PharmacoEconomics (2022) 40:1081–1093
https://doi.org/10.1007/s40273-022-01172-4

ORIGINAL RESEARCH ARTICLE

Is EQ-5D-5L Better Than EQ-5D-3L Over Time? A Head-to-Head Comparison of Responsiveness of Descriptive Systems and Value Sets from Nine Countries

Mathieu F. Janssen[1] · Ines Buchholz[2] · Dominik Golicki[3] · Gouke J. Bonsel[1,4]

Accepted: 5 July 2022 / Published online: 5 August 2022

Abstract

Objectives The aim of this study was to compare the responsiveness of EQ-5D-3L (3L) with EQ-5D-5L (5L) descriptive systems and value sets in two independent samples (rehabilitation and stroke patients).

Methods Descriptive system results were compared cross-sectionally, and descriptive responsiveness was tested by calculating changed level responses (‘moves’) from baseline to follow-up, proportion of improved patients, Paretian Classification of Health Change (PCHC), and probability of superiority (PS). Responsiveness of values based on nine country-specific value sets was assessed by standardized response mean (SRM) and standardized effect size (SES). Relative efficiency of 5L over 3L was assessed by calculating ratios of the SRM and SES statistics.

Results Descriptive comparisons confirmed earlier evidence and showed a consistent overestimation of health problems in 3L. Descriptive responsiveness improved with 5L in terms of moves per respondent, proportions of improved patients and PS, whereas PCHC showed mixed results. Better value responsiveness statistics were observed for 5L in rehabilitation patients for all value sets. In stroke patients, 3L showed better responsiveness statistics compared with 5L. Relative efficiency results were moderately to strongly better with 5L for rehabilitation, and slightly to moderately better with 3L for stroke.

Conclusions Descriptive results were the main driver of 3L–5L responsiveness differences. Responsiveness of 3L was influenced by the ‘confined to bed’ label and the overestimation bias of 3L, which affected all responsiveness results. This may impact quality-adjusted life-year (QALY) estimations, leading to over- or underestimations of QALYs gained, depending on the condition and condition severity. QALY calculations based on 5L data will result in more accurate estimates.

✉ Mathieu F. Janssen
mf.bas.janssen@gmail.com

Ines Buchholz
inesbuchholz@posteo.de

Dominik Golicki
dominik.golicki@wum.edu.pl

Gouke J. Bonsel
g.bonsel@erasmusmc.nl

[1] Section Medical Psychology and Psychotherapy, Department of Psychiatry, Erasmus MC, PO Box 2040, 3000 CA Rotterdam, The Netherlands

[2] EuroQol Group, Rotterdam, The Netherlands

[3] Department of Experimental and Clinical Pharmacology, Medical University of Warsaw, Warsaw, Poland

[4] Department of Public Health, Erasmus Medical Center, Rotterdam, The Netherlands

Key Points for Decision Makers

EQ-5D-3L systematically overestimates health problems when compared with EQ-5D-5L, leading to biased utilities.

This may impact quality-adjusted life-year (QALY) estimations when using EQ-5D-3L, leading to over- or underestimations of QALYs gained, depending on the condition and condition severity.

QALY calculations based on EQ-5D-5L data will result in more accurate estimates.

1 Introduction

The EQ-5D is a widely used preference-accompanied measure of health, developed to inform and improve decision making in health care [1]. EQ-5D value sets are a major feature of the EQ-5D instrument, facilitating the calculation of quality-adjusted life-years (QALYs) for use in economic evaluation of health care interventions. The original EQ-5D with three response levels ('3L') was revised and refined into the EQ-5D-5L ('5L') including five response levels [2]. Most studies comparing 3L with 5L demonstrated improved measurement properties with 5L, using cross-sectional data, assessing feasibility, distributional characteristics, inconsistencies, informativity and validity [3]. However, head-to-head comparisons of sensitivity to change (i.e., responsiveness) are rare [4–7] and use a variety of methods, leading to inconclusive results. Responsiveness to health changes is especially relevant in intervention and prevention studies where capturing improvement in health (or the absence of deterioration) is at the forefront.

