How should DCE with duration choice sets be presented for the valuation of health states?

Brendan Mulhern[1] (MRes), Richard Norman[2] (PhD), Koonal Shah[3] (PhD), Nick Bansback[4] (PhD), Louise Longworth[5] (PhD), Rosalie Viney[1] (PhD)

1 Centre for Health Economics Research and Evaluation, University of Technology Sydney, Sydney, Australia

2 Curtin University, Perth, Australia

3 Office of Health Economics, London, England

4 University of British Columbia, Vancouver, Canada

5 PHMR, London, England

Corresponding author:

Brendan Mulhern, Centre for Health Economics Research and Evaluation, University of Technology Sydney, 1 - 59 Quay St, Haymarket, NSW 2000.

E-mail: Brendan.mulhern@chere.uts.edu.au

Running head: Testing DCE presentation approaches

Earlier versions of this paper were presented at the EuroQol Group Plenary, Sept 2016, Berlin, and the International Academy of Health Preference Research, Sept 2016, Singapore

Word count: 4,686

Financial support for this study was provided by grants from the EuroQol Research Foundation and the Australian National Health and Medical Research Council (1065395). The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report. All of the authors are members of the EuroQol Research Foundation.

Abstract

Background: Discrete Choice Experiments including duration (DCE_{TTO}) can be used to generate utility values for health states from measures such as EQ-5D-5L. However methodological issues concerning the optimum way to present choice sets remain. The aim is to test a range of task presentation approaches designed to support the DCE_{TTO} completion process.

Methods: Four separate presentation approaches were developed to examine different task features including dimension level highlighting, and health state severity and duration level presentation. Choice sets included two EQ-5D-5L states paired with one of four duration levels, and a third 'immediate death' option. The same design including 120 choice sets (developed using optimal methods) was employed across all approaches. The online survey was administered to a sample of the Australian population who completed 20 choice sets across two approaches. Conditional logit regression was used to assess model consistency, and scale parameter testing investigated poolability.

Results: Overall 1,565 respondents completed the survey. Three approaches using different dimension level highlighting techniques produced mainly monotonic coefficients that resulted in a larger disutility as the severity level increased (excepting usual activities levels 2/3). The fourth approach using a level indicator to present the severity levels has slightly more non-monotonicity and produced larger ordered differences for the more severe dimension levels. Scale parameter testing suggested that the data cannot be pooled.

Conclusions: The results provide information regarding how to present DCE tasks for the purpose of health state valuation. The findings improve our understanding of the impact of different presentation approaches on valuation, and how DCE questions could be presented to be amenable to completion. However it is unclear if the task presentation impacts online respondent engagement.

Introduction

Health state values anchored on a 0 (dead) to 1 (full health) scale are used in the estimation of quality adjusted life years (QALYs), a key metric in the economic evaluation of health care interventions. These values can be derived from preference based measures such as the EQ-5D[1,2] and the SF-6D.[3,4] Discrete Choice Experiments (DCE) are widely used in health economics[5,6] and DCEs incorporating an attribute for duration (DCE_{TTO}) have been used to estimate health state values for generic and condition specific classification systems including the EQ-5D-3L,[7,8] EQ-5D-5L,[9-11] SF-6D[12] and the EORTC QLU C-10D.[13] The method is amenable to completion by respondents, and produces models that are generally logically ordered within dimensions. Recent methodological work has investigated different approaches to modelling the data, demonstrating that this has an impact on the position of health states relative to dead (which has a value of zero by definition) within the overall descriptive system, and the overall range of values produced.[14] A range of different design strategies have been tested, suggesting that Bayesian designs using uninformative (zero) priors may produce less difficult sets of DCE tasks, with generally larger differences between attribute levels, which may lead to more logically ordered dimension level models.[15]

An important methodological issue relating to DCE for health state valuation that has not been fully explored in the literature relates to the impact of the difficulty of the choice task on responses and the models produced. One way this can be mitigated is through strategies that facilitate comprehension, particularly through presentation of the choices. If respondents find the tasks difficult, or the presentation formats are not amenable to easy comprehension or assessment of alternatives, this may impact on the way in which respondents complete the tasks. This may just increase variability in responses, but there is also potential that it may result in systematic bias in responses (for example, respondents ignoring attributes to make the choice task manageable). This would lead to issues with the validity of the modelled values that might not be based on the full assessment of the choice sets as is assumed in the design of the study.

The optimal way to present the tasks for health state valuation, and the features of the task presentation that may support completion, has not been widely studied. Norman et al[16] presented tasks to value the EORTC QLU-C10D using an approach that highlighted the dimensions that differ within a choice set, and a 'text and table' format that included the dimensions that differ in the choice set table display, and those that did not in a separate

text section. The results suggested that the highlighting approach was more acceptable to respondents and yielded more logically ordered results. While these results suggest that highlighting may be a promising strategy, there are many other methods of presenting choice sets, including different highlighting approaches, that can be systematically tested to develop a method that will encourage full attention on the dimensions and valid completion of the tasks. There is also the possibility of using presentation to draw attention to the relative severity of levels as a means to avoid the 'preference reversals' that can occur with adjacent levels.[11] For example, Cole et al[17] presented health states in the context of the overall descriptive system and found that the number of logically inconsistent responses was reduced when respondents were given visual assistance. There are examples of similar work in contexts other than health state valuation. For example, Veldwijk et al[18] found that, when choosing between options for vaccination against rotavirus, options presented as words were preferred by survey respondents to graphical representations, and also resulted in more valid attribute estimates. Understanding of the attribute levels did not appear to differ under either approach. Similarly, Kenny et al[19] found in a study of general practitioner choice, that neither preferences nor the level of understanding of the attributes differed significantly across different presentation formats for frequency and quality rating attributes.

