CrossMark

ORIGINAL RESEARCH ARTICLE

Head-to-Head Comparison of EQ-5D-3L and EQ-5D-5L Health Values

Anna Selivanova[1] · Erik Buskens[1] · Paul F. M. Krabbe[1]

Abstract

Background The EQ-5D is a widely used preference-based instrument to measure health-related quality of life. Some methodological drawbacks of its three-level version (EQ-5D-3L) prompted development of a new format (EQ-5D-5L). There is no clear evidence that the new format outperforms the standard version.

Objective The objective of this study was to make a head-to-head comparison of the EQ-5D-3L and EQ-5D-5L in a discrete choice model setting giving special attention to the consistency and logical ordering of coefficients for the attribute levels and to the differences in health-state values.

Methods Using efficient designs, 240 pairs of EQ-5D-3L health states and 240 pairs of EQ-5D-5L health states were generated in a pairwise choice format. The study included 3698 Dutch general population respondents, analyzed their responses using a conditional logit model, and compared the values elicited by EQ-5D-3L and EQ-5D-5L for different health states.

Results No inconsistencies or illogical ordering of level coefficients were observed in either version. The proportion of severe health states with low values was higher in the EQ-5D-5L than in the EQ-5D-3L, and the proportion of mild/moderate states was lower in the EQ-5D-5L than in the EQ-5D-3L. Moreover, differences were observed in the relative weights of the attributes.

Conclusion Overall distribution of health-state values derived from a large representative sample using the same

measurement framework for both versions showed differences between the EQ-5D-3L and EQ-5D-5L. However, even small differences in the phrasing (language) of the descriptive system or in the valuation protocol can produce differences in values between these two versions.

Key Points for Decision Makers

Changes in phrasing and the use of a different valuation protocol, in combination with particular statistical models used to estimate the weights for the EQ-5D attributes, may explain the discrepancies between the 3L and 5L observed in earlier studies.

No inconsistencies or illogical ordering of level coefficients were observed for the EQ-5D-3L or for the EQ-5D-5L.

Differences in weights for the five EQ-5D health attributes have been observed: the highest weight for the EQ-5D-5L was for anxiety/depression; for the EQ-5D-3L, it was mobility.

✉ Anna Selivanova
  a.selivanova@umcg.nl

[1] Department of Epidemiology (FA40), University of Groningen, University Medical Center Groningen, PO Box 30.001, 9700 RB Groningen, The Netherlands

1 Introduction

Generic preference-based measures of health-related quality of life are frequently used to assess the impact of treatment or clinical pathways and to monitor population health [1–3]. Typically, preference-based measurement frameworks incorporate various independent attributes (notated for domains/dimensions) that jointly represent the

notion of health-related quality of life. The levels of these attributes are weighted to indicate the relative importance attributed to them by the respondents (expressed preferences). Weighted attribute levels are subsequently aggregated into a single number reflecting the quality or value of a health state [4]. To obtain such values, several instruments (e.g., EQ-5D, HUI-3, SF-6D, AQol) have been developed within a preference-based measurement framework.

The EuroQol Group (http://www.euroqol.org) developed the EQ-5D, a relatively simple, widely used instrument [5–9]. It comprises five health attributes in the descriptive system (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression) and a 20-cm visual analog scale. In the standard version (EQ-5D-3L), each of the attributes can take on three levels [10]. A considerable body of literature corroborates the sustainability of the instrument [11–15]. However, attention has been drawn to its limited sensitivity regarding small or moderate changes in patients' health states [16–19] and its considerable ceiling effects (i.e., almost no differentiation between mild health states), prompting an update of the instrument [20–23]. In the new version, the EQ-5D-5L, the number of levels used to classify health states increased from three to five. Testing its descriptive system performance in terms of its discriminatory power and sensitivity revealed a lower ceiling effect and a higher sensitivity [13, 19, 23–25]. Additionally, several studies noted that subtle differences in the phrasing of levels 4 (severe problems) and 5 (extreme problems) caused inconsistencies in elicited health-state values [26, 27].

Besides increasing the number of levels from three to five, the protocol to derive valuations was changed. For the EQ-5D-3L valuation protocol, originally the time trade-off (TTO) was chosen from among all possible health valuation techniques (standard gamble, TTO, rating/visual analog scale, person trade-off, and magnitude estimation). However, various shortcomings of this technique were identified [28–31], which encouraged the EuroQol Group to experiment with other methods, such as choice-based modeling. Choice models are grounded in modern measurement theory and are consistent with the random utility model in economic theory [32]. The applicability of choice models for health-state evaluations has been proposed and tested elsewhere [4, 33–35].

The association between the descriptive systems for the three- and five-level versions of the EQ-5D has been investigated extensively. Far less is known about the distribution of the values and the underlying weights for the levels of the attributes for both EQ-5D versions, which motivated the present study. This article presents a discrete choice (DC) study and a head-to-head comparison of the EQ-5D-3L and EQ-5D-5L with an emphasis on the consistency and logical ordering of attribute levels and the distributions of the estimated values.

2 Methods

2.1 Sample

Overall, 4036 persons participated in a self-completed computer-based assessment by Survey Sampling International (SSI) [Rotterdam, the Netherlands]. The sample is representative of age and sex for the general Dutch population based on the SSI panel of a working age of 18–65 years and recruitment during September–October 2016. Clear instructions were given to all participants, and those who fully completed the survey received a small financial compensation from SSI. The rewards were defined by the company's (SSI) internal agreements individually with the groups of respondents. Each one was randomly assigned to one of the 30 blocks of the survey. No limits on time for completion were imposed.

