Is dimension order important when valuing health states using Discrete Choice Experiments including duration?

Brendan Mulhern[1], Richard Norman[2], Paula Lorgelly[3], Emily Lancsar[4], Julie Ratcliffe[5], John Brazier[6], Rosalie Viney[1]

1 University of Technology Sydney, Centre for Health Economics Research and Evaluation, 1-59 Quay St, Haymarket, Sydney, NSW 2000, Australia
2 School of Public Health, Curtin University, Kent Street, Bentley, Perth, WA 6102, Australia
3 Office of Health Economics, Southside, 105 Victoria Street, London, SW1E 6QT, UK
4 Centre for Health Economics, Monash University, Building 75, 15 Innovation Walk, Melbourne, VIC 3800, Australia
5 Flinders Health Economics Group, Flinders University Adelaide, Adelaide, SA 5001, Australia
6 Health Economics and Decision Science, School of Health and Related Research, University of Sheffield, Regent Court, 30 Regent St, S1 4DA, UK

Corresponding author:
Brendan Mulhern
e-mail: Brendan.mulhern@chere.uts.edu.au
Tel no: +61 2 9514 4725
Fax no: +61 2 9514 4730

Acknowledgements:
This study was funded by the Australian National Health and Medical Research Council. Authors BM, RN and RV lead the development of the study design, the data collection and analysis, and drafted the first version of the manuscript. Authors JR, JB, EL and PL contributed to the development of the study design, supported the data analysis and interpretation of the results, and provided detailed revisions of earlier versions of the manuscript.

Abstract:

Background:

Discrete Choice Experiments with duration (DCE_{TTO}) can be used to estimate utility values for preference-based measures, such as the EQ-5D-5L. For self-completion, the health dimensions are presented in a standard order. However, for valuation, this may result in order effects. Thus, it is important to understand whether health state dimension ordering affects values. The aim of this study was to examine the importance of dimension ordering on DCE values using EQ-5D-5L.

Methods:

A choice experiment presenting two health profiles and a third immediate death option was developed. A three-arm study was used, with the same 120 choice sets presented online across each arm (n=360 per arm). Arm 1 presented the standard EQ-5D-5L dimension order, Arm 2 randomised order *between* respondents, and Arm 3 randomised *within* respondents. Conditional logit regression was used to assess model consistency, and scale parameter testing was used to assess model poolability.

Results:

There were minor inconsistencies across each arm, but the magnitude of the coefficients produced were generally consistent. Arm 3 produced the largest range of utility values (1 to -0.980). Scale parameter testing suggested that the models did not differ, and the data could be pooled. Follow up questions did not suggest variation in terms of difficulty.

Conclusions:

The results suggest that the level of randomisation used in DCE health state valuation studies does not significantly impact values, and dimension order may not be as important as other study design issues. The results support past valuation studies that use the standard order of dimensions.

Keywords: Discrete Choice Experiment, EQ-5D, Health state valuation, Quality Adjusted Life Year, Utilities

Key points for decision makers

- The development of EQ-5D-5L value sets for use in the estimation of Quality Adjusted Life Years is influenced by many of the methodological choices made, potentially including the order in which the dimensions are presented
- The results of this study suggest that values are similar across different three levels of dimension ordering (the 'standard' EQ-5D-5L order, and between and within respondent level randomisation)
- Therefore using the standard EQ-5D-5L order is reasonable, and the results of past studies presenting this order can be used with confidence

Introduction

Economic evaluations of health care interventions often use the Quality Adjusted Life Year (QALY) as the outcome measure. The QALY is a single index metric combining length and quality of life. The 'quality' weight (or utility score) is commonly derived using a generic preference based measure (PBM) such as the EQ-5D [1] or SF-6D [2,3]. PBMs use a health state classification system and utility scores are assigned to the health states described, which are anchored on a 1 (full health) to 0 (dead) scale.

Utility scores are derived using a preference elicitation task such as Time Trade Off (TTO; widely used for the EQ-5D [4]), Standard Gamble (SG; used for the SF-6D [2,3]) or Discrete Choice Experiments (DCE). DCEs require respondents to choose between two or more profiles described by the PBMs, and the resulting estimation of a model based on these choices produces values on a latent scale [5]. DCE including an attribute for duration (described as DCE_{TTO} [6]) allow the latent values to be anchored on the full health-dead utility scale. The approach has gained popularity in recent years, and has been shown to generate models that are consistent with other valuation methods, and potentially reduce administrative burden on respondents, and also the cost of data collection, Studies have been carried out online with general population samples for the three level EQ-5D (EQ-5D-3L) in Canada [6] and Australia [7], the five level EQ-5D (EQ-5D-5L) in Australia [8] and the UK [9,10] and the SF-6D in Australia [11].

A potential methodological issue for the use of DCE, as well as other valuation methods, is the order in which the dimensions are presented to respondents. In the majority of EQ-5D valuation studies, the dimensions are presented in the standard order seen by patients when self-completing the instrument (i.e. mobility (MO), self-care (SC), usual activities (UA), pain/discomfort (PD), and anxiety/depression (AD)). As well as consistency with the patient self-report version, for valuation it could be hypothesised that including the three functioning dimensions (MO-SC-UA) first allows respondents to generate a picture of aspects of health functioning before moving onto the more specific symptoms of PD and AD. However while there is an inherent logic to this ordering, it is possible that the order in which dimensions are presented may also influence the overall perception of the health state and also the resulting values which could be influenced by task completion strategies.

For example in a number of studies [4,7], mobility has been the main driver of utility (i.e. modelled to have the overall largest decrement), which might partially reflect the position of that dimension in the valuation task.

