Article:

Mulhern, B.J., Bansback, N., Norman, R. et al. (1 more author) (2020) Valuing the SF-6Dv2 classification system in the United Kingdom using a discrete-choice experiment with duration. Medical Care, 58 (6). pp. 566-573. ISSN 0025-7079

https://doi.org/10.1097/mlr.0000000000001324

OPEN

Valuing the SF-6Dv2 Classification System in the United Kingdom Using a Discrete-choice Experiment With Duration

Brendan J. Mulhern, MRes,† Nick Bansback, PhD,‡ Richard Norman, PhD,§ John Brazier, PhD,† and on behalf of the SF-6Dv2 International Project Group*

Objective: An updated version of the SF-6D Classification System (SF-6Dv2) has been developed, and utility value sets are required. The aim of this study was to test the development of a United Kingdom SF-6Dv2 value set, and address limitations of the existing SF-6D value set (which results in a narrow range of utilities). This was done using 2 discrete-choice experiment (DCE) tasks. Interactions and preference heterogeneity were also investigated.

Research Design and Subjects: An online sample of respondents (n = 3014) completed 10 DCE with duration choice sets from an efficient design of 300 (Design 1) and 2 DCE with duration choice sets including immediate death from a set of 60 (Design 2). Conditional logit regression was used to estimate value set models with and without interactions. We investigated preference heterogeneity using latent class models.

Results: Models including ordered coefficients within each dimension were developed, with the favored model including an additional interaction term when one dimension was at the most severe level. Value sets differed across Designs 1 and 2. Design 1 models had a wider utility range and a higher proportion of negative values. The most important dimensions were pain, mental health, and physical functioning. Preference heterogeneity was apparent, with a 2-class model describing the data.

Conclusions: We developed and applied a protocol to value the SF-6Dv2 using DCE. The results provide a provisional value set for use in resource allocation. The protocol can be applied internationally. Further work should investigate how to account for preference heterogeneity in value set production.

Key Words: SF-6D, utilities, discrete-choice experiments, valuation, quality-adjusted life year

(*Med Care* 2020;58: 566–573)

From the *Centre for Health Economics Research and Evaluation, University of Technology Sydney, Sydney, NSW, Australia; †Health Economics and Decision Science, School of Health and Related Research, University of Sheffield, Sheffield, UK; ‡School of Population and Public Health, Vancouver, BC, Canada; and §School of Public Health, Curtin University, Bentley, WA, Australia.

Presented at the International Academy of Health Preference Research and the Australian Health Economics Study Group in Brisbane, QLD, Australia, in October 2015.

Supported by royalties paid by users of version 1 of the SF-6D.

The authors declare no conflict of interest.

Correspondence to: Brendan J. Mulhern, MRes, Centre for Health Economics Research and Evaluation, University of Technology Sydney, 1-59 Quay Street, Haymarket, Sydney, NSW 2000, Australia. E-mail: brendan.mulhern@chere.uts.edu.au.

Supplemental Digital Content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website, www.lww-medicalcare.com.

ISSN: 0025-7079/20/5806-0566

The SF-6D[1,2] is a generic preference-based measure (PBM) used to estimate quality-adjusted life years (QALYs) in the economic evaluation of health technologies. PBMs provide a utility weight anchored on a full health (1) to dead (0) scale, with negative values equivalent to states worse than dead. This weight is usually generated from general population preferences using valuation methods such as standard gamble (SG),[1] time trade off (TTO),[3,4] or discrete-choice experiments (DCE).[5]

Version 1 of the SF-6D (hereon SF-6Dv1) was derived from the SF-36[6] and has been used widely to inform resource allocation.[7–12] It assesses health on 6 dimensions [Physical Functioning (PF), Role Functioning (RF), Social Functioning (SF), Pain (PA), Mental Health (MH), and Vitality (VT)] with 4–6 response levels. The UK valuation study was carried out using SG, producing a utility scale ranging from 0.29 to 1.[1,2,13]

Although used widely, the SF-6Dv1 has been criticized on both measurement and valuation grounds. A floor effect for the RF dimension means that patients score at the lowest possible level, and 4 levels collapse into 2 utility values, leading to insensitivity to change.[14–16] There is ambiguity between the intermediate severity levels of PF. The positively framed VT item contrasts with the negatively framed dimensions. The valuation resulted in high values for severe health states, and disordered levels. Ordering was forced by constraining levels to be the same. This reduced the number of values and impacted the sensitivity of the utility scale.[17] Due to these concerns, an updated classification system has been developed (SF-6Dv2).[18] A value set is now required.

Value sets have been developed using SG and TTO.[17,19] SG is grounded in expected utility theory, and respondents trade between a fixed state, and a probability of full health or death. SG has been criticized due to the complex nature of the probability trade off, and risk aversion tends to result in higher values.[17] TTO[20] involves trading in time and quality of life (QoL). Respondents may be able to trade time more easily than risk, but the iterative nature of the task, and the process for valuing states worse than dead, have been criticized.[21]

DCEs are based on random utility theory[22,23] and are used to estimate health state values.[24–26] Respondents choose between sets of health profiles, and a different cognitive process is required. However, the values are latent, and not on the utility scale. To anchor values, survival can be incorporated into the scenarios (described as DCE_{TTO}).[27] DCE_{TTO} has been used to value the EQ-5D-3L,[27,28] EQ-5D-5L,[29–31] and SF-6Dv1[32] internationally. Areas for further methodological consideration include comparing different DCE_{TTO} task formats,[33] the impact of interactions on values, and preference heterogeneity.

The aims of this study are 2-fold:

(1) To test the development of a UK value set for the SF-6Dv2 using a protocol including 2 DCE_{TTO} task formats.
(2) Explore DCE_{TTO} specific methodological issues including the impact of interactions and preference heterogeneity.

