HEALTH ECONOMICS
Health Econ. (2016)
Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/hec.3445

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

MARCEL F. JONKER[a,c*], ARTHUR E. ATTEMA[a], BAS DONKERS[b,c], ELLY A. STOLK[a,c,†] and MATTHIJS M. VERSTEEGH[d,†]

[a]*Institute of Health Policy and Management, Erasmus University Rotterdam, Rotterdam, The Netherlands*
[b]*Department of Business Economics, Erasmus University Rotterdam, Rotterdam, The Netherlands*
[c]*Erasmus Choice Modelling Centre, Erasmus University Rotterdam, Rotterdam, The Netherlands*
[d]*Institute for Medical Technology Assessment, Erasmus University Rotterdam, Rotterdam, The Netherlands*

ABSTRACT

Health state valuations of patients and non-patients are not the same, whereas health state values obtained from general population samples are a weighted average of both. The latter constitutes an often-overlooked source of bias. This study investigates the resulting bias and tests for the impact of reference dependency on health state valuations using an efficient discrete choice experiment administered to a Dutch nationally representative sample of 788 respondents. A Bayesian discrete choice experiment design consisting of eight sets of 24 (matched pairwise) choice tasks was developed, with each set providing full identification of the included parameters. Mixed logit models were used to estimate health state preferences with respondents' own health included as an additional predictor. Our results indicate that respondents with impaired health worse than or equal to the health state levels under evaluation have approximately 30% smaller health state decrements. This confirms that reference dependency can be observed in general population samples and affirms the relevance of prospect theory in health state valuations. At the same time, the limited number of respondents with severe health impairments does not appear to bias social tariffs as obtained from general population samples.

Received 16 July 2015; Revised 3 September 2016; Accepted 19 September 2016

KEY WORDS: health state valuation; prospect theory; discrete choice experiment; EQ-5D

Health state valuations provided by patients (i.e., experienced-based preferences) and the general public (i.e., social preferences) are generally not the same (De Wit *et al.,* 2000; Brazier *et al*., 2007). Differences can be attributed to many causes, ranging from aspects of adaptation to changes in internal standards (Menzel *et al.,* 2002; Ubel *et al.,* 2003; Dolan and Kahneman, 2008). Faced with these differences, the question arises whose values should be used and in what situations. For decisions based at the micro-level, such as in clinical care settings, there appears little controversy on the view that patients' own values should guide decision-making processes (Stamuli, 2011). However, at a more macro-level or societal level, where decisions concern the distribution of public resources, the views are more divided.

On the one hand, it is argued that the general public has no experience with specific impaired health states. As such, the general public is thought to underestimate adaptation to specific illnesses, whereas patients have

*Correspondence to: Institute of Health Policy and Management, Erasmus University Rotterdam, PO Box 1738, 3000DR Rotterdam, The Netherlands. E-mail: marcel@mfjonker.com

†Member of the EuroQol Group (Rotterdam, The Netherlands)

actual experience with impaired health, resulting in more reliable and relevant health state valuations. On the other hand, it is argued that patients may provide biased values because of vested interests, suppressed recognition of full health, and/or lowered expectations (Menzel *et al.,* 2002); additionally, public preferences are arguably more appropriate when public health care expenditures are concerned or when unbiased *ex ante* health state values are desired (Brazier *et al*., 2007).

The latter is based on the 'veil of ignorance' argument as put forward by Gold *et al.* (1996) and reflects the idea that the general public neither experiences the impaired health states under evaluation nor has any certainty that it will ever require health care for particular health problems. These conditions accommodate the measurement of unbiased *ex ante* health state values. However, upon closer examination, the Gold *et al.* (1996) 'veil of ignorance' argument actually supports the use of strictly healthy respondents, which is not the same as a general population sample. Indeed, a random sample from the general population is heterogeneous and comprises many respondents with impaired health. Unlike healthy respondents, these respondents have first-hand experience with particular impaired health states and consequently a unique vantage or reference point that precludes them from valuing health states as if from behind a 'veil of ignorance.'

The discrepancy between a random sample from the general population and a sample of healthy respondents is often not explicitly recognized, and even when it is acknowledged (e.g., in a footnote in Dolan and Kahneman, 2008), the difference is not further investigated. In this paper, we address the issue and explore the impact of using general population samples as a proxy for health state valuations within a sample of healthy respondents. First, we take into account that respondents from a random sample from the general population are likely to evaluate hypothetical health states from their own perspective, resulting in different evaluations depending on the respondents' own health state. This phenomenon, referred to as health state reference dependency, has a supporting evidence base (Bleichrodt *et al.,* 2001; Wakker, 2010) and is solidly grounded in prospect theory. Our first aim is hence to provide empirical evidence of the existence and relevance of reference dependency (c.q. prospect theory) in health state valuations. Second, we acknowledge that health state reference dependency can result in social tariffs that are upwards biased when assumed to be derived from strictly healthy respondents. Depending on the effect size of health state reference dependency and on the fraction of respondents with impaired health states, the difference between a standard general population tariff and one that is purged from reference dependency could be substantial. Our second aim is therefore to provide an empirical estimate of the size of this difference and to determine its relevance for existing and future health state valuations.

2. METHODS

2.1. EQ-5D health state values

The reference dependency hypothesis is tested using discrete choice experiment (DCE) health state valuations of the EuroQol five-dimensional five-level (EQ-5D-5L) instrument, which comprises the dimensions of 'mobility', 'self-care', 'usual activities', 'pain/discomfort', and 'anxiety/depression'. EQ-5D-5L health states are defined by selecting one level from each dimension (Table I), with 1–1–1–1–1 denoting perfect health and 5–5–5–5–5 the worst possible health state. The EQ-5D has a monotonic order in the attribute levels, implying that all levels should be valued equal to or worse than the previous level. In Figure 1, which shows two options to present and measure health state utility values, this is reflected by either a monotonic increasing decrement relative to perfect health (option A) or consistently negative estimates that capture the differences between two levels (option B), the latter being more informative about statistical efficiency but requiring the summation of all preceding levels to obtain the overall health state decrements.

