the proportions overall, when duration is the most important attribute to the majority of the sample, and the EQ-5D dimensions are not consistently ordered. In the original EQ-5D valuation study (Dolan, 1997), the weights derived from the TTO study indicate that Pain/Discomfort was the most important dimension followed by Mobility, Self Care, Anxiety/Depression and Usual Activities. Bansback et al. (2012) randomised the ordering of EQ-5D dimensions using DCE_{TTO} and still found that the mobility coefficient was largest. However the qualitative importance of the dimensions was not assessed, and the weights for EQ-5D-5L are not yet available. Further work conducting valuation tasks testing the possible ordering effect of the dimensions may be useful, particularly as this may influence the magnitude of disutility for each level of each dimension of the EQ-5D-5L tariff that will be developed in the near future. Furthermore, we have found that some respondents cannot tell the difference between certain response levels, in particular 'severe' and 'extreme'. These results, which have also been found elsewhere (Luo et al., 2011) may have implications for the sensitivity of the five level descriptive system and impact on the forthcoming EQ-5D-5L valuation study.

When considering the overall importance of all attributes included in the task, duration, either in full health or in the selected health state, is the most important attribute, and the majority of the sample indicate that they choose the longer duration in order to spend more time with others. This suggests that a range of duration values should be administered in binary choice health state valuation studies to test the importance of duration on responses. This can be done both quantitatively to assess the impact on utility values of varying duration, and qualitatively to investigate in detail why duration is the key attribute for respondents. Using a restricted set of durations means that the task use is deviating further from an iterative TTO. By having a richer set of durations we should be better able to model the group equivalent of the indifference point.

We found that a number of external factors and background characteristics may affect responses to the tasks. Marital status was an important factor, and this is in line with Dolan and Roberts (2002). At the beginning of valuation studies, respondents are not asked to consider how the state will impact on their lives beyond the health state attribute included in the scenarios under consideration. However, these results indicate that the majority of respondents do not consider the health states in isolation. Therefore certain background characteristics and external factors are influential in the health state valuation process. Theoretically it is possible to control for all background characteristics when carrying out studies, but in reality this is a complex process both in an online and face-to-face environment. Therefore at the minimum it is important to collect a range of background characteristics, and it might also be possible to ask respondents what they considered when answering, and investigate the results excluding those completing the task the 'wrong' way. Further research should continue to consider the importance of a range of external factors and how these might impact on choices made.

This study has a number of limitations. We used follow up probing questions to try to investigate reasons behind participants' responses, and although they were designed using a pilot study, it is possible that important factors about the questions or response behaviour were not captured. We could also not test in detail the reasons behind certain responses, for example why duration was consistently considered the most important attribute, as we did not have this capacity during the interview. To improve this aspect, further research should carry out a think aloud or cognitive interview study with respondents completing both iterative TTO and LT-TTO, and DCE_{TTO} building on the earlier think aloud work conducted by Ryan et al., (2000). This will be carried out by the project team in the near future. Furthermore, we only included a small number of health states and associated durations both in the health state and full health, which were designed to be a difficult choice between the two options. For DCE_{TTO} we attempted to also make the choice between the full health state profiles reasonably difficult. The frequencies of respondents choosing A and B reported in Table 1 indicate that we may not have made all of the choices as difficult as required, and this in turn may affect the answers to the follow up questions. A wider range of both EQ-5D health states, durations, and states produced by other generic and condition specific descriptive systems would test completion issues related to health state valuation tasks further.

In summary, there is a growing interest in the use of binary choice questions to conduct health state valuation exercises. However, little is understood about how respondents perceive the task and complete the exercise. We have tested three types of binary choice questions (TTO, LT-TTO and DCE_{TTO}) and found that the binary choice conceptualisation of both TTO tasks (i.e. those with less attributes that vary between tasks, and that only present time in full health as scenario B) may be more acceptable to respondents than DCE_{TTO} questions. We have also found that the incomplete conceptualisation of scenarios is a concern. There is also some evidence that certain attributes are more important than others which may be linked to an ordering effect. Furthermore a range of external factors may impact on responses. These results may inform the design of binary choice question valuation studies, and the next stage of this work is to carry out detailed interviews testing the completion of both iterative (TTO and LT-TTO) and binary choice valuation tasks and to develop

the methodology of designing and analysing a full valuation study for binary choice TTO and LT-TTO to produce utility weights.