Recent work has started to investigate the impact of dimension order on the valuation of health states. Tsuchiya and colleagues [12] presented EQ-5D-5L health states in four different dimension orders in a DCE with duration (DCE_{TTO}) task online and found differences in the magnitude of the dimension coefficients across the different orders, although the impact was not systematic. Mulhern and colleagues [13] tested TTO and DCE using face-to-face interviews and found some evidence that (for DCE) order impacted on the size of the coefficients reported, but again the pattern was unclear. Conversely, Norman and colleagues [14] found no impact of dimension order in a DCE_{TTO} study valuing a cancer specific preference based measure (EORTC QLU-C10D). Outside of health state valuation, Kjaer and colleagues [15] found that the position of the price attribute in study investigating patient preferences for psoriasis treatment was important to price sensitivity.

It is also worth noting that in testing for the impact of dimension order it is important to account for scale, which could be interpreted as choice consistency [16,17]. While there are several potential sources of scale heterogeneity, an intuitive one in this context is that ordering of the health state dimensions can impact respondents' ability to engage with and complete the choice task, leading to differing error variance (and scale) across order presentations. Different ordering could therefore systematically impact: (a) preferences only; (b) scale only; (c) both preferences and scale. The coefficients estimated from discrete choice models, often interpreted as preference parameters, are in fact confounded with scale. As such, where past work has qualitatively shown that the magnitude of the coefficients does not differ across different dimension orders, it could be that preferences do in fact differ but this is masked by the scale confound. Or conversely, where magnitudes of estimated preference parameters appear different, they may actually be the same after scale has been accounted for. As such, appropriate tests are required to disentangle such effects.

The ambiguous findings regarding the impact of dimension ordering on the valuation of a generic PBM such as EQ-5D-5L imply that it is important to test the issue systematically. The aim of this study was to assess the impact of health state dimension ordering on DCE_{TTO} models with a representative sample of the Australian general population. This develops the dimension order research into other valuation methods carried out elsewhere (with inconclusive results) to test DCE_{TTO} in a more systematic way using the generic EQ-5D-5L. DCETTO was used in this study as it is becoming more widely internationally, but various methodological issues including dimension order need to be tested to improve knowledge about the method. Therefore this study will provide further information about the relatively new technique, and may inform the level of dimension randomisation to apply in future health state valuation studies.

Methods

Task presentation

The DCE_{TTO} choice sets used the EQ-5D-5L [18] which classifies health states across five dimensions (mobility, self-care, usual activities, pain/discomfort and anxiety/depression) with five levels of severity (none, slight, moderate, severe, extreme/unable to). The DCE_{TTO} scenarios used in this study were based on methods used in previous studies [7,8,11] and consist of a choice between two EQ-5D-5L health states with associated duration (option A and option B) and also include a third option 'immediate death', which is fixed as option C in each choice task. Duration always appeared at the bottom of the choice sets, as the focus of this study is the potential variation in the health state dimensions included in the EQ-5D-5L. The four duration levels used were 2, 4, 8 and 16 years which were the same used in previous Australian EQ-5D DCE_{TTO} studies [7,8, 19]. Respondents were asked to choose the best and worst of the three options, providing a complete rank of the three options. Figure 1 presents a screenshot of the task with the dimensions in the 'standard' order.

Study design and state selection

A three arm study design, with the same choice sets administered across each arm, was used. Arm 1 presented the 'standard' order (MO-SC-UA-PD-AD-Duration), and acted as the control arm. Five items can be ordered a total of 120 possible ways, e.g. MO-SC-AD-PD-UA-Duration; SC-AD-PD-UA-MO-Duration and so on. Arm 2 randomised the 120 possible EQ-5D-5L orders *between* respondents, so that order differed across respondents but any individual respondent saw the same order for all choice sets. This randomisation method was included to allow each possible order to appear, whilst maintaining a similar level of difficulty for respondents as arm 1 given that only one order is seen. Arm 3 tested the impact of randomising the order *within* respondents, so each choice set had a different order that was randomly selected by the survey system, and was uncorrelated with the levels shown in each dimension of the choice set. This arm was included to assess whether respondents could validly complete DCETTO tasks where the order changed each time.

Comparing Arms 1 and 2, we might expect the first dimension that a respondent reads to be more important in determining their final choice than the middle ones. Therefore, in our case, the importance of mobility in Arm 2 may be less than in Arm 1, and the first dimension a person sees may have a greater weight in Arm 2 for a given order than it has in Arm 1 (given that it appears first, and notwithstanding the impact of variability). Arm 3 was expected to generate more inconsistent data and greater variability (in terms of both ordering and magnitude of the coefficients and also in terms of scale/error variance) in comparison to Arms 1 and 2 given the cognitive burden of having to complete choice tasks where the order changed each time. We aimed to collect a sample of approximately 360 respondents per arm to enable parameter estimation for each arm separately, and also for the pooled data, with approximately the same level of precision.

To estimate DCE_{TTO} models, the number of choice sets included should exceed the number of parameters that are being estimated in the model. In this case, the maximum number of parameters for DCE_{TTO} of EQ-5D-5L including five dimensions, continuous duration, and coefficient for each

dimension order position would be 37 (the sum of: the interaction of duration and each of the EQ-5D-5L main effects (5 dimensions x (5-1) levels x continuous duration (1) = 20); a linear duration term (1); the interaction of dimension order, position and level ((5-1) dimensions x (5-1) levels) = 16). Therefore including 120 choice sets in the design allowed for reliable parameter estimation and comparison across the arms. These were selected using a *d*-optimal procedure within the experimental design software NGene [20]. The same 120 choice sets were administered across each arm, with each respondent completing 10 choice sets that were randomly selected from the overall pool. This is in line with the choice set randomisation procedure carried out in DCE_{TTO} studies elsewhere [7,8] and gives approximately 90 observations per choice set (with approximately 30 per choice set per arm). This is within the range used in other DCE_{TTO} studies [8,9] and justifies including approximately 360 respondents in each.