This is the first study to value SF-6Dv2 using an online DCE_{TTO} protocol combining multiple formats, and updates the work by Brazier and colleagues[1,2] that developed SF-6Dv1. The protocol aims to improve SF-6D utilities by solving the existing issues including the constraining of severity levels leading to fewer overall values, and restriction of the range of the utilities produced. It is the first study to test heterogeneity for SF-6D dimensions using DCE_{TTO}, which is important, given the differing health experiences and life stages of general population respondents.

METHODS

The SF-6Dv2 Classification System

The SF-6Dv2 health state classification system (Appendix 1, Supplemental Digital Content 1, http://links.lww.com/MLR/B997) was developed using an established process for adapting existing QoL measures into classification systems.[34] This includes dimensionality assessment and item response theory to select items to represent each dimension.

The resulting classification system includes the same 6 dimensions as SF-6Dv1 (PF, RF, SF, PA, MH, VT).[18] The dimension descriptions have changed for all dimensions apart from SF. The classification system was derived from the SF-36v2 and was not restricted to items on both the SF-36 and SF-12.

DCE_{TTO} Task Format Designs

2 different choice set formats were designed. Design 1 displayed pairs of health profiles and 1 of 4 duration levels (1, 4, 7, and 10 y). Respondents chose which was better. An upper limit of 10 years was chosen for comparability with the time horizon used in the many TTO studies.[3,4] This format has been implemented in previous DCE_{TTO} valuation studies.[27,29,30,35]

Design 2 displayed pairs of SF-6Dv2 health profiles and duration as options A and B, and a third option of "Immediate death." Respondents provided a full ranking by indicating the best and the worst. This format has been used in the valuation of SF-6Dv1.[32] The dimension order within the choice set was randomized between respondents. We did this to counteract any impact of completion heuristics on the basis of dimension position.

Study Design

In past work with DCE_{TTO}, the number of choice sets in the design exceeds the number of parameters that the model is estimating[29] and we followed that approach here. The number of parameters estimated was 102 {100 interactions of dimension level and continuous duration [(25×4) = 100], 1 continuous duration and 1 extra term}. Design 1 included 300 choice sets divided into 30 blocks of 10 constructed using D-Optimal design methods in NGene.[36] Respondents were randomly allocated to a block. The choice set order within blocks was randomized.

Design 2 included extra 60 choice sets (2 per survey version). These were selected from the Design 1 choice sets based on the severity of the profiles, where more severe combinations were used. The immediate death option was appended to these choice sets. The Design 1 choice sets always appeared first, followed by Design 2. This was done as the tasks increase in complexity from presenting pairs to presenting triplets.

Recruitment and Survey Completion

Respondents representative of the UK population in age (18+) and sex were recruited from an online panel (Survey Sampling International), who randomly allocated individuals willing to take a survey at the time of data collection. Respondents read study information and consented. They then completed demographic questions, the SF-6Dv2 self-report version (Appendix 2, Supplemental Digital Content 1, http://links.lww.com/MLR/B997), read task instructions (Appendix 3, Supplemental Digital Content 1, http://links.lww.com/MLR/B997), and completed 10 Design 1 tasks, 2 Design 2 tasks, and the EQ-5D-5L.[37] Those completing the survey in > 2 minutes (the minimum time to be classified as a completer) were provided with an incentive. This process received approval from the University of Sheffield ethics committee.

Analysis—DCE_{TTO} Models

Conditional logit regression was the initial method used to model both Designs 1 and 2. We estimated coefficients for each level of each dimension interacted with continuous life years t, with the least severe level used as the baseline. The utility of profile j for individual i is as follows:

$$\mu_{ij} = \beta t_{ij} + \lambda' x_{ij} t_{ij} + \varepsilon_{ij}. \tag{1}$$

For Design 1, respondent i provided binary outcomes for choices between 2 profiles j. For Design 2, the respondent i provided data about which of the 3 profiles j is best or worst to provide a full ranking assuming independence of irrelevant alternatives. The coefficient β reflected the value of living in full health for 1 year and was expected to be positive; λ represented the disutility of the interaction between living with the SF-6Dv2 problems (x) for a duration of 1 year and was expected to be negative (indicating a decrement in utility in comparison with the baseline). The error term ε_{ij} was random.

The interacted value x_j was anchored on the health utility scale (V) using the coefficient β fixed at 1, and the adjusted disutility associated with each particular health state. This was the ratio of $\hat{\lambda}$ (dimension level coefficient) and $\hat{\beta}$ (duration coefficient) multiplied by the relevant terms in x_j:

$$V_j = 1 + \frac{\hat{\lambda}'}{\hat{\beta}} x_j. \tag{2}$$

Thus, for full health, this value is 1 (as all x_j terms are 0), but for nonfull health states, the effect of x_j is negative (representing a decrement). V_j can be negative, indicating a state worse than dead.

We also assessed the impact of including an interaction term (WORST1), which is included when a health state has ≥ 1 dimensions at the most severe level.

The results of the conditional logit regression are reported in terms of the "unanchored" (β and λ) and "anchored" coefficients (λ/β) that are on the utility scale and are comparable. To compare the estimates, we assessed the number of inconsistent coefficients (when an increase in severity leads to an increase rather than a decrease in utility), overall utility ranges, and proportion of states worse than dead. Model consistency is achieved by combining disordered levels. Model fit statistics tested include the log likelihood, and the Bayesian Information Criterion (BIC), which accounts for both the number of parameters and observations.

Exploration of Heterogeneity

Conditional logit assumes that all respondents share a common unobservable set of values. This may not be a realistic assumption, as the way people perceive health differs.

