
Article:
Mulhern, B., Bansback, N., Hole, A.R. et al. (1 more author) (2017) Using Discrete Choice Experiment with duration to model EQ-5D-5L health state preferences: Testing experimental design strategies. Medical Decision Making, 37 (3). pp. 285-297. ISSN 1552-681X

https://doi.org/10.1177/0272989X16670616

Using Discrete Choice Experiments with duration to model EQ-5D-5L health state preferences: Testing experimental design strategies

Brendan Mulhern[1,2] (MRes), Nick Bansback[3] (PhD), Arne Risa Hole[4](PhD), Aki Tsuchiya[2,4] (PhD),

1 University of Technology Sydney, Centre for Health Economics Research and Evaluation, Ultimo, Australia
2 School of Health and Related Research, University of Sheffield, UK
3 School of Population and Public Health, University of British Columbia, Canada
4 Department of Economics, University of Sheffield, UK

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: Designing DCE with duration

Earlier versions of this paper was presented at the UK Health Economists' Study Group conference, June 2014, Glasgow, the EuroQol Group Plenary, Sept 2014, Stockholm, and the International Academy of Health Preference Research, Amsterdam, November 2014.

Word count: 5,200

Financial support for this study was provided entirely by a grant from the EuroQol Group. The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report. The following authors are members of the EuroQol Group: Brendan Mulhern; Nick Bansback; Aki Tsuchiya.

Abstract

Background: Discrete choice experiments incorporating duration can be used to derive health state values for EQ-5D-5L. Methodological issues relating to the duration attribute and the optimal way to select health states remain. The aims of this study were to: test increasing the number of duration levels and choice sets where duration varies (aim 1); compare designs with zero and non-zero prior values (aim 2); and investigate a novel two-stage design to incorporate prior values (aim 3).

Methods: Informed by zero and non-zero prior values, two efficient designs were developed, each consisting of 120 EQ-5D-5L health *profile* pairs with one of six duration levels (aims 1 and 2). Another 120 health *state* pairs were selected, with one of six duration levels allocated in a second stage based on existing estimated utility of the states (aim 3). An online sample of 2,002 members of the UK general population completed 10 choice sets each. Differences across the regression coefficients from the three designs were assessed.

Results: The zero prior value design produced a model with coefficients that were generally logically ordered, but the non-zero prior value design resulted in a set of less ordered coefficients where some differed significantly. The two-stage design resulted in ordered and significant coefficients. The non-zero prior value design may include more "difficult" choice sets, based on the proportions choosing each profile.

Conclusions: There is some indication of compromised "respondent efficiency", suggesting that the use of non-zero prior values will not necessarily result in better overall precision. It is feasible to design discrete choice experiments in two stages by allocating duration values to EQ-5D-5L health state pairs based on estimates from prior studies.

Introduction

Generic preference based measures such as the EQ-5D[1] and SF-6D[2,3] can be used in the calculation of Quality Adjusted Life Years (QALYs) to inform the economic evaluation of health interventions. The 'quality' adjustment is based on the preferences of the general population for health states described by the measure, and anchored at 1 for full health and 0 for being dead. Traditionally, preferences have been derived using 'iterative' choice-based techniques such as the Time Trade Off (TTO),[4] which has been widely used to produce utility values for the EQ-5D-3L,[5] and the Standard Gamble (SG), which was used to value the SF-6D.[3]

More recently, studies have explored the use of Discrete Choice Experiments incorporating duration (DCE_{TTO}) to estimate utility values anchored on the full health - dead scale. DCE_{TTO} has been shown to produce logically consistent value sets for a range of descriptive systems in the UK, Canada and Australia, including the EQ-5D-3L,[6,7] the EQ-5D-5L[8-10] and the SF-6D.[11] For example, a methodological study based in the UK valued EQ-5D-5L online using DCE_{TTO} with 1,799 members of the general population completing 15 DCE_{TTO} choice sets s each (the PRET-AS study).[9,10] The results demonstrated generally logically ordered utility decrements within each dimension. The overall health state values ranged from 1 (for state 11111, i.e. no problems on any dimension) to -0.845 (for state 55555, i.e. worst health state with extreme problems on all dimensions) with approximately a third of the states modelled as worse than being dead. The distribution of values was unimodal, and the difference in health state value between the best (11111) and the next best (12111) state was small in comparison to, for example, the UK EQ-5D-3L TTO value set.[5]