2.2 Discrete Choice

Discrete choice modeling is a widely used technique to elicit personal and societal preferences in health-valuation studies [36]. The statistical literature classifies it within the modern framework of probabilistic DC models that are consistent with economic theory (i.e., the random utility model) [32, 37, 38]. All DC models establish the relative merit of one phenomenon based on its relative attractiveness. This technique requires participants to make choices among two or more presented scenarios (choice tasks) described by the means of specific attributes with certain levels.

2.3 Experimental Design and Selection of Health States

The EQ-5D-3L contains five attributes with three levels each, yielding $3^5 = 243$ possible health states. Health states were presented in pairs for comparison in the DC task. Thus, the number of potential pairs to be compared becomes 29,403. For EQ-5D-5L, the number of possible health states increases to $5^5 = 3125$, and the number of possible paired comparisons rises drastically to 4,881,250. Clearly, it is infeasible to present all possible pairs to the respondents, especially in the case of EQ-5D-5L. For both versions, therefore, health-state pairs had to be carefully selected to arrive at an informative set. Two important issues were taken into consideration in the selection: respondent fatigue and avoidance of dominance in the pairs.

The credibility of an individual's responses can be questionable when a person gets bored or fatigued, which could happen if the tasks are complex or numerous. Earlier studies suggested that up to 16 choice tasks are acceptable and do not affect the responses [31, 39, 40]. We offered each respondent a set of 16 choice tasks and reduced their complexity through a two-level overlap in the health-state descriptions for both versions of the EQ-5D. A two-level overlap implies fixing two of the five attributes at the same level and varying the other three.

Dominance is a common difficulty in health-state valuation exercises because all attributes are ordered, and people always prefer fewer health problems to more. Dominant pairs do not offer additional information, yet they reduce design efficiency. Therefore, it was decided to remove all combinations where every attribute of one health state in a pair was worse or the same (or better or the same) than every attribute of the other health state.

In view of the above solutions for the issues of fatigue and dominance, an approach to health-state selection was developed along similar lines, as set forth below for the EQ-5D-3L and the EQ-5D-5L. The set of non-dominant pairs for EQ-5D-3L was selected out of all possible 29,403 pairs, arriving at 14,580 pairs. Likewise, in EQ-5D-5L, the number of non-dominant pairs was reduced from 4,881,250 to 1,430,000 (Stata 14.0; StataCorp LLC, College Station, Texas, USA). Out of all non-dominant health-state pairs with a two-level overlap, we decided to select 240 pairs, which is considered sufficient to estimate regression coefficients for EQ-5D-5L attribute levels. It was decided to select the same number of pairs for the EQ-5D-3L. Therefore, 240 pairs in EQ-5D-5L and 240 pairs in EQ-5D-3L format were selected, using an efficient design routine programmed in Ngene software (ChoiceMetrics, Sydney, Australia) (the mnl model, taking 500 Bayesian draws, Halton sequence, modified Fedorov algorithm). All selected pairs were divided into 30 blocks with 16 choice tasks each, whereby 15 blocks contained all 16 tasks in EQ-5D-3L, and 15 blocks contained 16 tasks in EQ-5D-5L. The design was based on an iterative procedure, where designs are compared by their D-error (measure of statistical efficiency). After numerous iterations, the designs were checked for their D-errors and for the level balance. Level balance ensures all levels of all attributes appear with even frequency in the design. A perfectly even frequency of level balance can rarely be achieved; therefore, the fairly even distribution of levels was accepted. Finally, the design with the lowest D-error and better indicator of level balance was chosen. Efficient design in Ngene requires priors (approximations of the parameters), which were derived from an earlier EQ-5D-3L study [36] and from a multinational study of the EQ-5D-5L [4].

2.4 Response Tasks

The response task included two health-state descriptions comprising five attributes of the EQ-5D. The respondents had to decide which of the two health-state descriptions they preferred. Half of the blocks contained health-state descriptions defined by three levels of EQ-5D-3L (no problems, some problems, extreme problems), and half of the blocks contained health-state descriptions defined by five levels of EQ-5D-5L (no problems, slight problems, moderate problems, severe problems, extreme problems). The respondents were randomly assigned to one of the blocks, meaning that each person completed 16 response tasks only in the EQ-5D-3L format or (in the other block) only in the EQ-5D-5L format.

2.5 Analysis

2.5.1 EQ-5D-3L and EQ-5D-5L Values and Value' Distributions

The analysis of the data was performed using a DC conditional logit model (asclogit, Stata 15.0; StataCorp LLC, College Station, Texas, USA), which yields parameter estimates presented as regression coefficients. The main-effects value function included ten dummy variables for the EQ-5D-3L representing levels 2 and 3 for each of the five attributes: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. The main-effects model for the EQ-5D-5L included 20 dummy variables representing levels 2, 3, 4, and 5. The regression coefficients were checked for logical ordering and significance. In addition, we tested for the increments from one level to any other consecutive levels (post-hoc estimation, contrast, Stata SE 15.0; StataCorp LLC, College Station, Texas, USA) [41, 42].

Additionally, the values of all health states possible in EQ-5D-3L and EQ-5D-5L were calculated based on estimated coefficients. We used the original values derived with the choice model and rescaled them to the published results of the Dutch valuation studies for the 3L version and 5L version, respectively [43, 44]. For the EQ-5D-3L, the value range from the valuation studies was − 0.33 to 1.0, while for the EQ-5D-5L the value range was − 0.45 to 1.0. Finally, for both versions, the distributions of estimated values were compared. Kernel density graph and graphs of frequency distributions were produced for the EQ-5D-3L and the EQ-5D-5L (Stata SE 15.0). For comparison of value ranges in both graphs for EQ-5D-3L and the EQ-5D-5L, we provided distributions displaying the unscaled values and the values scaled to the Dutch tariffs.