The aim of this study is to develop and test a range of task presentation approaches designed to support the DCE_{TTO} completion process, and collect primary data to compare each approach. This study uses the EQ-5D-5L health state classification system as an example; however the methods could be applied to any preference based measure, and may have applications for DCE more generally. The various approaches will be compared in terms of both the models produced, and respondents' views about each.

Methods:

The EQ-5D-5L

In our DCE choice sets, the EQ-5D-5L classification system[2] is used to describe health. The EQ-5D-5L describes health related quality of life across five dimensions (mobility, self-care, usual activities, pain/discomfort and anxiety/depression) with five response levels (no problems, slight problems, moderate problems, severe problems and extreme problems/unable to). The use of DCE to elicit preferences for the EQ-5D-5L is widespread internationally, and it has been used to develop an value set based on the preferences of the

Australian population in a pilot study carried out by Norman et al.[11] It is also included in the recommended protocol for the valuation of the EQ-5D-5L internationally.[20]

Task presentation approaches

Four separate approaches to presenting the DCE task were developed to examine different features of a DCE_{TTO} task that in turn had been developed and used in previous research.[8,11-14] The DCE_{TTO} task format presents triplets including two health states with an associated duration and a third 'immediate death' option. The task features assessed included the use of dimension level highlighting and different methods to present the severity level of the health state and length of time spent in the health state. This aimed to support respondents' completion of the questions whilst including the full EQ-5D-5L descriptions. Including full EQ-5D-5L descriptions was important to maintain consistency with other DCE valuation work for EQ-5D-5L. [20] In addition, there is evidence suggesting that respondents completing DCE prefer words over graphics.[18] The different approaches were also set up to allow comparison across highlighting methods. Each approach presented choice sets including two EQ-5D-5L health states paired with one of four duration levels (2 years, 4 years, 8 years and 16 years). The third option 'immediate death' was included for all choice sets, and respondents chose the best and worst (thus giving a complete ranking across the three options). The same DCE design was employed across all four approaches. The presentation approaches are presented in Figure 1and are described below:

Approach A: An 'Alternate highlight' approach in which alternating grey and white shading is used to differentiate dimensions. This acted as the control arm. The rationale for highlighting in this way was to make it easier for respondents to read and focus on each dimension.

Approach B: A 'Yellow highlight' approach, where only the dimensions *differing* between options 1 and 2 were highlighted in yellow to allow easier comparisons between the options. The rationale is to encourage focus on the dimensions that differed for each choice set. This approach has been used in past DCE health state valuation work[16] and also in a DCE study investigating preferences for interventions prioritizing prevention or cure.[21]

Approach C: A 'Grey highlight' approach, where dimensions that were the *same* across options 1 and 2 were greyed out to allow easier identification of the dimensions that

differed.

Approach D: A 'Level highlight' approach, where a table displaying each of the attribute levels was included as a visual aid. The rationale for this approach was to test a more visual way of representing the similarities and differences of the levels across the dimensions. This approach represents a more visual way of presenting the levels as the severity indicators are shown in increasing severity order for each option, with the relevant level highlighted. Therefore the difference in severity between dimensions within the overall system can be seen. This is in contrast to approaches A, B and C, where just text is used.

Study design

The presentation approaches were tested online with the intention to recruit 1,500 members of the Australian general population, recruited from an online panel. This provided approximately 30,000 observations (250 observations per choice set) overall, leading to around 62 observations per choice set for each presentation approach. This was in line with other EQ-5D DCE valuation studies,[22-24] and was sufficient to allow for investigation of scale variability across different study arms.[25] Each respondent answered ten choice sets randomly selected from the overall design (without replacement) using one of the four approaches, then another ten choice sets again randomly selected from the overall design using one of the remaining three approaches. Thus, there were 12 survey versions with each approach appearing first and second (see Table 1 for more detail). This allowed for head to head comparisons between approaches, with each approach matched with each of the others, and appearing both first and second to ensure that there is no systematic bias linked to potential order effects relating to the influence of one type of task on another. The design included 120 choice sets that were selected using d-optimal methods within the DCE design software NGene.[26] No level of overlap (i.e. dimensions held at the same severity level within choice sets) was imposed on the design. DCE_{TTO} estimates coefficients for interactions between categorical health dimension levels and continuous duration, and therefore the number of parameters to estimate was 21 (interactions between EQ-5D-5L levels and continuous duration ((5-1) x 5 x 1=20), plus continuous duration (1)). Past work using DCE_{TTO} has used study designs with more choice sets than there are parameters to estimate[9] and this criterion was followed here.

We included task specific and comparative follow up questions about the approaches. The task specific questions included asking respondents about difficulty with completing the tasks. This included the difficulty of identifying the differences between states, the respondent's ability to imagine the health states, and whether they found the format helpful and considered the whole health state when completing the tasks. Comparative questions included rating the approaches on a 0 to 10 scale, asking which approach was easier, and asking whether the approach influenced their completion of the task.

Respondent recruitment and survey completion

Respondents who had previously opted in to an online panel (i-view), were recruited, and clicked a link to access the survey. They then answered screening questions to allocate them to a quota group based on age and gender, and were randomly allocated to one of the twelve survey versions. Following consent and further instructions, ten choice sets using one of the Approaches were completed followed by feedback questions specific to that Approach. A second set of ten choice sets from a different Approach was then completed followed by the same feedback questions for the second Approach. Further follow up questions comparing the Approaches were then completed, followed by detailed demographic information and a final free text question allowing respondents to comment generally about the survey. Respondents received a small incentive for completing the survey in the form of points that can be redeemed for a range of rewards. Data collection took place during March and April 2016.