References:

- Bansback N, Brazier J, Tsuchiya A, Anis A (2012), Using a discrete choice experiment to estimate societal health state utility values, *Journal of Health Economics*, forthcoming.
- Bosch JL, Kammit JK, Weinstein MC, Hunink MGM (1998), Estimating general-population utilities using one binary-gamble question per respondent, *Medical Decision Making*, 18:381-390
- Brooks R (1996). EuroQol: The current state of play. *Health Policy*, 37:53-72
- Devlin N, Tsuchiya A, Buckingham K, Tilling C (2011), A Uniform Time Trade Off Method for States Better and Worse than Dead: Feasibility Study of the 'Lead Time' Approach, *Health Economics*, 20(3): 348-361.
- Devlin N, Buckingham K, Shah K, Tsuchiya A, Tilling C, Wilkinson G, van Hout B (2010), *A comparison of alternative variants of the lead and lag time TTO*, Office of Health Economics Working Paper.
- Dolan P (1997) Modelling valuations for EuroQol health states. *Medical Care*, 35(11): 1095-1108
- Dolan P, Roberts J. (2002). To what extent can we explain time trade-off values from other information about respondents. *Social Science and Medicine*, 54: 919-929.
- De Bekker-Grob EW, Ryan M, Gerard K (2010). Discrete choice experiments in health economics: A review of the literature. *Health Economics*, Early View.
- Gudex C ed (1994) *Time Trade-Off User Manual: Props and Self-Completion Methods*: Centre for Health Economics, University of York
- Hakim Z, Pathak DS. (1999). Modelling the EuroQoL data: A comparison of discrete choice conjoint and conditional preference modelling. *Health Economics*, 8(2): 103-116.
- Herdman M, Gudex C, Lloyd A, Kind P, Parkin D,onsel G, Badia X (2011). Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L), *Quality of Life Research*, early view.
- Jansen SGT, Stigelbout AM et al. (2000). Unstable preferences: a shift in valuation or an effect of the elicitation procedure. *Medical Decision Making*, 20: 62-71.
- Luo N, Li M, Chevalier J, Lloyd A, Herdman M. (2011). A cross-cultural study of the scale labels in the EQ-5D-3l and EQ-5D-5L descriptive systems. *ISOQOL 2011, Denver*.
- Mulhern B, Tsuchiya A, Rowen D, Bansback N, Devlin N, Brazier J. (2011). Health state valuation of completion of self report health questions: Does mode of administration matter? *Health Economics Study Group Meeting; Bangor*.
- National Institute for Health and Clinical Excellence (2008). *Guide to the Methods of Technology Appraisal*. UK: NICE.
- Ratcliffe J, Couzner L, Flynn T, Sawyer M, Stevens K, Brazier J, Burgess L. (2011). Valuing Child Health utility 9D Health States with a Young Adolescent Sample: A feasibility study to compare best-worst scaling discrete choice experiment, standard gamble and time trade-off methods. *Applied Health Economics and Health Policy*, 9(1): 15-27.
- Robinson A, Dolan P, Williams A (1997). Valuing health states using VAS and TTO: what lies behind the numbers?, *Social Science and Medicine*, 45: 1289-1297.
- Robinson A, Spencer A (2006) Exploring challenges to TTO utilities: valuing states worse than dead. *Health Economics*, 15: 393-402.
- Rowen D, Mulhern B et al. (2011). Estimating preference based single index measures for dementia using DEMQOL and DEMQOL-Proxy. *Value in Health*, early view

- Ryan M, Farrar S (2000). Eliciting preference for healthcare using conjoint analysis. *British Medical Journal*, 320: 1530–3.
- Ryan M, Gerard K. (2003). Using discrete choice experiments to value health care: current practice and future prospects. *Applied Health Economics and Health Policy*, 2: 55-64.
- Ryan M, Watson V, Entwistle V. (2009). Rationalising the 'irrational': A think aloud study of discrete choice experiment responses. *Health Economics*, 18: 321-336.
- San Miguel F, Ryan M, Amaya-Amaya M. (2005). 'Irrational' stated preferences: a quantitative and qualitative investigation. *Health Economics*, 14: 307-322.
- Torrance GW, Thomas W, Sackett D. (1972). A utility maximization model for evaluation of health care programmes. *Health Services Research*, 7, 118-133.
- Tsuchiya A, Mulhern B (2011), Preparation for the Re-valuation of the EQ-5D Tariff (PRET) project: Overview of methods for project stages 1-3. *HEDS Discussion Paper*, 11/16
- Tsuchiya A, Mulhern B, Buckingham K, Devlin N. (2011). Testing Lead Time issues in an online environment. *EuroQol Plenary*, Oxford.
- Wittenburg E, Prosser LA. (2011). Ordering errors, objections and invariance in utility survey responses: A framework for understanding who, why and what to do. *Applied Health Economics and Health policy*, 9(4): 225-241.
- Yang Y, Brazier JE, Tsuchiya A, Young T. (2011). Estimating a preference based index for a 5-dimensional health state classification for asthma derived from the asthma quality of life questionnaire. *Medical Decision Making*, 31, 281-291.