Therefore, we investigated preference heterogeneity using latent class modeling.[38] The baseline utility function was adjusted to incorporate heterogeneity into the main coefficients for each individual respondent (i):

$$u_{ij} = \beta_i t_j + \lambda_i' x_j t_j + \varepsilon_{ij}. \tag{3}$$

Models including 2–6 classes were tested. The number of classes to extract was guided by the BIC, where, the model with the lowest value is preferred. To understand how preferences differ across the population, parameters indicating class membership of different demographic groups were estimated as binary dummy variables. These included age (18–45 and 46+), sex, having a long-term condition, and having children. Each of these may have different impacts in terms of health experiences, or external factors that may affect preferences. Health state values for class C were calculated as:

$$V_{Cj} = 1 + \frac{\hat{\lambda}_C'}{\hat{\beta}_C} x_j. \tag{4}$$

All analyses used Stata 15.[39]

RESULTS

Response Rate and the Sample

Overall, 5820 panel members were invited to take part, and 3948 (67.8%) accessed the survey. Of the responders, 429 (7.4%) were from an age and sex quota that was complete, and 459 (7.9%) started the survey, but did not complete it. This left 3014 (51.8%) completers. Table 1 reports the demographics of the 3000 respondents who provided full background information. The sample is matched to the UK general population in terms of age (18–25; 26–35; 36–45; 46–55; 56–65; 65+) and sex (49% men).[40]

TABLE 1. Demographic Characteristics

Characteristics	n (%)
Age	
Mean	46
Range	18–86
Male	1461 (49)
Married	1815 (60)
In employment	1582 (53)
Education > minimum age	2341 (78)
Have children	1442 (49)
Experience serious illness (self)	983 (33)
Experience serious illness (family)	1985 (67)
Experience serious illness (caring)	745 (25)
Have long-term condition	1493 (50)
SF-6Dv2	
At ceiling	164 (5.6)
At floor	9 (0.3)
Physical functioning	
Limited in vigorous activities *not at all*	1089 (36)
Limited in vigorous activities *a little*	1280 (43)
Limited in moderate activities *a little*	311 (10)
Limited in moderate activities *a lot*	224 (8)
Limited in bathing and dressing *a lot*	88 (3)
Role functioning	
Accomplish less than you would like *none of the time*	1168 (39)
Accomplish less than you would like *a little of the time*	835 (28)
Accomplish less than you would like *some of the time*	585 (20)
Accomplish less than you would like *most of the time*	281 (9)
Accomplish less than you would like *all of the time*	120 (4)
Social functioning	
Social activities are limited *none of the time*	1548 (52)
Social activities are limited *a little of the time*	612 (21)
Social activities are limited *some of the time*	493 (17)
Social activities are limited *most of the time*	211 (7)
Social activities are limited *all of the time*	103 (3)
Pain	
No pain	781 (26)
Very mild pain	869 (29)
Mild pain	622 (21)
Moderate pain	470 (16)
Severe pain	182 (6)
Very severe pain	61 (2)
Mental health	
Depressed or very nervous *none of the time*	1238 (41)
Depressed or very nervous *a little of the time*	900 (30)
Depressed or very nervous *some of the time*	509 (17)
Depressed or very nervous *most of the time*	268 (9)
Depressed or very nervous *all of the time*	75 (3)
Vitality	
Worn out *none of the time*	531 (18)
Worn out *a little of the time*	1095 (37)
Worn out *some of the time*	769 (26)
Worn out *most of the time*	431 (15)
Worn out *all of the time*	153 (5)

SF-6Dv2 indicates SF-6D Version 2 Classification System.

Unanchored DCE$_{TTO}$ Models—Design 1

Table 2 shows the unanchored models. Model 1 shows the unrestricted coefficients for Design 1. There is statistically significant disordering between levels 2 (worn out a little of the time) and 3 (worn out some of the time) of VT. These 2 levels were combined to generate Model 2 (a consistent model). Most of the coefficients at the more severe levels are significant at the 0.001 level both in comparison with the baseline and adjacent severity levels. Model 3 shows the coefficients for the ordered model including WORST1, which is negative, meaning that it leads to a further decrease in utility when applied. The standard errors, log likelihoods, and BICs were similar across the Design 1 models.

Anchored Models—Design 1

Table 3 shows the anchored health utility decrements for Models 2 and 3. The range of values produced for Model 2 is

TABLE 2. Unanchored Models (Designs 1 and 2)

