PharmacoEconomics
https://doi.org/10.1007/s40273-018-0694-6

ORIGINAL RESEARCH ARTICLE

Severity-Stratified Discrete Choice Experiment Designs for Health State Evaluations

Sesil Lim[1,2] · Marcel F. Jonker[1,3,4] · Mark Oppe[1,5] · Bas Donkers[1,2] · Elly Stolk[1,5]

Abstract

Background Discrete choice experiments (DCEs) are increasingly used for health state valuations. However, the values derived from initial DCE studies vary widely. We hypothesize that these findings indicate the presence of unknown sources of bias that must be recognized and minimized. Against this background, we studied whether values derived from a DCE are sensitive to how well the DCE design spans the severity range.

Methods We constructed an experiment involving three variants of DCE tasks for health state valuation: standard DCE, DCE-death, and DCE-duration. For each type of DCE, an experimental design was generated under two different conditions, enabling a comparison of health state values derived from current best practice Bayesian efficient DCE designs with values derived from 'severity-stratified' designs that control for coverage of the severity range in health state selection. About 3000 respondents participated in the study and were randomly assigned to one of the six study arms.

Results Imposing the severity-stratified restriction had a large effect on health states sampled for the DCE-duration approach. The unstratified efficient design returned a skewed distribution of selected health states, and this introduced bias. The choice probability of bad health states was underestimated, and time trade-offs to avoid bad states were overestimated, resulting in too low values. Imposing the same restriction had limited effect in the DCE-death approach and standard DCE.

Conclusion Variation in DCE-derived values can be partially explained by differences in how well selected health states spanned the severity range. Imposing a 'severity stratification' on DCE-duration designs is a validity requirement.

Key Points for Decision Makers

Unstratified efficient design algorithms cannot guarantee adequate coverage of the severity range.

If health state selection bias occurs in DCE-duration studies, the derived values may be too low.

Sampling choice task from different severity strata is a way to prevent skewed designs and biased values.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s40273-018-0694-6) contains supplementary material, which is available to authorized users.

✉ Sesil Lim
lim@ese.eur.nl

1 Erasmus Choice Modelling Centre, Erasmus University Rotterdam, Rotterdam, The Netherlands

2 Erasmus School of Economics, Erasmus University Rotterdam, Rotterdam, The Netherlands

3 Erasmus School of Health Policy & Management, Erasmus University Rotterdam, Rotterdam, The Netherlands

4 Duke Clinical Research Institute, Duke University, Durham, NC, USA

5 EuroQol Research Foundation, Rotterdam, The Netherlands

1 Introduction

The use of the discrete choice experiment (DCE) has attracted researchers' interest as an alternative to more conventional techniques, such as the time trade-off (TTO) method, to derive quality-adjusted life years (QALYs) in health state evaluation. One of the merits of using DCE methodologies is that they improve the feasibility of valuation studies. In contrast to TTO valuations, which require

Published online: 21 July 2018

organizationally complex and costly face-to-face interviews, DCE valuation surveys can be self-administered [1, 2]. However, the use of a DCE for valuation introduces an unusual requirement in the DCE, namely that the estimated values are anchored at 1 for full health and 0 for death. The validity of the proposed approaches to achieve that has yet to be established. Two proposed strategies include the DCE-death and the DCE-duration approaches. However, results obtained from initial applications of those methods were markedly different. For instance, Norman et al. [3, 4] showed that DCE-duration approaches consistently produce lower values than DCE-death approaches. Also compared to conventional health valuation methods, DCEs have produced discrepant results. Craig et al. [5] and Jonker et al. [6] reported a considerably longer value range derived from their DCE approaches (minimum values < − 1.5) than the range obtained with the conventional TTO for EQ-5D (− 0.594 to 1.000) [7]. Researchers now aim to understand why.

We aim to contribute to the body of knowledge of how best to implement DCE methods for health state valuation, with a focus on strategies for the development of the experimental design. In this area, methodological advancements made best practice somewhat of a moving target. On top of that, best practice for choosing a strategy for designs may well be context dependent [8, 9]. Whereas some general considerations always apply, such as the importance of identification and statistical efficiency, other demands can be application specific. The latter may be the case in the field of health state valuation.

A popular approach for the construction of experimental designs in DCEs is the (Bayesian) efficient design approach. These designs exploit prior information to arrive at a design that produces small asymptotic standard errors. Because of the direct link between standard errors and sample size requirement, this is a desirable property [10]. Efficient designs have been frequently used in DCEs for health valuation [2, 6, 11, 12]. However, these designs are not without problems. A potential problem is that designs purely optimized for statistical efficiency can produce more difficult choice sets [13]. As a result, respondents might not always have a clear preference for any of the options, or they may be tempted to use simplifying decision rules that obscure their true preferences and cause bias [14]. A current line of research is whether such concerns can be addressed by introducing constraints on the design generation algorithms for DCEs, for example, by forcing attribute-level overlap in the constructed choice sets [15, 16]. Another potential problem is that the choice sets will not be selected at random, but rather chosen to support estimation of a proposed utility function [8, 17]. The algorithm will favor choice tasks that clearly reveal attribute trade-offs and avoid strongly dominant alternatives [18]. As a consequence, each health state has a different probability to be included in the choice tasks [17]. This can cause bias if decisions derived from included health states do not predict decisions about health states that have a lower inclusion probability due to model misspecification.

Currently, it is unknown whether this bias is a problem in DCEs designed to capture the value of health, but we hypothesize that it might be. Because optimization algorithms consider the level of utility balance for better statistical efficiency [19], the fact that the DCE-death and the DCE-duration approaches present respondents with very different fixed alternatives can cause other health states to be favored in the different approaches. To investigate the issue, we set up an experiment featuring EQ-5D-5L health states. First, we examine whether the current best practice efficient DCE designs (i.e. '*unstratified*' efficient designs) tend to favor a particular type of health states in the context of various DCE formats. Second, we investigate the sensitivity of estimated health state values to the potentially skewed selection of health states by comparing estimates derived from unstratified designs with those from DCE designs that satisfy the requirement that the set of selected health states has to span the entire severity scale (i.e. '*severity-stratified*' designs).

Table 1 Overview of the study arms

	Unstratified	Severity-stratified
Standard DCE	1	4
DCE-death	2	5
DCE-duration	3	6

DCE discrete choice experiment

2 Methods

To investigate the issues mentioned above, we proposed a strategy for generating severity-stratified designs and compared the severity-stratified designs to unstratified efficient designs on (1) health state selection for inclusion in DCE tasks and (2) values derived from the DCE tasks. We did this in the context of three different DCE formats: standard DCE, DCE-death, and DCE-duration. Table 1 shows the overview of the six study arms used in this study.