2.5.2 Comparison of Differences in Weights for Health-State Attributes

The overall weights of each of the five EQ-5D attributes were calculated using the coefficient range method: the range between the coefficients of the individual levels was calculated and then converted to a proportion.

$$W_{\text{attribute}(i)} = \frac{\max C_i - \min C_i}{\sum_j (\max C_j - \min C_j)}, \quad (1)$$

where C_i represents the coefficients of the individual levels of attribute i and j represents the number of attributes.

3 Results

3.1 Sample

In total, 4036 respondents completed the survey. Out of this sample, 288 completed 16 choice tasks in less than 2 min, which was considered unrealistic and insufficient. In addition, responses of 50 individuals were deemed unreliable, given their pattern of choosing only the left (A) or only the right (B) alternative throughout the survey. Therefore, the forms of 338 respondents were disregarded. Finally, the analysis included 1824 respondents for the EQ-5D-3L and 1874 for the EQ-5D-5L (Table 1). An overall Chi-square test revealed significant differences between the samples completing EQ-5D-3L and EQ-5D-5L in terms of age groups: $p = 0.000$.

3.2 Comparison of EQ-5D-3L and EQ-5D-5L Coefficients and Overall Attribute Weights

No inconsistencies or illogical ordering of level coefficients were observed for the 3L and 5L versions. The spread of regression coefficients within each attribute consistently followed the same patterns across attributes: levels 2 and 3 lowered the values slightly, levels 4 and 5 even more so in the EQ-5D-5L. Moreover, the incremental differences between consecutive levels of each dimension were checked for significance, whereby it was observed that the move from level 5 to level 4 of severity had a smaller effect than the move from level 4 to level 3. All parameters in both models were statistically significant (Tables 2, 3).

Self-care was generally assigned less weight than the other four attributes in the EQ-5D-3L and in EQ-5D-5L (Table 2). Moreover, level 3 problems with mobility (confined to bed) appeared to have the largest effect on the values in the EQ-5D-3L format. Overall, the attribute mobility in the EQ-5D-3L version was assigned the highest relative weight. Regarding the EQ-5D-5L version, the respondents were more concerned about anxiety/depression and pain/discomfort than about problems with other attributes. Regarding the EQ-5D-3L version, we noted that pain/discomfort had more relative weight than anxiety/depression, while the opposite was noted for EQ-5D-5L.

3.3 Comparison of EQ-5D-3L and EQ-5D-5L Value Distributions

The original unscaled values of both EQ-5D versions were anchored to the values of the best and worst health states derived from the Dutch valuation studies [43, 44], and plotted as the frequency distribution of estimated values for 243 health states in the EQ-5D-3L and 3125 health states in the EQ-5D-5L (Fig. 1). The graph demonstrates that the distributions of values elicited with unscaled coefficients are similar to the distribution of the rescaled values because only the scale is changed, not the distribution of the values. These graphs and the kernel density graph (Fig. 2) demonstrate that EQ-5D-5L has more health states than EQ-5D-3L on the region with severe health states and fewer states on the region with milder states.

4 Discussion

4.1 Overall Discussion and Literature Review

This study contributes to the body of literature comparing the standard EQ-5D-3L and the new EQ-5D-5L. Here, the focus is on the logical ordering and differences in distributions of values for health states in these two versions. The health-state values were elicited from a sample of the general population applying a conventional DC approach. According to several earlier studies, the differences in EQ-5D-5L levels are subtle and may be hard to distinguish, which might have caused inconsistencies for some language versions (English) in the upper or lower levels of health attributes [26, 27, 44]. Eventually, such inconsistency would affect the validity of the estimated values. In the current Dutch study, we found that all coefficients for both versions of EQ-5D were logically ordered.

However, the results demonstrated that the overall weights for the attributes are different in the two EQ-5D versions. In the EQ-5D-5L, the highest weight was attributed to anxiety/depression followed by pain/discomfort; in the EQ-5D-3L, the highest weight was attributed to mobility. Larger weights of an attribute have larger effects on a health-state value: the negative changes in the levels of the most important attributes could overweigh the positive changes in the levels of the less important attributes, resulting in lower values.

Table 1 Respondents' characteristics

Characteristics	EQ-5D-3L (N = 1824)	EQ-5D-5L (N = 1874)
Male, n (%)	797 (44)	876 (47)
Age, years, mean (SD)	45.5 (14.3)	51.2 (13.4)
Age group, years, n (%)		
18–24	101 (13)	74 (8)
25–34	90 (11)	47 (5)
35–44	134 (17)	95 (11)
45–54	214 (27)	172 (20)
Over 55	258 (32)	488 (56)
Female, n (%)	1,027 (56)	998 (53)
Age, years, mean (SD)	42.8 (13.8)	44.9 (15.1)
Age group, years, n (%)		
18–24	145 (14)	179 (18)
25–34	175 (17)	108 (11)
35–44	192 (19)	121 (12)
45–54	276 (27)	224 (22)
Over 55	239 (23)	366 (37)
Diseases, n (%)		
No diseases	701 (38)	705 (33)
Neck and back pain	440 (24)	459 (25)
Pain (abdomen, migraine, chronic)	231 (13)	208 (11)
Sleep problems	258 (141)	281 (15)
Fatigue	337 (19)	360 (19)
Diabetes mellitus	132 (7)	163 (9)
Heart disease	94 (5)	140 (7)
Hearing or vision loss	149 (8)	182 (10)
Asthma/COPD	177 (10)	163 (9)
Eczema	126 (7)	145 (8)
Mental health problems	171 (9)	179 (10)
Stroke	16 (1)	37 (2)
Rheumatism (osteoarthritis, arthritis)	186 (10)	195 (10)
Cancer	27 (2)	46 (2)
Epilepsy	20 (1)	14 (0.5)
Lung disease	38 (2)	37 (2)
Gastrointestinal disease	63 (4)	64 (3)