	Model 1: Design 1 (Inconsistent)			Model 2: Design 1 Model 1 (Consistent)			Model 3: Design 1 WORST1 Term (Consistent)			Model 4: Design 2 (Inconsistent)			Model 5: Design 2 Model 4 (Consistent)			Model 6: Design 2 WORST1 Term (Consistent)		
Parameters	Coef.[†]	Sig (bet)[‡]	SE	Coef.	Sig (bet)	SE	Coef.	Sig (bet)	SE	Coef.	Sig (bet)	SE	Coef.	Sig (bet)	SE	Coef.	Sig (bet)	SE
PF2×LY[§]	−0.006		0.004	−0.006		0.005	−0.005		0.005	−0.007		0.004	−0.007		0.005	−0.007		0.005
PF3×LY	−0.010**	0.394	0.004	−0.010**	0.341	0.004	−0.010*	0.317	0.004	−0.011*	0.363	0.004	−0.011*	0.387	0.004	−0.012	0.349	0.004
PF4×LY	−0.029***	< 0.001	0.004	−0.030***	< 0.001	0.005	−0.027***	< 0.001	0.005	−0.029***	< 0.001	0.004	−0.029***	< 0.001	0.005	−0.027***	0.001	0.005
PF5×LY	−0.064***	< 0.001	0.004	−0.064***	< 0.001	0.005	−0.055***	< 0.001	0.005	−0.062***	< 0.001	0.004	−0.062***	< 0.001	0.005	−0.052***	< 0.001	0.005
RF2×LY[‖]	−0.013***		0.004	−0.012***		0.004	−0.012**		0.004	−0.012*		0.004	−0.007		0.005	−0.009*		0.005
RF3×LY	−0.014***	0.773	0.004	−0.014***	0.680	0.004	−0.016***	0.308	0.004	−0.002	0.024	0.004	−0.007	NA	0.005	−0.009*	NA	0.005
RF4×LY	−0.030***	< 0.001	0.004	−0.030***	< 0.001	0.005	−0.030***	0.003	0.004	−0.018***	< 0.001	0.004	−0.018***	0.004	0.004	−0.019***	0.012	0.004
RF5×LY	−0.038***	0.068	0.004	−0.038***	0.068	0.005	−0.030***	0.904	0.004	−0.029***	0.013	0.004	−0.029***	0.015	0.004	−0.019***	0.983	0.004
SF2×LY[¶]	−0.001		0.005	−0.001		0.005	−0.002		0.005	−0.001		0.005	−0.001		0.004	−0.002		0.004
SF3×LY	−0.008**	0.107	0.004	−0.008**	0.118	0.004	−0.009*	0.193	0.004	−0.006	0.125	0.004	−0.007	0.156	0.004	−0.007	0.241	0.005
SF4×LY	−0.030***	< 0.001	0.005	−0.030***	< 0.001	0.005	−0.031***	< 0.001	0.005	−0.029***	< 0.001	0.004	−0.030***	< 0.001	0.005	−0.031***	< 0.001	0.005
SF5×LY	−0.049***	< 0.001	0.004	−0.050***	< 0.001	0.005	−0.041***	0.034	0.005	−0.046***	< 0.001	0.004	−0.048***	< 0.001	0.005	−0.036***	0.246	0.004
PA2×LY[#]	−0.023***		0.005	−0.023***		0.005	−0.023***		0.005	−0.014*		0.005	−0.013*		0.004	−0.012*		0.004
PA3×LY	−0.029***	0.229	0.005	−0.029***	0.184	0.005	−0.029***	0.189	0.005	−0.025***	0.024	0.005	−0.026***	0.014	0.004	−0.026***	0.006	0.005
PA4×LY	−0.042***	0.006	0.005	−0.043***	0.005	0.005	−0.042***	0.010	0.005	−0.030***	0.303	0.005	−0.032***	0.186	0.005	−0.030***	0.484	0.005
PA5×LY	−0.135***	< 0.001	0.005	−0.136***	< 0.001	0.006	−0.137***	< 0.001	0.005	−0.125***	< 0.001	0.005	−0.126***	< 0.001	0.004	−0.128***	< 0.001	0.005
PA6×LY	−0.195***	< 0.001	0.005	−0.195***	< 0.001	0.006	−0.185***	< 0.001	0.006	−0.191***	< 0.001	0.005	−0.191***	< 0.001	0.005	−0.179***	< 0.001	0.004
MH2×LY[††]	−0.008*		0.004	−0.009*		0.004	−0.008		0.005	−0.007		0.005	−0.006		0.005	−0.005		0.004
MH3×LY	−0.026***	< 0.001	0.004	−0.025***	< 0.001	0.004	−0.026***	< 0.001	0.005	−0.019***	0.007	0.004	−0.019***	0.008	0.004	−0.018***	0.006	0.005
MH4×LY	−0.073***	< 0.001	0.004	−0.074***	< 0.001	0.005	−0.071***	< 0.001	0.005	−0.076***	< 0.001	0.004	−0.075***	< 0.001	0.004	−0.071***	< 0.001	0.004
MH5×LY	−0.106***	< 0.001	0.004	−0.105***	< 0.001	0.005	−0.097***	< 0.001	0.005	−0.104***	< 0.001	0.004	−0.103***	< 0.001	0.004	−0.091***	< 0.001	0.005
VT2×LY[‡‡]	−0.011**		0.004	−0.005		0.004	−0.005		0.004	−0.019***		0.004	−0.006		0.005	−0.006		0.004
VT3×LY	0.001	0.010	0.004	−0.005	NA	0.004	−0.005	NA	0.004	0.005	< 0.001	0.005	−0.006	NA	0.004	−0.006	NA	0.004
VT4×LY	−0.026***	< 0.001	0.004	−0.026***	< 0.001	0.005	−0.024***	< 0.001	0.005	−0.019***	< 0.001	0.005	−0.019***	< 0.001	0.005	−0.018***	0.002	0.004
VT5×LY	−0.044***	< 0.001	0.004	−0.044***	< 0.001	0.005	−0.036***	0.007	0.005	−0.042***	< 0.001	0.005	−0.042***	< 0.001	0.005	−0.032***	0.001	0.005
LY[§§]	0.290***		0.008	0.290***		0.009	0.298***		0.009	0.338***		0.008	0.340***		0.009	0.352****		0.009
WORST1_LY[‖‖]							−0.025		0.006							−0.035		0.006
No. observations	30,140			30,140			30,140			39,182			39,182			39,182		
Log likelihood	−18,419			−18,422			−18,413			−23,635			−23,654			−23,631		
BIC	37,124			37,120			37,112			52,696			52,735			52,700		

†Coefficient estimate.
‡Significance between dimension levels.
§Interactions of Physical Functioning dimension levels and duration.
‖Interactions of Role Functioning dimension levels and duration.
¶Interactions of Social Functioning dimension levels and duration.
#Interactions of Pain dimension levels and duration.
††Interactions of Mental Health dimension levels and duration.
‡‡Interactions of Vitality dimension levels and duration.
§§Duration (life years).
‖‖Interaction when ≥1 dimension is at the worst level.
BIC indicates Bayesian Information Criterion; NA, not available.
*Significant at 0.05.
**Significant at 0.01.
***Significant at 0.001.