Although there is evidence for the feasibility of DCE_{TTO}, important methodological and design issues remain. One issue relates to the duration attribute. DCE_{TTO} data are modelled by examining the interaction between the level in each dimension with duration (see Bansback et al[6] for details), and therefore the accuracy of this parameter affects the overall precision of the anchored value set. The standard error of the duration coefficient observed in the PRET-AS study above was relatively large. Since only three levels of duration (1, 5, and 10 years) were used (generating six unique ratios combining different levels of duration), and with only 15% of pairs having different durations across the options, there is

scope to explore this dimension further. Increasing the number of levels and combinations of duration may allow for a wider range of variation across scenarios.

A second issue relates to the efficiency of the DCE design. Previous designs have assumed no prior knowledge regarding the value of the parameters (i.e. it used 'zero prior' values). However, these previous studies provide some information about the parameter values, and therefore could be used to improve the statistical efficiency of experimental designs, requiring fewer respondents or responses for each respondent.[12] This could be approached in two ways. One approach would be to include the coefficients from an existing study as 'non-zero prior values' to select the health *profile* pairs (i.e. pairs of scenarios described in terms of EQ-5D-5L states with specified durations) in an efficient design. The expectation associated with the incorporation of prior information in the design concerns the efficiency of the design (i.e. the precision of the coefficient estimates), rather than the estimated coefficients themselves. At the same time it should be noted that a more statistically efficient design may mean that each choice is more difficult for the respondent, which may lower respondent efficiency (see Johnson et al[12] for a discussion of statistical versus respondent efficiency). The other approach would be to generate the design in two stages: first to select pairs of EQ-5D-5L *states* using zero-priors in an efficient design, and then to use predicted utility values from previous studies to combine the states with duration values to form the health state *profile.* The duration values are chosen so that the expected distribution of responses to each health state profile pair will fall around a given proportion.[12,13]

Based on these issues, this paper reports a study that seeks to improve the precision of future DCE_{TTO} designs by addressing three aims:

1. To increase the number of levels of the duration attribute of the DCE_{TTO} design, and hence the proportion of choice sets where duration varies (in comparison to the earlier UK EQ-5D-5L study, PRET-AS).
2. To compare the impact of using a design with zero prior values and a design with non-zero prior values on the preciseness of the coefficient estimates and the extent to which they are logically ordered.

3. To examine an alternative two-stage approach to DCE_{TTO} design (allocating duration based on the estimated value of the states), and compare this with the (one-stage) designs that use zero and non-zero prior values.

Methods

Choice set design

The DCE_{TTO} choice sets used in this study are based on EQ-5D-5L which was developed from the EQ-5D-3L instrument to improve the sensitivity of the descriptive system and standardise the wording across the dimensions.[14] EQ-5D classifies health states across five dimensions (mobility, self-care, usual activities, pain/discomfort and anxiety/depression). In contrast to the three-level version, the EQ-5D-5L adds two intermediate levels of severity (none, slight, moderate, severe, extreme/unable). The DCE_{TTO} profiles used in this study consist of "you" living in a particular EQ-5D-5L 'health scenario' for one of six levels of duration T (where T = 6 months, 1, 2, 4, 7 and 10 years) followed by death. This generates 20 unique ratios of duration which allows for trading off across a wider range of more subtle and less dramatic than with the six ratios possible in PRET-AS. Respondents were asked which health profile they prefer. Figure 1 displays a screenshot of a choice set.