COPD chronic obstructive pulmonary disorder, *SD* standard deviation

Mobility, especially level 3 (confined to bed), had the most significant impact in the EQ-5D-3L. It is clear that 'confined to bed' has a different phrasing format for level 3 than it has in the other attributes. In later versions of the EQ-5D, namely the version for youth (EQ-5D-Y) and the EQ-5D-5L, the formulation of the worst levels was changed into 'unable to walk' [19, 44–47]. In the EQ-5D-5L version, with the most severe level formulated as 'unable to', the effect of mobility on the health-state values declined. Changing the phrasing from 'confined to bed' to 'unable to walk' is likely responsible for the shift in the level of importance. 'Confined to bed' seems to imply isolation and dependence, while 'unable to walk' may be interpreted as a less serious limitation.

A large multinational study based on DC modeling for the EQ-5D-5L [4] showed greater importance assigned to pain/discomfort and anxiety/depression attributes for the Dutch population, while for the US population the attribute mobility had the greater importance. The Dutch valuation study for the EQ-5D-5L confirmed that the greatest importance was assigned to pain/discomfort and anxiety/depression [44]. Mulhern et al. [48], in their study using DC modeling, observed that the attribute pain/discomfort also showed the largest effect.

Table 2 Regression coefficients for the EQ-5D-3L and EQ-5D-5L based on a discrete choice model

EQ-5D-3L (the five attributes with their overall weights)	β (SE)	EQ-5D-5L (the five attributes with their overall weights)	β (SE)
Mobility (0.248)		Mobility (0.172)	
No problems (level 1)	Reference	No problems (level 1)	Reference
Some problems (level 2)	− 0.323 (0.02)	Slight problems (level 2)	− 0.138 (0.04)
Confined to bed (level 3)	− 1.550 (0.03)	Moderate problems (level 3)	− 0.290 (0.03)
		Severe problems (level 4)	− 0.968 (0.04)
		Unable to (level 5)	− 1.267 (0.04)
Self-care (0.146)		Self-care (0.156)	
No problems (level1)	Reference	No problems (level 1)	Reference
Some problems (level 2)	− 0.318 (0.02)	Slight problems (level 2)	− 0.098 (0.04)
Unable to (level 3)	− 1.044 (0.03)	Moderate problems (level 3)	− 0.297 (0.03)
		Severe problems (level 4)	− 0.938 (0.04)
		Unable to (level 5)	− 1.123 (0.04)
Usual activities (0.178)		Usual activities (0.175)	
No problems (level 1)	Reference	No problems (level 1)	Reference
Some problems (level 2)	− 0.172 (0.02)	Slight problems (level 2)	− 0.150 (0.04)
Unable to (level 3)	− 1.055 (0.03)	Moderate problems (level 3)	− 0.228 (0.03)
		Severe problems (level 4)	− 0.969 (0.03)
		Unable to (level 5)	− 1.302 (0.04)
Pain/discomfort (0.237)		Pain/discomfort (0.237)	
None (level 1)	Reference	None (level 1)	Reference
Moderate (level 2)	− 0.247 (0.02)	Slight* (level 2)	− 0.076 (0.04)
Extreme (level 3)	− 1.423 (0.03)	Moderate (level 3)	− 0.262 (0.04)
		Severe (level 4)	− 1.150 (0.04)
		Extreme (level 5)	− 1.636 (0.04)
Anxiety/depression (0.191)		Anxiety/depression (0.259)	
None (level 1)	Reference	None (level 1)	Reference
Moderate (level 2)	− 0.379 (0.03)	Slight (level 2)	− 0.253 (0.04)
Extreme (level 3)	− 1.324 (0.03)	Moderate (level 3)	− 0.543 (0.04)
		Severe (level 4)	− 1.347 (0.04)
		Extreme (level 5)	− 1.957 (0.04)
Log likelihood	− 16979.542	Log likelihood	− 16477.634

Table 2 continued

EQ-5D-3L (the five attributes with their overall weights)	β (SE)	EQ-5D-5L (the five attributes with their overall weights)	β (SE)
Wald Chi-square	4874.59	Wald Chi-square	5988.72

All variables were statistically significant at the 99% confidence level, $p < 0.01$, except $*p = 0.037$

SE standard error

Table 3 Estimations for the EQ-5D-3L and EQ-5D-5L increments for consecutive levels