Health state preferences for the EQ-5D are derived by asking a sample of respondents to repeatedly evaluate two or more hypothetical health states and indicate which one they prefer. The observed discrete choices are then used to infer the latent preference weights that respondents use to trade off the various attribute levels of each EQ-5D health state. In the standard health state valuation model, it is assumed that

DOI: 10.1002/hec

Table I. The EuroQol five-dimensional five-level instrument

	Mobility		Pain/discomfort
MO1	I have *no* problems in walking about	PD1	I have *no* pain or discomfort
MO2	I have *slight* problems in walking about	PD2	I have *slight* pain or discomfort
MO3	I have *moderate* problems in walking about	PD3	I have *moderate* pain or discomfort
MO4	I have *severe* problems in walking about	PD4	I have *severe* pain or discomfort
MO5	I am *unable* to walk about	PD5	I have *extreme* pain or discomfort
	Self-care		Anxiety/depression
SC1	I have *no* problems in washing or dressing myself	AD1	I am *not* anxious or depressed
SC2	I have *slight* problems in washing or dressing myself	AD2	I am *slightly* anxious or depressed
SC3	I have *moderate* problems in washing or dressing myself	AD3	I am *moderately* anxious or depressed
SC4	I have *severe* problems in washing or dressing myself	AD4	I am *severely* anxious or depressed
SC5	I am *unable* to wash or dress myself	AD5	I am *extremely* anxious or depressed
	Usual activities		
UA1	I have *no* problems doing my usual activities		
UA2	I have *slight* problems doing my usual activities		
UA3	I have *moderate* problems doing my usual activities		
UA4	I have *severe* problems doing my usual activities		
UA5	I am *unable* to do my usual activities		

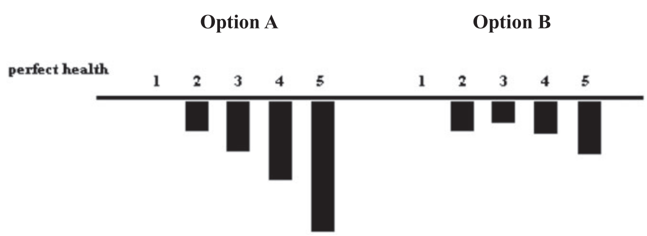

Figure 1. Graphical presentation of EuroQol five-dimensional health state values. Options A and B reflect identical health state preferences. In option A, preferences for levels 2–5 are all depicted relative to perfect health (i.e., level 1), which readily shows the overall decrement of each level. In option B, preferences for levels 2–5 are depicted relative to the preceding level, which is more convenient in highlighting whether adjacent levels have statistically significantly different decrements. Option B is implemented in this paper

respondents do not take their own health into account when valuing health states. However, in case of reference dependency, respondents' latent preference weights are allowed to depend on their own health. This can result in a value set where some of the health state decrements are mitigated (or aggravated) based on the respondents' health.

A straightforward approach to modeling health state reference dependency would be to include a full set of reference dependency dummies in the health state valuation model. These dummies would capture, for all levels equal to or higher than the respondents' own health in a particular dimension, an increment or decrement in the respondents' latent preference weights. However, this approach effectively doubles the number of parameters in the health state valuation model, whereas the identification of reference dependency requires statistically significant estimates of effects that are a fraction of the standard EQ-5D preference weights (Figure 2). Accordingly, this approach would require substantially larger survey sample sizes than typically available.

In contrast, a more feasible and parsimonious test for reference dependency only includes a single additional parameter per EQ-5D dimension. This parameter can capture a common increment or decrement per health state level (i.e., an additive specification) or a common multiplication factor (i.e., a multiplicative specification) as depicted in Figure 2. Because the overall reference dependency effect is expected to depend on the effect size

DOI: 10.1002/hec

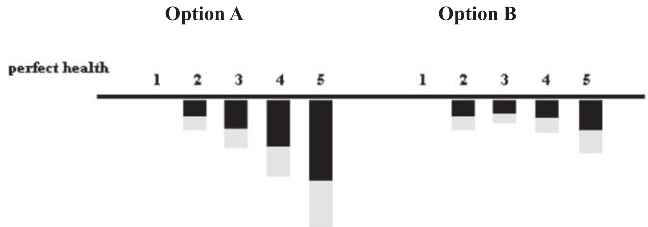

Figure 2. Graphical presentation of EuroQol five-dimensional health-state values with multiplicative reference dependency. Options A and B reflect identical health state preferences, either relative to perfect health (option A) or relative to the preceding level (option B). Option B is implemented in this paper. In both options, the effect of reference dependency is depicted in gray and taken as a proportion of the overall health state decrements. This allows the overall effect of reference dependency to be larger in level 5 than, for example, in levels 2 or 3. The black preferences are for those whose own health is equal to or worse than the health state levels under evaluation, whereas the combined areas (i.e., gray and black surfaces) depict the preferences of respondents whose own health is better than the respective health state levels

of the standard, non-reference dependent health state valuations, the multiplicative specification with dummy coding relative to the previous level is considered more realistic and therefore used in this paper. Note that the exact model specification and details about the estimation procedure are included in Section 2.4.

2.2. DCE design

2.2.1. Imposed constraints. In the DCE design used to elicit health state values, preferences are derived using choice tasks that require respondents to choose between living longer in a relatively bad health state or shorter in a relatively good health state. This is an efficient format for obtaining health state preferences and allows for the endogenous scaling of health state preferences on the quality-adjusted life year (QALY) scale (Flynn, 2010), but it is also relatively complicated for respondents. To reduce the task complexity for respondents and to obtain more informative and reliable trade-offs, two types of constraints are imposed: (i) those that improve the internal consistency of health states and (ii) those that aim to simplify the trade-offs between health states. The constraints that aim to improve the internal consistency of health states ensure that hypothetical health states do not combine zero problems with usual activities (i.e., UA level 1) with extreme problems with pain/discomfort (i.e., PD level 5) and/or extreme problems with anxious or depressed (i.e., AD level 5). These constraints make the health states more realistic and easier for respondents to evaluate. At the same time, the number of constraints is deliberately minimized to avoid too much reduction in the statistical power of the DCE design.