2.1 The Discrete Choice Experiment (DCE) Choice Tasks

Figure 1 provides an example of the three DCE formats. The health states were defined by the five dimensions of the EQ-5D-5L instrument: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. For each dimension, five

Fig. 1 Presentation of choice tasks: **a** standard DCE; **b** DCE-death; **c** DCE-duration. *DCE* discrete choice experiment

levels are used to describe the severity of impairment in monotonic order from 'no problems' (level 1) to 'extreme problems/unable' (level 5).

The *standard DCE* was a forced choice paired comparison between two health states where respondents were asked to choose between 10 years in health state A and 10 years in health state B. This task focused on the direct trade-off between the health state attributes and produces values on a latent scale.

In the *DCE-death* format, each choice task had three alternatives, A, B, and C, and respondents compared A to B and B to C using a so-called 'matched pairwise choice task' [6, 15, 20]. The A–B comparison resembled the standard DCE described above. The next question was a forced choice between 10 years in health state B versus immediate death (i.e., B–C comparison). Each choice task thus comprises two pairwise comparisons so that the number of observations will be twice as high as in the standard DCE. However, the cognitive burden is only marginally increased, because option B appears in both comparisons and option C is fixed and easy to imagine.

In the *DCE-duration* format, each choice task also had three alternatives, A, B, and C, that were compared using a matched pairwise choice task. Respondents were first asked to choose between 10 years in health state A and 10 years in health state B (i.e., A–B comparison), followed by the B–C comparison, where option C was always health state 11111 (i.e., no problems in any EQ-5D dimension) with a duration shorter than that of option B. Length of life in the perfect

health was restricted to 12 levels: 2, 4, 6 months and 1, 2, …, 9 years.[1]

In order to reduce task complexity and respondent burden, all choice tasks for the A–B comparisons included attribute-level overlap [6, 15]. For each pair of choices A and B, a minimum of two out of five dimensions were presented at the same level. In addition, combinations of the first level (no problem) of usual activities with the fifth level (extreme problems) of pain/discomfort and/or anxiety/depression were avoided to make health states easier to imagine and evaluate. Lastly, intensity color coding was used to further reduce task complexity. Imposing attribute-level overlap and color coding as well as excluding implausible states is currently best practice considering the reduced dropout rate and improved respondents' attribute attendance in DCE [15].

2.2 Experimental Designs With/Without Severity-Stratified Restriction

We implemented heterogeneous DCE design algorithms to create for each study arm a unique experimental design comprising 168 choice tasks, distributed over eight sub-designs [21]. The algorithm optimizes for Bayesian D error for the total design, while simultaneously optimizing for the Bayesian D errors of each of the eight sub-designs. In essence, this strategy produces a blocked design with eight blocks, where the design within each block is optimized in addition to the optimization of the overall design across blocks. A Latin hypercube sample optimized for maximum minimum distance between points and a greedy optimization algorithm was used to optimize the weighted averaged Bayesian D error with one-third of the weight assigned to the aggregated efficiency and two-thirds on the individual efficiencies of the sub-designs. Note that the design algorithm controlled for left–right randomization of the two states by including both options A and B in comparison with option C in the Bayesian design criterion, even though only one of the two choice options was presented (in random order) to the survey respondents.

To obtain an identifiable DCE design at the individual level, each sub-design contained 21 choice tasks, that is, the number of parameters to be estimated in a main effects model. As Bliemer and Rose [22] suggested, a DCE design optimized for a standard conditional logit model performs well for estimating panel mixed logit models. Therefore, the design was optimized for a conditional logit model, which reduced the computational burden substantially.

Whereas the full candidate set of all possible EQ-5D-5L health states (excluding 225 implausible health states) was used to optimize the unstratified DCE designs, the severity-stratified DCE designs used different candidate sets for each choice task. The creation of severity-stratified designs involved the following steps:

1. Informative priors were used to predict latent utility values for all health states, which were subsequently used to divide the health states into 21 severity strata (i.e., 3125/21 = ~ 148 states per stratum for each DCE format, thus comprising as many severity strata as there were choice tasks in each DCE design).
2. A total of 225 implausible health states were removed from the full set of 3125 health states and from each of the 21 strata.
3. For each stratum, candidate sets were constructed by creating all possible combinations of health states in the stratum with all other possible health states (i.e., $148 \times 2899/2 = \sim 0.2$ million).
4. The design algorithm created a DCE design that included exactly one choice task from each candidate set in each sub-design.

Prior values used for the DCE design optimization (and thus also in step 1) were obtained from previous research (based on an unstratified DCE design; unpublished to date), which contains 350 Dutch respondents for each DCE format. The design algorithm was implemented in Julia [23].

2.3 Statistical Analysis

To analyze the health state preferences, a mixed logit model[2] was used. For the standard DCE, the utility of the respondent i for the health state j in the choice task t was specified as:

$$U_{ijt} = X_{ijt}\beta_i + \epsilon_{ijt} \quad (1)$$

where X_{ijt} consists of 20 dummies for EQ-5D-5L instruments assuming the level 1 (no health problem) as the reference category for each dimension. The error ϵ_{ijt} is assumed independent and identically distributed with an extreme value

[1] Experimental designs to make respondents compare two health states within a vast gap of life years (i.e., 10 years in impaired health state vs. 2 months in the perfect health) may make the model sensitive to potential non-attendance to duration then inflate the impact of severely impaired health states. To clear this concern, we re-analyzed parameter values for the DCE-duration format excluding choice tasks containing 2 months of duration in perfect health (i.e., choice tasks with the biggest differences in duration). Our findings were qualitatively the same and quantitatively also almost identical.

[2] We used the mixed logit rather than other discrete choice models (i.e., conditional logit model) because it (1) does not exhibit the independence from irrelevant alternatives (IIA) property and the restrictive substitution pattern, (2) allows the correlation among coefficients, and (3) can take potential correlated responses across observations from the same individual into account in the repeated choices situation [24].

distribution, and the vector of individual-specific coefficients β_i is assumed to follow a multivariate normal distribution with the population mean μ and covariance matrix $\sum$, that is, $\beta_i \sim \text{MVN}(\mu, \sum)$. The same utility function was applied to the DCE-death approach; however, now X_{ijt} includes a dummy indicating death options.

For the DCE-duration approach, the utility was specified as the function of the product of the number of life years (T_{ijt}) and its observed characteristics (X_{ijt}) and their corresponding coefficients as follows:

$$U_{ijt} = \left(T_{ijt}X_{ijt}\right)\beta_i + \epsilon_{ijt} \quad (2)$$

Note that X_{ijt} consists of dummies for the EQ-5D-5L instrument and an intercept with the value 1, and the coefficient for the duration main effect represents the value respondent i assigns to living in perfect health for 1 year.