Analysis:

Table 1: Health state valuation question combinations and responses to examples

Question Type	Scenario A			% choosing A	Scenario B		% choosing B	
	EQ-5D-5L Health state	LT in FH	Duration		Health state	Duration		
Type I (TTO)	Example 1	12332	n/a	10 years	72.8	FH	7.25 years	27.2
	Example 2	34243	n/a	5 years	63.4	FH	2.5 years	36.6
	Example 3	43554	n/a	1 year	69.8	FH	10 weeks	30.2
Type II (LT-TTO)	Example 1	12332	10 yrs	10 years	44.6	FH	17.25 years	55.4
	Example 2	34243	10 yrs	5 years	25.5	FH	12 years	74.5
	Example 3	43554	2 yrs	1 year	35.3	FH	1.5 years	64.7
Type III (DCE_{TTO})	Example 1	12332	n/a	10 years	50.5	21323	10 years	49.5
	Example 2	34243	n/a	5 years	72.3	43344	10 years	27.7
	Example 3	43554	n/a	1 year	73.8	55355	5 years	26.2

LT: Lead Time; FH: Full health

Table 2: Sample demographics and interview setting

Demographic	N (%)
Version number	
1 (Type III/Type II)	53 (17.3)
2 (Type II/Type III)	51 (16.7)
3 (Type III/Type I)	50 (16.3)
4 (Type I/Type III)	52 (17.0)
5 (Type II/Type I)	50 (16.3)
6 (Type I/Type II)	50 (16.3)
Male	152 (49.7)
Age (m(sd))	46.46 (17.88)
Age range	
18-24	47 (15.4)
25-34	50 (16.3)
35-44	56 (18.3)
45-54	54 (17.6)
55-64	42 (13.7)
65+	57 (18.6)
Marital status	
Married/partner	193 (63.1)
Other	113 (36.9)
Employment status	
Employed or self employed	168 (54.9)
Student	8 (2.6)
Not working	130 (42.5)
Children aged under 18?	116 (37.9)
Dependents aged 18+?	18 (5.9)
Education	
Beyond minimum age	159 (52.0)
Degree level	66 (21.6)
Self reported health	
EQ-5D	
Index score (m(sd))	0.821 (0.29)
In best health state (11111)	145 (47.4)
Health status	
Good health	268 (87.6)
Poor health	38 (12.4)
Satisfied with health	
Yes (6-10)	254 (83.0)
No (0-5)	52 (17.0)
Satisfied with life	
Yes (6-10)	265 (86.6)
No (0-5)	41 (13.4)
Interviewer information	
Understanding of task	
Good	241 (79.3)
Moderate	61 (20.1)
Completion of task	
Concentrated very hard	232 (76.3)
Concentrated fairly hard	72 (23.7)
Interview environment	
Quiet with no distraction	244 (80.3)
Some background distraction	47 (15.5)
Disruptions and interruptions	13 (4.3)

Table 3: Which question type is easiest (n(%))?

		Question			
		Type I	Type II	Type III	Both the same (overall)
N completing question		202	204	206	306
<i>Question easiest</i>					
Overall		51 (25.2)	67 (32.8)	28 (13.8)	160 (52.2)
When question appears first		31 (60.8)	35 (52.2)	16 (57.1)	n/a
When question appears second		20 (39.2)	32 (48.8)	12 (42.9)	n/a
<i>Compared to other question</i>					
Easier than type I		n/a	18 (18)	18 (17.6)	n/a
Easier than type II		14 (14)	n/a	10 (9.6)	n/a
Easier than type III		37 (36.2)	49 (47.1)	n/a	n/a

Table 4: Relative importance of all question attributes by type

Dimension	Type I		Type II		Type III	
	N (%)	Rank	N (%)	Rank	N (%)	Rank
Duration	44 (21.8)	2	41 (26.6)	2	77 (37.4)	1
Duration in full health	59 (29.2)	1	44 (28.6)	1	N/A	N/A
Mobility	28 (13.9)	3	15 (9.7)	3	31 (15.0)	2
Self care	28 (13.9)	3	9 (5.8)	6	26 (12.6)	4
Usual activities	16 (7.9)	6	14 (9.1)	4	17 (8.3)	6
Pain/discomfort	20 (9.9)	5	21 (6.9)	5	24 (11.7)	5
Anxiety/depression	5 (2.5)	7	8 (2.6)	7	31 (15.0)	2