Data analysis

The DCE_{TTO} data were analysed using conditional logit regression taking into account repeated observations within respondents, which is described in detail elsewhere.[7,11] We focused on homogeneous models as these have been widely used in the generation of utility value sets from DCE_{TTO} data[7-13,15,16], and therefore are the primary comparison point with past research. We received a full ranking for the three options included in each choice set, but in the analysis reported here we excluded the data for option C (immediate death) and focused on the choice between health states with duration (Options A and B) presented. Therefore if a respondent chooses 'immediate death' as the best and option A [B] as the worst, then scenario B [A] is modelled as the preferred. This has been done in previous work by the authors.[8,11,25] For each approach we estimated coefficients for an interaction between each EQ-5D-5L attribute x, where x includes four levels reflecting movements from full health in each dimension and life years t:

$$u_{ij} = \beta t_{ij} + \boldsymbol{\lambda}' \boldsymbol{x}_{ij} t_{ij} + \varepsilon_{ij} \quad (1)$$

This estimates the unanchored model (i.e. not anchored on the utility scale) where the coefficient β is the value of living in full health for one year and λ′ represents the disutility of living with the specified set of EQ-5D-5L health problems (x) for one year. The error term εij is a random term, and assumed to be independent and identically distributed extreme value type I.

To estimate the model anchored on the full health – dead utility scale (V), the estimate for the interactions of each EQ-5D-5L dimension level and duration were divided by the estimate of the coefficient for duration:

$$V_j = 1 + \frac{\hat{\boldsymbol{\lambda}}'}{\hat{\beta}} \boldsymbol{x}_j \quad (2)$$

To calculate a health state value, each level of each dimension was then calculated as a movement away from full health (anchored at 1). Negative values (i.e. states modelled as worse than dead) were possible.

Comparison of approaches

For each approach we compared the ordering and magnitude of the unanchored models, the characteristics of the anchored coefficients and the utility scales produced. We also modelled the data based on whether the approach was presented first or second. Across all versions the first set of ten pairs and the second set of ten pairs were modelled separately (for example Approach 1 first compared to Approach 1 second) to assess whether the characteristics of the models differed.

Poolability testing

We tested the null hypothesis that values do not differ across the four approaches using a Swait and Louviere[27] test. This examines the variability between the systematic and random components of a DCE design. If the null hypothesis is accepted the data can be pooled for analysis. Firstly, a grid search approach (following the method outlined by Viney et al[28]) was used to identify the relative scale parameters of each of the experimental approaches (B, C and D) in comparison to the control approach (A) which was anchored at one. The aim was to identify the maximum log likelihood for the experimental arms. A restricted pooled model scaled using the parameters identified (Lμ) was then estimated along with an unrestricted pooled (Lp) and individual models for each arm (La, Lb, Lc, Ld) using conditional

logit regression as outlined in the data analysis section. To assess differences, the likelihood ratio (LR) test outlined in equation 3 was initially used to compare the models. If the LR statistic is above a critical value (calculated as the difference between the number of parameters in the unrestricted (84) and restricted (25) model, in this case 59 degrees of freedom) then the null hypothesis can be rejected. If the null hypothesis cannot be rejected then a second more restrictive test (equation 4) is carried out. If this test also rejects the null then the data can be pooled and models can be estimated with no scale parameter adjustment.

$$LR = -2[L\mu - (La + Lb + Lc + Ld)] \quad (3)$$

$$LR = -2(L\mu - Lp) \quad (4)$$

Follow up questions

Responses to the follow up questions used to assess and compare each approach separately were analysed descriptively. The ratings provided for each approach were compared using one-way ANOVA difference tests.

Results

Sample and survey completion

In total 2,226 people responded to the invitation to participate, with 1,565 (70.3%) of these fully completing the survey. Of those not completing, 71 (3.2%) belonged to a quota that was full so were immediately excluded, 122 (5.5%) exited at the information page, 217 (9.7%) at the DCE description and introduction page, 213 (9.6%) during the DCE tasks and 38 (1.7%) later in the survey. The dropout rate amongst the 251 respondents who reached the DCE tasks ranged from 13% to 16% across the survey versions. Any DCE data collected from respondents who dropped out was not included in the modelling. The demographic characteristics of the sample are reported in Table 2. There was a higher proportion of younger female respondents in the sample than the Australian general population due to the panel provider leaving this quota group open after it had been filled. However as this is a methodological study, and we are not assessing differences in preferences based on gender or using the results to develop a representative value set, the full sample was used. Age and gender did not significantly differ across the four approaches.

In line with expectations, each approach was completed by between 780 and 787 respondents, with between 249 and 268 observations per choice set overall. This meant 62 to 67 observations per choice set for each presentation approach. The median time taken to complete each of the 12 survey versions ranged from 13.5 to 15.9 minutes. The mean time taken is 33 mins, but this is skewed by 9 respondents who each recorded a completion time of more than 1.5 hours.

Comparison of approaches – DCE models

Across the approaches, the proportions of respondents choosing Option C (immediate death) as the best option (A: 15.3%; B: 12.3%; C: 14.0%; D: 12.5%) and the worst option (A: 43.0%; B: 43.4%; C: 41.5%; D: 38.6%) across all choice sets were similar. Table 3 reports the unanchored models for each approach, including the statistical significance of each dimension level in comparison to the baseline, and also between adjacent severity levels. Approach A (control) and approaches B and C (which used different presentational techniques to highlight differences within choice sets) produced logically ordered coefficients that resulted in a larger disutility as the severity level increases across almost all levels of all dimensions. The exceptions to this were usual activities levels 2 and 3 which were disordered across all models such that level 3 produces a smaller decrement than level 2 (however only Approach C was significant). This reversal is not uncommon, and may reflect the fact that in a DCE the difference between slight and moderate may not be large when people only see one level of the dimension. Approach D (Level indicator) had two further non-statistically significant inconsistencies between levels 1 and 2 of pain/discomfort and anxiety/depression, but produced larger ordered differences for levels 4 and 5. The difference between the adjacent severe and extreme levels for anxiety/depression was only significant for approach D. The models based on whether the approach style appeared first or second, with gender groups modelled separately, and also excluding respondents who took more than 1.5 hours to complete the survey, were similar (results available from the authors on request).