The specified models were estimated using the Bayesian Markov chain Monte Carlo (MCMC) methods as implemented in the *R* package *bayesm* [25]. Gibbs sampling was used to update μ and $\sum$, and a Metropolis–Hastings algorithm was used to update β_i. A multivariate normal prior (with a mean of zero and a variance of $100{\cdot}I$) was used for μ, and an inverse Wishart prior (with the dimension of $\sum$ plus 3 degrees of freedom, i.e., ν, and a location parameter νI) was used for $\sum$. Mean posterior estimates and 95% credible intervals were calculated by thinning the MCMC draws every fifth iteration for a total of 100,000 iterations. Convergence was established using visual inspection of chains and the convergence diagnostics as implemented in the R package *CODA* [26].

For testing hypotheses, the values for health states derived from the DCE-death and DCE-duration approaches were rescaled on the QALY scale where death has a value of 0 and full health a value of 1. To rescale the values, we divided the EQ-5D-5L parameters by the absolute value of the parameter for 'death' for DCE-death, and by the parameter value for 'duration' for DCE-duration for each draw of the posterior distribution of parameters. Next, the hypotheses that efficient design algorithms for the DCE-death and DCE-duration approaches tend to choose health states in skewed severity ranges was tested by comparing the distribution of values between designs with and without the severity stratification. For the hypothesis regarding the sensitivity of extrapolated health state values to the selection of health states, differences in values for the same health states between the designs with and without severity stratification were examined.

As DCEs aim to predict the choice probabilities of alternatives among given choice sets, we compared the predictive performance of estimates from the severity-stratified designs with those without that restriction using the mean errors (MEs), that is, the average deviation of predicted choice probability of a health state from the observed choice probabilities in each study arm. We used MEs to examine the direction of the bias that the estimates of each study arm produced.[3] Specifically, when comparing the impaired health states with the death or perfect health states, positive (negative) MEs regarding the impaired health state suggests that the predicted model of the study arm is likely to undervalue (overvalue) the disutility of impaired states so that it over-predicts (under-predicts) the choice probability of living in the impaired health condition compared with the actual observation. Cross validation of the MEs was done by applying the valuation function obtained in one study arm to the data of the other study arm of the same DCE format. The posterior predictive choice probability distribution was obtained by simulating mixed logit probabilities for each sample of the parameters in the posterior distribution, from which the distribution of MEs was inferred. Whether MEs were significantly different from zero was determined based on the 95% level credible intervals of the distribution of MEs.

2.4 Data Collection

The fieldwork was undertaken by Survey Sampling International (SSI) through an online platform during 2 weeks in December 2015. The target sample size was 3000 respondents (i.e., 500 respondents per study arm) representative of the Dutch general population regarding age, gender, and education. Respondents were recruited from SSI's online panel that contains representative panelists of the population aged 15–65 years and as many panelists aged over 65 years to resemble a nationally representative sample as much as possible. All respondents who gave consent for participation were asked about their demographics to enable stratification of the sample and were randomly assigned by SSI's survey management software to one of the six study arms and to one of eight sub-designs within that arm. After receiving the information regarding EQ-5D-5L instruments, respondents completed the 21 choice tasks in a random order. A total of 693 respondents who did not complete the tasks were excluded from the analysis. The average response time of respondents was 27 min (50% of respondents completed within 10 min).

[3] We also examined mean squared errors (MSE) by study arms (see the electronic supplementary material) to provide a full insight for the effect of imposing the severity-stratified restriction.

Table 2 Descriptive statistics of respondents

Characteristics	Subgroup	Value	
Overall sample vs. Netherlands population			
		Overall sample ($N=3122$)	Dutch population[a]
Age	15–20	197 (6.3%)	7.3%
	20–40	1006 (32.2%)	29.5%
	40–65	1390 (44.5%)	41.3%
	65–80	472 (15.1%)	16.8%
	Over 80	57 (1.8%)	5.2%
Gender	Female	1530 (49.0%)	54.9%
	Male	1592 (51.0%)	45.2%
Education	Low	1026 (32.9%)	30.1%[b]
	Medium	1402 (44.9%)	39.8%
	High	694 (22.2%)	30.1%
Self-rated health		0.819 ± 0.218	0.869 ± 0.170[c]
Standard DCE			
		Unstratified ($N=526$)	Severity-stratified ($N=520$)
Age	15–20	31 (5.9%)	30 (5.8%)
	20–40	165 (31.4%)	186 (35.8%)
	40–65	229 (43.5%)	223 (42.9%)
	65–80	88 (16.7%)	68 (13.1%)
	Over 80	13 (2.5%)	13 (2.5%)
Gender	Female	259 (49.2%)	254 (48.8%)
	Male	267 (50.8%)	266 (51.2%)
Education	Low	167 (31.7%)	161 (31.0%)
	Medium	235 (44.7%)	233 (44.8%)
	High	124 (23.6%)	126 (24.2%)
Self-rated health		0.815 ± 0.214	0.827 ± 0.218
DCE-death			
		Unstratified ($N=520$)	Severity-stratified ($N=518$)
Age	15–20	42 (8.1%)	36 (6.9%)
	20–40	171 (32.9%)	158 (30.5%)
	40–65	232 (44.6%)	235 (45.4%)
	65–80	67 (12.9%)	83 (16.0%)
	Over 80	8 (1.5%)	6 (1.2%)
Gender	Female	257 (49.4%)	267 (51.5%)
	Male	263 (50.6%)	251 (48.5%)
Education	Low	174 (33.5%)	177 (34.2%)
	Medium	232 (44.6%)	235 (45.4%)
	High	114 (21.9%)	106 (20.5%)
Self-rated health		0.821 ± 0.219	0.812 ± 0.221
DCE-duration			
		Unstratified ($N=521$)	Severity-stratified ($N=517$)
Age	15–20	25 (4.8%)	33 (6.4%)
	20–40	164 (31.5%)	162 (31.3%)
	40–65	238 (45.7%)	233 (45.1%)
	65–80	86 (16.5%)	80 (15.5%)
	Over 80	8 (1.5%)	9 (1.7%)
Gender	Female	252 (48.4%)	241 (46.6%)
	Male	269 (51.6%)	276 (53.4%)
Education	Low	174 (33.4%)	173 (33.5%)
	Medium	233 (44.7%)	234 (45.3%)

Table 2 (continued)

Characteristics	Subgroup	Value	
	High	114 (21.9%)	110 (21.3%)
Self-rated health		0.835 ± 0.203	0.807 ± 0.232