TABLE 3. Anchored Value Set Models

Parameter	Model 2	Model 3	Model 5	Model 6
PF1†	0	0	0	0
PF2	−0.021	−0.019	−0.022	−0.021
PF3	−0.036**	−0.034*	−0.033*	−0.033*
PF4	−0.102***	−0.092***	−0.087***	−0.076***
PF5	−0.221***	−0.186***	−0.183***	−0.147***
RF1‡	0	0	0	0
RF2	−0.042**	−0.039**	−0.020	−0.025*
RF3	−0.049***	−0.055***	−0.020	−0.025*
RF4	−0.104***	−0.099***	−0.053***	−0.053***
RF5	−0.132***	−0.102***	−0.085***	−0.054***
SF1§	0	0	0	0
SF2	−0.004	−0.008	−0.001	−0.005
SF3	−0.029**	−0.029*	−0.020	−0.020
SF4	−0.102***	−0.103***	−0.087***	−0.087***
SF5	−0.171***	−0.137***	−0.140***	−0.103***
PA1∥	0	0	0	0
PA2	−0.079***	−0.076***	−0.040*	−0.035*
PA3	−0.102***	−0.097***	−0.076***	−0.074***
PA4	−0.148***	−0.139***	−0.095***	−0.084***
PA5	−0.469***	−0.460***	−0.371***	−0.364***
PA6	−0.673***	−0.620***	−0.565***	−0.507***
MH1¶	0	0	0	0
MH2	−0.030*	−0.026*	−0.018	−0.015
MH3	−0.089***	−0.086***	−0.055***	−0.051***
MH4	−0.253***	−0.236***	−0.222***	−0.204***
MH5	−0.361***	−0.324***	−0.303***	−0.259***
VT1#	0	0	0	0
VT2	−0.017	−0.015	−0.018	−0.017
VT3	−0.017	−0.015	−0.018	−0.017
VT4	−0.089***	−0.080***	−0.057***	−0.051***
VT5	−0.150***	−0.121***	−0.123***	−0.091***
WORST > 1††		−0.084***		−0.100
Value set characteristics				
Range	1 to −0.709	1 to −0.574	1 to −0.399	1 to −0.261
SWD (%)‡‡	15.0	15.2	4.3	3.8
Overall coefficient magnitude	Pain—Mental Health—Physical Functioning—Social Functioning—Vitality—Role Functioning			

†Physical Functioning.
‡Role Functioning.
§Social Functioning.
∥Pain.
¶Mental Health.
#Vitality.
††Interaction estimate.
‡‡Percentage of states valued as worse than dead.
*Significant at 0.05.
**Significant at 0.01.
***Significant at 0.001.

from 1 (111111) to −0.708 (555655) and 15% of all 18,750 states are negative. The health state dimension coefficients for Model 3 are smaller overall, but the WORST1 term leads to an extra decrement. This results in a smaller utility range (1 to −0.574), with a similar percentage (15.2%) worse than dead. Figure 1 shows a density plot including Models 2 and 3, and SF-6Dv1. The SF-6Dv2 models have a similar smooth distribution across the utility range, but differ markedly from the SF-6Dv1, where the most values are clustered between 0.4 and 0.75.

Unanchored Models—Design 2

Table 2 shows the unanchored Design 2 models. Model 4 shows the unrestricted coefficients. The coefficients for RF levels 2 (accomplished less a little of the time) and 3 (accomplished less some of the time) and VT levels 2 and 3 are disordered (with the disordering significant). Model 5 shows the consistent model. Model 6 includes WORST1, which has the same pattern as Design 1. The majority of the coefficients significantly differ from the dimension baseline, and adjacent severity levels. As with Design 1, the SEs, log likelihoods, and BICs for the Design 2 models were similar.

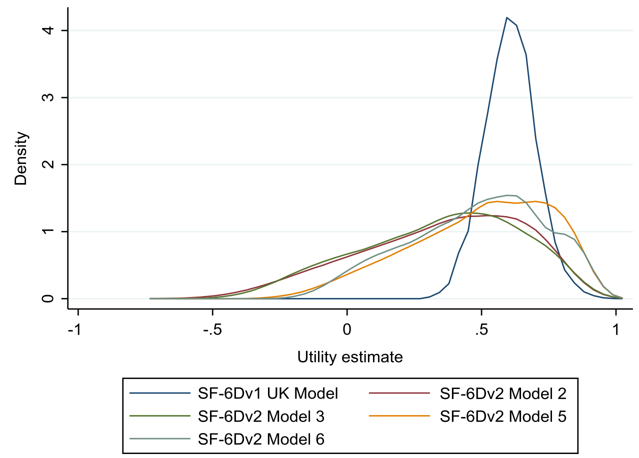

FIGURE 1. Density plots of the estimated value sets and SF-6D Classification System (SF-6Dv1). full color online

Anchored Models—Design 2

Table 3 shows the anchored coefficients for Design 2. Model 5 demonstrates that introducing the Design 2 choice sets reduces the utility range (1 to −0.399) and the percentage of negative states (to 4.3%). A similar pattern as for Design 1 applies when WORST1 is included. The dimension level coefficients for Model 6 are smaller than Model 5, but the extra term leads to a further decrement. Figure 1 suggests that the Design 2 models have a higher density of values between 0.5 and 1 and a lower density <0.5 than Design 1 and differ from the SF-6Dv1.

Assessing Heterogeneity

The BIC was the lowest for the latent class model with 2 classes (Table 4). Across both designs, class 1 includes respondents (42% and 52%, respectively) who display a strong preference for longer duration and avoiding pain, but less on the other 5 dimensions. Class 2 (58% Design 1; 48% Design 2) includes respondents who place more weight on 3 health state dimensions (PF, PA, and MH). Class 1 is more likely to include older respondents and those with children, and less likely to include respondents with a long-term condition.