Table 5: External factors and impact of background characteristics on response

	Response (n,%)				
	Always	Often	Sometimes	Rarely	Never
Impact of feelings about health and life	134 (43.8)	106 (34.6)	54 (17.6)	4 (1.3)	8 (2.6)
Impact of health state on life and financial situation	72 (23.5)	75 (24.5)	81 (26.5)	27 (8.8)	51 (16.7)
Choose longer duration in order to spend more time with others	59 (19.3)	54 (17.6)	71 (23.2)	34 (11.1)	88 (28.8)

Figure 1:

Question type I: Binary choice TTO

	<i>HEALTH SCENARIO A</i>	<i>HEALTH SCENARIO B</i>
	You live for 5 years with the following: Slight problems in walking about No problems washing or dressing yourself Moderate problems doing your usual activities Slight pain or discomfort Not anxious or depressed	You live in full health for 4 years
	----- Then you die	----- Then you die
<i>Which scenario do you think is better?</i>	<input type="checkbox"/>	<input type="checkbox"/>

Question type II: Binary choice LT-TTO

	<i>HEALTH SCENARIO A</i>	<i>HEALTH SCENARIO B</i>
	You live in full health for 5 years Then you live for 5 years with the following: Slight problems in walking about No problems washing or dressing yourself Moderate problems doing your usual activities Slight pain or discomfort Not anxious or depressed	You live in full health for 9 years
	----- Then you die	----- Then you die
<i>Which scenario do you think is better?</i>	<input type="checkbox"/>	<input type="checkbox"/>

Question type III: Binary choice DCE_{TTO}

	<i>HEALTH SCENARIO A</i>	<i>HEALTH SCENARIO B</i>
	You live for 5 years with the following: Slight problems in walking about No problems washing or dressing yourself Moderate problems doing your usual activities Slight pain or discomfort Not anxious or depressed	You live for 5 years with the following: No problems in walking about Slight problems washing or dressing yourself Slight problems doing your usual activities Moderate pain or discomfort Not anxious or depressed
	----- Then you die	----- Then you die
<i>Which scenario do you think is better?</i>	<input type="checkbox"/>	<input type="checkbox"/>