| EQ-5D-3L	β (SE)	EQ-5D-5L	β (SE)
Mobility		Mobility	
Some → problems (level 2 → level 1)	0.323 (0.02)	Slight → no problems (level 2 → level 1)	0.138 (0.03)
Confined to bed → some problems (level 3 → level 2)	1.227 (0.03)	Moderate → slight problems (level 3 → level 2)	0.152 (0.03)
		Severe → moderate problems (level 4 → level 3)	0.678 (0.03)
		Unable → severe problems (level 5 → level 4)	0.298 (0.03)
Self-care		Self-care	
Some → no problems (level 2 → level 1)	0.318 (0.02)	Slight → no problems (level 2 → level 1)	0.098 (0.04)
Unable → some problems (level 3 → level 2)	0.726 (0.02)	Moderate → slight problems (level 3 → level 2)	0.199 (0.04)
		Severe → moderate problems (level 4 → level 3)	0.641 (0.04)
		Unable → severe problems (level 5 → level 4)	0.185 (0.04)
Usual activities		Usual activities	
Some → no problems (level 2 → level 1)	0.172 (0.02)	Slight → no problems (level 2 → level 1)	0.150 (0.04)
Unable → some problems (level 3 → level 2)	0.884 (0.03)	Moderate → slight problems* (level 3 → level 2)	0.079 (0.04
		Severe → moderate problems (level 4 → level 3)	0.741 (0.04)
		Unable → severe problems (level 5 → level 4)	0.333 (0.04)
Pain/discomfort		Pain/discomfort	
Moderate → none (level 2 → level 1)	0.247 (0.02)	Slight → none* (level 2 → level 1)	0.076 (0.04)
Extreme → moderate (level 3 → level 2)	1.176 (0.03)	Moderate → slight (level 3 → level 2)	0.186 (0.04)
		Severe → moderate (level 4 → level 3)	0.888 (0.04)
		Extreme → severe (level 5 → level 4)	0.486 (0.03)
Anxiety/depression		Anxiety/depression	
Moderate → none (level 2 → level 1)	0.379 (0.03)	Slight → none (level 2 → level 1)	0.253 (0.04)
Extreme → moderate (level 3 → level 2)	0.945 (0.02)	Moderate → slight (level 3 → level 2)	0.289 (0.03)
		Severe → moderate (level 4 → level 3)	0.804 (0.03)
		Extreme → severe (level 5 → level 4)	0.610 (0.04)

All variables were statistically significant at the 99% confidence level, $p < 0.01$, except $*p < 0.05$

SE standard error

Overall, we observed differences in the health-state distributions for severe and mild/moderate states derived from the EQ-5D-3L and EQ-5D-5L. Our findings are not in line with those of Mulhern et al. [49], who observed the opposite. However, it may be attributed to the fact that the EQ-5D-3L UK value set has a larger range of values than the EQ-5D-5L UK value set. In addition, the samples analyzed in that study were recruited differently (England,

UK) and different valuation methods were used (TTO, visual analog scale). Overall, the distributions of health states in the current study showed a somewhat lower proportion of severe states in the EQ-5D-5L than in the EQ-5D-3L. These findings are not in line with the findings published in the Dutch tariff [44], demonstrating the values for all attainable health states to be higher in the 3L version for the severe health states and higher for the 5L version for
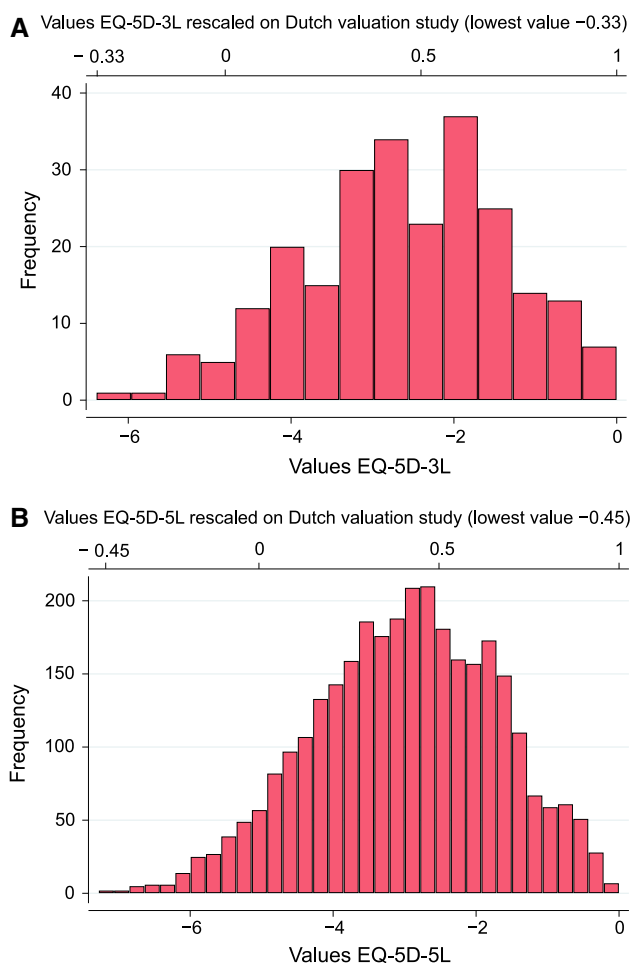
Fig. 1 Frequency distribution of **a** all 243 EQ-5D-3L health-state values and rescaled values; **b** all 3125 EQ-5D-5L health-state values and rescaled values
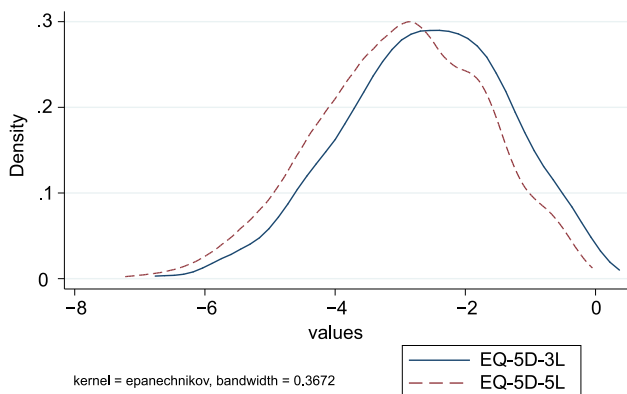
Fig. 2 Kernel density plot for EQ-5D-3L and EQ-5D-5Lvalues

moderate and mild health states (on the value range 0.35–0.75). Again, such a discrepancy may be caused by differences in the conceptual and valuation approaches used. The current study is based only on DC estimations, while the Dutch tariff is based on the composite TTO and

tasks for valuing worse-than-death states were included. In the Dutch tariff study, DC results were used to identify the appropriate TTO modeling techniques, but not to estimate health-state values.