Recruitment and data collection

The survey was administered to an online panel (Pure Profile), representative of the Australian population in terms of age and gender. Members of the online panel accessed an invitational weblink, and completed demographic and self-reported health questions, 10 DCE_{TTO} tasks, and then a series of follow-up questions assessing the difficulty of the choice tasks, difficulty in imagining the states, and whether they considered all or only part of the descriptions provided when answering. The study was approved by the University of Technology Sydney Human Research Ethics Committee.

Analysis

Estimating and comparing individual models

Conditional logit regression was used to generate unanchored and anchored estimates across the three arms using the model:

$$u_j = \beta t_j + \boldsymbol{\lambda}' \boldsymbol{x}_j t_j + \varepsilon_j \qquad (1)$$

where β is the utility of living in full health for duration t, $\boldsymbol{\lambda}$ is the utility of living in state $\boldsymbol{x}$ for duration t, and $\boldsymbol{\varepsilon}$ is the error term. The coefficients for the interactions between health state dimension and duration are anchored to the utility scale by assuming that the utility of living in full health for a shorter duration is equal to living in a sub-optimal state for a longer duration. This means that the anchored values for each level of each dimension are produced by dividing the interaction coefficients by the coefficient for duration (both estimated from the conditional logit regression displayed in equation 1) as shown in equation 2:

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

Further information on the method can be found in Bansback et al [7]. In the analysis presented here, we model the choices between the two health state and duration combinations, and exclude the immediate death option data. This means that in the analysis reported here we use the respondent's

choice of best and worst of the scenarios to rank all three of those presented in each choice set, and then assume IID to model the data from the two scenarios where the respondent will have provided an indication of which they think is best. This approach to modelling similar data has been done elsewhere [7,8] and is repeated here as the focus of this paper is the scenarios where the dimension order is manipulated. Also, it has been shown that the way in which the 'dead' data is modelled affects the coefficient values [19]. Across the arms we compared the consistency of the dimension level coefficients (in terms of the amount of disordered levels, where the magnitude of an ordered coefficient increases as the severity level increases). The characteristics of the EQ-5D-5L value sets were also compared. This included directly comparing the range of utility values and the overall magnitude of the dimension level coefficients (which controls for scale as all magnitudes are relative to the respective coefficient on duration). Comparing the magnitude of the coefficients across the arms in comparison to where they appear allows inferences about the impact of fixed and variable ordering

Assessing differences between the arms

We tested the null hypothesis that values do not differ across the three arms using a Swait and Louviere [14] test, which examines the variability between the systematic and random components of a DCE design. If we fail to reject the null hypothesis the data can be pooled. The test is carried out by using a grid search [17] to identify the relative scale parameters of the experimental Arms (2 and 3), with the control Arm (1) normalised to one, that result in the maximum log likelihood for Arms 2 and 3. A restricted pooled model scaled using the parameters identified ($L\mu$) is estimated along with an unrestricted pooled (L_p) and individual models for each arm (L_1, L_2 and L_3) using conditional logit. To assess differences, two likelihood ratio (LR) tests are used to compare the models (see equations 3 and 4). If the LR statistic is within a critical value then the null hypothesis cannot be rejected and the data can be pooled and models estimated with no scale parameter adjustment.

$$LR = -2[L\mu - (L_1 + L_2 + L_3)] \quad (3)$$

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

Testing the impact of dimension order

We modelled the impact of dimension order, where alongside the dimension parameter we estimated interactions between dimension position and severity level. This is because the effect of the dimension level which is expressed as a coefficient decrement (e.g. the disutility of poor mobility) cannot be separated from any dimension order effect (the focus on the first dimension) which is included within the decrement. If the first dimension is given particular attention, then the coefficients for mobility will be artificially inflated. By randomising the dimension order we were able to separately identify dimension effects (the coefficients for each level) from order effects (coefficients based on the order of appearance), as each of the arms included the same choice sets.

The model to test dimension order was:

$$u_j = \beta t_j + \boldsymbol{\lambda}' \boldsymbol{x}_j t_j + P' \boldsymbol{x}_j t_j + \varepsilon_j \quad (5)$$

Where $\boldsymbol{\lambda'}$ is the marginal effect of dummies relating to interactions between duration (***t***) and movements away from level one in each of the five health state dimensions (***x***) (therefore 20 coefficients). P' is a dummy for the health state dimension appearing in order position 2, 3, 4 or 5 which is interacted with the health state dimension severity level (***x***) and duration (***t***)(therefore16 coefficients). Dimension order effect parameters are not fitted to the dimension in the first position in the choice set) or to level 1 in each dimension to avoid over-specification. As an example of this, if we use a fixed ordering of dimensions, the coefficient for the MO2 and duration interaction also currently includes a coefficient for the effect of the first dimension being at level 2. By identifying this independently a value for the health state dimension without the effect of the dimension level and position can be obtained. These models were run using only responses from those who saw a non-standard ordering of dimensions, so data from arm 1 was excluded. To account for clustering of responses within respondents, we adjusted the standard errors using a clustered sandwich estimator, and then ran an F-test to explore the joint significance of the order effect terms. We also carried out descriptive analysis of the follow up questions across the arms, using chi squared tests. Analysis was carried out using Stata version 13 [21].