Figure 2 shows the anchored models for each of the approaches. Approach D produced the largest overall decrements for all dimensions except self-care, but with inconsistencies and generally smaller decrements (apart from for mobility) at the mild end of the severity scale. This is reflected in the overall range of utility values modelled which was substantially larger for approach D (Approach A: 1 to -0.688; Approach B: 1 to -0.714; Approach C: 1 to -0.754;

Approach D: 1 to -1.089). The approaches based on highlighting all found that pain/discomfort had the largest overall decrement (a proxy for importance based on the worst severity level). Approach D found that mobility had the largest overall decrement. Usual activities had the smallest decrement across all four approaches.

Comparison of approaches – Poolability testing

With the control approach A anchored to one, the scale parameters were 1.158 (approach B), 1.065 (approach C) and 1.068 (approach D). Therefore the arm with the most variability in comparison to the control approach is approach B. The scaled restricted model estimated using the scale parameters (Lμ) and the unrestricted pooled model (Lp) are included in Table 4. The log likelihood statistics were -19127.7 (Lμ) and -19132.3 (Lp). The log likelihood statistics of each individual model (reported in Table 3) were -4828 (approach A), -4694 (approach B), -4758 (approach C) and -4801 (approach D). Therefore, the results of the first likelihood ratio test are as follows:

$$LR = -2[-19128 - (-4828 + -4694 + -4758 + -4801)] = 94$$

This is greater than the critical value and therefore the null hypothesis that the values do not differ across the approaches is rejected and the data cannot be pooled. Given that the null is rejected, the second likelihood ratio test statistic is not required.

Comparison of approaches – Follow up questions

Figure 3 and Table 5 display the responses to the feedback questions. Approximately 30% of respondents found the tasks difficult, with around 50% agreeing that the presentation format helped them to complete the exercise (and this was generally consistent across approaches). Approximately 75% reported considering the whole description whilst answering the question, and this did not differ across the presentation formats. Nearly 40% agreed that the task presentation influenced completion (with another 40% neutral about the influence).

Table 6 shows the proportions of respondents who rate each approach as the easier of the two they completed, and descriptive statistics relating to their rating of the tasks. Overall, 43.8% of those completing approach D found it easier than the other approach presented which was the highest overall; whereas only 25% reported that approach C was the easier. Across the approaches, around 40% of respondents were neutral regarding whether one approach was more difficult than the other. The rating of approaches D and B were

significantly higher than approach C ($p = 0.002$ and $p < 0.001$ respectively). Coupled with the results regarding ease of completion, approach C seems to be the least favoured by respondents. Approach D had a higher proportion of both low (0 to 4) and high (9 to 10) ratings.

Discussion

We examined four different approaches to presenting DCE tasks for use in the valuation of health state classification systems such as the EQ-5D-5L. The aim of this study was to investigate whether the task format can make it easier for the respondent to understand the task and to provide the answer that best reflects their preferences. The results suggest that there is not a large impact of highlighting dimensions that differ in each task, though sufficient differences exist suggesting the data cannot be pooled. However, respondent feedback suggests that certain methods of highlighting make the task easier to complete. For example, approach C (Grey highlight) was reported to be the most difficult and had a significantly lower rating than other approaches based on highlighting so it may be prudent to avoid using a similar presentation style in future DCE studies.

The level indicator approach (D) produces a model with characteristics that contrast interestingly with the other models. The coefficients for levels 4 and 5 of the anxiety/depression dimension are statistically significant different which is not the case with the other approaches. The level indicator approach, therefore, may help remove the commonly found preference reversal[29] between the descriptors severe and extreme, as the format helps the respondent identify which is 'worse'. This approach also has a larger overall utility scale than the others, and a different dimension with the largest overall decrement as a proxy for importance (mobility over pain/discomfort). If a more visual table based format was adopted for further use, then this could have implications for the overall utility gain, and the gain related to change on certain dimensions. There is still the preference reversal for UA across all models but it is not statistically significant for Approach D. However, for milder problems, Approach D fared less well and, unlike the other approaches, did not significantly differentiate between levels 1 and 2 on the PD and AD dimensions.

Approach D also has the highest rating overall, and is reported to be easier than the other approaches by the largest number of respondents. These results may indicate that many respondents complete online DCE questions using visual cues to help them (potentially not fully taking in all of the textual information presented to them), and therefore it is worth

considering development of more visual methods of presenting health state dimensions in more detail.

The presentation formats used in this study are largely consistent with the previous approaches to DCE_{TTO} used for the valuation of EQ-5D-5L, and this was done to test the impact of format within the constraints of choice tasks previously used for the DCEs involving the EQ-5D-5L. Options for further work could include testing the application of more diverse changes, for example around the use of graphics, or colours to represent different severity levels. This would allow an assessment of how different formats potentially encourage the use of different task completion strategies and heuristics.

This work has a number of limitations. Firstly, beyond the self-reported feedback questions and the time taken, we cannot fully assess respondent engagement and how this differs across the approaches. Time taken may be a useful proxy for engagement as those taking a very short or long time may not be fully paying attention to the survey. Those taking a long time may complete a certain amount of the survey before returning to complete it later. However excluding these respondents does not alter the models across the approaches. Further work could test how different presentation methods impact respondent attention using, for example, eye tracking, which has been used to attempt to understand how people complete DCE tasks.[30] However, sample size requirements would preclude model estimation). Secondly, each respondent saw only two approaches so it was not possible to get their feedback across all approaches. In terms of areas for future development, approach D seems promising, but will need systematic testing of the change in both how the levels are presented within the table, and how they are worded, as an important feature of approach D is that both of these aspects differ from the usual format. It would also be useful to test the further applicability of the approaches with other preference based measures that differ in structure to the EQ-5D such as the SF-6D .[3,4] One area for further research could be to ask respondents to pick their preferred task format at the beginning of the survey. However if there is bias introduced by one of the approaches, and each approach is not completed equally, then this could be problematic for any value sets produced. Thirdly, we used an opt in online panel of respondents who may not be representative of the population in unmeasurable characteristics, and are incentivized to complete surveys for a reward. Also, health literacy is important in the interpretation of the impacts of complex information such as that presented in a DCE,[19,31] and may therefore

influence choices made, but we do not know the literacy level of the sample. In recent work it was found that health literacy was associated with logical inconsistencies in health state valuations elicited using TTO[32] However it should be noted that many different panels have been used internationally for the completion of DCE health state valuation studies[7,8,15,33] and have produced logical findings overall.