The inconclusive evidence suggested that 5L does not necessarily result in better responsiveness than the 3L. In a sample of stroke patients, Golicki et al. found 5L values to be slightly less responsive compared with 3L [4], while Jia et al. found 5L and 3L to be equally responsive in a sample of hepatitis B patients [5]. Rather than using values, Buchholz et al. compared the responsiveness of 3L and 5L focusing on the descriptive systems, indicating 5L performed better within all comparisons [6]. The inconclusive findings when comparing values could be due to the use of interim ('crosswalk') values in both studies, based on a mapping function that converts 5L descriptive system scores into values based on 3L value sets [4, 5]. Empirical evidence showed that values based on mapping functions are differently distributed than values derived from valuation studies [8]. The only study available using value sets derived from standard 5L valuation studies demonstrated 5L to be more responsive over 3L in cataract surgery patients [7].

EQ-5D consists of five one-item dimensions of health: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression, with 3L dimensions describing levels of problems as 'no', 'some/moderate', and 'unable to/extreme' problems ('confined to bed' for mobility), and 5L as 'no', 'slight', 'moderate', 'severe', and 'unable to/extreme' problems. Value sets, providing values for each health profile, are available for many countries for both 3L and 5L, reflecting the societal preferences for the corresponding country [9].

Our study aim was to perform a more robust and comprehensive comparison of 3L versus 5L responsiveness, by comparing 3L with 5L using two independent patient samples, applying nine country-specific value set pairs. We built on a recently developed framework [10], separating the potential contribution of descriptive and value components in responsiveness. There, a bias in 3L was detected, with 3L systematically overreporting health problems when compared with 5L in nine study samples across six countries. Under the assumption that more levels lead to less measurement bias, while also taking the validity of 5L level descriptors into account [2], we concluded that 5L leads to more accurate estimations of health on the individual and group level. For example, we found that most respondents scoring the middle-level descriptor on 3L ('some/moderate') did not select the middle level on 5L ('moderate') but the milder level ('slight'). For aggregate results, this led to an overestimation of self-reported health problems on 3L in comparison with 5L, translating into an underestimation of aggregate values (utilities).

2 Methods

2.1 Paired 3L–5L Descriptive Data

Two longitudinal datasets were used: a German sample of inpatient rehabilitation patients (n = 225) and a Polish sample of stroke patients (n = 112) [4, 6]. The rehabilitation sample was tested at baseline and at the end of rehabilitation (follow-up), while the stroke sample was tested 1 week (baseline) and 4 months (follow-up) post stroke. Respondents were asked to complete both 3L and 5L as part of a larger paper survey (Table 1). In rehabilitation patients, the order of 3L and 5L was randomized, and that sequence was maintained across time points. In the stroke sample, the order was fixed, always starting with 5L. Only data of patients who fully completed both 3L and 5L at both time points were included.

The two different patient samples represent different patterns of health and potential health change. This may have an effect on responsiveness when a large part of observations would be at the 'tipping point' between two levels in 3L, but not in 5L.

2.2 Pairs of 3L–5L Value Sets

Nine pairs of 3L and 5L country-specific value sets were included: Canada, China, England/United Kingdom, Germany, Japan, The Netherlands, Korea, Poland, and Spain [11–28]. The same value sets were selected as for our previous study [10], with the addition of Germany and Poland, appropriate for the study samples. Most 3L valuation studies followed similar protocols, although there were differences in the sampling of respondents (affecting representation), sample size and health state selection [29, 30]. With the introduction of 5L, the EuroQol Valuation Technology Platform (EQ-VT) was developed—a standardized valuation protocol for uniform data procurement [31]. In addition to standardization of a computer-assisted personal interview

Table 1 Patient characteristics of the two study samples

	German rehabilitation sample	Polish stroke sample
Sample characterization		
Total sample size		
Baseline	230	112
Follow-up	224	112
Age: mean (SD)	57 (12.0)	70.6 (11.0)
Gender: *n* (%) female	160 (69.6)	58 (51.8)
Diseases: (ICD-10), *n* (%)		
Diseases of the spine and the back	58 (25.2)	
Arthropathies	40 (17.4)	
Affective disorders	38 (16.5)	
Neurotic, somatoform and stress-related disorders	11 (4.8)	
Inflammatory arthroplasties	40 (17.4)	
I61 (intracerebral hemorrhage)		8 (7.1)
I63 (cerebral infarction)		104 (92.9)
Respondent: *n* (%) baseline		
Patient	230 (100)	91 (81.3)
Proxy	0 (0)	21 (18.7)
Study design		
Type of study	Multicenter longitudinal study	Single-center observational longitudinal cohort
Baseline	Baseline, at the beginning of rehabilitation	Baseline during index hospitalization, before discharge
Follow-up	Approx. 4 (orthopaedics) to 6 (psychosomatics) week later, before discharge (at the end of rehabilitation)	Post-stroke recovery phase about 4 months later in outpatient clinics, neurological rehabilitation department or patients own home
Measures	EQ-5D-3L, EQ-5D-5L, SF-36 V1 or V2	Barthel Index, EQ-5D-3L, EQ-5D-5L, modified Rankin Scale
Mode of administration	Paper-and-pencil	Paper-and-pencil
Classifications of change (baseline to follow-up)[a]		
mRS-based: *n* (%)		
Improved		43 (38.4)
Stable		50 (44.6)
Deteriorated		19 (17.9)
Barthel Index-based: *n* (%)		
Improved		37 (33.0)
Stable		60 (53.6)
Deteriorated		15 (13.4)
Self-rated health-based: *n* (%)		
Improved	87 (39.2)	
Stable	121 (54.5)	
Deteriorated	14 (6.3)	