The constraints aimed at simplifying the trade-off between health states constitute a choice format referred to as ‘matched pairwise choice tasks’. First, all choice tasks are restricted to pairwise comparisons rather than more efficient (and more complex) comparisons between several health states simultaneously. Second, all choice tasks are constrained to not simultaneously contain a comparison between different impaired health states and varying lengths of life. These are replaced by two types of choice tasks: (i) pairwise comparisons between impaired health states with equal length of life and (b) pairwise comparisons between a shorter length of life in perfect health and a longer length of life with impaired health. Third, both types of pairwise comparisons are linked together by imposing that one of the impaired health states in the first comparison is equal to the impaired health state in the second comparison (Figure 3). This takes advantage of the respondents’ familiarity with a given choice set to include an additional discrete choice task with perfect health at very limited additional effort. Fourth, three of the five EQ-5D attributes are constrained to having the same value (cf. Maddala *et al.,* 2003; Louviere *et al.,* 2008). This attribute level overlap implies that respondents can focus on the direct trade-off between the other health state attributes. Finally, the quantity of life in the choice tasks is restricted to 10 years for the pairwise comparisons with equal length of life, and some smaller value for the perfect health state. The latter is restricted to integer values (i.e., whole years) or 1, 3, 6, 9, or 18 months when smaller than 2 years.

DOI: 10.1002/hec

Figure 3. Visual presentation of the matched pairwise choice tasks

2.2.2. Visual presentation. Figure 3 shows how the matched pairwise choice tasks were presented to the survey respondents. In the first choice task, respondents are asked to indicate which option (A or B) they prefer. As explained, the remaining life expectancy is 10 years for both health states. Accordingly, the choice between A and B is only determined by the relative attractiveness of the health state attribute levels. In the second 'matched' choice task, respondents are again asked to indicate which option (B or C) they prefer. Option B is always the same as in the first choice task, and option C always represents perfect health with a shorter remaining life expectancy.

The visual format of the matched pairwise choice tasks is used to further clarify the trade-offs for respondents: the non-active choice option is blurred, black lines are used to indicate the active choice options, and check boxes are only placed below the options that respondents can actually select. Furthermore, to allow respondents to already consider the trade-off between options B and C when evaluating the differences between options A and B, the quantity of life attribute of option C is not blurred in the first pairwise choice task. The same effect is achieved in a paper-and-pencil questionnaire by depicting both choice tasks one above the other on a single page, as in Figure 3. Additionally, color coding is used to aid respondents in comparing the differences between health states. A 'color blind' coding system is used, which is based on different shades of purple to denote the various attribute levels. This color coding system is specifically optimized to signal differences in the attribute levels for individuals with red-green color blindness (the most prevalent form of color blindness) while keeping the text readable for respondents who suffer from other forms of color blindness. Also, respondents randomly received the survey either in the standard color coding or in reverse color-coding order, as shown in Figure 3.

DOI: 10.1002/hec

2.2.3. Efficient design generation. To be able to implement the imposed constraints on the DCE choice tasks while ensuring sufficient statistical efficiency, the DCE design has been constructed using efficient design algorithms. These algorithms improve upon the efficiency of orthogonal design algorithms by taking prior information about average respondents' preferences into account. This prior information was obtained from a pilot study of Bansback *et al.* (2010).

To account for uncertainty in the priors and to avoid a locally optimal design that (rapidly) loses efficiency when the implemented priors deviate from the true population preferences, the DCE design was created using a custom-programmed Bayesian algorithm. Bayesian design optimization algorithms are computationally more demanding but superior in handling uncertainty in the priors (e.g., Sándor and Wedel, 2001). Given the large number of parameters in the EQ-5D design, a substantial number of draws from the prior parameter distributions are required to differentiate reliably between efficiencies of alternative designs. A Latin hypercube sample of 2500 draws, optimized by maximizing the minimum distance between points using the genetic algorithm as described by Stocki (2005), was used in the optimizations.

Further robustness was introduced by simultaneously optimizing eight different DCE sub-designs. As shown by Sándor and Wedel (2005), the implementation of a heterogeneous DCE design comes at no additional cost to respondents because each individual respondent receives randomly only one of the possible sub-designs. However, the use of several sub-designs significantly increases the robustness and efficiency of the overall DCE design, particularly when the true population preferences are uncertain and the number of attributes and attribute levels increases. The intuition behind this is simple: by spreading the attribute levels optimally not only within subjects, as is performed in homogeneous designs, but also across subjects, as is performed in heterogeneous DCE designs, the increased variation in the choice tasks enables the variation in the dependent variable to be captured more effectively. However, the individual-level efficiency is considered more important than the combined (i.e., population) efficiency, especially when measuring reference dependency at the individual level. Accordingly, the design optimization was based on the minimization of the weighted average Bayesian D-efficiency for 12 matched pairwise choice tasks (i.e., 24 choice tasks) per respondent with one-third of the weight assigned to the population efficiency and two-thirds of the weight assigned to the individual efficiencies of the sub-designs.

Given the large number of parameters, the inclusion of eight sub-designs, and the large number of draws to evaluate the Bayesian distributions, the DCE design was optimized for the standard multinomial logit model. This model is less computationally demanding to optimize for, and evidence from Bliemer and Rose (2010) indicates that DCE designs that are optimized for the multinomial logit model also perform well for estimating panel mixed logit (MIXL) models. The DCE design optimizations were programmed in MATLAB. Following Sándor and Wedel (2005), a so-called greedy approach was used to optimize the design wherein the simultaneous optimization of the eight sub-designs and quantity and quality of life interactions is replaced with a sequence of partial optimizations, each of which is considerably faster than the simultaneous optimization. In our implementation, the optimization alternated between a row-based algorithm that randomly replaced health state comparisons in one of the eight sub-designs while adhering to the constraints as described in Section 2.2.1 and a column-based algorithm that optimized the trade-off between quality and quantity of life in each sub-design by looping over all possible durations of life.

2.3. Survey administration

The DCE survey has been administered in two separate waves. To obtain the first sample, the DCE survey was administered as an online questionnaire to 436 respondents from the Longitudinal Internet Studies for the Social Sciences (LISS) panel as described by Scherpenzeel and Das (2010). The LISS is a large, nationally representative household panel subsidized by the Dutch government for scientific research. Panel members complete online questionnaires every month for about 15–30 min in total, for which they receive a small financial compensation. To obtain the second sample, the DCE survey was included as an additional paper-and-pencil questionnaire administered under supervision by trained interviewers to approximately one-third of

DOI: 10.1002/hec

the respondents who participated in the official Dutch nationally representative EuroQol EQ-5D valuation study of Versteegh *et al.* (2016). Respondents received a small financial compensation for participating in the EQ-5D valuation study, although participation in the *ex post* paper-and-pencil DCE study was voluntary and could be declined without effect on the payment.