Education: low = primary and junior secondary education including both general and vocational schools; medium = senior secondary education including general and vocational schools, and pre-university; high = bachelor's, master's and doctoral degree. Self-rated health: average values of respondents' self-rated EQ-5D-5L health state that converted on QALY scale using a Dutch tariff [27]

DCE discrete choice experiment, *QALY* quality-adjusted life year

[a]Population rates in the Netherlands in 2017 were retrieved from the Statistics Netherlands (CBS) website. The distribution of age and gender were given for the population over 15 years old, while the distribution according to the level of education was available only for the population between 15 and 75 years old

[b]The population with unknown educational level (1.5%) was included. Also, the population with level 1 diploma of the senior secondary vocational school was included, while respondents with that characteristic belonged to the middle education group in the study

[c]Reference values for the Dutch general population based on 979 respondents [27]. Note that this paper did not collect data stratified by respondents' health state

3 Results

Table 2 shows the background characteristics of respondents. Respondents' characteristics are comparable to those in the Dutch population, and no significant imbalance in respondents' characteristics between unstratified and severity-stratified designs was found.

Figure 2 and Table 3 show the distributions of the modeled values for all EQ-5D-5L health states. As shown in Fig. 2, the distribution of states included in the design more closely followed the distribution for all health states when the severity stratification was applied for all DCE formats. It is most apparent for the DCE-duration format, where the unstratified design has a much more skewed distribution than the severity-stratified design.

For DCE-duration, the mean and the standard deviation (SD) of the distribution of health state values included in the unstratified design (i.e., the black bars) were 0.31 and 0.44, respectively, whereas those in the severity-stratified design were 0.09 and 0.35. A similar effect was hypothesized to exist in the DCE-death approach, but no strong evidence was found (mean 0.41 and SD 0.30 for the unstratified design; mean 0.40 and SD 0.26 for the severity-stratified design).

Table 4 shows the parameter estimates and corresponding 95% credible intervals for all six study arms. Almost all estimates are statistically significant, and all models resulted in logically consistent parameter estimates in the sense that worse levels of impairment are associated with larger utility decrements.

For the standard DCE, estimates in Table 4 are expressed on the latent utility scale, and therefore the obtained parameter estimates cannot be directly compared to the ones obtained in the other arms. However, the difference in scale between the unstratified and severity-stratified designs is very small, as can be seen from the values for state 55555 (the worst EQ-5D-5L state) and the fact that the 95% credible intervals overlap for all parameters when comparing the models from both designs. Similar results were observed for the DCE-death estimates on the QALY scale. For DCE-duration, estimated values for state 55555 are different and 95% credible intervals for several parameters (i.e., level 4 of 'Mobility' and level 5 of 'Self-care' and 'Anxiety/depression') do not overlap when comparing the unstratified design with the severity-stratified design.

Figure 3 shows scatter plots for each DCE format, comparing the values obtained by the designs with and without severity stratification. For the standard DCE and DCE-death formats, estimated values based on the severity-stratified design are close to those of the unstratified design. However, for the DCE-duration format, health state values of the severity-stratified design are higher than those of the unstratified design, especially on the range of states that are worse than death. The proportion of health states considered worse than death among 3125 health states was 56.0% for the unstratified design versus 42.8% for the severity-stratified design.

Table 5 shows MEs by study arm to compare the in-sample and out-of-sample forecasting accuracy of the severity-stratified design with those of the unstratified design. That is, column 4 shows the 'unstratified' model predicting the 'unstratified' observed choice probabilities; column 5 shows the 'severity-stratified' model predicting the 'unstratified' observed choice probabilities; column 6 shows the 'unstratified' model predicting the 'severity-stratified' observed choice probabilities; column 7 shows the 'severity-stratified' model predicting the 'severity-stratified' observed choice probabilities.

For DCE-death and DCE-duration, MEs were computed by separating health states into severity ranges: bad, medium, and better health state for QALY ≤ 0,