CONCLUSIONS

This article describes a study using a DCE_{TTO}-based protocol including 2 task formats to estimate a value set for the SF-6Dv2 classification system (derived from the SF-36v2). The results generally reflect the monotonic nature of the instrument, where the magnitude of the utility increases as the severity of the health dimension also increases. This is a key requirement of value sets for use in QALY estimation. The protocol used also explores a number of important aspects of design and analysis. The addition of Design 2 tasks including immediate death reduces the overall utility range and frequency of states worse than dead.

TABLE 4. Latent Class Models With 2 Classes

	Design 1				Design 2			
	Class 1		Class 2		Class 1		Class 2	
Parameters	Coef.*	Utility	Coef.	Utility	Coef.	Utility	Coef.	Utility
PF2×LY†	–0.012	–0.021	–0.010	–0.056	0.006	0.011	–0.021	–0.097
PF3×LY	–0.033	–0.057	–0.006	–0.033	–0.027	–0.048	–0.013	–0.060
PF4×LY	–0.014	–0.024	–0.042	–0.233	–0.003	–0.005	–0.050	–0.230
PF5×LY	–0.088	–0.151	–0.063	–0.350	–0.071	–0.126	–0.069	–0.318
RF2×LY‡	–0.021	–0.036	–0.016	–0.089	–0.022	–0.039	–0.002	–0.009
RF3×LY	–0.027	–0.046	–0.010	–0.056	–0.009	–0.016	0.014	0.065
RF4×LY	–0.058	–0.100	–0.018	–0.100	–0.034	–0.060	0.008	–0.037
RF5×LY	–0.095	–0.164	–0.021	–0.117	–0.053	–0.094	–0.015	–0.069
SF2×LY§	0.013	0.022	–0.009	–0.050	0.016	0.028	–0.012	–0.055
SF3×LY	0.010	0.017	–0.017	–0.094	0.020	0.035	–0.027	–0.124
SF4×LY	–0.031	–0.053	–0.031	–0.172	–0.031	–0.055	–0.027	–0.124
SF5×LY	–0.074	–0.127	–0.039	–0.217	–0.053	–0.094	–0.048	–0.221
PA2×LY‖	–0.047	–0.081	–0.019	–0.106	–0.030	–0.053	–0.009	–0.041
PA3×LY	–0.061	–0.105	–0.024	–0.133	–0.040	–0.071	–0.020	–0.092
PA4×LY	–0.073	–0.126	–0.031	–0.172	–0.067	–0.119	–0.011	–0.051
PA5×LY	–0.253	–0.435	–0.094	–0.522	–0.239	–0.423	–0.060	–0.276
PA6×LY	–0.349	–0.601	–0.142	–0.789	–0.393	–0.696	–0.077	–0.355
MH2×LY¶	0.011	0.019	–0.015	–0.083	0.038	0.067	–0.028	–0.129
MH3×LY	–0.009	–0.015	–0.034	–0.189	–0.007	–0.012	–0.031	–0.143
MH4×LY	–0.089	–0.153	–0.075	–0.417	–0.147	–0.260	–0.045	–0.207
MH5×LY	–0.113	–0.194	–0.115	–0.639	–0.171	–0.303	–0.069	–0.318
VT2×LY#	–0.010	–0.017	–0.010	–0.056	–0.012	–0.021	–0.017	–0.078
VT3×LY	0.016	0.028	–0.004	–0.022	0.010	0.018	0.007	–0.032
VT4×LY	–0.021	–0.036	–0.031	–0.172	–0.034	–0.060	–0.013	–0.060
VT5×LY	–0.092	–0.158	–0.024	–0.133	–0.094	–0.166	–0.016	–0.074
LY**	0.581		0.180		0.565		0.217	
Range		1 to –0.329		1 to –1.245		1 to –0.479		1 to –0.355
Class share	0.421		0.579		0.520		0.480	
Demographics (baseline class 2)								
Age	0.667		0		0.785		0	
Sex	0.019		0		0.340		0	
Have long-term condition	0.047		0		0.065		0	
Have children	0.326		0		0.266		0	
BIC		63,255				63,525		

*Coefficient estimate.
†Interactions of Physical Functioning dimension levels and duration.
‡Interactions of Role Functioning dimension levels and duration.
§Interactions of Social Functioning dimension levels and duration.
‖Interactions of Pain dimension levels and duration.
¶Interactions of Mental Health dimension levels and duration.
#Interactions of Vitality dimension levels and duration.
**Duration (life years).
BIC indicates Bayesian Information Criterion.

Model 3 is recommended for use in the estimation of QALYs from SF-6Dv2. This model is ordered within dimensions, where increasing severity leads to a decrease in utility, and is based on an efficient design developed using established experimental design procedures. The addition of Design 2 was methodological in nature; thus, using the core design developed using efficient procedures is preferred. The utilities estimated from the SF-36 differ in a number of ways (Appendix 4, Supplemental Digital Content 1, http://links.lww.com/MLR/B997). The classification system has been improved by simplifying the dimension descriptions and changing the direction of all dimensions to negative framing. The value set evidenced a wider range with more possible values, given less disordering than version 1, which will improve the sensitivity of utilities to change in health. Appendix 5 (Supplemental Digital Content 1, http://links.lww.com/MLR/B997) describes how to calculate health state values using Model 3.

In all of the unrestricted models, there is a small reversal between levels 2 and 3 of VT. This could be linked to the overall severity of the dimension where "worn out" could be perceived as a nonsevere health issue for the general population. The response levels used "a little of the time" and "some of the time" and respondents may not be able to tell which is worse.

The value sets produced using DCE differ from those for SF-6Dv1,[1] and this has implications for decision-making. In comparison to Model 3, the most striking difference is the larger range, with the minimum value calculated as –0.574 compared with 0.29. This will have implications for the magnitude of QALYs estimated using SF-6D. Explicitly, it will lead to relative prioritization of treatments that benefit QoL as the utility values will result in a larger QALY gain. The value set includes negative values (states modeled as worse than dead), which was not the case for SF-6Dv1. This is in part due to the valuation method used, because SG generates higher values.[17] Changes in the descriptive system, particularly in terms of PA, which has a larger range of severity,[18] and the introduction of 5 levels for role functioning, also contribute to the increased scale.