The online and paper-and-pencil surveys were similarly structured. In both samples, the survey was first briefly introduced. In the online survey, the introduction was followed by a self-rating question in which respondents were asked to rate their current health in terms of the EQ-5D health dimensions. This procedure familiarized the online respondents with the format of the EQ-5D health states that are used in the DCE questions and resulted in the health state measurements needed for the reference dependency evaluation. In the paper-and-pencil survey, the self-rating question was omitted because it was already included in the official EQ-5D valuation protocol. Next, both surveys included two warm-up questions that carefully explained the layout and the required trade-offs for the DCE duration questions. Then, the actual design of 12 matched pairwise choice tasks was shown, resulting in a total of 24 pairwise comparisons for each respondent.

2.4. Statistical analyses

To estimate EQ-5D health state preferences, a standard MIXL specification was used. In the model without reference dependency, the utility specification is defined as the product of the quantity and quality of life. More specifically, the utility for alternative *a* in choice task *t* for respondent *i* is modeled as the product of the health state characteristics (X_{ita}) and health state preference parameters (β_i) multiplied by the number of life years *T* of each health state alternative:

$$u_{ita} = (\beta_i . X_{ita}) . T_{ita} + \varepsilon_{ita} , \tag{1}$$

with the error term ε_{ita} assumed independent and identically distributed (i.i.d.) with an extreme value distribution. This specification implies proportional time preferences and zero utility for health states with zero length of life, which anchors death at 0 on the latent utility scale. Also, conform standard MIXL assumptions, the respondent-specific β-parameters are assumed multivariate normal distributed with a common population mean μ and covariance matrix Σ:

$$\beta_i \sim \text{Multivariate Normal}(\mu, \Sigma) . \tag{2}$$

This allows for the measurement of average population health state preferences (μ) while accounting for preference heterogeneity. Finally, with the health state characteristics dummy coded and with the best levels used as reference categories, the model's intercept captures the utility of perfect health and defines the highest possible utility for the included EQ-5D health states. This accommodates the calculation of standardized health state decrements on the QALY scale by dividing all elements in the vector μ by the intercept, that is, the first element of μ. These transformed estimates are health state decrements on a QALY scale that is anchored at 0 (death) and 1 (perfect health) and can be directly compared across different samples and model specifications.

In the model with reference dependency, the utility function is slightly expanded to incorporate an additional set of dummy variables Z_{itaj}. These dummy variables indicate for each of the corresponding X_{itaj} variables whether the health state levels in the choice tasks are lower than the respondents' self-assessed health state level (i.e., 0) or equal to or higher than the respondents' health state level (i.e., 1). The reference dependent utility decrement for a health state attribute level, which we denote by β^*_{ij}, is then given by $\beta_{ij}\,(1+ Z_{itaj}\,\gamma_j)$, with γ_j denoting the reference dependency scaling factor for the EQ-5D dimension of attribute-level j. Note that γ_j is the same for all attribute levels of a given EQ-5D dimension, but allowed to be different across EQ-5D dimensions. The reference dependency model is then defined as:

$$u_{ita} = (\beta^*_i . X_{ita}) . T_{ita} + \varepsilon_{ita} \tag{3}$$

with the error term ε_{ita} still i.i.d. extreme value. This specification is identical to the standard MIXL specification in equation 1 when the reference dependency dummies are 0, or it involves the scaling of the β_i-parameters with a EQ-5D dimension-specific scaling factor when some of the reference dependency dummies are 1.

DOI: 10.1002/hec

The models with and without reference dependency are programmed in the BUGS language and estimated in OpenBUGS using Bayesian Markov chain Monte Carlo (MCMC) methods. This entails selecting prior distributions for the unknown parameters and updating these via the likelihood of the observed data. Uninformative multivariate normal priors (i.e., with a mean of zero and standard deviation of 10) were assigned to μ and γ and a Wishart prior with an identity scale matrix and 21 degrees of freedom to Σ. Standard Gibbs update steps were used to update μ and Σ, and a Metropolis-within-Gibbs update step was implemented in the OpenBUGS software to reliably update γ and β using directional antithetic sampling, as described by Bai (2009) and Bédard *et al.* (2014). All estimations used 15,000 MCMC iterations to let the chains converge from divergent starting points and 30,000 MCMC iterations (with a thinning of 10) to reliably approximate the mean of the posterior distributions on the latent and QALY scale. Convergence was evaluated based on a visual inspection of the chains and the full set of convergence diagnostics as implemented in the R CODA package (Plummer *et al.,* 2006).

3. RESULTS

In the online sample, 636 panel members aged 16 years and older were invited to participate in the survey; 430 completed the survey (69%), 6 dropped out during the survey (1%), and 30% did not respond. In the paper-and-pencil study, 371 participants of the Dutch EQ-5D valuation study were administered the additional DCE duration survey questions after having completed the main valuation study. Of these 371 respondents, nine (2%) omitted one or more DCE questions and four respondents had missing health state information (<1%), resulting in a net sample of 358 respondents.

Table II contains descriptive statistics and Table III an overview of the respondents' self-rated health state in the online, paper, and combined survey sample in terms of the EQ-5D health state dimensions and levels. With both samples nationally representative in terms of age, sex, and education, the observed distribution of health states is very similar. As can also be seen, the combined sample is relatively small with few respondents in the worst health state levels of the EQ-5D dimensions. This is particularly noticeable in the self-care dimension, in which more than 95% of the respondents report zero problems with self-care, and provides an important rationale to combine the paper and online samples.

Prior to combining the online and paper-based survey samples, their respective EQ-5D estimates on the latent utility and QALY scale are shown in Table IV. First, all estimates are statistically significant, as indicated by the 95% credible intervals that do not contain 0. Second, all estimates are logically consistent with a clear monotonic order of the attribute levels, which is indicated by the consistently negative estimates that capture decrements relative to the preceding health state levels. Third, even though there is no guarantee that the estimates of both samples are directly comparable on the latent utility scale, the Bayesian 95% credible intervals of the corresponding parameters already show considerable overlap. As mentioned, the theoretically correct

Table II. Descriptive statistics, by data source

Data source	Variable	N	Mean	Standard deviation	Minimum	Maximum
Paper	Age	358	47	18	18	84
	Sex[a]	358	0.52	0.50	0	1
	Education (low)	358	0.35	0.48	0	1
	Education (medium)	358	0.35	0.48	0	1
	Education (high)	358	0.30	0.46	0	1
Online	Age	430	49	17	16	88
	Sex[a]	430	0.50	0.50	0	1
	Education (low)	430	0.34	0.48	0	1
	Education (medium)	430	0.33	0.47	0	1
	Education (high)	430	0.32	0.47	0	1

[a]Female = 0; male = 1.