The report by the National Institute for Health and Care Excellence [50] suggested that the 5L instrument showed higher mean utility scores than the 3L, meaning that the improvements in health are slightly less in the 5L than in the 3L, which results in interventions being considered as less cost effective if based on the 5L. This may lead policy makers to give due consideration to the choice of a version: EQ-5D-5L may produce smaller benefits of innovations for severe health states, according to our study, which may discourage end users from using this version. These findings raise challenges about the choice of the EQ-5D version to be used: for particular interventions, end users are likely to prefer the EQ-5D-3L, indicating higher benefits of interventions. However, the studies included in the National Institute for Health and Care Excellence document are not based on valuations. In fact, the analysis underlying that document used self-reported health assessments scored according to the EQ-5D descriptive system. Therefore, the comparison between the current study and the study of the National Institute for Health and Care Excellence should be taken with caution.

4.2 Limitations

It is worth mentioning the following limitation of our study: there is a difference in the age group proportions of the two samples. We tried to reach the comparability of the representativeness and sample sizes for 3L and 5L versions; however, significant age differences were observed according to the Chi-square test. One might argue that such differences would bias the estimated results. However, an additional analysis with inclusion of age groups as a separate predictor into the choice model did not reveal any statistically significant effect of age on the estimated coefficients.

By their nature, health-state values derived with choice models cannot be interpreted as absolute (cardinal) numbers because of two reasons. First, the best health state (full health) is dominant and cannot be used in the choice model as an anchor. Second, the location of death is unknown because a 'death' option was not included. Consequently, DC models position health states on a scale between the best and the worst health states. Therefore, one of the main problems with choice models is normalizing its scale to a death-full health (0.0–1.0) scale. To solve this problem, a task extension or additional tasks should be included in the design, such as death questions, duration on the health states, or an accompanied TTO task. We did not use either of these techniques. Instead, we used the published Dutch

valuation studies [43, 44] as an anchor for the values elicited with the DC model. By doing so, the rescaling limitation remains but anchor points are based in current evidence.

Recent studies using different valuation frameworks for quality-adjusted life-year calculations showed smaller differences between the same health states in the EQ-5D-5L version in comparison with the original EQ-5D-3L, which raised concerns among end users (e.g., pharmaceutical companies) [44, 49, 50]. In a recent UK study estimating a value function for the EQ-5D-5L, the composite TTO was introduced as a new valuation technique. That innovation is a derivate of the conventional TTO based on a combination of lead-time TTO [51] and standard TTO as used in the 3L. This UK study applied a rescaling for the states 'worse than death' (negative utilities) that differs from the rescaling used in the original EQ-5D-3L [1]. In addition, the UK study [52] analyzed TTO responses and DC responses together in a hybrid model incorporating several other analytical procedures (e.g., censoring, additional parameter for heterogeneity of respondents, forcing consistency in levels of attributes) [53]. Moreover, the authors of the Dutch tariff [44] admitted that the similarities between the EQ-5D-3L and EQ-5D-5L are not necessarily expected owing to differences in the phrasing and valuation methods used. Therefore, the divergence between the 3L and 5L version, if based on the official EuroQol protocol, is likely to be a combined effect of the differences in the way individuals respond to the changed descriptive system and because a totally new and different valuation framework has been introduced [54]. The present study did not use a TTO technique. Instead, we used DC for both versions of EQ-5D, which resulted in certain differences in the weights and overall distributions of the EQ-5D-3L and the EQ-5D-5L health-state values.

Values derived with DC seem to be more robust and less effected by possible framing effects, as the judgmental DC task is more straightforward and simple than the TTO variants. However, it needs to be stated that the design strategy of selecting equal amount of DC pairs for both versions may have had an impact on the estimated values. Specifically, selecting 240 DC pairs for the EQ-5D-3L would enable broader coverage of the health states than selecting 240 pairs for the EQ-5D-5L because the EQ-5D-5L comprises more health states.

Consequently, such a design setting would result in more precise estimates for the EQ-5D-3L than for the EQ-5D-5L. However, based on earlier studies [4, 44, 48], having 240 pairs for the EQ-5D-5L is highly sufficient to get precise estimates. Moreover, the standard deviations of the coefficients, which reflect precision of an estimated coefficient, showed that the difference is minor (the maximum standard deviation in the model for EQ-5D-3L is 0.3,

while the maximum standard deviation in the model for EQ-5D-5L is 0.4).

4.3 Strengths

The present study has several strengths. First, a large representative sample of the Dutch general population has been achieved. Second, it used the same valuation method (DC) and the same statistical analysis for both EQ-5D versions. Third, an efficient design was applied to maximize the precision of estimated regression coefficients, while respondent fatigue was prevented by applying a two-level overlap. Overall, this is the first head-to-head DC study to compare health-state values derived from EQ-5D-3L and EQ-5D-5L using large samples.

5 Conclusion

The distributions of health states suggested that the proportion of severe health states with low values in the EQ-5D-5L was slightly higher than in the EQ-5D-3L, and the proportion of mild/moderate states was lower in the EQ-5D-5L than in the EQ-5D-3L. Additionally, the overall weights of the attributes in the EQ-5D-3L and the EQ-5D-5L are different. We suggest that even small differences in the phrasing of the descriptive system or in the valuation protocol may affect individual responses and thereby the elicited values. Finally, it needs to be emphasized that the applied valuation framework in combination with particular statistical models used to estimate the weights for the attributes and their levels may explain the substantial discrepancies between the 3L and 5L observed in earlier studies.