Figure 2: Proportions of the sample reporting difficulty completing the task

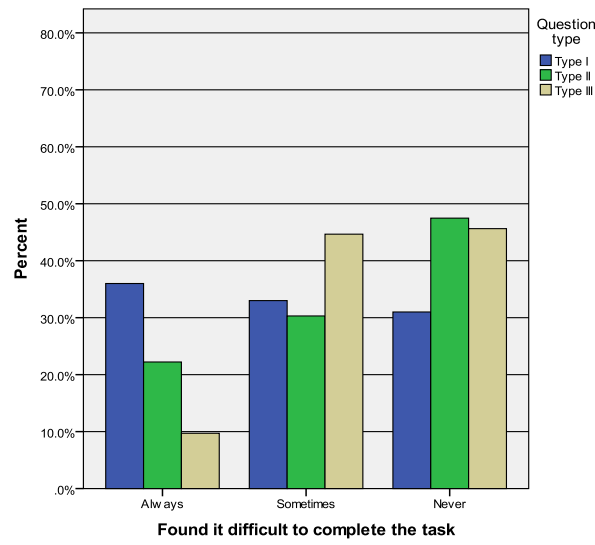


Figure 3: Attendance to attributes across the question types

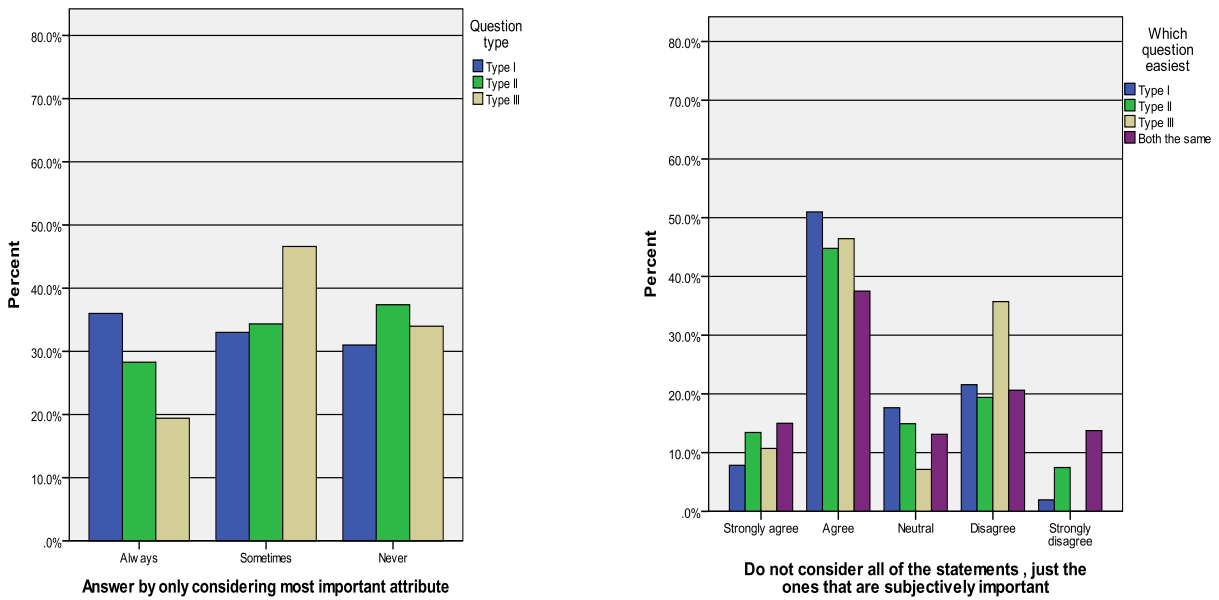


Figure 4: Importance of EQ-5D-5L dimensions overall (all question types pooled)

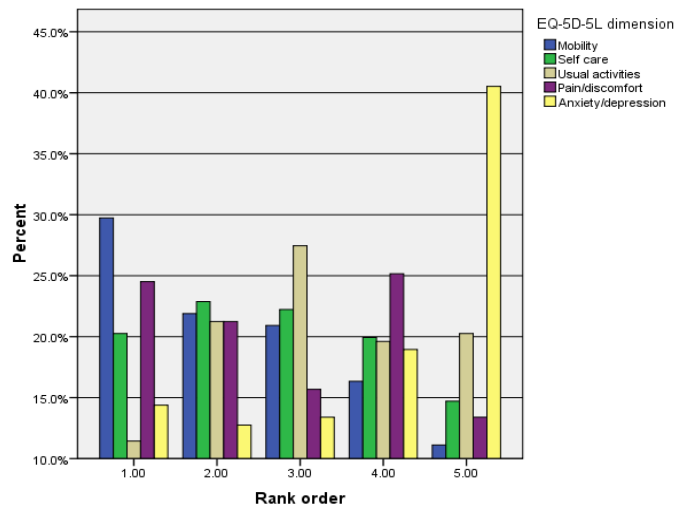


Figure 5: Importance of EQ-5D dimensions by question type

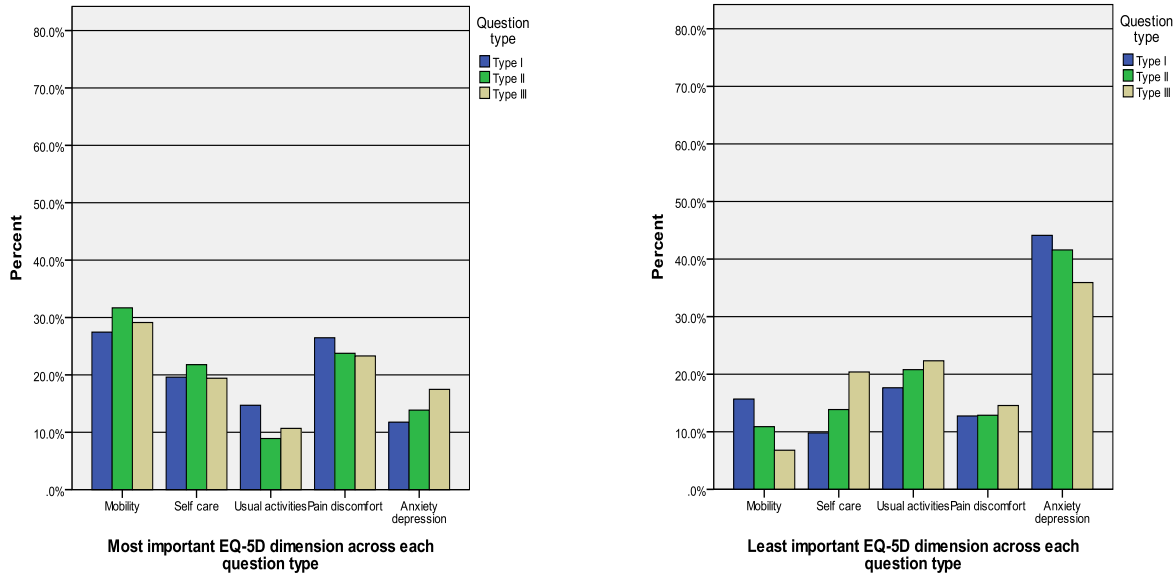


Figure 6: Proportions of sample that could distinguish between EQ-5D-5L response levels