mRS modified Rankin scale, *ICD-10* International Classification of Diseases, Tenth Edition, *SF-36* Short-Form 36, *SD* standard deviation, *3L* EQ-5D-3L, *5L* EQ-5D-5L

[a] mRS: improvement of at least one level (improved), no change (stable), worse at least one level (deteriorated). Barthel Index: improvement of at least 9.25 points (improved), deterioration of at least − 9.25 points (deteriorated), deterioration of more than − 9.25 points, no change or improvement of < 9.25 points (stable). Self-rated health: response follow-up better (improved), no change (stable), response follow-up worse (deteriorated)

mode of administration, health state selection, and valuation methodology, a protocol of interviewer training and quality control during data collection was implemented [32]. For the United States (US), instead of using the recommended (separately developed) national value sets, 3L and 5L value sets were included that were derived for methodological purposes via EQ-VT from a common same sample [33], eliminating any potential effects induced by different protocols, study sample, valuation technique, or interviewers. The US values allowed for further assessment of the separate impact of descriptive results and values on responsiveness.

2.3 Analysis

2.3.1 Descriptive Cross-Sectional Analysis

Descriptive 3L and 5L statistics were calculated on the cross-sectional data (baseline and follow-up separately). The number of unique health profiles was determined for 3L and 5L in both patient samples. Next, we compared level sum scores (LSS) between 3L and 5L, by dimension. Recoding was applied to arrive at commensurability across levels: no problems = 0 (3L/5L), slight problems = 1 (5L), some/moderate problems = 2 (middle level 3L/5L), severe problems = 3 (5L), and extreme problems/unable to = 4 (most severe level 3L/5L). Dimension-specific LSS differences were 'standardized' by dividing absolute differences between 3L and 5L dimensions by sample size and the maximum possible level value (i.e., 4). The overall difference was calculated by summing the differences across dimensions and additionally dividing by the number of dimensions (i.e., 5). The resulting values (for both dimension-specific and overall standardized differences) range from – 1 to 1, with 0 meaning no difference and – 1 (or 1) meaning maximum difference of reported health problems between 3L and 5L. All 3L–5L dimension differences were statistically compared using Wilcoxon signed-rank tests.

2.3.2 Descriptive Responsiveness

First, inconsistencies in change between 3L and 5L were calculated within patients; an inconsistency exists if a dimension in 3L improves, while the same dimension in 5L deteriorates, or vice versa. Second, the absolute and average number of reported level changes from baseline to follow-up by respondent (here, 'moves') were calculated as a key descriptive indicator of responsiveness (e.g., moving from level 4–2 involves two moves). Third, the percentage of improved, stable, and deteriorated patients by dimension, and the percentage of improved patients according to the Paretian Classification of Health Change (PCHC) [34] were calculated and compared for 3L and 5L. According to PCHC, a health profile is considered to be 'better' if it is better on at least one dimension and not worse on any other dimension, and vice versa for 'worse'. Health profiles are considered 'the same' if there is no change on any dimension, and 'mixed' if a health profile is better in at least one dimension and worse in at least one dimension. Finally, a non-parametric effect size measure (probability of superiority [PS]) was calculated [6, 35] by dividing for each dimension the number of patients with positive changes by the total sample size. Ties (persons with no changes) were accounted for by adding half the number of ties in the numerator. The percentage of improved patients by dimension, the PCHC and the PS were interpreted as effect measures of descriptive responsiveness.