One concern with using DCE_{TTO} without the immediate death option to value health states is that it does not confront the respondent directly with whether any given state is better or worse than dead, but imputes this from their responses. To test this, we included DCE with duration choice sets that also present an immediate death option. The results suggest that including the choice sets incorporating immediate death reduces the overall utility range and frequency of states worse than dead. Other studies have collected data to value PBMs using 1 of the 2 tasks. Bansback et al[29] used the pair structure to value EQ-5D-5L in the United Kingdom, whereas Norman et al[31] valued EQ-5D-5L in Australia using the triplet structure. Comparisons of the value sets are difficult, given differences in study design and populations, and this is the first study to compare both types of task to some extent. However, the addition of Design 2 was not part of the efficient design process, and, therefore, we recommend a model based on Design 1.

In the models reported, pain has the largest overall decrement, followed by mental health and physical functioning, with social functioning, vitality, and role functioning being smaller. This is the same pattern as was observed for SF-6Dv1, indicating that overall preferences for the dimensions are similar. However, the magnitude of the decrements in utilities compared with the baseline differs markedly. SF-6Dv1 includes an interaction that is included if PF is reported at 1 of the 3 most serious levels or the other dimensions are reported at the 2 most serious levels. We have included an extra coefficient term (WORST1) that is added if any dimension is at the worst level. This has the effect of decreasing the overall range of utility values in comparison with the model without interactions.

The valuation protocol developed for this study can be used internationally to develop country-specific value sets. The development and use of DCE_{TTO} to generate country-specific values has a range of benefits in comparison with other iterative valuation methods as the studies can be carried out relatively cheaply and quickly using online panels. In some developing countries, online use is not as widespread. If this is the case, then recruitment and data collection could mix methods to achieve sufficient coverage. Although we have developed a study design that can be applied internationally, we recognize that the modeling approach used should be adapted to fit country-specific data.

Further comparisons of the SF-6Dv2 with SF-6Dv1 in existing data to understand the change in utilities produced are required. It is also important to compare the values to other PBMs to assess the impact on the QALY values estimated. The EQ-5D-5L now has a number of international value sets, including in England,[3,41] and comparing the new descriptive systems and value sets of the most widely used generic measures internationally will be informative.

This study has a number of limitations and areas for further work. We did not fully measure the level of respondent engagement in the task. We do set a minimum completion time for inclusion in the survey, and the models are relatively stable for subgroups of completers on the basis of time taken (Supplementary Appendix 6, Supplemental Digital Content 1, http://links.lww.com/MLR/B997). There may also be certain unobservable characteristics of respondents who opt into online panels. Recent studies have found differing levels of test-retest reliability using DCE, and testing how reliable DCE methods are for eliciting stable preferences is another area for investigation.

Due to the design structure in DCE, many assumptions are made that will result in models with reasonable face validity. The aim of efficient designs process is to generate a design that allows for comparisons of all severity levels within and across dimensions. However, the addition of duration as a continuous attribute complicates the design process, and potentially the efficiency. The addition of duration does allow for a comparison of value set characteristics. However, there is no gold standard, or revealed preferences against which to compare.

The latent class model shows that there are groups of respondents with different responses. The link between demographic group and class characteristics is also informative, and the finding that older individuals and those with children prefer a longer duration supports qualitative work testing valuation methods. Given variance in the estimates of the classes within the overall model, further investigation of the demographic heterogeneity, and how responses could be combined into a single value set for use in decision-making, is important. This could establish whether coefficients can be weighted on the basis of the proportion of the sample in each class, and how variance could be taken into account.

In conclusion, we have used a DCE protocol to value the SF-6Dv2. The results provide a provisional value set for calculating QALYs and the protocol can be applied internationally to develop country-specific SF-6Dv2 value sets.

ACKNOWLEDGMENTS

The authors acknowledge the input of the SF-6Dv2 international project team: Jordi Alonso, Beate Bestmann, Jakob Bjorner, Luciane Cruz, Rajabali Daroudi, Lara Ferreira, Pedro Ferreira, Shunichi Fukuhura, Barb Gandek, Lewis Kazis, Thomas Kohlmann, Maria Knoph Kvamme, Cindy Lam, Clara Mukuria, Brendan Mulhern, Jan Abel Olsen, Julie Ratcliffe, Antonio Rosello, Donna Rowen, Akbari Sari, Rick Sawatsky, Elly Stolk, Dong Suh, Gemma Vilagut, John Ware, David Whitehurst, Carlos Wong, Jing Wu, and Yosuke Yamamoto. They also thank the respondents for taking part in the study, Epigenesys for designing and managing the survey, and Survey Sampling International for providing the respondents.

REFERENCES

1. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. *J Health Econ*. 2002;21:271–292.
2. Brazier JE, Roberts J. Estimating a preference-based index from the SF-12. *Med Care*. 2004;42:851–859.
3. Devlin N, Feng Y, Shah K, et al. Valuing health-related quality of life: an EQ-5D-5L value set for England. *Health Econ*. 2018;27:7–22.
4. Dolan P. Modeling valuations for EuroQol health states. *Med Care*. 1997;35:1095–1108.
5. King MT, Viney R, Pickard AS, et al. Australian utility weights for the EORTC QLU-C10D, a multi-attribute utility instrument derived from the