References

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

2. Feeny D, Furlong W, Torrance GW, Goldsmith CH, Zhu Z, DePauw S, Denton M, Boyle M. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. Med Care. 2002;40(2):113–28.

3. Hamming JF, De Vries J. Measuring quality of life. Br J Surg. 2007;94:923–4.

4. Krabbe PFM, Devlin NJ, Stolk EA, Shah KK, Oppe M, van Hout B, Quik EH, Pickard AS, Xie F. Multinational evidence of the applicability and robustness of discrete choice modeling for deriving EQ-5D-5L health-state values. Med Care. 2014;52(11):935–43.

5. Hurst N, Kind P, Ruta D, Hunter M, Stubbings A. Measuring health-related quality of life in people with rheumatoid arthritis: validity, responsiveness and reliability of the EuroQoL (EQ-5D). Br J Rheumatol. 1997;36:551–9.

6. Rabin R, de Charro F. EQ-5D: a measure of health status from the EuroQol Group. Ann Med. 2001;33:337–43.

7. Russell RT, Feurer ID, Wisawatapnimit P, Pinson CW. The validity of EQ-5D US preference weights in liver transplant candidates and recipients. Liver Transpl. 2009;15:88–95. https://doi.org/10.1002/lt.21648.

8. Xu R, Insinga RP, Golden W, Hu XH. EuroQol (EQ-5D) health utility scores for patients with migraine. Qual Life Res. 2011;20(4):601–8.

9. Devlin NJ, Brooks R. EQ-5D and the EuroQol Group: past, present and future. Appl Health Econ Health Policy. 2017;15:127. https://doi.org/10.1007/s40258-017-0310-5.

10. Brooks R. EuroQol: the current state of play. Health Policy. 1996;37(1):53–72.

11. Johnson JA, Coons SJ. Comparison of the EQ-5D and SF-12 in an adult US sample. Qual Life Res. 1998;7:155–66.

12. Johnson JA, Pickard AS. Comparison of the EQ-5D and SF-12 health surveys in a general population survey in Alberta, Canada. Med Care. 2000;38(1):115–21.

13. Pickard AS, De leon MC, Kohlmann T, Cella D, Rosenbloom S. Psychometric comparison of the standard EQ-5D to a 5 level version in cancer patients. Med Care. 2007;45:259–63.

14. Dyer MT, Goldsmith KA, Sharples LS, Buxton MJ. A review of health utilities using the EQ-5D in studies of cardiovascular disease. Health Qual Life Outcomes. 2010;8:1–13.

15. Janssen MF, Lubetkin EI, Sekhobo JP, Pickard AS. The use of the EQ-5D preference-based health status measure in adults with type 2 diabetes mellitus. Diabet Med. 2011;28:395–413.

16. Myers C, Wilks D. Comparison of Euroqol EQ-5D and SF-36 in patients with chronic fatigue syndrome. Qual Life Res. 1999;8:9. https://doi.org/10.1023/A:1026459027453.

17. Wu AW, Jacobson KL, Frick KD, Clark R, Revicki DA, Freedberg KA, Scott-Lennox J, Feinberg J. Validity and responsiveness of the EuroQol as a measure of health-related quality of life in people enrolled in an AIDS clinical trial. Qual Life Res. 2002;11:273–82.

18. Macran S, Weatherly H, Kind P. Measuring population health: a comparison of three generic health status measures. Med Care. 2013;41(2):218–31.

19. Janssen MF, Pickard AS, Golicki D, Gudex C, Niewada M, Scalone L, Swinburn P, Bussbach J. Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: a multi-country study. Qual Life Res. 2013;22:1717–27.

20. Badia X, Herdman M, Kind P. The influence of ill-health experience on the valuation of health. Pharmacoecononics. 1998;13:687–96.

21. Brazier J, Roberts J, Tsuchiya A, Busschbach J. A comparison of the EQ-5D-3L and SF-6D across seven patient groups. J Health Econ. 2004;13(9):873–84.

22. Sullivan PW, Lawrence WF Jr, Ghushchyan V. A national catalogue of preference-based scores for chronic conditions in the U.S. Med Care. 2005;43:736–49.

23. Scalone L, Ciampichini R, Fagiuoli S, Gardini I, Fusco F, Gaeta L, Del Prete A, Cesana G, Mantovani LG. Comparing the performance of the standard EQ-5D 3L with the new version EQ-5D 5L in patients with chronic hepatic disease. Qual Life Res. 2013;22:1707–16.

24. Janssen MF, Birnie E, Haagsma JA, Bonsel GJ. Comparing the standard EQ-5D three-level system with a five-level version. Value Health. 2008;11:275–84.

25. Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, Bonsel G, Badia X. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res. 2011;20(10):1727–36.

26. Xie F, Pullenayegum E, Gaebel K, Oppe M, Krabbe PFM. Eliciting preferences to the EQ-5D-5L health states: discrete choice experiment or multiprofile case of best-worst scaling? Eur J Health Econ. 2014;15(3):281–8.

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

28. van Osch SM, Wakker PP, van den Hout WB, Stiggelbout AM. Correcting bases in standard gamble and time tradeoff utilities. Med Decis Mak. 2004;24(5):511–7.

29. van der Pol M, Roux L. Time preference bias in time trade-off. Eur J Health Econ. 2005;6(2):107–11.