It is difficult to judge the validity of each approach without a 'gold standard' value set. This issue is tackled by looking at the ordering of responses and the characteristics of the value set in comparison to an approach that mimics that used in many DCE_{TTO} studies. However it is unclear which approach best reflects respondent preferences. Also, it remains open to interpretation what an 'easy' DCE task is, and whether that is the most valid format to use. The ideal approach would not make the task too easy, but makes it easy for the respondent to understand and pay attention to what they are being asked to do. Further work could attempt to understand the extent to which a range of valuation methods such as DCE and TTO are cognitively challenging, and how the different tasks require different cognitive processes for valid completion. This would help develop tasks that can be answered validly during the study design process.

In conclusion, the results provide useful information regarding how to present DCE tasks for the purpose of health state valuation, where innovative presentation methods used for DCE outside of its application for health state valuation (for example using pictures and graphics to explain attributes) are not the standard. The findings suggest that format changes may impact EQ-5D-5L values, and the results provide a basis for further research developing and testing different aspects of task presentation in more detail.

Acknowledgements

This study was funded by the EuroQol Research Foundation and the Australian National Health and Medical Research Council. Earlier versions were presented at the EuroQol 33rd Scientific Plenary Meeting (Berlin, 2016) and the International Academy of Health Preference Research, Sept 2016, Singapore. The views expressed do not necessarily reflect the views of the EuroQol Research Foundation or the National Health and Medical Research Council. Ethics approval was obtained from the Centre for Health Economics Research and Evaluation, University of Technology Sydney Program Ethics Process. We are grateful to Liv

Ariane Augestad for her comments on an earlier draft during the EuroQol Plenary Meeting, and all the respondents who took part. The usual disclaimers apply.

References

1 Brooks R. EuroQol: The current state of play. Health Policy. 1996;37:53-72.

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

3 Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ. 2002;21:271-92.

4 Brazier JE, Roberts J. Estimating a preference-based index from the SF-12. Med Care. 2004;42(9):851-59.

5 De Bekker-Grob E, Ryan M, Gerard K. Discrete Choice Experiments in Health Economics: A review of the literature. Health Econ. 2010; 21(2):145-172.

6 Clark M, Determann D, Petrou S et al. Discrete Choice Experiments in Health Economics: A review of the literature. Pharmacoeconomics. 2014;32:883-902.

7 Bansback N, Brazier J, Tsuchiya A, Anis A. Using a discrete choice experiment to estimate societal health state utility values. J Health Econ. 2012;31(1):306-18.

8 Viney R, Norman R, Brazier J, et al. An Australian discrete choice experiment to value EQ-5D health states. Health Econ. 2013;23:729-42.

9 Bansback N, Hole AR, Mulhern B, Tsuchiya A. Testing a discrete choice experiment including duration to value health states for large descriptive systems: Addressing design and sampling issues. Soc Sci Med. 2014;114:38-48.

10 Mulhern B, Bansback N, Brazier J et al. Preparatory study for the re-valuation of the EQ-5D tariff: Methodology report. Health Technol Assess. 2014;18:12.

11 Norman R, Cronin P, Viney R. A pilot discrete choice experiment to explore preferences for EQ-5D-5L health states. Applied Health Economics and Health Policy. 2013;11(3):287-98.

12 Norman R, Viney R, Brazier J, et al. Valuing SF-6D Health States Using a Discrete Choice Experiment. Med Decis Mak. 2014;34(6):773-86.

13 Norman R, Kemmler G, Viney R et al. Order of presentation of dimensions did not systematically bias utility weights from a discrete choice experiment. Value Health. 2016;19(8):1033-38.

14 Norman R, Mulhern B, Viney R. 2016. The impact of different DCE-based approaches when anchoring utility scores. PharmacoEconomics, in press.

15 Mulhern B, Bansback N, Hole AR, Tsuchiya A. Using Discrete Choice Experiments with duration to model EQ-5D-5L health state preferences: Testing experimental design strategies. Med Decis Mak, in press.
16 Norman R, Viney R, Aaronson N et al. Using a discrete choice experiment to value the QLU-C10D: feasibility and sensitivity to presentation format. Qual Life Res. 2016;25(3):637-49.
17 Cole A, Shah K, Mulhern B, Feng Y, Devlin N. Valuing EQ-5D-5L health states 'in context' using a discrete choice experiment. European Journal of Health Economics. 2017; DOI 10.1007/s10198-017-0905-7
18 Veldwijk J, Lambooij MS, van Til JA et al. Words or graphics to present a Discrete Choice Experiment: Does it matter? Patient Education and Counseling. 2015;98(11):1376-84.
19 Kenny P, Goodall S, Street DJ, Greene J. Choosing a Doctor: Does Presentation Format Affect the Way Consumers Use Health Care Performance Information? The Patient. 2017;doi 10.1007/s40271-017-0245-9.
20 Oppe M, Devlin NJ, van Hout B, Krabbe PFM, de Charro F. A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. Value Health. 2014;17:445-53.
21 Luyten J, Kessels R, Goos P, Beutels P. Public preferences for prioritizing preventive and curative health care interventions: A discrete choice experiment. Value Health 2015;18:224-33.
22 Stolk EA, Oppe M, Scalone L, Krabbe P. Discrete choice modeling for the quantification of health states: the case of the EQ-5D.Value Health 2010;13(8):1005-13.
23 Ramos-Goni JM, Rivero-Arias O, Errea M, Stolk EA, Herdman M, Cabasés JM. Dealing with the health state 'dead' when using discrete choice experiments to obtain values for EQ-5D-5L heath states. Eur J Health Econ 2013;14(1):S33-42.
24 Scalone L, Stalmeier PF, Milani S, Krabbe PF. Values for health states with different life durations. Eur J Health Econ. 2015;16(9):917-25.
25 Mulhern B, Norman R, Lorgelly P, Lancsar E, Ratcliffe J, Brazier J, Viney R. Is dimension order important when valuing health states using Discrete Choice Experiments including duration? Pharmacoeconomics. 2016;35(4):439-51.
26 Choice Metrics. Ngene [software for experimental design]. NGene 2012.
27 Swait J, Louviere J. The role of the scale parameter in the estimation and comparison of multinomial logit models. Journal of Marketing Research. 1993;30(3):305-14.