Experimental design

The EQ-5D-5L describes 3,125 possible health states, and combining these with six levels of duration amounts to 18,750 possible health profiles and therefore over 350m possible profile pairs. The general guidance for pairwise DCE choice design is that the minimum number of pairs required is the number of parameters to be estimated. As described by Bansback and colleagues,[9] DCE_{TTO} models the pairwise choice data in terms of interactions between the health state levels (categorical) and duration (continuous), and therefore the number of parameters is 21 (interactions between each of the EQ-5D-5L level dummies and continuous duration ((5-1) x 5 x 1=20), plus continuous duration). However, further analyses involving EQ-5D-5L main effects and quadratic duration (the above 21, plus EQ-5D-5L main effects 5x(4-1)=20, interactions between these and duration squared 20x1=20, and duration squared) would require 62 parameters.

Three designs are used in this paper. The first two are based on selecting pairs of health *profiles* in a single step. Type Ia was designed to address aim 1, using zero prior values for the parameter values. Type Ib used non-zero prior values taken from the PRET-AS study, and was compared to Type Ia to address aim 2. Both designs were based on the D-efficiency criterion, which is a summary measure of the precision that would be achieved from a given design given the prior values. The Type Ia and Ib designs include 69 (58%) and 86 (72%) pairs respectively where duration differs across the profiles.

The Type II design addressed aim 3, and involved firstly generating 120 EQ-5D-5L health *state* pairs and secondly selecting the duration (either 6mth, 1yr, 4yr, 7yr or 10yr) for each profile. This was done using coefficients generated from the PRET-AS study in an optimization procedure to select the durations which resulted in an expected split of 70% vs 30% in the choice between health profiles. These values are chosen as they lie within the range identified by Kanninen[13] for the optimal probability split for DCE designs (see also Fowkes and Wardman[16]). In total, 111 (93%) of the choice sets achieved a predicted probability between 66% and 74%. The remaining nine choice sets had EQ-5D-5L scenarios which were too different to achieve a 70%-30% split, and were allocated either a matched duration of 10 years, or different durations of 10 years and 6 months. In total, 100 (80%) of the choice sets were assigned different durations. Again, no restrictions were applied.

For the Type I designs the experimental design programme Ngene[15] was used to select 120 pairs and allocate each pair to one of 12 blocks of 10 for each survey version. Stata was used to select the 120 pairs for the Type II design.

There are concerns that some EQ-5D-5L states are more difficult to imagine than others and some states may even appear to be "implausible".[5] However, since there is no agreed measure of how difficult a given health state is to imagine or any threshold along this measure beyond which states become unimaginable, no health state combinations were excluded from any of the designs.

Survey design, recruitment and the sample

Respondents were recruited from an existing commercial internet panel (IPSOS Observer), and were selected according to quotas based on the UK general population for age (across

five age groupings) and gender. First, potential respondents were invited by e-mail and accessed the survey webpage, where they read detailed project information and consented to take part. Those consenting to participate then completed questions on demographic background and self-reported health status, the Office of National Statistics (ONS) wellbeing questions,[17] and EQ-5D-5L for their own health. They were then presented with information about the DCE_{TTO} tasks including details about the EQ-5D-5L health dimensions, and instructions to imagine: that they would experience each health state for the period shown without relief or treatment; that death would be very swift and completely painless; and that they would have no other health problems besides what was indicated. This was followed by ten DCE_{TTO} choice sets. Respondents were screened out if they completed the survey in less than the minimum time of 2 minutes.

Analysis

DCE_{TTO} modelling

We followed the analysis described in detail in Bansback et al (2012).[6] Conditional logit regression[18] was used to estimate the coefficients of a utility function μ defined by a vector of four dummy variables for each EQ-5D-5L attribute $\mathbf{x}$ (no problems is the reference for each attribute) and life years t:

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

The data are a series of binary outcomes indicating respondent i's choices between profiles j = 1, 2. The coefficient β represents the value of living in full health for one year and is expected to be positive; λ represents the disutility of living with the specified set of EQ-5D-5L health problems ($\mathbf{x}$) for one year and thus is expected to be negative. The error term ε_{ij} is a random term which is assumed to be standard Type I extreme value distributed. In this paper we assume respondents do not discount future health and trade time in a constant proportion and we treat the life years attribute as continuous.