DOI: 10.1002/hec

Table III. Respondents' self-rated EQ-5D health state, by data source[a]

EQ-5D	Data source	EQ-5D level[a] 1	2	3	4	5
Mobility	Paper	260 (73%)	62 (17%)	24 (6%)	11 (3%)	1 (0.3%)
	Online	336 (78%)	62 (15%)	19 (4%)	12 (3%)	1 (0.2%)
	Combined	596 (76%)	124 (16%)	43 (5%)	23 (3%)	2 (0.3%)
Self-care	Paper	341 (95%)	13 (4%)	3 (1%)	0 (0%)	1 (0.3%)
	Online	413 (96%)	14 (3%)	3 (1%)	0 (0%)	0 (0%)
	Combined	754 (96%)	27 (3%)	6 (1%)	0 (0%)	1 (0.1%)
Usual activities	Paper	253 (71%)	73 (20%)	27 (8%)	3 (1%)	2 (1%)
	Online	322 (75%)	67 (15%)	35 (8%)	6 (1%)	0 (0%)
	Combined	575 (73%)	140 (18%)	62 (8%)	9 (1%)	2 (0.3%)
Pain/discomfort	Paper	186 (52%)	112 (31%)	51 (14%)	8 (2%)	1 (0.3%)
	Online	226 (53%)	143 (34%)	50 (12%)	7 (2%)	4 (1%)
	Combined	412 (53%)	255 (32%)	101 (13%)	15 (2%)	5 (1%)
Anxiety/depression	Paper	285 (80%)	55 (15%)	14 (4%)	5 (1%)	1 (0.3%)
	Online	331 (77%)	73 (17%)	20 (5%)	5 (1%)	1 (0.2%)
	Combined	616 (78%)	126 (16%)	34 (4%)	10 (1%)	2 (0.3%)

EQ-5D, EuroQol five-dimensional instrument.
[a]EQ5D levels range from 1 for no problems, to 5 for severe problems.

Table IV. EQ-5D estimates of the separate samples [a]

	Latent utility scale		QALY scale	
	Online sample	Paper sample	Online sample	Paper sample
Full health	2.61 (2.27, 2.97)	2.31 (2.07, 2.57)	1.00 (n/a)	1.00 (n/a)
Mobility 2	−0.16 (−0.20, −0.13)	−0.12 (−0.16, −0.08)	−0.06 (−0.08, −0.05)	−0.05 (−0.07, −0.03)
Mobility 3	−0.10 (−0.14, −0.06)	−0.05 (−0.09, −0.01)	−0.04 (−0.05, −0.02)	−0.02 (−0.04, −0.00)
Mobility 4	−0.46 (−0.54, −0.37)	−0.39 (−0.45, −0.33)	−0.17 (−0.21, −0.14)	−0.17 (−0.19, −0.14)
Mobility 5	−0.41 (−0.50, −0.32)	−0.18 (−0.25, −0.11)	−0.16 (−0.19, −0.13)	−0.08 (−0.11, −0.05)
Self-care 2	−0.16 (−0.19, −0.12)	−0.11 (−0.15, −0.08)	−0.06 (−0.08, −0.05)	−0.05 (−0.06, −0.03)
Self-care 3	−0.05 (−0.09, −0.01)	−0.09 (−0.14, −0.05)	−0.02 (−0.03, −0.00)	−0.04 (−0.06, −0.02)
Self-care 4	−0.40 (−0.46, −0.34)	−0.25 (−0.31, −0.20)	−0.15 (−0.18, −0.13)	−0.11 (−0.13, −0.09)
Self-care 5	−0.36 (−0.44, −0.30)	−0.06 (−0.13, −0.01)	−0.14 (−0.17, −0.11)	−0.03 (−0.06, −0.00)
Usual activities 2	−0.15 (−0.19, −0.12)	−0.15 (−0.19, −0.11)	−0.06 (−0.08, −0.05)	−0.07 (−0.08, −0.05)
Usual activities 3	−0.09 (−0.13, −0.05)	−0.05 (−0.09, −0.01)	−0.03 (−0.05, −0.02)	−0.02 (−0.04, −0.00)
Usual activities 4	−0.48 (−0.55, −0.42)	−0.43 (−0.50, −0.37)	−0.19 (−0.21, −0.16)	−0.19 (−0.21, −0.17)
Usual activities 5	−0.39 (−0.45, −0.33)	−0.24 (−0.31, −0.18)	−0.15 (−0.18, −0.13)	−0.11 (−0.13, −0.08)
Pain/discomfort 2	−0.18 (−0.22, −0.15)	−0.17 (−0.21, −0.14)	−0.07 (−0.09, −0.06)	−0.08 (−0.09, −0.06)
Pain/discomfort 3	−0.11 (−0.14, −0.07)	−0.06 (−0.10, −0.02)	−0.04 (−0.05, −0.03)	−0.02 (−0.04, −0.01)
Pain/discomfort 4	−0.69 (−0.79, −0.60)	−0.76 (−0.85, −0.67)	−0.27 (−0.31, −0.23)	−0.33 (−0.36, −0.30)
Pain/discomfort 5	−0.64 (−0.73, −0.55)	−0.49 (−0.59, −0.40)	−0.25 (−0.28, −0.21)	−0.21 (−0.25, −0.18)
Anxiety/depression 2	−0.22 (−0.26, −0.19)	−0.27 (−0.33, −0.22)	−0.09 (−0.10, −0.07)	−0.12 (−0.14, −0.10)
Anxiety/depression 3	−0.18 (−0.22, −0.14)	−0.07 (−0.12, −0.01)	−0.07 (−0.09, −0.05)	−0.03 (−0.05, −0.01)
Anxiety/depression 4	−0.59 (−0.70, −0.50)	−0.50 (−0.58, −0.43)	−0.23 (−0.27, −0.19)	−0.22 (−0.25, −0.19)
Anxiety/depression 5	−0.76 (−0.87, −0.67)	−0.69 (−0.80, −0.59)	−0.30 (−0.34, −0.26)	−0.30 (−0.34, −0.26)
N	430	358	430	358

EQ-5D, EuroQol five-dimensional instrument; n/a, not applicable; QALY, quality-adjusted life year.
[a]Mean posterior estimates of μ with 95% credible intervals in parenthesis.