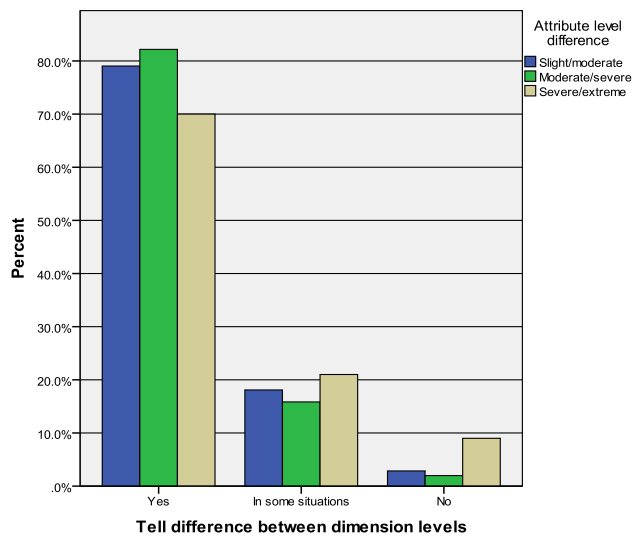


Figure 7: Influence of external factors on response

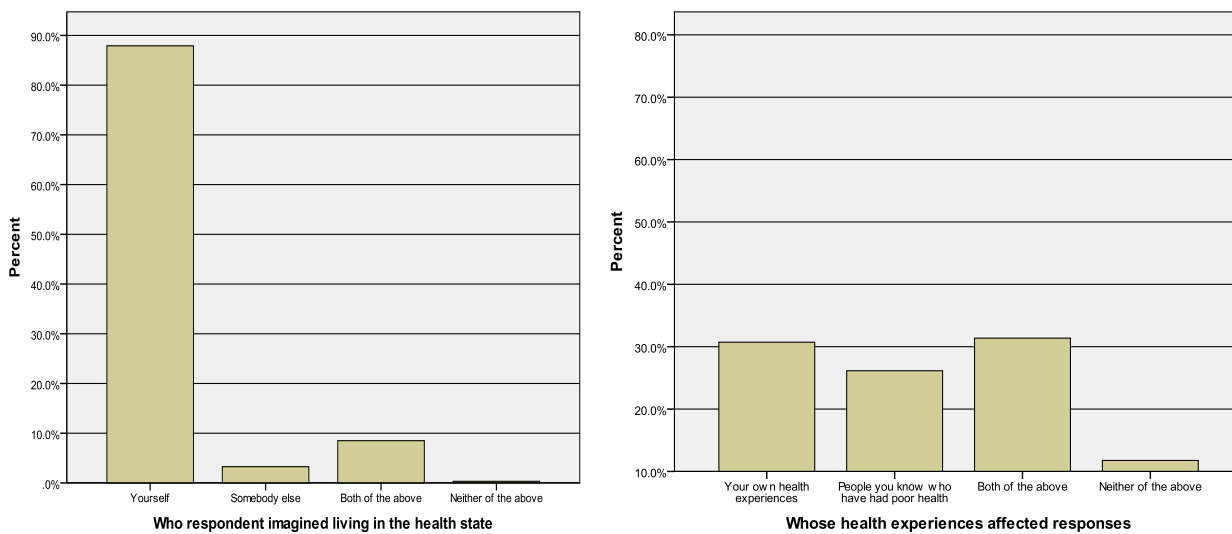


Figure 8: External factor – influence on others

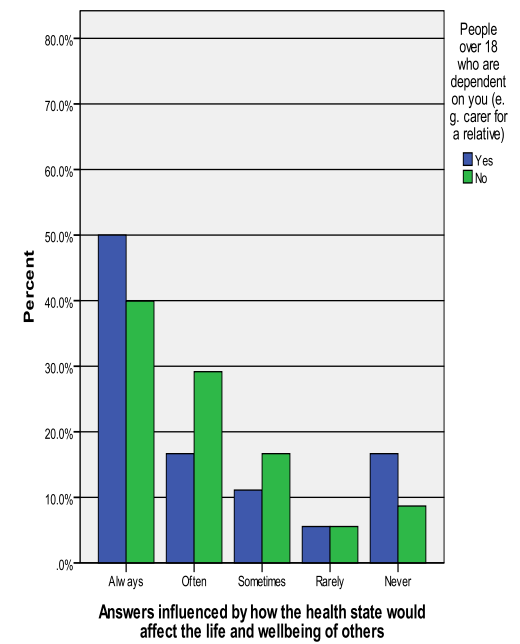