As is shown in Bansback et al,[6] the value for each health state $\mathbf{x_j}$ can be anchored on the health utility scale (V) from the estimates as the additive value of full health (fixed at 1) and

the disutility associated with each particular health state. This can be expressed as the ratio of $\hat{\lambda}$ and $\hat{\beta}$, multiplied by $\mathbf{x_j}$, such that:

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

Thus, for full health this value is 1 (since when $\mathbf{x}$ = 0 the disutility is 0), whilst for other health states the effect of $\frac{\hat{\lambda}\prime}{\hat{\beta}}\boldsymbol{x}_j$ is negative (i.e. representing a decrement from full health), and can even be less than -1 (or, in other words the value of the health state, V_j, can be negative) indicating a state worse than being dead.

The results are reported in terms of the 'unanchored' coefficients (β and λ) which are on a latent scale and therefore their magnitudes are not directly comparable across models, and the 'anchored' coefficients (λ/β) which are on the scale with 1 for full health and 0 for being dead and are therefore comparable across models.

Assessing the duration coefficient and comparing the designs

To assess the duration attribute (aim 1) we examined the sign and ordering of the coefficients using the Type Ia results. The sign of the duration coefficient should be positive (as utility increases with the time spent living in full health), while the sign of the interaction coefficients on the levels of each dimension are expected to be negative since they are all worse than the baseline (which is level 1, no problems). Furthermore, the levels in each dimension should have a logical ordering, whereby more severe levels should have larger decrements from the baseline. This was compared to the model from the PRET-AS study.[9,10]

To assess the impact of non-zero prior values (aim 2) and the two-stage design (aim 3), we first compared the models produced by the three Types of design, by comparing differences in the sign, ordering and significance of the coefficients across the models, and the standard errors (of the anchored coefficients only). We next examined differences in the anchored scales by assessing the predicted values of six select EQ-5D-5L states (five very mild states and the very worst state).

To gauge response efficiency, the difficulty of choice sets within each design was considered. The difficulty of a choice set was proxied by the distribution of respondents

across the two options within the pair: an easy choice set is one where one option emerges as the clear majority choice by a wide margin (e.g. 90% of respondents choose the majority choice), while a difficult choice set is one where the majority choice has a much narrower margin (e.g. 55% of respondents choose it). The distribution of the 120 choice sets across different levels of difficulty was compared across the three Types.

The three designs are compared in two further analyses. First, we used the approach proposed by Swait and Louviere[19] to test the null hypothesis that preferences are heterogeneous across the three sub-samples. The likelihood-ratio test statistic is given by $LR = -2(LL_R - LL_U)$ where LL_R is the log-likelihood of a model estimated on the pooled sample which allows for scale differences but assumes that the value of living in full health for a specified duration (β) and the disutility of an EQ-5D-5L health state for a certain duration (λ) do not vary across the sub-samples (this is defined as the restricted model). LL_U is the sum of the log likelihoods of three conditional logit models estimated on the sub-samples, which together form the unrestricted model (which allows for variation in preferences across the three sub-samples). The restricted model is estimated using the *clogithet* Stata module.[20,21]

Furthermore, we examined non-trading with respect to duration, defined as respondents consistently choosing the health profiles with the longest duration. While such lexicographic behaviour may reflect "genuine" preferences,[22] it may also indicate a simplifying heuristic is being used which results in non-trading across dimensions. By Type, we examined the number of times the profile with the longer duration was chosen, and for each Type and survey version, the number of pairs where duration differs, and the number of respondents always selecting the option with a longer duration when available to them in the version they completed. Respondents who never chose the profile with the shorter duration were defined as "duration-based non-traders" and their demographic characteristics were compared to the rest of the sample. DCE_{TTO} models excluding non-traders were compared to the full sample models.