DOI: 10.1002/hec

Table V. VEQ-5D estimates of the combined sample[a]

	Latent utility scale		QALY scale	
	Without reference dependency	With reference dependency	Without reference dependency	With reference dependency
Full health	2.24 (2.06, 2.44)	2.21 (2.03, 2.41)	1.00 (n/a)	1.00 (n/a)
Mobility 2	−0.13 (−0.15, −0.10)	−0.13 (−0.16, −0.11)	−0.06 (−0.07, −0.05)	−0.06 (−0.07, −0.05)
Mobility 3	−0.07 (−0.10, −0.05)	−0.08 (−0.10, −0.05)	−0.03 (−0.04, −0.02)	−0.03 (−0.05, −0.02)
Mobility 4	−0.37 (−0.42, −0.33)	−0.37 (−0.42, −0.33)	−0.17 (−0.19, −0.15)	−0.17 (−0.19, −0.15)
Mobility 5	−0.26 (−0.32, −0.21)	−0.26 (−0.31, −0.21)	−0.12 (−0.14, −0.10)	−0.12 (−0.14, −0.10)
Self-care 2	−0.12 (−0.15, −0.10)	−0.12 (−0.15, −0.10)	−0.06 (−0.06, −0.05)	−0.06 (−0.07, −0.05)
Self-care 3	−0.06 (−0.09, −0.04)	−0.06 (−0.08, −0.04)	−0.03 (−0.04, −0.02)	−0.03 (−0.04, −0.02)
Self-care 4	−0.28 (−0.32, −0.25)	−0.28 (−0.32, −0.25)	−0.13 (−0.14, −0.11)	−0.13 (−0.14, −0.11)
Self-care 5	−0.20 (−0.25, −0.16)	−0.20 (−0.24, −0.15)	−0.09 (−0.11, −0.07)	−0.09 (−0.11, −0.07)
Usual activities 2	−0.13 (−0.16, −0.11)	−0.14 (−0.16, −0.11)	−0.06 (−0.07, −0.05)	−0.06 (−0.07, −0.05)
Usual activities 3	−0.07 (−0.10, −0.05)	−0.07 (−0.10, −0.05)	−0.03 (−0.04, −0.02)	−0.03 (−0.04, −0.02)
Usual activities 4	−0.41 (−0.45, −0.37)	−0.41 (−0.45, −0.37)	−0.18 (−0.20, −0.17)	−0.18 (−0.20, −0.17)
Usual activities 5	−0.29 (−0.33, −0.25)	−0.29 (−0.33, −0.25)	−0.13 (−0.15, −0.11)	−0.13 (−0.15, −0.11)
Pain/discomfort 2	−0.16 (−0.18, −0.13)	−0.18 (−0.21, −0.16)	−0.07 (−0.08, −0.06)	−0.08 (−0.10, −0.07)
Pain/discomfort 3	−0.08 (−0.10, −0.05)	−0.08 (−0.11, −0.06)	−0.03 (−0.04, −0.02)	−0.04 (−0.05, −0.03)
Pain/discomfort 4	−0.67 (−0.73, −0.61)	−0.66 (−0.72, −0.60)	−0.30 (−0.32, −0.27)	−0.30 (−0.33, −0.28)
Pain/discomfort 5	−0.55 (−0.62, −0.49)	−0.55 (−0.61, −0.49)	−0.25 (−0.27, −0.22)	−0.25 (−0.27, −0.22)
Anxiety/depression 2	−0.22 (−0.24, −0.19)	−0.23 (−0.26, −0.20)	−0.10 (−0.11, −0.08)	−0.10 (−0.12, −0.09)
Anxiety/depression 3	−0.12 (−0.15, −0.10)	−0.12 (−0.15, −0.10)	−0.06 (−0.07, −0.04)	−0.06 (−0.07, −0.04)
Anxiety/depression 4	−0.50 (−0.56, −0.45)	−0.50 (−0.55, −0.44)	−0.22 (−0.25, −0.20)	−0.22 (−0.25, −0.20)
Anxiety/depression 5	−0.69 (−0.76, −0.62)	−0.68 (−0.75, −0.61)	−0.31 (−0.34, −0.28)	−0.31 (−0.34, −0.28)
γ Mobility	n/a	−0.24 (−0.38, −0.08)	n/a	n/a
γ Self-care	n/a	−0.45 (−0.80, −0.04)	n/a	n/a
γ Usual activities	n/a	−0.25 (−0.39, −0.09)	n/a	n/a
γ Pain/discomfort	n/a	−0.34 (−0.44, −0.23)	n/a	n/a
γ Anxiety/depression	n/a	−0.31 (−0.45, −0.14)	n/a	n/a
N	788	788	788	788

EQ-5D, EuroQol five-dimensional instrument; n/a, not applicable; QALY, quality-adjusted life year.
[a]Mean posterior estimates of μ and γ with 95% credible intervals in parenthesis

comparison between both samples is on the QALY scale. Only two parameters differ in magnitude on the QALY scale (i.e., those of the most severe levels of the mobility and self-care domains), whereas the other parameters have overlapping 95% credible intervals. Based on these results and the sensitivity analysis presented in the Supporting Information, both samples are safely pooled for the reference dependency analysis.

Table V contains the EQ-5D estimation results for the combined sample. Starting with the results of the specification that does not take reference dependency into account, the estimates in Table V are very similar to those presented in Table IV. As before, all estimates are logically consistent with the same clearly monotonic order of the attribute levels, and the estimations have sufficient statistical power to differentiate between the EQ-5D levels. Turning to the results of the specification that takes reference dependency into account, all reference dependency estimates are negative; this indicates that respondents who currently experience an impaired health state level assign less disutility to this level than respondents who currently experience a better health state level. Furthermore, all reference dependency estimates are statistically significant, that is, have 95% credible intervals that do not contain 0. Hence, we find clear evidence that respondents value EQ-5D health state levels below their own health differently than those equal to or higher than their own health state level. The observed difference is approximately 30%, with somewhat different estimates for the various EQ-5D health state dimensions. However, the accompanying Bayesian 95% credible intervals overlap and show little evidence that the reference dependency effect varies by dimension.