Results

Response rate and demographics

Overall, 2,710 respondents accessed the survey. Of these, 1,576 (58%) were classified as "over quota" (i.e. belonging to a particular age/gender quota group that had already been completed) so were excluded before completing the survey, 61 (2%) dropped out during completion, and 1,073 (40%) fully completed the survey (Arm 1 n=366 (34%); Arm 2 n=346 (32%); Arm 3 n=361 (34%)). The mean time taken to complete was 11.7 minutes (range 2.3 mins to 63.4 mins), and this did not differ across the arms (p = 0.724).

Table 1 reports the demographic characteristics and self-reported health status of the sample. The sample is generally representative of the Australian general population in terms of age and gender, and there were no significant differences in any of the background characteristics across the arms.

DCE_{TTO} models

Table 2 reports the unanchored interaction coefficients (with the disordered coefficients bolded) and the value of the anchored coefficients for each arm (Figure 2 also presents the anchored values). Across all three arms, the majority of the modelled decrements are ordered as expected (17 of 20 for Arms 1/2, and 19 of 20 for Arm 3). Regarding disordering, the coefficient for the usual activities dimension level 3 is smaller than level 2 meaning that 'moderate' problems results in a smaller decrement than 'slight' problems, but the difference between the levels is not significant. Relative to Arm 3, Arms 1 and 2 each have an additional two disordered coefficients. Arm 1 has evidence of small non-significant inconsistencies at level two of pain/discomfort and level five of

anxiety/depression. Arm 2 has evidence of small non-significant inconsistencies at level five of usual activities and anxiety/depression. In terms of dimension 'importance', Arms 1 and 3 have the same order (MO – SC – PD – AD – UA) and the order of the first three differs for Arm 2 (PD – MO – SC – AD – UA). The range of utility values (i.e. the values assigned to 11111 and 55555) is smallest for the 'fixed' order (Arm 1; 1 to -0.785) and increases for the between (Arm 2; 1 to -0.795) and within (Arm 3; 1 to -0.980) randomisation levels.

Hypothesis testing for preference and scale differences

The scale parameters identified from the grid search following the methodology of Swait and Louviere [16] were 0.938 (Arm 2) and 0.998 (Arm 3). The scale parameter of Arm 1 is fixed at 1. The difference between the Arm 1 fixed value and the Arm 2 scale parameter is larger than for Arm 3, and therefore Arm 2 has slightly more variability. Table 3 displays the scaled restricted ($L\mu$, estimated using the scale parameters) and unrestricted (L_p) pooled models and the log likelihood statistics for these. The log likelihood statistics for the three individual models (L_1-L_3) are reported in Table 2. The results of the likelihood ratio tests indicate that the difference in the log likelihoods between the scaled and individual models (LR = 43), and scaled and unrestricted model (LR = 2) mean that the null hypothesis of no difference across the arms cannot be rejected and the data can be pooled.

Testing the impact of dimension order

The analysis with order effect terms estimated is reported in Table 4. Of the 16 additional coefficients estimated, only one was statistically significant at the 5% level, and the magnitude of the coefficients was small relative to the coefficients relating to specific EQ-5D dimension-level combinations. Running an F-test on the joint significance of the additional 16 order effect coefficients fails to reject the null hypothesis of no order effect (p=0.0606).

Follow up questions about task difficulty and completion strategies

Table 5 reports the results of follow up questions focusing on completion strategies across each arm. Overall, 632 respondents (58%) reported focusing on a particular part of the health description, although this did not differ across the arms (p=0.122). The pain dimension was focused on the most, and this did not vary across the arms. There were no differences in the consistency of the models produced across the arms (in terms of the coefficient decrements) based on whether the respondent reported focusing on all or part of the description. In total, 251 respondents (23%) reported using no strategy to complete the DCE, 452 (42%) reported only focusing on 'few' or 'most' aspects, and 370 (34%) reported focusing on 'all' aspects, and the completion strategies reported did not differ across the arms.

Figure 3 reports the results of the follow up questions focusing on task difficulty. The majority of the respondents reported that the task was clear. There was no difference across the arms in terms of respondents reporting that the task was difficult overall, in terms of telling the difference between the descriptions or in imagining the scenarios. There was also no difference in the frequency of

respondents reporting that the order in which the dimensions were presented was confusing, where the majority agreed that the order was not confusing.

Discussion

This study tested the impact of different levels of EQ-5D-5L dimension order randomisation on the magnitude and consistency of the modelled health dimension coefficients using the DCE_{TTO} valuation method. This included using the standard order of dimensions (used widely in other valuation studies), and two different levels of randomisation: one arm varied dimension order between subjects, and the other varied dimension order within subjects. The analysis found that the order of the magnitude of coefficients on levels within an attribute, while at times inconsistent, was similar across the arms. Importantly, the non-significant differences in the modelled results across the arms suggest that the order in which the dimensions are presented does not have a major impact on the magnitude of the modelled coefficients or the overall dimensions when using a DCE with duration to value health states. The findings are in line with the results of Norman and colleagues [14] who found no impact of dimension order on the valuation of a cancer specific PBM (EORTC QLU-C10D) Therefore the evidence suggests that dimension order may not affect the values produced as much as other DCE_{TTO} study design issues. These include the use of prior values in the design of the study [22], the range and number of the duration attribute levels used, or the way in which the choice sets are presented [23].