Results

Response rate and demographics

For Type Ia, approximately 12,000 panel members were invited to take part by e-mail, with 1,618 (13%) accessing the survey. Of these, 340 (21%) were screened out due to full quotas, 433 (27%) did not pass the information and consent pages, 41 (3%) dropped out during the survey, two (0.001%) completed in less than 2 minutes, with 802 (50%) fully completing the survey. Each of the 12 survey versions including ten choice sets was completed by between 48 (6%) and 81 (10%) respondents.

Similar figures apply for Type Ib. Approximately 11,000 members of the online panel were invited to take part, and 1,567 (14%) accessed the survey. Of these, 340 (22%) dropped out due to full quotas, 377 (24%) did not consent to take part (or did not pass the information page), 50 (3%) dropped out during the survey, and 800 (51%) respondents fully completed the survey (none were excluded for taking less than 2 minutes). Between 56 (7%) and 77 (10%) completed each of the 12 survey versions.

For Type II, approximately 3,800 members of the panel were invited to take part by e-mail, and 643 (17%) accessed the survey. Of these, 222 (35%) did not consent to take part (or did not pass the information page), 21 (3%) dropped out during the survey, and 400 (62%) respondents fully completed the survey (none were excluded on the basis of taking less than 2 minutes). Between 30 (8%) and 40 (10%) completed the 12 survey versions.

The demographics of the samples are reported in Table 1, and are similar across all survey versions. Type Ib has significantly more respondents in the best EQ-5D-5L health state (11111) than Type Ia but self-reported health is not significantly different. There are no other significant differences in demographic characteristics across the groups. The characteristics of the PRET-AS sample are included for comparison where available.[10]

DCE_{TTO} models

Table 2 reports the unanchored coefficients for the Types Ia, Ib, II and the PRET-AS samples, and Figure 2 displays the anchored coefficients for the same designs. The significance levels of the coefficients in Table 2 are in comparison to level 1 (reference), and to the level directly before (one level milder). For the Type Ia design with zero prior values there are small non-significant inconsistencies between levels 2 and 3 of the pain/discomfort and

anxiety/depression dimensions. This indicates that slight and moderate problems in these dimensions are valued on average as being no different in terms of severity.

The Type Ib design with non-zero prior values has more non-significant inconsistencies between mobility levels 4 and 5; and usual activities levels 1, 2 and 3 (where the coefficients for levels 2 and 3 are both positive leading to a non-significant increase in utility as health level decreases). Levels 1, 2 and 3 of the pain/discomfort dimension are also inconsistent, where level 2 has a non-significant positive coefficient. Level 3 has a non-significant negative coefficient as expected, but the difference between the two coefficients is significant.

For Type II, mobility level 2 has the 'wrong' sign but is non-significant. Regarding the earlier PRET-AS study, the design includes non-significant coefficients for mobility level 2 and self-care level 2. All the remaining coefficients in PRET-AS are logically ordered and significant.

The standard errors of the anchored model coefficients differ across the designs. The Type Ia design standard errors (range 0.021 to 0.025) are approximately half the size of the Type Ib standard errors (range 0.039 to 0.049) indicating increased precision. The Type II standard errors are in the same range as the Type Ia design (from 0.021 to 0.026) with approximately half the sample size, and the PRET-AS errors range from 0.011 to 0.015 (with a substantially larger number of observations).

The bottom rows of Table 2 display the predicted anchored utility values for six health states for the three designs. The Type Ib and Type II model produces utility values above 1 (that are not significantly different to 1), but this is not the case for Type Ia. PRET-AS also produces an estimate above 1 (not significant). Regarding the relative ranking of the five mildest states, there is no clear pattern across the four models. The value for the worst state (55555) ranges from -0.852 (Type Ib) to -0.706 (Type Ia).

Figure 2 compares the anchored utility decrements produced for each design. The similarity of the decrements across designs varies across the dimensions.