Cancer-Specific Quality of Life Questionnaire, EORTC QLQ-C30. *Pharmacoeconomics*. 2018;36:225–238.
6. Ware JE, Sherbourne CD. The MOS 36-Item Short-Form Health Survey (SF-36): I. Conceptual framework and item selection. *Med Care*. 1992;30: 473–483.
7. Canadian Agency for Drugs and Technologies in Health. *Guidelines for the Economic Evaluation of Health Technologies: Canada*, 3rd ed. Ottawa, ON, Canada: Canadian Agency for Drugs and Technologies in Health; 2006.
8. College voor zorgverzekeringen. *Guidance for Outcomes Research for the Assessment of the Cost Effectiveness of In-Patient Medicines*. Diemen, The Netherlands: College voor zorgverzekeringen; 2008.
9. Health Information and Quality Authority (HIQA). Guidelines for the economic evaluation of health technologies in Ireland; 2010.
10. Norwegian Medicines Agency. *The National System for the Introduction of New Health Technologies Within the Specialist Health Service*. Oslo, Norway: Norwegian Medicines Agency; 2014.
11. Pharmaceutical Benefits Advisory Committee. *Guidelines for Preparing Submissions to the Pharmaceutical Benefits Advisory Committee (PBAC)*. Canberra, Australia: Commonwealth of Australia; 2015.
12. SMC Scottish Medicines Consortium. What we do (remit); 2014. Available at: www.scottishmedicines.org.uk/About_SMC/What_we_do/Remit. Accessed May 3, 2018.
13. Kharroubi SA, Brazier JE, Roberts J, et al. Modelling SF-6D health state preference data using a nonparametric Bayesian method. *J Health Econ*. 2007;26:597–612.
14. Brazier J, Roberts J, Tsuchiya A, et al. A comparison of the EQ-5D and SF-6D across seven patient groups. *Health Econ*. 2004;13:873–884.
15. Ferreira L, Ferreira P, Pereira L, et al. An application of the SF-6D to create heath values in Portuguese working age adults. *J Med Econ*. 2008;11:215–233.
16. Longworth L, Bryan S. An empirical comparison of EQ-5D and SF-6D in liver transplant patients. *Health Econ*. 2003;12:1061–1067.
17. Brazier J, Ratcliffe J, Salomon J, et al. *Measuring and Valuing Health Benefits for Economic Evaluation*. Oxford, UK: Oxford University Press; 2017.
18. Brazier J, Mulhern B, Bjorner J, et al. Developing a new version of the SF-6D health state classification system: SF-6Dv2. *Med Care*. 2020. [Epub ahead of print]..
19. Pinto-Prades JL, Attema A, Sánchez-Martínez FI. Measuring health utility in economics. Oxford Research Encyclopedias; 2019.
20. Torrance GW. Measurement of health state utilities for economic appraisal. *J Health Econ*. 1986;12:39–53.
21. Oppe M, Rand-Hendriksen K, Shah K, et al. EuroQol protocols for time trade-off valuation of health outcomes. *Pharmacoeconomics*. 2016;34:993–1004.
22. Thurstone LL. A law of comparative judgment. *Psychol Rev*. 1927;34: 273–286.
23. McFadden D. Conditional logit analysis of qualitative choice behavior. In: Zarembka P, ed. *Frontiers in Econometrics*. New York, NY: Academic Press; 1974:105–143.
24. Stolk E, Oppe M, Scalone L, et al. Discrete choice modelling for the quantification of health states: the case of the EQ-5D. *Value Health*. 2010;13:1005–1013.
25. Krabbe P, Devlin N, Stolk E, et al. Multinational evidence of the applicability and robustness of discrete choice modelling for deriving EQ-5D-5L health state values. *Med Care*. 2014;52:935–943.
26. Craig BM, Rand K. Choice defines QALYs: a US valuation of the EQ-5D-5L. *Med Care*. 2018;56:529–536.
27. Bansback N, Brazier J, Tsuchiya A, et al. Using a discrete choice experiment to estimate societal health state utility values. *J Health Econ*. 2012;31:306–318.
28. Viney R, Norman R, Brazier J, et al. An Australian discrete choice experiment to value EQ-5D health states. *Health Econ*. 2013;23: 729–742.
29. Bansback N, Hole AR, Mulhern B, et al. 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.
30. Mulhern B, Bansback N, Brazier J, et al. Preparatory study for the re-valuation of the EQ-5D tariff: methodology report. *Health Technol Assess*. 2014;18:12.
31. Norman R, Cronin P, Viney R. A pilot discrete choice experiment to explore preferences for EQ-5D-5L health states. *Appl Health Econ Health Policy*. 2013;11:287–298.
32. Norman R, Viney R, Brazier J, et al. Valuing SF-6D health states using a discrete choice experiment. *Med Decis Making*. 2014;34:773–786.
33. Norman R, Mulhern B, Viney R. The impact of different DCE-based approaches when anchoring utility scores. *Pharmacoeconomics*. 2016;34: 805–814.
34. Brazier J, Rowen D, Mavranezouli I, et al. Developing and testing methods for deriving preference-based measures of health from condition-specific measures (and other patient-based measures of outcome). *Health Technol Assess (Rockv)*. 2012;16:1–114.
35. Mulhern B, Bansback N, Hole AR, et al. Using discrete choice experiment with duration to model EQ-5D-5L health state preferences: testing experimental design strategies. *Med Decis Making*. 2017;37:285–297.
36. Choice Metrics. Ngene [software for experimental design]; 2012.
37. Herdman M, Gudex C, Lloyd A, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res*. 2011;20:1727–1736.
38. Greene WH, Hensher DA. A latent class model for discrete choice analysis: contrasts with mixed Logit. Sydney, NSW, Australia: Institute for Transport Studies, University of Sydney: Working Paper ITS-WP-02-08; 2002.
39. StataCorp. *Stata Statistical Software: Release 15*. College Station, TX: StataCorp LLC; 2017.
40. Office of National Statistics. *UK census 2011*. London, UK: Office of National Statistics; 2011.
41. Feng Y, Devlin N, Shah K, et al. New methods for modelling EQ-5D-5L value sets: an application to English data. *Health Econ*. 2018;27:23–38.