We found small non-systematic differences across the arms. The results do not support the hypothesis that particular patterns of dimension orders may lead to particular patterns of modelled coefficients. An example of a response pattern may be respondents focusing on the first or last dimensions presented whilst answering the task rather than comparing all of the dimensions. However there is the indication of differences in the overall range of the values and also the magnitude of the coefficients (and therefore utility decrements), and this may have implications if the values were used to estimate QALYs. For example Arm 2 would place more weight on improvements in pain/discomfort. The larger range of Arm 3 would lead to a relative emphasis on interventions that improve quality of life. However it is unclear if these patterns would be the same using a larger sample as would be recruited to produce a nationally representative value set for use in cost utility analysis. It is worth noting that the pooled sample coefficient order (MO-SC-PD-AD-UA) and range (1 to -0.857) differs to that found in the earlier pilot valuation of EQ-5D-5L in Australia that used the standard order (where the magnitude of the coefficients was AD-PD-MO-SC-UA and range 1 to -0.723 for the most comparable model) [8]. The findings from Arm 1 show that the first dimension presented (MO) had the largest decrement. However this is difficult to interpret as mobility is also largest for Arm 3, but only appears first approximately 20% of the time meaning that the importance of mobility may be an indication of genuine preferences. It would be useful to repeat Arm 1 on a separate sample to compare the magnitude of the coefficients and test the number of inconsistencies.

These findings are generally in line with other work testing the impact of EQ-5D-5L dimension order on the values produced for TTO and DCE which also did not find specific patterns based on a number of set dimension orders [12,13]. The previous work presented a small number of fixed orders between respondents. This study adds to the past work by including two levels of randomisation where all EQ-5D-5L orders are possible either between (Arm 2) or within (Arm 3) respondents. Arm 3 was expected to generate more inconsistent data in comparison to Arms 1 and 2 given the cognitive burden of having to complete choice tasks where the order changed each time. The poolability assessment shows that this was not the case, and the results support previous findings in this area [7,8]

An issue for consideration is the minor coefficient disordering found across the arms. A similar level of disordering is common in DCE studies with and without duration. For example disordering between anxiety/depression levels 4 (severe) and 5 (extreme) was found for Arm 1 in this study, and this was apparent in an earlier DCE_{TTO} study valuing EQ-5D-5L in Australia [8], and it is worth noting that both used the 'standard' dimension order. In other DCE_{TTO} work comparing zero and non-zero prior design strategies Mulhern and colleagues [22] found different levels of disordering across the designs between self-care levels 1 (none) and 2 (slight), usual activities levels 2 (slight) and 3 (moderate) - a disordering also found in this study, and pain/discomfort levels 2 and 3. This demonstrates that disordering is common in DCE studies, but there is no clear pattern. There are a number of reasons why disordering may occur including the study design and choice set selection methods used, the format of the DCE task where respondents may not be clear about the overall ordering of the levels, and respondent perception of the level descriptors. For example, in qualitative work and modelling studies, difficulty distinguishing 'severe' and 'extreme' has been reported [24,25] both semantically and in terms of severity, where a maximum trade-off is reached at severe as both levels are as bad as each other. Disordering at the lower severity levels may be due to dominant attribute heuristics leading respondents to put less weight on trade-offs between lower severity levels of less important dimensions. To deal with disordering it is possible to impose ordering, or test other models that may help to order the coefficients (for example including interactions). However as this paper is a methodological comparison of dimension ordering, we have not done this as the level of disordering is a useful part of the comparative analysis.

The standard order of EQ-5D health state dimension presentation has been used in a wide range of TTO [26] and DCE [5,9] valuation studies. For EQ-5D the standard order was originally used as it was hypothesised to allow respondents to build up a picture of the health state starting with the functional dimensions (MO, SC and UA). However in DCE, order effects might be important because of heuristics. However the results of this study suggest that the models are similar irrespective of whether the study design employs the standard order, or uses randomisation, and there was no difference across the arms in terms of respondents reporting that the task was difficult overall. Therefore for DCE_{TTO} we do not find evidence to suggest that the inherent logic hypothesised in the standard order of the state is an important factor in the task completion process. Future studies

should consider using the standard order, or imposing between subjects randomisation (i.e. arm 2), if there is concern about the residual possibility of bias which can be controlled for whilst presenting each respondent with only one dimension order as has been the standard in all previous DCETTO studies [6-9].

Using a dimenison order that makes the valuation process as easy as possible would appear important, however respondents in this study did not report that a particular level of randomisation was more difficult than the others when asked about imagining the states or task completion. DCE completed online may be prone to strategic completion and the use of heuristics, and the follow up questions used here are a way of eliciting useful data in this area. This may not be the case with other valuation methods such as Time Trade Off where respondents typically consider a single holistic impairment profile in that case against full health.

This study has a number of limitations. We cannot fully understand respondent engagement with and concentration on the task beyond measuring the time taken, and their perception of the dimension orders presented, as the survey was online. This, however, is an issue for all valuation work conducted using computer based methods. The results of the follow up questions, and the small amount of disordered coefficients (i.e. one or three out of 20), suggests that respondents had a reasonable level of engagement. This is because disordering across the descriptive system and within each dimension would be expected if a large group of respondents answered at random. Eye tracking methods could be utilised to further understand the process used to complete the tasks and the dimensions focused on by respondents, and this has been done elsewhere for DCE [27]. We also have not tested the impact of altering the position of the duration attribute as the focus of this study was the dimensions included in the EQ-5D, something which would be valuable future research. In our analysis we were also unable to identify effects within a dimension when the ordering was changed (for example what is the impact on preferences for pain/discomfort when different dimensions appear in the first position in the order). Regarding our model testing the impact of dimension order, we chose to apply an additive model in line with other work in this area [14], and did not include a multiplicative model which could be an equally valid way to test order effects. This is an area for further investigation. We also did not allow for any modelling of preference heterogeneity due to the size of the sample. However as the aim is to assess the impact of dimension order the study design and modelling carried out was sufficient to answer this question.