2.3.3 Value Responsiveness

Responsiveness of values was assessed using anchor-based approaches based on standardized response mean (SRM) and standardized effect size (SES), which are commonly used responsiveness statistics in patient-reported outcomes and the most commonly used in studies focused on EQ-5D [36, 37]. SRM was calculated as the ratio of the mean change to the standard deviation (SD) of that change. SES was calculated by dividing the mean change by the SD of the baseline measurement (originally introduced as Glass's

Table 2 Baseline and follow-up 3L versus 5L LSS by dimension[a] and patient sample, and standardized differences[b]

	Mobility			Self care			Usual activities			Pain/discomfort			Anxiety/depression			Sum
	3L	5L	*Δ*	3L	5L	*Δ*	3L	5L	*Δ*	3L	5L	*Δ*	3L	5L	*Δ*	*Δ*
Rehabilitation																
Baseline	222	248	*– 0.029*	74	83	*– 0.010*	360	320	*0.044*	456	383	*0.081*	256	234	*0.024*	*0.022*
Follow-up	168	186	*– 0.021*	58	61	*– 0.003*	268	234	*0.039*	370	294	*0.087*	182	149	*0.038*	*0.028*
Stroke																
Baseline	226	216	*0.022*	210	201	*0.020*	248	231	*0.038*	188	173	*0.033*	200	174	*0.058*	*0.034*
Follow-up	164	166	*– 0.004*	144	123	*0.047*	200	175	*0.056*	178	165	*0.029*	174	141	*0.074*	*0.040*

All 3L and 5L differences were statistically significant (Wilcoxon signed-rank test)

LSS level sum score, *3L* EQ-5D-3L, *5L* EQ-5D-5L

[a] Recoded: no problems = 0; 3L and 5L on the same scale. For 3L, level 2 = 2 and level 3 = 4; and for 5L, level 2 = 1, level 3 = 2, level 4 = 3 and level 5 = 4

[b] Calculated as the proportion of maximum possible reported health problems, with 0 meaning no differences between 3L and 5L, and – 1 or 1 meaning maximum possible difference between 3L and 5L (*Δ* = 3L minus 5L)

Delta [38]). External anchors that classified patients into change categories (improved, stable and deteriorated) were based on the five-level self-rated general health (SRH) question (item 1 of the SF-36: poor, fair, good, very good, excellent) for the rehabilitation sample, and on the modified Rankin Scale (mRS) and the 10-item version of the Barthel Index (BI) for the stroke sample (Table 1). The mRS and BI are widely used validated outcome measures in stroke with good psychometric properties [39]. Change categories were defined as follows for improved, stable and deteriorated, respectively; for mRS: improvement of at least one level; no change; worse at least one level; for BI (based on earlier published minimal clinically important differences [40]): more than or equal to 9.25 points; less than 9.25 points and more than − 9.25 points; less than or equal to − 9.25 points; SRH: response follow-up better; no change; response follow-up worse. Resulting SES and SRM statistics were interpreted using general benchmarks for effect size: 0.2–0.49 was interpreted as a small magnitude of effect; 0.5–0.79 was interpreted as a medium effect; and ≥0.8 was interpreted as a large effect [41].

Finally, to compare responsiveness for the nine value sets between 5L and 3L directly, we computed the 5L/3L ratio of the SRM and SES statistics as a measure of relative efficiency, so that a ratio higher than 1.0 indicated that 5L was more responsive than 3L [42]. For all comparisons, 95% confidence intervals (CIs) of SES, SRM and ratios were calculated using 1000 bootstrap samples.

Statistical significance was achieved when the values were different from 0 for SRM and SES, and different from 1.0 for the ratios.

As additional analysis, we investigated descriptive results for the improved subsamples by calculating LSS changes and 3L–5L differences to assess which dimensions account for the largest impact on responsiveness.