28 Viney R, Savage E, Louviere J. Empirical investigation of experimental design properties of discrete choice experiments in health care. Health Econ. 2005;14(40):349-62.
29 Santos M, Monteiro AL, Herdman M, Craig BM. Further Evidence on EQ-5D-5L Preference Inversion: A Brazil/US Collaboration. Qual Life Res. 2017; doi: 10.1007/s11136-017-1591-8.
30 Uggeldahl K, Jacobsen C, Lundhede TH, Olsen SB. Choice certainty in Discrete Choice Experiments: Will eye tracking provide useful measures? Journal of Choice Modelling. 2016;20(1):35-48.31 McCaffery KJ, Dixon A, Hayen A, Jansen J, Smith S, Simpson JM. The influence of graphic display format on the interpretations of quantitative risk information among adults with lower education and literacy: a randomized experimental study. Med Decis Mak. 2012;32:532-44.
32 Al Sayah F, Johnson J, Ohinmaa A, Xie F, Bansback N. Health literacy and logical inconsistencies in valuations of hypothetical health states: results from the Canadian EQ-5D-5L valuation study. Qual Life Res, 2017;26(6):1483-1492.
33 Jonker M, Attema A, Donkers B, Stolk E, Versteegh M. Are health state valuations from the general public biased? A test of health state reference dependency using self-assessed health and an efficient discrete choice experiment. Health Econ. 2016;DOI:10.1002/hec.3445.

Table 1: Study design

Arm	First approach	Second approach
Version 1	Alternate highlight	Yellow highlight
Version 2	Alternate highlight	Grey highlight
Version 3	Alternate highlight	Level highlight
Version 4	Yellow highlight	Alternate highlight
Version 5	Yellow highlight	Grey highlight
Version 6	Yellow highlight	Level highlight
Version 7	Grey highlight	Alternate highlight
Version 8	Grey highlight	Yellow highlight
Version 9	Grey highlight	Level highlight
Version 10	Level highlight	Alternate highlight
Version 11	Level highlight	Yellow highlight
Version 12	Level highlight	Grey highlight

Table 2: Sample demographics

Demographic characteristic		N (%)
Total sample		1,565
Approach completion		
	Alternate highlight	783 (25.0)
	Yellow highlight	780 (24.9)
	Grey highlight	780 (24.9)
	Level highlight	787 (25.1)
Time taken (Mins)		
	Mean (SD)	34 (102)
	Median	15
	Median range across versions	13.5 to 15.9
Age (m,sd)		
	18 - 29	597 (38.1)
	30 – 44	351 (22.4)
	45 – 59	311 (19.9)
	60 – 74	212 (13.5)
	75+	94 (6.0)
Male		618 (39.5)
Country/region of birth		
	Australia/New Zealand	1273 (81.4)
	Europe	132 (8.4)
	Asia	124 (7.9)
	North America	14 (0.9)
	South America	6 (0.4)
	Africa	15 (1.0)
Education		
	Primary/secondary	454 (29.0)
	Trade Cert/Diploma	466 (29.8)
	Bachelor's degree or higher	645 (41.2)
Currently studying		352 (22.5)
Gross income		
	$50,000 or less	680 (43.5)
	$50,001 – $100,000	457 (29.3)
	$100,000 +	195 (12.5)
	No information	233 (14.9)
Married/de facto		899 (57.4)
General health	**Excellent**	208 (13.3)
	Very good	600 (38.3)
	Good	523 (33.4)
	Fair	183 (11.7)
	Poor	51 (3.3)
Self-reported health condition		644 (41.2)
Condition requiring hospitalisation in last 5y		411 (26.3)