Conclusion

The impact of dimension order appears minimal which supports past valuation work in this area that has used the standard order. The results suggest that the level of randomisation used in DCE health state valuation studies does not significantly impact valuations, and dimension order may not be important as other design issues in the completion of DCE_{TTO} studies. Therefore it is reasonable to use the standard presentation of dimensions employed in the self-complete version.

Compliance with Ethical standards:

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional research committee (University of Technology Sydney Human Research Ethics Committee (Program Approval Number 2015000135). The following authors are members of the EuroQol Research Foundation (the copyright holders of EQ-5D-5L): Brendan Mulhern, Richard Norman, John Brazier, and Rosalie Viney. There are no other conflicts of interest.

References:

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

2. 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.

3. Brazier J, Roberts J. Estimating a preference-based index from the SF-12. Med Care. 2002;42(9):851-9.

4. Dolan P. Modeling valuations for EuroQol health states. Med Care. 1997;35(11):1095-108.

5. 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.

6. 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.

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

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

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, Buckingham K, Cairns J, Devlin N, et al. Preparatory study for the revaluation of the EQ-5D tariff: Methodology report. Health Technol Assess. 2014;18:12.

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

12. Tsuchiya A, Mulhern B, Bansback N, Hole AR. Using DCE with duration to examine the robustness of preferences across the five dimensions of the EuroQol instrument: The second paper from the FEDEV project. EuroQol Group Plenary Proceedings. 2014.

13. Mulhern B, Shah K, Janssen MF, Longworth L. Valuing health using Time Trade Off and Discrete Choice methods: Does dimension order impact on health state values? Value Health. 2016;19(2):210-7.

14 Norman R, Kemmler G, Viney R, Pickard AS, Gamper E, Holzner B, Nerich V, King M. Order of Presentation of Dimensions Does Not Systematically Bias Utility Weights from a Discrete Choice Experiment. Value Health, in press.15. Kjaer T, Bech M, Gyrd-Hansen D, Hart-Hansen K. Ordering effect and price sensitivity in discrete choice experiments: need we worry? Health Econ. 2006;15(11):1217-28.

16. 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.

17. Viney R, Savage E, Louviere J. Empirical investigation of experimental design properties of discrete choice experiments in health care. Health Econ. 2005;14:349-62.

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

19. Norman R, Mulhern B, Viney R. The impact of different DCE-based approaches when anchoring utility scores. Pharmacoeconomics 2016;34(8):805-14.

20. ChoiceMetrics. Ngene [software for experimental design]. NGene. 2012.

21. StataCorp. Stata Statistical Software: Release 13. College Station, TX: StataCorp LP. 2013.

22. Mulhern B, Bansback N, Hole AR, Tsuchiya A. Using Discrete Choice Experiment with duration to model EQ-5D-5L health state preferences: Testing experimental design strategies. Med Decis Mak. 2016; in press.

23 Viney R, Mulhern B, Norman N, Shah K, Bansback N, Longworth L. Using DCE with duration to value EQ-5D-5L: Investigating task presentation. EuroQol Group Plenary, Berlin, 2016.

24. Mulhern B, Bansback N, Brazier J, Buckingham K, Cairns J, Devlin N, et al. Preparatory study for the revaluation of the EQ-5D tariff: Methodology report. Health Technol Assess. 2014;18:12.

25. Craig BM, Pickard AS, Rand-Hendriksen K. Do health preferences contradict ordering of EQ-5D labels? Qual Life Res. 2015;24(7):1759-65.

26. 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.

24. Krucien N, Ryan M, Hermens F. Using Eye-tracking methods to inform decision making processes in Discrete Choice Experiments. Health Economist's Study Group. 2014.

Table 1: Demographics of the sample across each arm

Demographic	Arm 1 (N,%)	Arm 2 (N,%)	Arm 3 (N,%)	Significance
N	366 (34.1)	346 (32.3)	361 (33.6)	
Age group				0.414
18-29	93 (25.4)	80 (23.1)	79 (21.8)	
30-44	95 (26.0)	105 (30.3)	86 (23.8)	
45-59	86 (23.5)	76 (22.0)	103 (28.5)	
60-74	61 (16.7)	60 (17.3)	67 (18.6)	
75+	31 (8.5)	25 (7.2)	26 (7.2)	
Male	189 (51.6)	162 (46.8)	172 (47.6)	0.384
Country of birth				0.437
Australia	292 (79.8)	263 (76.0)	277 (76.7)	
Other	74 (20.2)	83 (24.0)	84 (23.3)	
Highest education level				0.656
Primary	20 (5.5)	19 (5.5)	15 (4.2)	
Secondary	109 (29.7)	97 (28.0)	119 (55.1)	
Trade cert/diploma	120 (32.8)	117 (33.8)	111 (30.7)	
Bachelors	78 (21.3)	85 (24.6)	88 (24.3)	
Higher	39 (10.7)	28 (8.1)	28 (7.8)	
Currently studying	59 (16.1)	30 (8.7)	46 (12.7)	0.212
Has child up to 16	32 (8.7)	26 (7.5)	24 (6.6)	0.565
Self-rated health status				0.443
Excellent	43 (11.7)	48 (13.9)	46 (12.7)	
Very good	141 (38.5)	114 (32.9)	130 (36.0)	
Good	118 (32.2)	103 (29.8)	102 (28.2)	
Fair	48 (13.1)	63 (18.2)	68 (18.8)	
Poor	16 (4.4)	18 (5.2)	15 (4.2)	
Chronic condition	150 (41.0)	146 (42.2)	150 (41.6)	0.948
Condition reported				
Tiredness/fatigue	51 (14.0)	53 (15.3)	52 (14.4)	0.869
High blood pressure	69 (18.9)	53 (15.3)	72 (19.9)	0.249
Pain	70 (19.1)	76 (22.0)	80 (22.2)	0.533
Heart disease	15 (4.1)	14 (4.0)	19 (5.3)	0.672
Insomnia	29 (7.9)	37 (10.7)	38 (10.5)	0.369
Osteoarthritis	46 (12.6)	35 (10.1)	50 (13.9)	0.306
Anxiety/nerves	65 (17.8)	55 (15.9)	71 (19.7)	0.424
Stroke	9 (2.5)	2 (0.6)	4 (1.1)	0.086
Depression	68 (18.6)	62 (17.9)	61 (16.9)	0.837
Cancer	9 (2.5)	7 (2.0)	10 (2.8)	0.811
Diabetes	25 (6.8)	28 (8.1)	26 (7.2)	0.804
Breathing problems	38 (10.4)	36 (10.4)	29 (8.0)	0.464