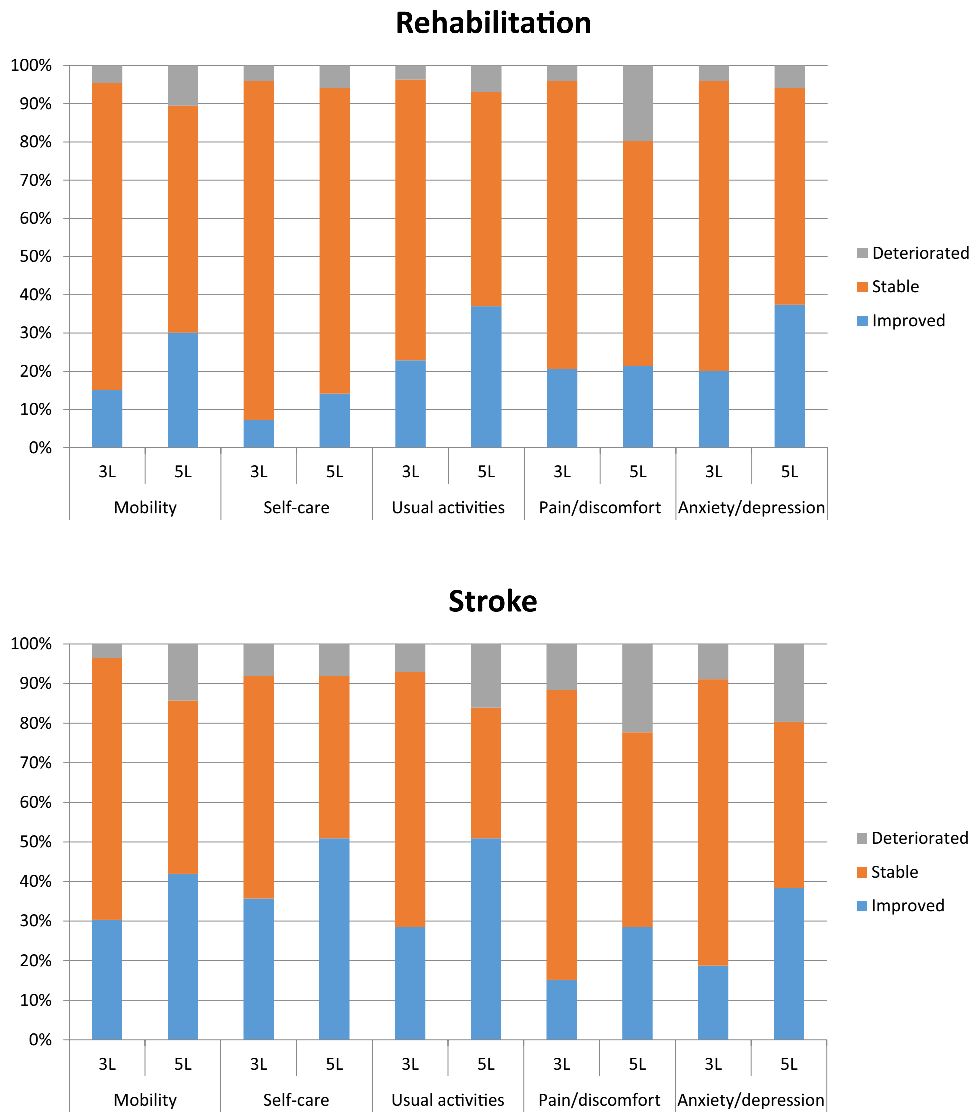

Fig. 1 Sensitivity to change (t1–t2) for rehabilitation ($n = 219$) and stroke ($n = 112$) samples. Deteriorated, stable and improved percentages were calculated according to individual-level responses that were worse at follow-up (more reported problems), stayed the same, or were better (less reported problems), respectively. *3L* EQ-5D-3L, *5L* EQ-5D-5L