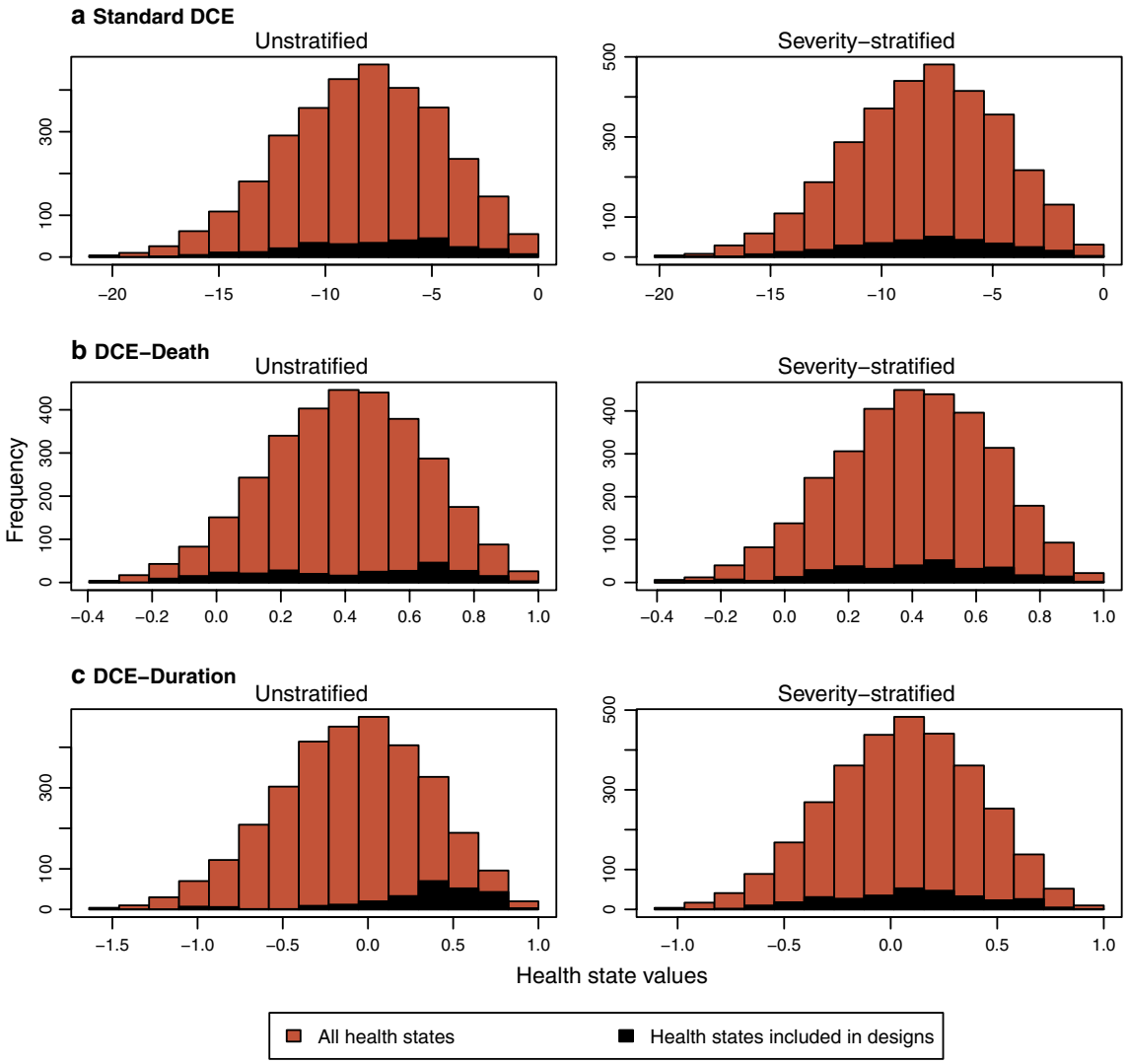

Fig. 2 Comparison of distributions of health state values between designs with and without severity-stratification. Distribution of modeled values for all possible EQ-5D health states (red bars) and modeled values for EQ-5D health states included in the designs (black bars). Health state values are on latent utility scales for the standard DCE (**a**), while they are on QALY scales for DCE-death (**b**) and DCE-duration (**c**). *DCE* discrete choice experiment, *QALY* quality-adjusted life year

Table 3 Distribution of health states selected for the designs over severity strata

	Standard DCE		DCE-death		DCE-duration	
	Unstratified	Severity-stratified	Unstratified	Severity-stratified	Unstratified	Severity-stratified
Number of unique health states	284 (100%)	319 (100%)	275 (100%)	319 (100%)	256 (100%)	310 (100%)
Better health state	–	–	127 (46.2%)	112 (35.1%)	88 (34.4%)	44 (14.2%)
Medium health state	–	–	114 (41.5%)	188 (58.9%)	125 (48.8%)	145 (46.7%)
Bad health state	–	–	33 (12%)	19 (6.0%)	43 (16.8%)	121 (39.0%)

Bad health states for QALY ≤ 0, medium health state for $0 <$ QALY ≤ 0.5, and better health state for $0.5 <$ QALY. Because the standard DCE produces values on a latent scale, the division in three severity strata was omitted for those designs

DCE discrete choice experiment, *QALY* quality-adjusted life year

Table 4 EQ-5D parameter estimates with 95% credible intervals on QALY scales for 6 study arms

Perfect health	Standard DCE[a]				DCE-death				DCE-duration			
	Unstratified (arm 1)		Severity-stratified (arm 4)		Unstratified (arm 2)		Severity-stratified (arm 5)		Unstratified (arm 3)		Severity-stratified (arm 6)	
	0.00	(N/A)	0.00	(N/A)	1.00	(N/A)	1.00	(N/A)	1.00	(N/A)	1.00	(N/A)
Mobility 2	− 0.19	(− 0.38, 0.03)	− 0.21	(− 0.43, − 0.01)	− 0.01	(− 0.03, 0.00)	− 0.03	(− 0.04, − 0.01)	− 0.08	(− 0.10, − 0.06)	− 0.06	(− 0.09, − 0.04)
Mobility 3	− 0.56	(− 0.77, − 0.33)	− 0.62	(− 0.86, − 0.38)	− 0.06	(− 0.07, − 0.04)	− 0.06	(− 0.07, − 0.04)	− 0.15	(− 0.17, − 0.13)	− 0.11	(− 0.14, − 0.09)
Mobility 4	− 2.34	(− 2.65, − 2.01)	− 2.23	(− 2.56, − 1.92)	− 0.15	(− 0.17, − 0.14)	− 0.17	(− 0.19, − 0.15)	− 0.33	(− 0.37, − 0.29)	− 0.26	(− 0.30, − 0.22)
Mobility 5	− 3.81	(− 4.24, − 3.37)	− 3.50	(− 3.98, − 3.07)	− 0.22	(− 0.25, − 0.20)	− 0.23	(− 0.25, − 0.20)	− 0.45	(− 0.51, − 0.40)	− 0.37	(− 0.42, − 0.32)
Self-care 2	− 0.07	(− 0.28, 0.15)	− 0.47	(− 0.69, − 0.26)	− 0.02	(− 0.04, − 0.01)	− 0.03	(− 0.04, − 0.01)	− 0.08	(− 0.10, − 0.06)	− 0.07	(− 0.10, − 0.05)
Self-care 3	− 0.82	(− 1.05, − 0.58)	− 0.90	(− 1.11, − 0.69)	− 0.08	(− 0.10, − 0.06)	− 0.05	(− 0.07, − 0.04)	− 0.13	(− 0.16, − 0.11)	− 0.13	(− 0.16, − 0.10)
Self-care 4	− 2.25	(− 2.55, − 1.96)	− 2.01	(− 2.31, − 1.73)	− 0.14	(− 0.16, − 0.13)	− 0.16	(− 0.18, − 0.14)	− 0.31	(− 0.35, − 0.27)	− 0.27	(− 0.31, − 0.23)
Self-care 5	− 2.89	(− 3.26, − 2.53)	− 2.86	(− 3.24, − 2.50)	− 0.19	(− 0.21, − 0.17)	− 0.18	(− 0.21, − 0.16)	− 0.40	(− 0.46, − 0.36)	− 0.31	(− 0.36, − 0.27)
Usual activities 2	− 0.41	(− 0.62, − 0.21)	− 0.57	(− 0.78, − 0.36)	− 0.03	(− 0.05, − 0.02)	− 0.04	(− 0.05, − 0.02)	− 0.05	(− 0.07, − 0.03)	− 0.08	(− 0.11, − 0.06)
Usual activities 3	− 0.90	(− 1.11, − 0.69)	− 0.92	(− 1.17, − 0.69)	− 0.09	(− 0.11, − 0.07)	− 0.07	(− 0.09, − 0.06)	− 0.09	(− 0.12, − 0.07)	− 0.13	(− 0.15, − 0.10)
Usual activities 4	− 2.49	(− 2.79, − 2.19)	− 2.55	(− 2.88, − 2.24)	− 0.20	(− 0.22, − 0.18)	− 0.18	(− 0.19, − 0.15)	− 0.27	(− 0.31, − 0.23)	− 0.25	(− 0.29, − 0.21)
Usual activities 5	− 3.25	(− 3.62, − 2.86)	− 3.63	(− 4.06, − 3.22)	− 0.25	(− 0.27, − 0.23)	− 0.25	(− 0.27, − 0.22)	− 0.42	(− 0.48, − 0.37)	− 0.35	(− 0.41, − 0.31)
Pain/discomfort 2	− 0.29	(− 0.50, − 0.08)	− 0.46	(− 0.67, − 0.24)	− 0.03	(− 0.04, − 0.01)	− 0.04	(− 0.06, − 0.03)	− 0.08	(− 0.10, − 0.06)	− 0.11	(− 0.13, − 0.09)
Pain/discomfort 3	− 1.08	(− 1.31, − 0.84)	− 1.14	(− 1.39, − 0.89)	− 0.10	(− 0.12, − 0.09)	− 0.10	(− 0.12, − 0.08)	− 0.16	(− 0.18, − 0.13)	− 0.15	(− 0.17, − 0.12)
Pain/discomfort 4	− 3.27	(− 3.70, − 2.85)	− 3.28	(− 3.70, − 2.89)	− 0.24	(− 0.27, − 0.22)	− 0.26	(− 0.29, − 0.23)	− 0.42	(− 0.47, − 0.37)	− 0.37	(− 0.42, − 0.32)
Pain/discomfort 5	− 5.31	(− 5.89, − 4.74)	− 5.08	(− 5.68, − 4.54)	− 0.37	(− 0.40, − 0.34)	− 0.39	(− 0.43, − 0.36)	− 0.66	(− 0.74, − 0.58)	− 0.53	(− 0.61, − 0.47)
Anxiety/depression 2	− 0.76	(− 0.98, − 0.54)	− 0.71	(− 0.91, − 0.50)	− 0.05	(− 0.07, − 0.04)	− 0.06	(− 0.08, − 0.05)	− 0,09	(− 0.12, − 0.07)	− 0.09	(− 0.12, − 0.07)
Anxiety/depression 3	− 1.37	(− 1.66, − 1.08)	− 1.12	(− 1.40, − 0.86)	− 0.11	(− 0.13, − 0.09)	− 0.10	(− 0.12, − 0.08)	− 0,19	(− 0.22, − 0.16)	− 0.17	(− 0.21, − 0.14)
Anxiety/depression 4	− 3.66	(− 4.12, − 3.21)	− 2.96	(− 3.40, − 2.55)	− 0.26	(− 0.29, − 0.24)	− 0.24	(− 0.27, − 0.21)	− 0,40	(− 0.46, − 0.35)	− 0.37	(− 0.42, − 0.32)
Anxiety/depression 5	− 5.84	(− 6.53, − 5.16)	− 5.15	(− 5.81, − 4.55)	− 0.36	(− 0.40, − 0.33)	− 0.36	(− 0.39, − 0.32)	− 0,70	(− 0.78, − 0.62)	− 0.54	(− 0.62, − 0.47)
State 55555	− 21.1	(− 23.4, − 19.2)	− 20.2	(− 22.1, − 18.2)	− 0.40	(− 0.48, − 0.32)	− 0.41	(− 0.50, − 0.32)	− 1.64	(− 1.93, − 1.38)	− 1.11	(− 1.37, − 0.89)