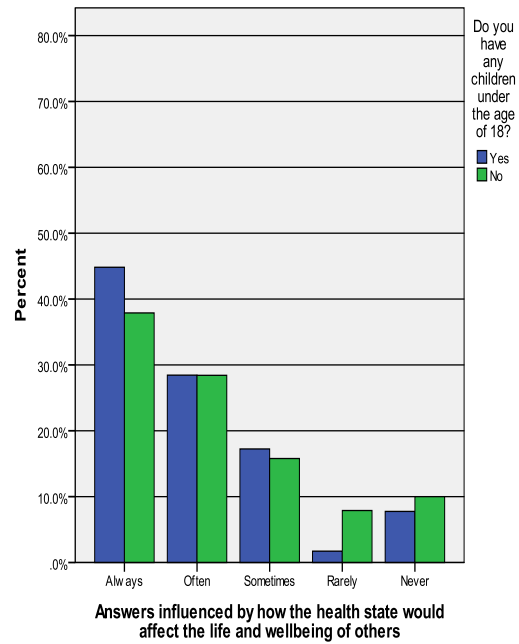
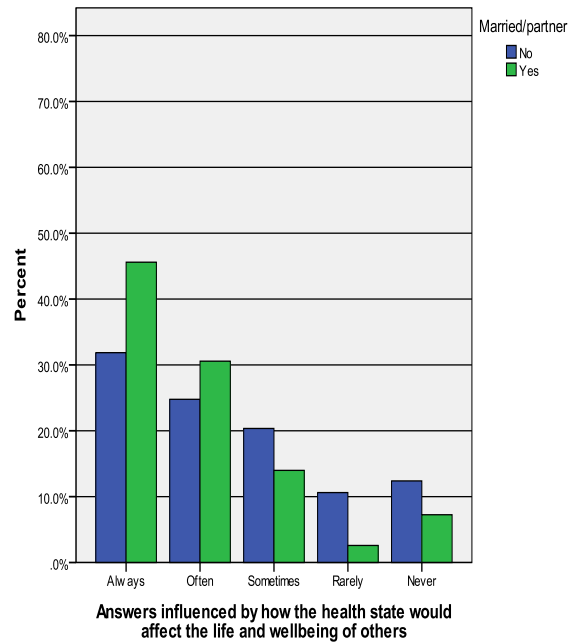
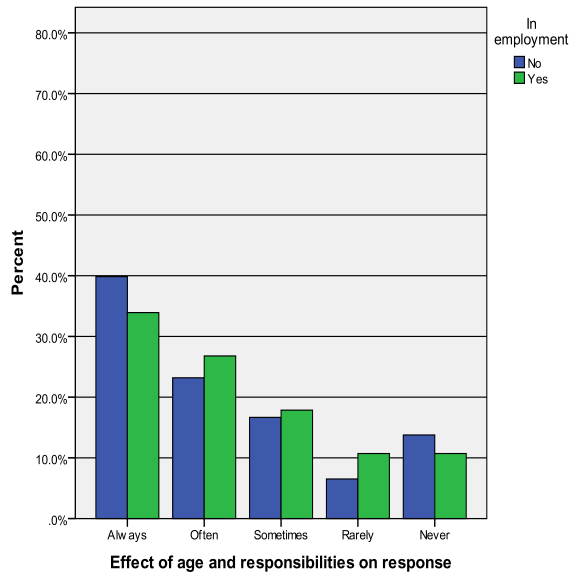
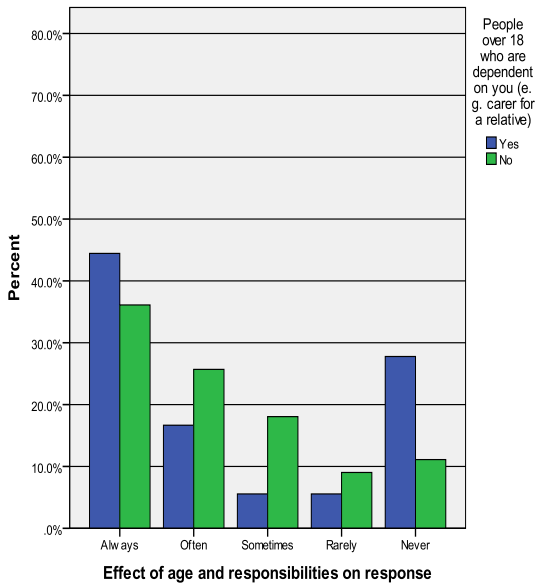
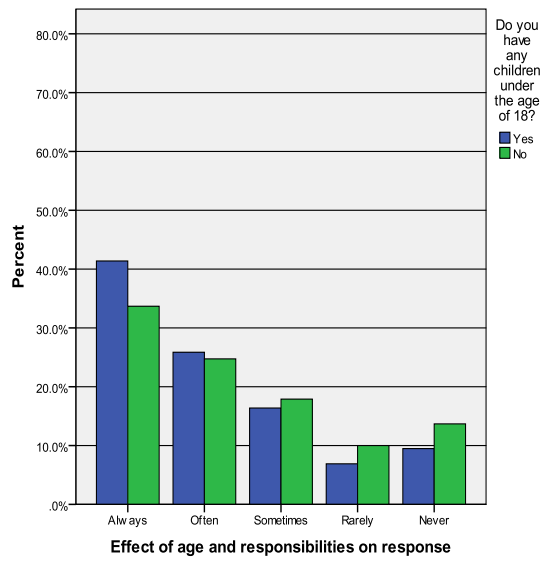
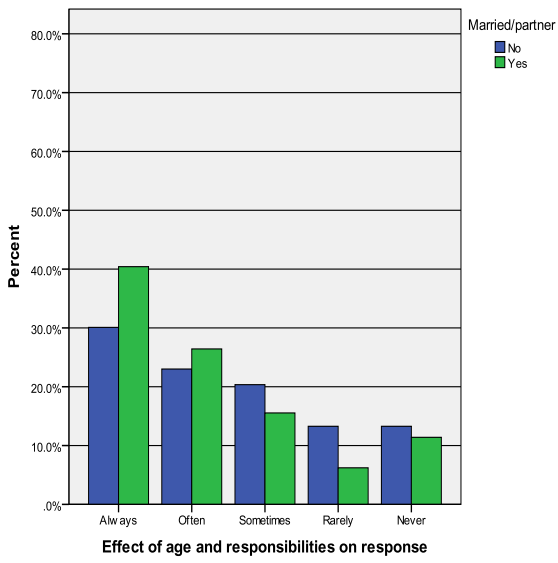