<u>Objective (1) – Investigating the duration attribute</u>

Table 2 illustrates that for the Type Ia design, increasing the number of duration levels and the pairs where duration varies (in comparison to the PRET-AS design) still produces a model with generally logically ordered coefficients. There are differences with the PRET-AS model, where the difference between the slight and moderate severity levels is smaller and the anchored coefficients have smaller standard errors. However, note that the PRET-AS model is based on more than three times the number of observations.

Objectives (2) and (3) – Comparing the different DCE design approaches

Table 2 shows that the Type Ib design has more evidence of disordering than Type Ia. For example, there is a disordering between the coefficients for levels 1 and 2 of the pain/discomfort dimension where level 2 has a positive coefficient. This would mean an increase from no pain/discomfort to slight pain/discomfort would lead to an increase in utility. However, none of the disordering is statistically significant. The standard errors of the anchored coefficients (which are directly comparable) are larger for Type Ib than Type Ia. The Type II model displays a high level of ordering with half the number of observations of Types Ia and Ib.

Figure 3 is a set of histograms illustrating the distribution of the 120 choice sets across five classes of difficulty, ranging from where 90-100% of respondents choose the majority choice to where 50-59% of respondents do. For example in the Type Ia dataset, 98.6% of respondents preferred the health state 32131 for 4 years over 55354 for 4 years. This was defined as easier than the choice between 25111 for one year and 33451 for 7 years, where 52% and 48% preferred each profile respectively. Panel (a) is for Type Ia with zero prior values, panel (b) is for Type Ib with non-zero prior values, and panel (c) for Type II. The patterns across the three Types are clearly different: Type Ia has a relatively uniform distribution of choice sets across the five classes of difficulty; Type Ib has a distinctly upward sloping pattern, with 40% of the choice sets in the most difficult class; and Type II is clearly unimodal, but the mode at 80%> is less difficult than the intended 70%-30% split.

Further analyses – heterogeneous preferences across sub-samples

Table 3 reports the model estimated on the pooled (Types Ia, Ib and II) dataset, which allows for scale differences but assumes that the value of living in full health for a specified

duration (β) and the disutility of an EQ-5D-5L health state for a certain duration (λ) do not vary across the samples. The likelihood ratio statistic is 134.85, which implies that the null hypothesis that the parameters are equal across the groups is rejected. The results suggest that the scale is lowest for Type Ib, followed by Type Ia and Type II. This indicates that the choices are more "noisy" for Type Ib, with Type II the least noisy.

Further analyses – duration-based non-trading

The Type Ia design includes 68 (56.7%) choice sets where duration varies resulting in 4567 observations. Of these, 2847 (62.3%) result in the respondent choosing the profile with the longer duration. The equivalent numbers for Type Ib are 86 (71.7%) choice sets with 5785 observations, leading to 2987 (51.6%) with the longer duration chosen. For Type II, 101 (84.2%) choice sets differ, with 3326 observations, and 2126 (63.9%) with the longer duration chosen. Table 4 reports, for each Type and survey version: the number of pairs where duration differs; and across each block and overall, the number of duration-based non-traders: i.e. respondents never selecting the option with a shorter duration when available. At the individual respondent level, for Type Ia, 130 (16.2%) are non-traders: i.e. always chose the option with the longer duration, for Type Ib this was 68 (8.5%), and for Type II was 38 (9.5%). The proportions of non-traders differs significantly between Type Ia and Type Ib ($p<0.000$) but not between Type Ib and Type II ($p=0.648$).

The demographic characteristics of the duration-based non-traders and the rest do not differ. Analysis comparing the full sample models with those excluding non-traders indicates limited differences, with the same pattern of coefficient ordering (and disordering) within each dimension demonstrated. There are minor differences in the anchored value ranges (with the models excluding non-traders having a value for 55555 between 0.1 and 0.2 lower for each of the designs). (Details available from corresponding author.)

Discussion

This paper presents the results of a study addressing three design issues relating to using DCE_{TTO} to value health states from descriptive systems such as the EQ-5D-5L: the effect of

the duration attribute (aim 1); the effect of using non-zero prior values in the design (aim 2); and the effect of a two-stage design (aim 3).