DCE discrete choice experiment, *N/A* not applicable, *QALY* quality-adjusted life year

[a]For standard DCE, parameter estimates and 95% credible intervals are reported on latent utility scales

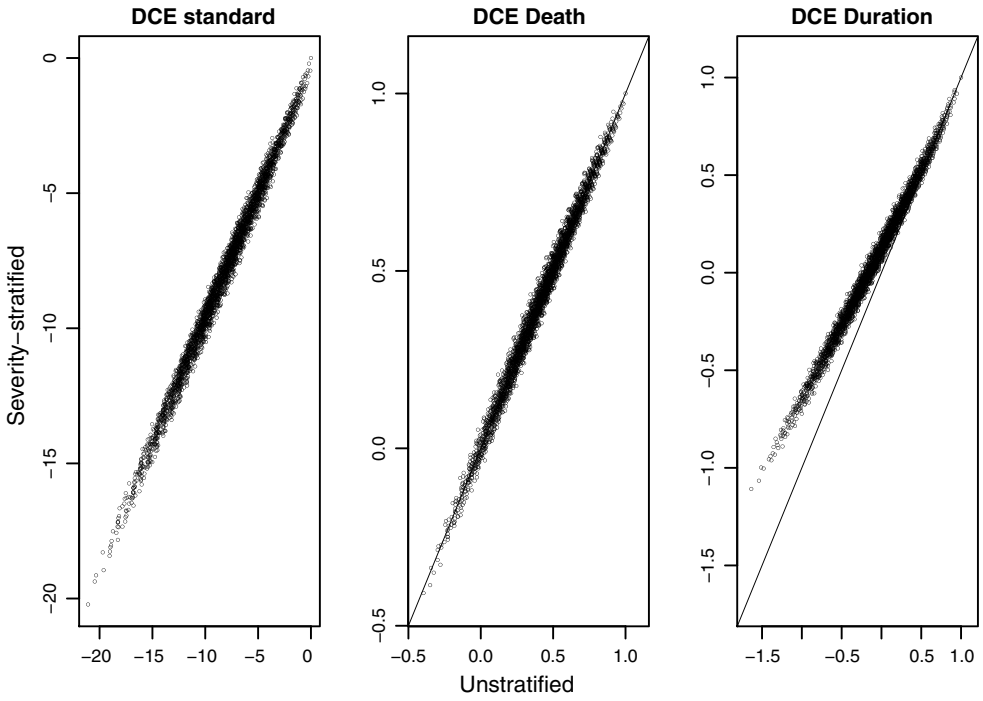

Fig. 3 Comparison of values for all EQ-5D health states between designs with and without severity-stratification. The 45° line is omitted from the graph on the left, which shows the impact of the severity-stratified restriction in the standard DCE choice task, because both sets of values are on a latent scale and adding a 45° line might be misleading as a basis for comparison. *DCE* discrete choice experiment

$0 < \text{QALY} \leq 0.5$, and $\text{QALY} > 0.5$, respectively. In addition, comparisons of the choice tasks were separately included: choice probabilities of impaired health states in A-B comparison tasks and B-C comparison tasks.

When we computed MEs across all health states, all six study arms produced insignificant MEs that were very close to zero because positive and negative errors offset each other. However, when we divided health states into severity ranges, some errors were found to be significantly different from zero. An expected result was that the out-of-sample predictions are more likely to show significant errors than the in-sample predictions, regardless of whether the severity stratification was applied. Beyond that, we found few noticeable differences between designs in most cases. However, for the DCE-duration, we found that the unstratified design produced significant negative errors for bad health states (i.e., column 4, italicized) while errors in the severity-stratified design were not significant (i.e., column 7, italicized), especially on B-C tasks. Also, for B-C tasks, the out-of-sample predictions produced by the severity-stratified design (0.0028) were much better than the unstratified design (− 0.0559) suggesting that the latter overestimated the willingness to trade-off life years to avoid bad health states significantly. These results suggest that the skewed health states selection for the DCE-duration introduced a downward bias on estimated values.