Figure 9: External factor - age and level of responsibilities to others



Appendix 1: Follow up questions used

Probing questions

1. When answering the questions, which part of the health state was most important to you?
 - Length of time spent in the health state
 - Problems walking about
 - Problems washing or dressing yourself
 - Problems with usual activities
 - Level of pain or discomfort
 - Level of anxiety or depression

2.

	Yes	Sometimes	No
There is too much information included in these scenarios so I just look at the bit that is most important to me			
I found it difficult to answer these questions			
When answering these questions, I chose the scenario with the fewest number of severe health areas			
Unless the state is severe, the number of years that you live for is the most important part of the scenario			
It is not clear what full health means			
It is difficult to imagine changing from full health to a poor health state so suddenly			
It is not realistic that time in full health is always shorter			

3. Please rank, from 1-5, which areas of health were most important when answering the questions?
 - Problems walking about
 - Problems washing or dressing yourself
 - Problems with usual activities
 - Level of pain
 - Level of anxiety/depression
4. When answering the questions, who did you imagine living in the health state?
 - Yourself
 - Somebody else
 - Both of the above
 - Neither of the above
5. Whose health experiences had an effect on your responses to the questions?
 - My own health experiences
 - People I know who have had poor health
 - Both of the above
 - Neither of the above
- 6.

Which set of questions...	First set	Second set	Both the same
Did you find the easiest			
Made you think the most about the effect of the health scenario on the other people around me (e.g. family)?			

7.

Question	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
When answering, I do not consider all of the statements , just the ones that are important to me					
The layout of the questions means that they can be answered easily					
It is difficult to imagine what it would actually be like to live in the scenarios					
The scenarios are not realistic					
There is too much to think about to give a credible answer					

8.

Question	Yes	In some situations	No
I can tell the difference between slight and moderate problems for each health area			
I can tell the difference between moderate and severe problems for each health area			
I can tell the difference between severe and extreme problems for each health area			

9.

Question	Yes	In some situations	No
I can tell the difference between slight and moderate problems for each health area (VERSIONS 1 AND 4 ONLY)			
I can tell the difference between moderate and severe problems for each health area (VERSIONS 2 AND 5 ONLY)			
I can tell the difference between severe and extreme problems for each health area (VERSIONS 3 AND 6 ONLY)			

10.

Question	Yes	No
It is hard to believe that I would be left without relief or treatment by doctors and other health professionals		
It is possible that my answers would change if I was asked the same questions in a week's time		

11.

Question	Always	Often	Sometimes	Rarely	Never
My answers were influenced by how the health state would affect the life and wellbeing of those around me (for example my children, parents or partner)					
My age and my responsibilities to others had an effect on how I answered the questions					
How severe the health scenario is does not matter, I would choose to live in the scenario with a longer duration to spend time with the people close to me					
The impact that living in the each health state would have on my life and my financial situation was an important consideration					
How I would feel about my health and life when living in the scenarios is an important consideration					