Table 2: Unanchored and anchored DCE model coefficients

Parameter	Arm 1			Arm 2			Arm 3		
	Coef.[a]	Sig[b]	Disutility[c]	Coef.	Sig	Disutility	Coef.	Sig	Disutility
MO2 x T[d]	-0.011	0.269	-0.050	-0.026	0.008	-0.122	-0.031	0.001	-0.153
MO3 x T	-0.026	0.006	-0.119	-0.035	<0.001	-0.164	-0.033	0.001	-0.163
MO4 x T	-0.087	<0.001	-0.397	-0.075	<0.001	-0.352	-0.072	<0.001	-0.355
MO5 x T	-0.103	<0.001	-0.470	-0.101	<0.001	-0.474	-0.106	<0.001	-0.522
SC2 x T	-0.032	0.001	-0.146	-0.025	0.012	-0.117	-0.012	0.211	-0.059
SC3 x T	-0.034	0.001	-0.155	-0.037	<0.001	-0.174	-0.036	<0.001	-0.178
SC4 x T	-0.074	<0.001	-0.338	-0.075	<0.001	-0.352	-0.077	<0.001	-0.379
SC5 x T	-0.095	<0.001	-0.434	-0.080	<0.001	-0.376	-0.102	<0.001	-0.502
UA2 x T	-0.021	0.019	-0.096	-0.017	0.059	-0.080	-0.004	0.633	-0.020
UA3 x T	**-0.012**	0.183	**-0.055**	**-0.009**	0.360	**-0.042**	**0.004**	0.657	**0.020**
UA4 x T	-0.050	<0.001	-0.228	-0.049	<0.001	-0.230	-0.028	0.003	-0.138
UA5 x T	-0.055	<0.001	-0.251	**-0.045**	<0.001	**-0.212**	-0.033	0.002	-0.163
PD2 x T	**0.004**	0.703	**0.018**	-0.016	0.119	-0.075	-0.012	0.221	-0.059
PD3 x T	-0.014	0.140	-0.064	-0.039	<0.001	-0.183	-0.014	0.155	-0.069
PD4 x T	-0.051	<0.001	-0.233	-0.080	<0.001	-0.376	-0.067	<0.001	-0.315
PD5 x T	-0.071	<0.001	-0.324	-0.101	<0.001	-0.475	-0.093	<0.001	-0.458
AD2 x T	-0.020	0.039	-0.091	-0.016	0.112	-0.075	-0.009	0.378	-0.044
AD3 x T	-0.040	<0.001	-0.183	-0.027	0.012	-0.127	-0.024	0.019	-0.118
AD4 x T	-0.070	<0.001	-0.320	-0.075	<0.001	-0.352	-0.065	<0.001	-0.320
AD5 x T	**-0.067**	<0.001	**-0.306**	**-0.055**	<0.001	**-0.258**	-0.068	<0.001	-0.335
T	0.219	<0.001		0.213	<0.001		0.203	<0.001	
Number of observations	3646			3458			3603		
Log Likelihood	-2226			-2153			-2203		
Range (11111 to 55555)			1 to -0.785			1 to -0.795			1 to -0.980
Coefficient magnitude order (largest level 5 value first)	MO – SC – PD – AD – UA			PD – MO – SC – AD – UA			MO – SC – PD – AD - UA		

[a] Coefficient for the decrement of each level of each health state dimension from the baseline level 1 (no problems)

[b] Significance of each coefficient value in comparison to the baseline level 1

[c] The anchored coefficient for each level of each dimension used to calculate utility values for each health state

[d] Each parameter is listed as an interaction of the health state dimension level (MO2, MO3, etc) and duration (T). MO: mobility; SC: self-care; UA: usual activities; PD: pain/discomfort; AD: anxiety/depression

Values in bold: Inconsistent coefficients

Table 3: Unanchored and anchored pooled DCE model coefficients for Swait and Louviere (1993) test