4 Discussion

This paper investigated the effect of imposing the severity stratification on Bayesian D-efficient DCE designs created for valuing health. We found that imposing severity stratification on DCE-duration was required to ensure that the selected set of health states covered the severity range well. The model estimates derived from the severity-stratified design also demonstrated better predictive performance than unstratified designs, especially regarding the choice probability of bad health states, preventing a downward bias on the values for poor health states. In the other investigated DCE types, we find less evidence of favoritism in the selection of health states, and imposing severity stratification had no substantial effect on values. The results suggest that efficient design algorithms need to be implemented carefully in the contexts of DCE-duration studies for health valuation.

It is instructive to reflect on the reasons why it matters so much to impose severity stratification on an efficient design algorithm used to construct a DCE with duration for health valuation. The low accuracy of predicted

Table 5 Mean signed errors for predicting choice probability

Parameter estimates used	Choice sets predicted			
	Unstratified choice sets		Severity-stratified choice sets	
	Unstratified	Severity-stratified	Unstratified	Severity-stratified
Standard DCE	0.3446×10^{-18}	0.1850×10^{-18}	-0.9972×10^{-18}	-1.3217×10^{-18}
DCE-death				
All health states				
All	0.7730×10^{-18}	1.0810×10^{-18}	0.1181×10^{-18}	1.1725×10^{-18}
A–B	0.8205×10^{-18}	0.6095×10^{-18}	0.5434×10^{-18}	1.2958×10^{-18}
B–C	0.7217×10^{-18}	1.6624×10^{-18}	-0.3774×10^{-18}	1.0441×10^{-18}
All	0.7730×10^{-18}	1.0810×10^{-18}	0.1181×10^{-18}	1.1725×10^{-18}
Bad health states				
All	0.0056	0.0130[a]	− 0.0039	− 0.0098
A–B	0.0062[a]	0.0142[a]	0.0043[a]	0.0012
B–C	0.0045	0.0097	− 0.0259	− 0.0331
Medium health states				
All	0.0015	− 0.0030	0.0103[a]	0.0049
A–B	− 0.0025[a]	− 0.0025[a]	0.0021[a]	0.0020[a]
B–C	0.0097	− 0.0044	0.0289[a]	0.0110
Better health states				
All	− 0.0017	− 0.0087[a]	0.0022	− 0.0028
A–B	− 0.0005	− 0.0016[a]	− 0.0044[a]	− 0.0037[a]
B–C	− 0.0067	− 0.0265[a]	0.0178	− 0.0008
DCE-duration				
All health states				
All	-0.6185×10^{-18}	-1.2218×10^{-18}	0.1516×10^{-18}	0.2327×10^{-18}
A–B	-1.0474×10^{-18}	-2.0947×10^{-18}	0.4598×10^{-18}	1.2122×10^{-18}
B–C	-0.1624×10^{-18}	-0.2443×10^{-18}	-0.1796×10^{-18}	-0.7710×10^{-18}
Bad health states				
All	− *0.0122*[a]	0.0015	− 0.0200[a]	− *0.0052*
A–B	− *0.0012*	0.0009	− 0.0029[a]	− *0.0012*[a]
B–C	− *0.0347*[a]	0.0028	− 0.0559[a]	− *0.0135*
Medium health states				
All	0.0028	0.0151[a]	− 0.0086	0.0035
A–B	0.0038[a]	0.0060[a]	0.0048[a]	0.0060[a]
B–C	0.0007	0.0345[a]	− 0.0365[a]	− 0.0017
Better health states				
All	0.0041	0.0049	0.0032	0.0038
A–B	− 0.0041[a]	− 0.0075[a]	− 0.0074[a]	− 0.0162[a]
B–C	0.0228	0.0355[a]	0.0309[a]	0.0449[a]

All = choice probabilities of impaired health states (regardless of comparison tasks); A–B = choice probabilities of impaired health states in A–B comparison tasks; B–C = choice probabilities of impaired health states in B–C comparison tasks

Note, bad health states for QALY ≤ 0, medium health state for 0 < QALY ≤ 0.5, and better health state for 0.5 < QALY. Because the standard DCE produces values on a latent scale, the division in three severity strata was omitted for those designs

DCE discrete choice experiment, *QALY* quality-adjusted life year

[a]Significant at 5% level

values of poor states based on a pro-mild set of health states reveals an extrapolation issue. Extrapolation per se does not cause a bias; it only does so when the model is misspecified. Hence, our findings indicate that the model was misspecified and that we can mitigate this problem by better spreading the data, thus ensuring that the resulting

QALY tariffs are less affected by extrapolation. In particular, the DCE-duration model seems to be sensitive to the assumptions made regarding duration preferences, as immediate death is not included so that the anchor point for the QALY scale is completely defined by extrapolation. The efficient optimization of the DCE design with a fixed (perfect health) comparator has aggravated the extrapolation problem, because it is efficient to include a skewed selection of relatively healthy health states. This reflects the special characteristic of DCE-duration models that utilities are derived using a multiplicative utility function with life years acting as a multiplier of the health state utility. Issues with utility dominance may arise in this context more easily than in standard applications of DCEs.

A limitation of this study is that it was beyond its scope to explore the extent to which our results are specific to the matched pairwise choice format that was used in this study. Having full health as a fixed alternative and the relatively long duration (i.e., 10 years) assumed for the impaired health states might have exaggerated issues that led to skewed selection of health states. Furthermore, we have not considered the merit of efficient designs in this context relative to other design generating approaches that do not require the implementation of strategies to enhance the spread of the data. The need to impose severity stratification makes construction of efficient designs for DCE-duration studies more difficult, and hence may influence the trade-offs between pros and cons of efficient versus other designs. Third, we did not find evidence of health state selection in the DCE-death approach, but we do not know if this result holds when valuing health states derived from other descriptive systems (e.g., disease-specific ones, where the mass of health states may be on a different location on the full health–dead scale). Fourth, assuming the normal distribution for parameters' distribution may be inappropriate to specify the monotonic attribute-level effect due to its unbounded nature. Using more flexible distribution with a fixed bound can be considered to avoid the potential violation of monotonicity. Last, we measured respondents' preference on the length of life using both months and years as the temporal unit in the perfect health state and converted months to years in the analysis. However, respondents may not treat values in months and the equivalent amount of years in the same way when valuing health states; thus, be cautious in further study [28].