30. Doctor JN, Bleichrodt H, Lin JH. Health utility bias: a systematic review and meta-analytic evaluation. Med Decis Mak. 2010;30:58–67.

31. Viney R, Norman R, Brazier J, Cronin P, King MT, Ratcliffe J, Street D. An Australian choice experiment to value EQ-5D health states. J Health Econ. 2014;23:729–42.

32. Arons MMA, Krabbe PFM. Probabilistic choice models in health-state valuation research: background, theories, assumptions and applications. Expert Rev Pharmacoecon Outcomes Res. 2013;13(1):93–108.

33. McKenzie L, Cairns J, Osman L. Symptom-based outcome measures for asthma: the use of discrete choice methods to assess patient preferences. Health Policy. 2001;57:193–204.

34. Ratcliffe J, Brazier J, Tsuchiya A, Symonds T, Brown M. Using DCE and ranking data to estimate cardinal values for health states for deriving a preference-based single index from the sexual quality of life questionnaire. J Health Econ. 2009;18:1261–76.

35. Bansback N, Brazier J, Tsuchiya A, Anis A. Using a discrete choice experiment to estimate societal health state utility values. J Health Econ. 2012;31:306–18.

36. Stolk EA, Oppe M, Scalone L, Krabbe PFM. Discrete choice modeling for the quantification of health states: the case of the EQ-5D. Value Health. 2010;13:1005–13.

37. Krabbe PFM. Thurstone scaling as a measurement method to quantify subjective health outcomes. Med Care. 2008;46(4):357–65.

38. Louviere JJ, Lancsar E. Choice experiments in health: the good, the bad, the ugly and toward a brighter future. Health Econ Policy Law. 2009;4:527–46.

39. Coast J, Flynn TN, Salisbury C, Louviere J, Peters TJ. Maximising responses to discrete choice experiments: a randomised trial. Appl Health Econ Health Policy. 2006;5:249–60.

40. Hall J, Fiebig DG, King MT, Hossain I, Louviere JJ. What influences participation in genetic carrier testing? Results from a discrete choice experiment. J Health Econ. 2006;25:520–37.

41. Ramos-Goñi JM, Craig BM, Oppe M, Ramallo-Fariña Y, Pinto-Prades JL, Luo N, Rivero-Arias O. Handling data quality issues to estimate the Spanish EQ-5D-5L value set using a hybrid interval regression approach. Value Health. 2017. https://doi.org/10.1016/j.jval.2017.10.023.

42. Finn JD. The selection of contrast. In: A general model for multivariate analysis. New York: Holt, Rinehart and Winston; 1974.

43. Lamers LM, McDonnell J, Stalmeier PF, Krabbe PF, Busschbach JJ. The Dutch tariff: results and arguments for an effective design for national EQ-5D valuation studies. Health Econ. 2006;15(10):1121–32.

44. Versteegh MM, Vermeulen KM, Evers SMAA, de Wit GA, Prenger R, Stolk EA. Dutch tariff for the five-level version of EQ-5D. Value Health. 2016;19(4):343–52.

45. van Hout B, Janssen MF, Feng YS, Kohlmann T, Busschbach J, Golicki D, Lloyd A, Scalone L, Kind P, Pickard AS. Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. Value Health. 2012;15:708–15.

46. Wang P, Luo N, Tai ES, Thumboo J. The EQ-5D-5L is more discriminative than the EQ-5D-3L in patients with diabetes in Singapore. Value Health Reg Issues. 2016;9:57–62.

47. Burström K, Bartonek A, Broström EW, Sun S, Egmar A-C. EQ-5D-Y as a health-related quality of life measure in children and adolescents with functional disability in Sweden: testing feasibility and validity. Acta Paediatr. 2014;103:426–35.

48. Mulhern B, Bansback N, Hole AR, Tsuchiya A. Using discrete choice experiments with duration to model EQ-5D-5L health state preferences. Med Decis Mak. 2017;37(3):285–97.

49. Mulhern B, Feng Y, Shah K, van Hout B, Janssen B, Herdman M, Devlin N. Comparing the UK EQ-5D-3L and English EQ-5D-5L value sets: a report by the Centre for Health Economics Research and Evaluation. 2017. Available from: https://www.ohe.org/publications/comparing-uk-eq-5d-3l-and-english-eq-5d-5l-value-sets. Accessed 4 July 2017.

50. Wailoo A, Alava MH, Grimm S, Pudney S, Gomes M, Sadique Z, Meads D, O'Dwyer J, Barton G, Irvine L. Comparing the EQ-5D-3L and 5L versions: what are the implications for cost effectiveness estimates? Report by the decision support unit. 2017. Available from: http://scharr.dept.shef.ac.uk/nicedsu/wp-content/uploads/sites/7/2017/05/DSU_3L-to-5L-FINAL.pdf. Accessed 17 Aug 2017.

51. Janssen BMF, Oppe M, Versteegh MM, Stolk EA. Introducing the composite time trade-off: a test of feasibility and face validity. Eur J Health Econ. 2013;14(1):5–13.

52. Devlin N, Shah K, Feng Y, Mulhern B, van Hout B. Valuing health-related quality of life: an EQ-5D-5L value set for England. Health Econ. 2018;27(1):7–22.

53. Feng Y, Devlin NJ, Shah KK, Mulhern B, van Hout B. New methods for modelling EQ-5D-5L value sets: an application to English data. Health Econ. 2018;27(1):23–38.

54. Oppe M, Devlin NJ, van Hout B, Krabbe PFM, de Charro F. A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. Value Health. 2014;17(4):445–53.