Table 3: Unanchored DCE models for each presentation type

Parameter	Alternate highlight				Yellow highlight				Grey highlight				Level highlight			
	Coef.	P (btwn)[1]	95% CI		Coef.	P (btwn)	95% CI		Coef.	P (btwn)	95% CI		Coef.	P (btwn)	95% CI	
MO2 x LY	-0.021*		-0.034	-0.008	-0.032*		-0.045	-0.018	-0.016**		-0.029	-0.002	-0.026*		-0.039	-0.013
MO3 x LY	-0.037*	0.006	-0.050	-0.025	-0.043*	0.071	-0.056	-0.029	-0.031*	0.010	-0.044	-0.018	-0.041*	0.009	-0.054	-0.028
MO4 x LY	-0.075*	<0.001	-0.088	-0.063	-0.086*	<0.001	-0.099	-0.072	-0.076*	<0.001	-0.090	-0.063	-0.085*	<0.001	-0.098	-0.072
MO5 x LY	-0.080*	0.385	-0.094	-0.067	-0.106*	0.002	-0.121	-0.091	-0.080*	0.561	-0.094	-0.066	-0.107*	<0.001	-0.122	-0.093
SC2 x LY	-0.026*		-0.039	-0.014	-0.033*		-0.047	-0.020	-0.027*		-0.040	-0.014	-0.016**		-0.029	-0.003
SC3 x LY	-0.034*	0.199	-0.048	-0.021	-0.038*	0.484	-0.052	-0.024	-0.032*	0.482	-0.045	-0.019	-0.039*	<0.001	-0.052	-0.025
SC4 x LY	-0.078*	<0.001	-0.090	-0.065	-0.077*	<0.001	-0.090	-0.064	-0.068*	<0.001	-0.081	-0.055	-0.066*	<0.001	-0.079	-0.054
SC5 x LY	-0.084*	0.288	-0.097	-0.070	-0.099*	<0.001	-0.113	-0.085	-0.094*	<0.001	-0.107	-0.080	-0.074*	0.200	-0.087	-0.061
UA2 x LY	**-0.028***		-0.040	-0.016	**-0.024***		-0.036	-0.011	**-0.034***		-0.047	-0.022	**-0.008**		-0.020	0.004
UA3 x LY	**-0.017****	0.068	-0.030	-0.004	**-0.019****	0.479	-0.033	-0.006	**-0.016****	0.003	-0.029	-0.003	**-0.006**	0.769	-0.019	0.007
UA4 x LY	-0.046*	<0.001	-0.059	-0.033	-0.049*	<0.001	-0.063	-0.036	-0.047*	<0.001	-0.060	-0.034	-0.037*	<0.001	-0.049	-0.024
UA5 x LY	-0.054*	0.299	-0.067	-0.040	-0.055*	0.456	-0.069	-0.040	-0.054*	0.350	-0.068	-0.039	-0.067*	<0.001	-0.081	-0.053
PD2 x LY	-0.018**		-0.032	-0.005	-0.015**		-0.029	-0.001	-0.004		-0.018	0.009	**0.004**		-0.010	0.017
PD3 x LY	-0.045*	<0.001	-0.058	-0.032	-0.032*	-0.012	-0.045	-0.018	-0.034*	<0.001	-0.047	-0.021	-0.023*	<0.001	-0.036	-0.009
PD4 x LY	-0.085*	<0.001	-0.099	-0.072	-0.090*	<0.001	-0.104	-0.076	-0.086*	<0.001	-0.099	-0.072	-0.063*	<0.001	-0.077	-0.050
PD5 x LY	-0.103*	0.005	-0.117	-0.090	-0.107*	0.008	-0.121	-0.093	-0.105*	0.002	-0.119	-0.092	-0.094*	<0.001	-0.108	-0.081
AD2 x LY	-0.006		-0.019	0.007	-0.007		-0.021	0.007	-0.017**		-0.030	-0.004	**0.003**		-0.011	0.016
AD3 x LY	**-0.073***	<0.001	-0.047	-0.019	-0.031*	<0.001	-0.046	-0.017	-0.036*	<0.006	-0.050	-0.021	-0.017	0.003	-0.031	-0.003
AD4 x LY	**-0.071***	<0.001	-0.083	-0.058	-0.086*	<0.001	-0.100	-0.073	-0.082*	<0.001	-0.096	-0.070	-0.059*	<0.001	-0.072	-0.046
AD5 x LY	-0.079*	0.173	-0.092	-0.067	-0.088*	0.741	-0.102	-0.075	-0.083*	0.931	-0.095	-0.069	-0.087*	<0.001	-0.100	-0.074
LY	0.237*		0.212	0.262	0.265*		0.239	0.291	0.236*		0.211	0.262	0.206*		0.182	0.230
N obs.	7,830				7,800				7,800				7,870			
Log likelihood	-4,828				-4,694				-4,758				-4,801			

*= sig at 0.001; ** = sig at 0.01; ; 1 = significance between adjacent levels; coefficients in bold are disordered between levels

Table 4: Swait and Louviere poolability testing

Parameter		Scaled model	Unrestricted
MO2 x LY		-0.022	-0.023
MO3 x LY		-0.035	-0.038
MO4 x LY		-0.075	-0.080
MO5 x LY		-0.087	-0.093
SC2 x LY		-0.024	-0.026
SC3 x LY		-0.033	-0.036
SC4 x LY		-0.067	-0.072
SC5 x LY		-0.081	-0.087
UA2 x LY		-0.022	-0.023
UA3 x LY		-0.013	-0.014
UA4 x LY		-0.041	-0.044
UA5 x LY		-0.053	-0.057
PD2 x LY		-0.008	-0.009
PD3 x LY		-0.031	-0.034
PD4 x LY		-0.075	-0.081
PD5 x LY		-0.095	-0.102
AD2 x LY		-0.006	-0.007
AD3 x LY		-0.027	-0.029
AD4 x LY		-0.069	-0.074
AD5 x LY		-0.078	-0.084
LY		0.219	0.235
Scale	**Alternate highlight**	1.000	
	Yellow highlight	1.158	
	Grey highlight	1.065	
	Level highlight	1.068	
Log Likelihood		-19128	-19132
Obs		62,600	62,600

Table 5: Feedback questions about the tasks overall

	Disagree	Neutral	Agree
The presentation made the tasks easier to complete	133 (8.5)	476 (30.4)	956 (61.1)
The presentation made the descriptions easy to understand	109 (6.9)	515 (32.9)	941 (60.1)
The presentation influenced the way the task was completed	306 (19.5)	651 (41.6)	608 (38.8)