5 Conclusion

We conclude that differences in how well selected health states span the severity range can explain part of the differences in values across DCE (duration) studies. Imposing 'severity stratification' on DCE-duration designs ensures robustness of the results against extrapolation from a misspecified model. Until we know how widespread associated extrapolation issues are in reported value sets, we need to be careful in the use of DCE-derived health state values.

Data Availability Statement The datasets generated for and/or analyzed during the current study are available from the corresponding author on reasonable request.

Acknowledgements Sesil Lim conducted the statistical analyses and drafted the first version of the manuscript. Marcel F. Jonker conceptualized the severity stratification approach, developed the study design, managed the data collection, and helped review and edit the manuscript. Mark Oppe, Bas Donkers, and Elly Stolk supported the development of the study design, analysis of the data, and interpretation of results as well as helped in the detailed review and editing of the manuscript. The views expressed in this work are those of the individual authors and do not necessarily reflect the views of the EuroQol Group.

Compliance with Ethical Standards

Funding This work has received financial support from the EuroQol Research Foundation grant number 2015300.

Conflict of Interest Marcel F. Jonker, Mark Oppe, and Elly Stolk are members of the EuroQol Group. Elly Stolk and Mark Oppe are employees of the EuroQol Research Foundation. Sesil Lim and Bas Donkers have no conflict of interest to declare.

References

1. Bijlenga D, Birnie E, Bonsel GJ. Feasibility, reliability, and validity of three health-state valuation methods using multiple-outcome vignettes on moderate-risk pregnancy at term. Value Health. 2009;12(5):821–7.
2. Stolk EA, Oppe M, Scalone L, Krabbe PF. Discrete choice modeling for the quantification of health states: the case of the EQ-5D. Value Health. 2010;13(8):1005–13.
3. Norman R, Viney R, Brazier J, Burgess L, Cronin P, King M, Ratcliffe J, Street D. Valuing SF-6D health states using a discrete choice experiment. Med Decis Making. 2014;34(6):773–86.
4. Norman R, Mulhern B, Viney R. The impact of different DCE-based approaches when anchoring utility scores. Pharmacoeconomics. 2016;34(8):805–14.
5. Craig BM, Greiner W, Brown DS, Reeve BB. Valuation of child health-related quality of life in the United States. Health Econ. 2016;25(6):768–77.
6. Jonker MF, Attema AE, Donkers B, Stolk EA, Versteegh MM. 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. 2017;26(12):1534–47.
7. Dolan P. Modeling valuations for EuroQol health states. Med Care. 1997;35:1095–108.
8. Bridges JF, Hauber AB, Marshall D, Lloyd A, Prosser LA, Regier DA, Johnson FR, Mauskopf J. Conjoint analysis applications in health–a checklist: a report of the ISPOR good research practices for conjoint analysis task force. Value Health. 2011;14(4):403–13.
9. Johnson FR, Lancsar E, Marshall D, Kilambi V, Mühlbacher A, Regier DA, Bresnahan BW, Kanninen B, Bridges JF. Constructing experimental designs for discrete-choice experiments: report of the ISPOR conjoint analysis experimental design good research practices task force. Value Health. 2013;16(1):3–13.
10. Rose JM, Bliemer MC. Constructing efficient stated choice experimental designs. Transp Rev. 2009;29(5):587–617.
11. 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.
12. Mulhern B, Bansback N, Hole AR, Tsuchiya A. Using discrete choice experiments with duration to model EQ-5D-5L health state preferences. Med Decis Making. 2016;37(3):285–97.
13. Flynn TN, Bilger M, Malhotra C, Finkelstein EA. Are efficient designs used in discrete choice experiments too difficult for some respondents? A case study eliciting preferences for end-of-life care. Pharmacoeconomics. 2016;34(3):273–84.
14. Dellaert BGC, Donkers B, Van Soest AHO. Complexity effects in choice experiment–based models. J Marketing Res. 2012;49(3):424–34.
15. Jonker MF, Donkers B, De Bekker-Grob EW, Stolk EA. The effect of level overlap and color coding on attribute non-attendance in discrete choice experiments. Value Health. 2017. https://doi.org/10.1016/j.jval.2017.10.002.
16. Norman R, Viney R, Aaronson NK, et al. Using a discrete choice experiment to value the QLU-C10D: feasibility and sensitivity to presentation format. Qual Life Res. 2016;25:637–49.
17. Walker JL, Wang Y, Thorhauge M, Ben-Akiva M. D-efficient or deficient? A robustness analysis of stated choice experimental designs. Theory Decis. 2018;84(2):215–38.
18. Bliemer MCJ, Rose JM. Experimental design influences on stated choice outputs: an empirical study in air travel choice. Transport Res A-Policy Pract. 2011;45:63–79.
19. Huber J, Zwerina K. The importance of utility balance in efficient choice designs. J Marketing Res. 1996;33(3):307–17.
20. Jonker MF, Donkers B, De Bekker-Grob EW, Stolk E. Advocating a paradigm shift in health-state valuations: the estimation of time-preference corrected QALY tariffs. Value Health. 2018. https://doi.org/10.1016/j.jval.2018.01.016.
21. Sándor Z, Wedel M. Heterogeneous conjoint choice designs. J Marketing Res. 2005;42(2):210–8.
22. Bliemer MC, Rose JM. Construction of experimental designs for mixed logit models allowing for correlation across choice observations. Transport Res B-Meth. 2010;44(6):720–34.
23. Bezanson J, Karpinski S, Shah VB, Edelman A. Julia: a fast dynamic language for technical computing. 2012. https://arxiv.org/abs/1209.5145v1. Accessed 31 May 2018.
24. Hensher DA, Greene WH. The mixed logit model: the state of practice. Transportation. 2003;30:133–76.
25. Rossi P, McCulloch R. bayesm: Bayesian inference for marketing/micro-econometrics. R package version 3.1-0.1. 2017. https://cran.r-project.org/web/packages/bayesm/bayesm.pdf. Accessed 8 Feb 2018.
26. Plummer M. Package 'coda'. R package version 0.91-1. 2016. https://cran.r-project.org/web/packages/coda/coda.pdf. Accessed 1 July 2018.
27. Versteegh MM, Vermeulen KM, Evers SMAA, De Wit GA, Prenger R, Stolk EA. Dutch tariff for the five-level version of EQ-5D. Value Health. 2016;19(4):343–52.
28. Jakubczyk M, Craig BM, Barra M, Groothuis-Oudshoorn CGM, Hartman JD, Huynh E, Ramos-Goñi JM, Stolk EA, Rand K. Choice defines value: a predictive modeling competition in health preference research. Value Health. 2018;21(2):229–38.