DOI: 10.1002/hec

4. DISCUSSION

4.1. Hypotheses

Reference dependency in health state valuations refers to the tendency of respondents in valuation surveys to take their own health state into account when valuing health states. Using a combined sample of 788 respondents who completed a DCE duration valuation study for the Dutch EQ-5D-5L, we found clear evidence of reference dependency: respondents with impaired health lower than or equal to the health state levels under evaluation were estimated to have, on average, approximately 30% smaller health state decrements. This confirms our hypothesis that reference dependency can be observed in a random sample from the general population, although it should be noted that the estimate for the EQ-5D 'self-care' dimension is relatively imprecise because of the vast majority of respondents without any impairment in this particular dimension (i.e., 96%; Table III). Despite the evidence of reference dependency, we found no support for our second hypothesis, namely, that correcting for respondents' own health results in a social tariff with more severe health state decrements. Instead, we found that the health state decrements before and after the reference dependency correction are almost identical. There are some minor differences for the EQ-5D 'usual activities' and 'pain/discomfort' dimensions, but these are well within the Bayesian 95% credible intervals of the estimates. Accordingly, we conclude that reference dependency does not bias the QALY estimates and that a general population sample can be considered as a reliable proxy for health state valuations by a sample of healthy respondents.

The finding that social values are hardly affected by reference dependency, in spite of the latter's large effect on the individual-level estimates, is somewhat surprising. Possibly, this result is contingent upon the use of a MIXL model, which already accommodates general respondent heterogeneity. The subsequent inclusion of reference dependency, which identifies a systematic part of the general preference heterogeneity, then does not alter the choice model outcomes. Hence, it is not guaranteed that the same results would have been obtained if a standard conditional logit model (which does not accommodate respondent heterogeneity) would have been used. However, because the i.i.d. assumption is not met, this model should not be used for a standard health state valuation model anyway (Train, 2003, pp. 52–53) as it can result in biased estimates.

4.2. Strengths and weaknesses

The presented analyses have at least two important limitations. The first limitation is that the presented analyses are based on a random sample from the general population that contains relatively few respondents with impaired health states. The latter implies that the reference dependency estimates are predominantly based on reference dependency in mild health states. Indeed, with a larger sample and with more respondents with impaired health states, the reference dependency estimates would have been more reliable. Furthermore, it would have been possible to determine whether reference dependency actually differs by EQ-5D dimension, which was not possible with the current sample. With a larger sample, it would also have been possible to accommodate interactions and nonlinear reference dependency (e.g., by allowing for stronger effects for more severe EQ-5D levels), which provides an interesting specification that was also beyond the scope of the available sample. The second important limitation is that the presented reference dependency estimates are solely based on respondents' current health status, which ignores the potential impact of respondents' past experiences and the health experience of people in respondents' direct vicinity (e.g., friends and relatives). Previous research based on visual analog scale measurements has determined that it is current health and not past experiences with impaired health states that have an effect on health state valuations (Dolan, 1996). Still, further research is required to establish whether these findings also hold for health state evaluations using DCEs.

Our analyses also have some important strengths, which are highlighted by the fact that evidence of reference dependency in health state valuations is actually found. First, a very efficient design is required to obtain sufficient statistical power to discriminate between the levels of the EQ-5D instrument. Existing valuation studies often have insufficient power to discriminate between adjacent health state levels (although this is often masked by only reporting statistical significance relative to perfect health rather than the preceding), which

DOI: 10.1002/hec

makes it impossible to identify an effect that is even smaller and modeled as a fraction of the difference between adjacent levels. Second, to separate reference dependency from other sources of respondent heterogeneity, a substantial amount of information for individual respondents is required. In the implemented matched pairwise DCE design, all EQ-5D parameters are identified at the individual level (i.e., information about all parameters is obtained for each respondent), and the heterogeneous Bayesian optimization procedure ensures a highly efficient DCE design at capturing respondent heterogeneity. In contrast, existing DCE valuation studies often use homogeneous designs and are based on fewer choice tasks per respondent and/or on the blocking of a single larger design (Oppe *et al.,* 2014; Bansback *et al.,* 2014; Norman *et al.,* 2013). The same issues are present in health state valuation studies based upon different valuation techniques, such as TTO and SG (e.g., Lamers *et al.,* 2006).

4.3. Implications

The presented results contribute towards an improved understanding of the differences between population subgroups in their valuation of health states. As shown, health state values are better described by models that include reference dependency than by standard expected utility models that overlook the relation between respondents' current health and their health state valuations. Interestingly, the practical implications of our findings depend on the position that one takes in the debate about whose values count. If health state values should reflect actually experienced health state conditions, the presented results imply a major shift in health state utility weights, which involves a significant reduction in health state decrements and in turn affects priority decisions based on QALY estimates – very similar to using patient values.

In most jurisdictions, however, social health state values are preferred. Irrespective of the reason why social values are requested (for democratic reasons, from an insurance perspective, or the Gold *et al.,* 1996 'veil of ignorance' argument), our study supports the current practice of using average population preferences derived from a sample from the general population, even when a sample of healthy respondents would have been preferred based on theoretical grounds. The latter is an important implication because, for example, Gold *et al.* (1996) and Brazier *et al*. (2007) argue, based on opposing theoretical arguments, that one sampling strategy should be preferable to the other. In contrast, the results presented in this paper clearly show that both strategies result in empirically indistinguishable tariffs, even though we can confirm that the sampling strategies themselves are quite different. Indeed, one would need to drop almost half (42%) of our combined general population sample to obtain a sub-sample of strictly healthy respondents.

An avenue for future research might involve investigating to what extent reference dependency explains the differences between patient values and social values. If the main difference between patients and a general population sample is their reference point, a desire to obtain experienced-based health state values (as expressed, for example, in Sweden) may not necessitate the involvement of patient samples. Instead, we hypothesize that a reference dependent QALY tariff that closely resembles patients' preferences can be derived from a general population sample, provided that the health states are described in generic terms and a sufficient number of respondents with impaired health are included.

Yet another interesting avenue for future research would be to determine the extent to which nonlinear time preferences (e.g., Doyle, 2013), interactions between EQ-5D attributes and levels (e.g., Stolk *et al.,* 2010), and the framing of the DCE tasks (e.g., Dolan *et al.,* 2003) have an impact on the presented results. In all of these cases, an efficient DCE design, such as the matched pairwise choice tasks as implemented in this paper, would be an important methodological consideration.