Regarding aim 1, the results demonstrate that increasing the number of duration levels and pairs where duration varies (in comparison to an earlier DCE_{TTO} study with EQ-5D-5L in the UK) is feasible, and produces a model with coefficients that are generally logically ordered. Other DCE_{TTO} studies have also used more than three levels and found generally consistent models.[6-8,11] It is possible that using more levels will increase the validity of responses as it enables for smaller differences between life years in choice sets. Using more levels of duration that have closer ratios may make it less likely that respondents will choose on the basis of duration alone (i.e. lexicographic decision making). Even with the closer ratios between duration levels designed in this study (including 10 years vs 7 years; and 7 years vs 4 years with the smallest ratios), the results suggest that there is minority of respondents who choose based on duration only throughout, and this varies across the designs.

Although the results were compared with the earlier UK EQ-5D-5L PRET-AS study, it should be noted that the duration attribute is not the only difference across the two studies. Other differences include the DCE design (for PRET-AS we used the modified Fedorov algorithm implemented in Stata while the Type I designs in the current study used the Ngene swapping algorithm), the sample size (1,799 for a single design in PRET-AS; 2,002 across three designs for the current study), and the internet panel used to provide the sample. Respondents from different internet panels may exhibit different behaviours based on the criteria for panel membership and recruitment strategies. The number of DCE_{TTO} choice sets per respondent was also different (15 PRET-AS vs. 10 current study). Since the results of PRET-AS suggested that models using ten pairs were more consistent than those using 15,[9,10] the current study uses ten. Other studies in this area have used similar numbers of choice sets and produced logically consistent models.[6-8,11]

Regarding aim 2, the model estimated on the non-zero prior value design data was more inconsistent than the model estimated using the zero prior value data, with less precise coefficient estimates. To our knowledge this is the first comparison of zero and non-zero prior value designs using DCE_{TTO}, and the findings are unexpected, since the introduction of the non-zero prior values should, in theory, improve the precision of the parameter

estimates. While it is the case that the use of prior values that differ substantially from the true parameter values would lead to an inefficient design, our prior values are taken from an earlier published study (PRET-AS)[9,10] which also used DCE_{TTO} to estimate values for the EQ-5D-5L in the UK. On the other hand, it seems to be the case that the non-zero design resulted in a larger number of choice sets where the proportions of respondents across the options is closer. It is reasonable to assume that such choice sets are more difficult and challenging for the respondent to complete, resulting in larger estimation errors. The inclusion of some choice sets that contribute less towards improving the precision of the parameter estimates, but allow the respondents to remain focused by not overburdening them, may be the reason that the Type 1a design is found to be the most efficient overall. Nevertheless, it should also be noted that the drop-out rate from the survey amongst those who started it is no higher for Type Ib across the two types (both approximately 3%).

Regarding aim 3, the results demonstrate that the two-stage design of allocating duration based on the estimated health state severity is feasible, as it produces a generally logical and ordered model using the same number of pairs as Type Ia and Ib, but with half the sample size. This two-stage design is in effect very similar to what is proposed by Kanninen,[13] where the levels of a key attribute, in our case duration, is manipulated to achieve the required proportions based on the results from an earlier round of data collection. The positive results that we obtain mean that the two-stage design approach has potential, and could be tested further in comparison to other more established design methods.

Non-trading will impact on the DCE_{TTO} process and therefore the validity of the models produced. The analysis of duration-based non-trading suggests that some respondents may choose based on duration alone. Overall 236 (12%) of the 2,002 respondents always choose the option with longer duration, although the proportion differs across the designs indicating that the preferences for duration in comparison to the health dimensions may differ across the blocks. However, removing duration-based non-traders does not change the characteristics of the models. It is important to note that lexicographic behaviour may reflect genuine preferences. For example, always choosing the profile with longer duration could be a genuine preference based on religion or maximising time with family as

suggested by qualitative work.[10] This could also apply to the EQ-5D-5L dimensions (where respondents always choose the option with the least severe level on one dimension, and only consider a second dimension if there is a tie): these may be genuine preferences for that particular aspect of health which helps the respondent to decide between the profiles presented.