Parameter	Unrestricted pooled model			Restricted pooled model		
	Coef.[a]	Sig[b]	Disutility[c]	Coef.	Sig	Disutility
MO2 x T[d]	-0.023	<0.001	-0.110	-0.023	<0.001	-0.107
MO3 x T	-0.031	<0.001	-0.148	-0.032	<0.001	-0.150
MO4 x T	-0.079	<0.001	-0.376	-0.080	<0.001	-0.374
MO5 x T	-0.103	<0.001	-0.490	-0.105	<0.001	-0.491
SC2 x T	-0.023	<0.001	-0.110	-0.023	<0.001	-0.107
SC3 x T	-0.036	<0.001	-0.171	-0.036	<0.001	-0.168
SC4 x T	-0.075	<0.001	-0.357	-0.077	<0.001	-0.360
SC5 x T	-0.092	<0.001	-0.438	-0.094	<0.001	-0.439
UA2 x T	-0.014	0.008	-0.067	-0.014	0.008	-0.065
UA3 x T	**-0.005**	0.356	**-0.024**	**-0.005**	0.359	**-0.023**
UA4 x T	-0.042	<0.001	-0.200	-0.043	<0.001	-0.201
UA5 x T	-0.043	<0.001	-0.205	-0.044	<0.001	-0.206
PD2 x T	-0.008	0.185	-0.038	-0.008	0.194	-0.037
PD3 x T	-0.023	<0.001	-0.110	-0.023	<0.001	-0.107
PD4 x T	-0.065	<0.001	-0.310	-0.066	<0.001	-0.308
PD5 x T	-0.088	<0.001	-0.419	-0.089	<0.001	-0.416
AD2 x T	-0.015	0.007	-0.071	-0.016	0.007	-0.075
AD3 x T	-0.031	<0.001	-0.148	-0.031	<0.001	-0.145
AD4 x T	-0.070	<0.001	-0.333	-0.072	<0.001	-0.336
AD5 x T	**-0.064**	<0.001	**-0.305**	**-0.065**	<0.001	**-0.304**
T	0.210	<0.001		0.214	<0.001	
Number of observations	10707			10707		
Log Likelihood	-6604			-6603		
Range (11111 to 55555)			1 to -0.857			1 to -0.856
Coefficient magnitude order (largest level 5 value first)	MO-SC-PD-AD-UA			MO-SC-PD-AD-UA		

[a] Coefficient for the decrement of each level of each health state dimension from the baseline level 1 (no problems)

[b] Significance of each coefficient value in comparison to the baseline level 1

[c] The anchored coefficient for each level of each dimension used to calculate utility values for each health state

[d] Each parameter is listed as an interaction of the health state dimension level (MO2, MO3, etc) and duration (T). MO: mobility; SC: self-care; UA: usual activities; PD: pain/discomfort; AD: anxiety/depression

Values in bold: Inconsistent coefficients

Table 4: Modelling order effects

Parameter	Coefficient[a]	Significance[b]
MO2 x T[c]	-0.032	0.001
MO3 x T	-0.039	<0.001
MO4 x T	-0.077	<0.001
MO5 x T	-0.109	<0.001
SC2 x T	-0.022	0.019
SC3 x T	-0.040	<0.001
SC4 x T	-0.078	<0.001
SC5 x T	-0.096	<0.001
UA2 x T	-0.015	0.115
UA3 x T	**-0.006**	0.529
UA4 x T	-0.042	<0.001
UA5 x T	-0.042	<0.001
PD2 x T	-0.017	0.065
PD3 x T	-0.033	<0.001
PD4 x T	-0.077	<0.001
PD5 x T	-0.104	<0.001
AD2 x T	-0.016	0.090
AD3 x T	-0.030	0.002
AD4 x T	-0.073	<0.001
AD5 x T	**-0.068**	<0.001
T	0.211	<0.001
2nd position x level 2 x T[d]	-0.006	0.516
2nd position x level 3 x T	-0.005	0.585
2nd position x level 4 x T	-0.003	0.752
2nd position x level 5 x T	-0.007	0.484
3rd position x level 2 x T	0.013	0.174
3rd position x level 3 x T	0.015	0.129
3rd position x level 4 x T	0.023	0.016
3rd position x level 5 x T	0.019	0.067
4th position x level 2 x T	0.010	0.282
4th position x level 3 x T	0.012	0.196
4th position x level 4 x T	0.003	0.739
4th position x level 5 x T	0.007	0.525
5th position x level 2 x T	-0.001	0.880
5th position x level 3 x T	-0.001	0.917
5th position x level 4 x T	-0.011	0.239
5th position x level 5 x T	0.004	0.675

[a] Coefficient for the decrement of each level of each health state dimension from the baseline level 1 (no problems)

[b] Significance of each coefficient value in comparison to the baseline level 1

[c] Each parameter is listed as an interaction of the health state dimension level (MO2, MO3, etc) and duration (T). MO: mobility; SC: self-care; UA: usual activities; PD: pain/discomfort; AD: anxiety/depression

[d] Each parameter is listed as an interaction of the position the dimension appears in the choice set, and the health state dimension level and duration

Values in bold: Inconsistent coefficients

Table 5: Summary of follow up questions about task completion

	Arm 1	Arm 2	Arm 3	Sig[a]
Answer focusing on only part of choice set?				0.122
Yes	200 (54.6)	213 (61.6)	219 (60.7)	
No	166 (45.4)	133 (38.4)	142 (39.3)	
Which part?				0.649
Mobility	44 (17.4)	33 (13.0)	38 (14.7)	
Self-care	39 (15.4)	53 (20.9)	42 (16.3)	
Usual activities	26 (10.3)	26 (10.2)	21 (8.1)	
Pain/discomfort	67 (26.5)	67 (26.4)	74 (28.7)	
Anxiety/depression	30 (11.9)	34 (13.4)	42 (16.3)	
Duration	47 (18.6)	41 (16.1)	41 (15.9)	
Strategy used to answer the question				0.776
No strategy	85 (23.2)	76 (22.0)	90 (24.9)	
Focus on few aspects	59 (16.1)	48 (13.9)	57 (15.8)	
Focus on most aspects	102 (27.9)	91 (26.3)	95 (26.3)	
Focus on all aspects	120 (32.8)	131 (37.9)	119 (33.0)	

[a] Significance across arms using one way ANOVA