Table 6: Comparing completion ease and rating

Approach	Easier (%)	More difficult (%)	Rating of each approach from 0 (low) to 1 (high)				
			M (SD)	0 to 4	5 to 6	7 to 8	9 to 10
Alternate highlight	30.8	32.8	6.59 (2.04)	97 (12.4)	263 (33.6)	292 (37.3)	131 (16.7)
Yellow highlight	33.7	33.6	6.75 (2.23)	96 (12.3)	222 (28.5)	294 (37.7)	168 (21.5)
Grey highlight	25.5	37.7	6.35 (2.11)	108 (13.8)	299 (38.3)	265 (34.0)	108 (13.9)
Level highlight	43.8	29.9	6.80 (2.38)	120 (15.2)	180 (22.9)	287 (36.5)	200 (25.4)

Figure 1: The four presentation approaches

Approach A: Alternate highlight

	Scenario A	Scenario B	Scenario C
Mobility	Slight problems in walking about	Moderate problems in walking about	
Self Care	Severe problems washing or dressing yourself	Slight problems washing or dressing yourself	
Usual activities	Unable to do your usual activities	No problems doing your usual activities	Immediate Death
Pain/discomfort	Severe pain or discomfort	Slight pain or discomfort	
Anxiety/depression	Severely anxious or depressed	Extremely anxious or depressed	
Years before death	2 years	2 years	
Which is best?	○	○	○
Which is worst?	○	○	○

Approach B: Yellow highlight

	Scenario A	Scenario B	Scenario C
Mobility	Moderate problems in walking about	Moderate problems in walking about	
Self Care	Unable to wash or dress yourself	Unable to wash or dress yourself	
Usual activities	Severe problems doing your usual activities	Moderate problems doing your usual activities	Immediate Death
Pain/discomfort	Slight pain or discomfort	Severe pain or discomfort	
Anxiety/depression	Extremely anxious or depressed	Extremely anxious or depressed	
Years before death	16 years	4 years	
Which is best?	○	●	○
Which is worst?	●	○	○

Approach C: Grey highlight

	Scenario A	Scenario B	Scenario C
Mobility	**No problems in walking about**	**Slight problems in walking about**	
Self Care	Moderate problems washing or dressing yourself	Moderate problems washing or dressing yourself	
Usual activities	**Unable to do your usual activities**	**Slight problems doing your usual activities**	**Immediate Death**
Pain/discomfort	**Moderate pain or discomfort**	**Extreme pain or discomfort**	
Anxiety/depression	**Not anxious or depressed**	**Slightly anxious or depressed**	
Years before death	**4 years**	**16 years**	
Which is best?	●	○	○
Which is worst?	○	●	○

Approach D: Level highlight

	Scenario A	Scenario B	Scenario C
Problems with walking about	None **Slight** Moderate Severe Unable to	None Slight Moderate **Severe** Unable to	
Problems with washing or dressing	**None** Slight Moderate Severe Unable to	None Slight **Moderate** Severe Unable to	
Problems doing usual activities	None Slight **Moderate** Severe Unable to	**None** Slight Moderate Severe Unable to	Immediate Death
Pain/discomfort	None **Slight** Moderate Severe Extreme	None Slight Moderate **Severe** Extreme	
Anxiety/depression	None Slight Moderate **Severe** Extreme	None **Slight** Moderate Severe Extreme	
Years before death	4	16	
Which is best?	○	○	○
Which is worst?	○	○	○

Figure 2: Anchored DCE coefficients for each presentation type

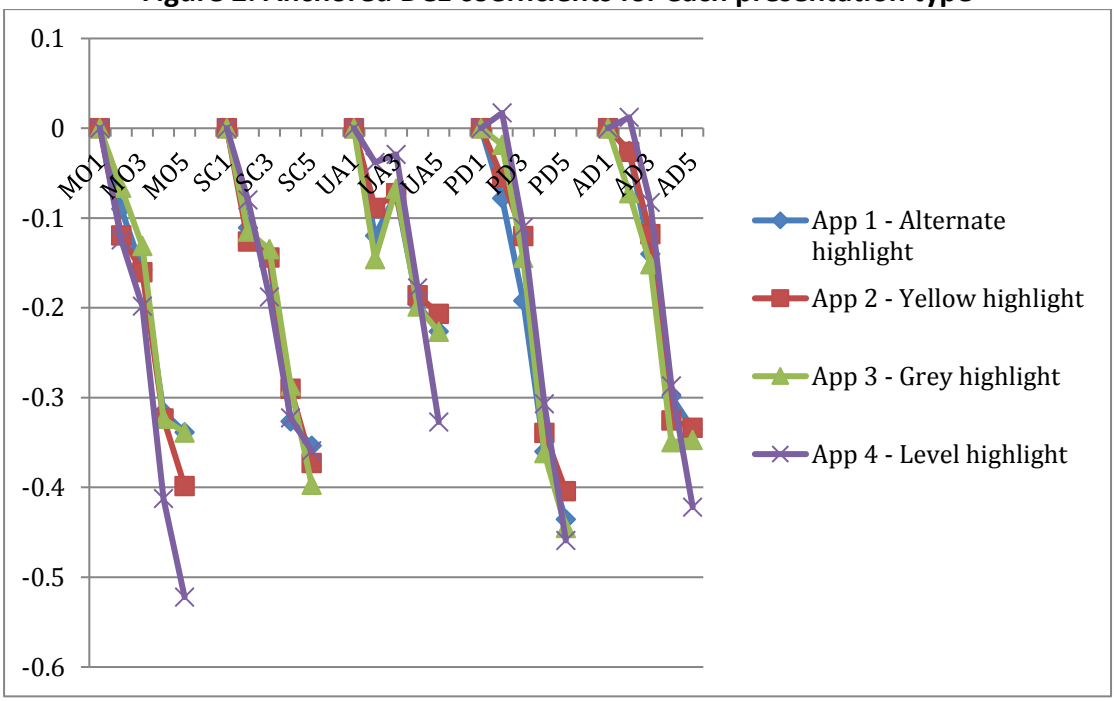

Figure 3: Feedback questions by approach