This study is not without drawbacks, which could be areas for future research. We cannot fully gauge respondent behaviour whilst completing the choice sets. Further work to try and understand how respondents complete the choice sets, and the impact of heuristics and response strategies such as those found in other DCE studies[10,23,24] would prove useful. Regarding the study design process, we designed one set of states for each of the two Type I designs, but we could have developed a range of designs and selected those with more similar numbers of pairs that displayed different durations (which the lexicographic choice analysis suggests is an important aspect of the process). We also do not know to what extent the non-zero prior values used were indicative of the preferences of the population that the tasks were administered to. This may limit the inferences that can be drawn about the use of prior values in DCE_{TTO} studies per se. However, the prior values were taken from an earlier valuation study of EQ-5D-5L using DCE_{TTO}, with a similar sample in terms of observable characteristics, using the same mode of administration, all of which would suggest high comparability. Testing other prior values from different studies, in different countries, and for different descriptive systems, may help establish the extent to which the use of non-zero prior values impacts on the consistency of the models produced and whether they are useful in the design of health state valuation studies using DCE_{TTO}.

Acknowledgements

This study was funded by the EuroQol Research Foundation. The views expressed do not necessarily reflect the views of the EuroQol Research Foundation. Ethics approval was obtained from the University of Sheffield Research Ethics Committee. We are grateful to Ben van Hout, Julie Ratcliffe, Elly Stolk, Richard DeAbreu Lourenco, and the three anonymous referees for *Medical Decision Making* for their comments, 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 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 JE, Roberts J. Estimating a preference-based index from the SF-12. Med Care. 2004;42(9):851-59.

4 Gudex C ed. Time Trade-Off User Manual: Props and Self-Completion Methods. CHE Occasional Papers 20, Centre for Health Economics, University of York;1994.

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

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, 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 Economics and 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 et al. Preparatory study for the re-valuation of the EQ-5D tariff: Methodology report. Health Technol Assess. 2014;18:12.

11 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.

12 Johnson FR et al. Constructing experimental designs for discrete-choice experiments: Report of the ISPOR Conjoint Analysis Experimental Design Good Research Practise Task Force. Value Health. 2013; 16:3-13.

13 Kanninen BJ. Optimal Design for Multinomial Choice Experiments. Journal of Marketing Research. 2002; 39(2):214-27.

14 Herdman M, Gudex C, Lloyd A 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.

15 Choice Metrics. Ngene [software for experimental design]. NGene 2012.

16 Fowkes AS, Wardman M. The design of stated preference travel choice experiments. J Transport Econ Policy 1988;13:27–44.

17 Dolan P, Metcalfe R. Measuring Subjective Wellbeing: Recommendations on Measures for use by National Governments. Journal of Social Policy. 2012;41(2):409-27.

18 McFadden DL. Conditional logit analysis of qualitative choice behavior. In Frontiers in Econometrics, Zarembka P (ed.). Academic Press: New York: 1974.

19 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.

20 Hole AR. CLOGITHET: Stata module to estimate heteroscedastic conditional logit models. Statistical Software Components S456737, Boston College Department of Economics;2006.

21 Hole AR. Small-sample properties of tests for heteroscedasticity in the conditional logit model. Economics Bulletin. 2006;3(18):1-14.

22 Lancsar E, Louviere J. Deleting 'irrational' responses from discrete choice experiments: a case of investigating or imposing preferences? Health Econ. 2006;15(8):797-811.

23 Hole AR, Norman R, Viney R. Response patterns in health state valuation using endogenous attribute attendance and latent class analysis. Health Economics. 2014; DOI: 10.1002/hec.3134.

24 Ryan M, Watson V, Entwistle V. Rationalising the ‘irrational’: A think aloud study of discrete choice experiment responses. Health Econ. 2009;18:321-36.