To conclude, the current paper quantified the effect of reference dependency on health state values and promoted understanding of how health state values differ between population subgroups. Strong reference dependence was found, suggesting substantial differences between healthy and unhealthy individuals. However, as a result of the low incidence of bad health states in the general population and the MIXL methodology that already accounts for general heterogeneity in preferences, we find that health state valuations from the general public and those from strictly healthy respondents result in indistinguishable QALY tariffs.

DOI: 10.1002/hec

CONFLICT OF INTEREST

The authors have no conflict of interest.

ACKNOWLEDGEMENTS

This work has relied on LISS data that were collected by CentERdata (Tilburg University, The Netherlands) through its MESS project, and has made use of supercomputer facilities that were sponsored by the National Computing Facilities Foundation (NCF), both of which are funded by the Netherlands Organization for Scientific Research (NWO). This work has also received financial support from the EuroQol Research Foundation, but does not necessarily reflect the opinion of the EuroQol Group.

REFERENCES

Bai, Y. An Adaptive Directional Metropolis-within-Gibbs Algorithm. *Preprint* (2009).

Bédard M, Douc R, Moulines E. 2014. Scaling analysis of delayed rejection MCMC methods. *Methodology and Computing in Applied Probability* **16**(4): 811–838.

Brazier JE, Ratcliffe J, Tsuchiya A, Salomon J. 2007. *Measuring and Valuing Health Benefits for Economic Evaluation*, Oxford University Press: Oxford.

Bansback N, Brazier J, Tsuchiya A, Anis A. Using discrete choice experiments to value EQ-5D health states. EuroQol plenary paper, Greece (2010).

Bansback N et al. 2014. Testing a discrete choice experiment including duration to value health states for large descriptive systems: addressing design and sampling issues. *Social Science & Medicine* **114**: 38–48.

Bleichrodt H, Pinto JL, Wakker PP. 2001. Making descriptive use of prospect theory to improve the prescriptive use of expected utility. *Management Science* **47**(11): 1498–1514.

Bliemer MCJ, Rose JM. 2010. Construction of experimental designs for mixed logit models allowing for correlation across choice observations. *Transportation Research Part B: Methodological* **44**(6): 720–734.

Dolan P. 1996. The effect of experience of illness on health-state valuations. *Journal of Clinical Epidemiology* **49**(5): 551–564.

Dolan P, Kahneman D. 2008. Interpretations of utility and their implications for the valuation of health. *The Economic Journal* **118**(525): 215–234.

Dolan P, Olsen JA, Menzel P, Richardson J. 2003. An inquiry into the different perspectives that can be used when eliciting preferences in health. *Health Economics* **12**(7): 545–551.

Doyle JR. 2013. Survey of time preference, delay discounting models. *Judgment and Decision Making* **8**(2): 116.

Flynn TN. 2010. Using conjoint analysis and choice experiments to estimate QALY values. *PharmacoEconomics* **28**(9): 711–722.

Gold ME, Russell LB, Siegel JE, Weinstein MC (Eds). 1996. *Cost-effectiveness in Health and Medicine*, Oxford University Press: New York.

Lamers LM et al. 2006. The Dutch tariff: results and arguments for an effective design for national EQ-5D valuation studies. *Health Economics* **15**(10): 1121–1132.

Louviere JJ et al. 2008. Designing discrete choice experiments: do optimal designs come at a price? *Journal of Consumer Research* **35**(2): 360–375.

Maddala T, Phillips KA, Reed Johnson F. 2003. An experiment on simplifying conjoint analysis designs for measuring preferences. *Health Economics* **12**(12): 1035–1047.

Menzel P et al. 2002. The role of adaptation to disability and disease in health-state valuation: a preliminary normative analysis. *Social Science & Medicine* **55**(12): 2149–2158.

Norman R, Cronin P, Viney R. 2013. A pilot discrete choice experiment to explore preferences for EQ-5D-5L health states. *Applied Health Economics and Health Policy* **11**(3): 287–298.

Oppe M et al. 2014. A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. *Value in Health* **17**(4): 445–453.

Plummer M, Best N, Cowles K, et al. 2006. CODA: convergence diagnosis and output analysis for MCMC. *R News* **6**: 7–11.

Sandor Z, Wedel M. 2001. Designing conjoint choice experiments using managers' prior beliefs. *Journal of Marketing Research* **38**(4): 430–444.

Sándor Z, Wedel M. 2005. Heterogeneous conjoint choice designs. *Journal of Marketing Research* **42**(2): 210–218.
Scherpenzeel AC, Das M. 2010. "True" longitudinal and probability-based internet panels: evidence from the Netherlands. In *Social and Behavioral Research and the Internet: Advances in Applied Methods and Research Strategies*, Das M, Ester P, Kaczmirek L (eds.), Taylor & Francis: Boca Raton; 77–104.
Stamuli E. 2011. Health outcomes in economic evaluation: who should value health? *British Medical Bulletin* **97**(1): 197–210.
Stocki R. 2005. A method to improve design reliability using optimal Latin hypercube sampling. *Computer Assisted Mechanics and Engineering Sciences* **12**(4): 393.
Stolk EA, Oppe M, Scalone L, Krabbe PF. 2010. Discrete choice modeling for the quantification of health states: the case of the EQ-5D. *Value in Health* **13**(8): 1005–1013.
Train K. 2003. *Discrete Choice Methods with Simulation*(Second edn), Cambridge University Press: Cambridge.
Ubel PA, Loewenstein G, Jepson C. 2003. Whose quality of life? A commentary exploring discrepancies between health-state evaluations of patients and the general public. *Quality of Life Research* **12**(6): 599–607.
Versteegh MM, Vermeulen KM, Evers SM, de Wit GA, Prenger R, Stolk EA. 2016. Dutch Tariff for the five-level version of EQ-5D2016. *Value in Health* . DOI:10.1016/j.jval.2016.01.003.
Wakker PP. 2010. *Prospect Theory: For Risk and Ambiguity*, Cambridge University Press: Cambridge.
De Wit GA, Busschbach JJV, De Charro F. 2000. Sensitivity and perspective in the valuation of health status: whose values count? *Health Economics* **9**(2): 109–126.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's website.

DOI: 10.1002/hec