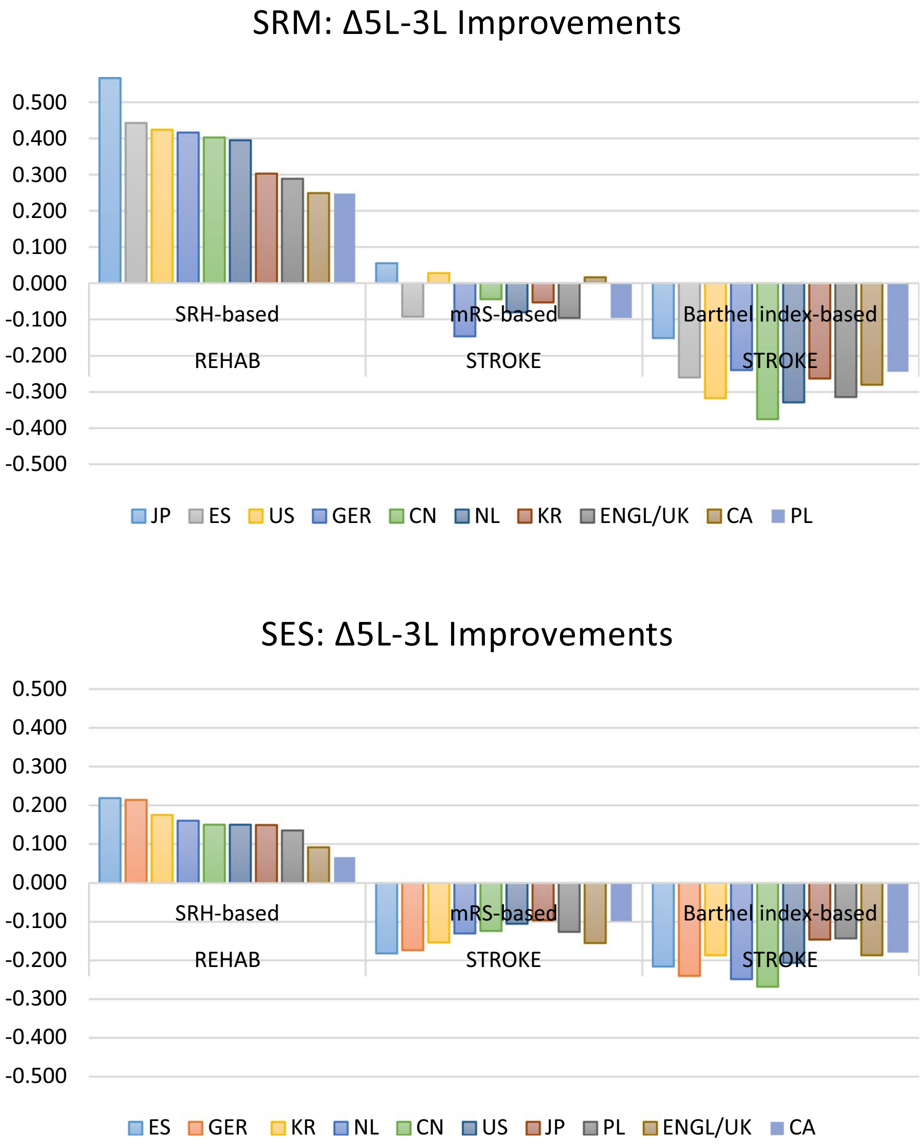

Fig. 2 SRM and SES differences in responsiveness for the improved subsamples. US value sets were derived using an identical methodological approach within the same sample of respondents. Δ = 5L – 3L; 5L shows better responsiveness over 3L if $\Delta_{\text{improvements}}$ are positive. *CA* Canada, *CN* China, *ENGL/UK* England/United Kingdom, *ES* Spain, *GER* Germany, *JP* Japan, *KR* Korea, Republic (South Korea), *NL* The Netherlands, *PL* Poland, *US* United States, *SRM* standardized response mean, *SES* standardized effect size, *3L* EQ-5D-3L, *5L* EQ-5D-5L, *SRH* self-rated health, *mRS* modified Rankin scale, *rehab* rehabilitation

2.4 Hypotheses

In line with previous evidence, we expected 5L descriptive cross-sectional results to reflect a higher number of different profiles and to show an overestimation of reported 3L health problems compared with 5L, with a possible exception of mobility (due to the 'confined to bed' level descriptor that is rarely scored). For a detailed analysis and description of the 3L bias, and our related claim on superior accuracy of 5L, we refer to our earlier study [10]. Overestimation of 3L was expected to be highest at the mild part of the severity spectrum [10]. In terms of descriptive responsiveness, we hypothesized that the number of moves will increase substantially with 5L, and PS will increase slightly to moderately (note that PS was previously reported for the rehabilitation sample [6]).

For the rehabilitation patients, we expected better value responsiveness for 5L. As, on average, rehabilitation patients moved from moderately impaired health states to mildly impaired health states [6], 3L overestimation might increase from baseline to follow-up (as we previously observed overestimation to be higher in mild conditions), leading to a reduced mean 3L difference from baseline to follow-up, and hence reduced responsiveness. The stroke patients generally moved from severe/moderate to moderately impaired health. Here, it is difficult to predict what to expect due to the mixed evidence of 3L overestimation in the moderate to severe spectrum [10].

Table 3 Observed relative efficiency of 5L over 3L values using the SRM and SES statistic ratio (improved subsamples only)[a]

	Rehabilitation						Stroke											
	SRH-based SRM			SRH-based SES			mRS-based SRM			mRS-based SES			BI-based SRM			BI-based SES		
	SRM ratio	[95% CI]		SES ratio	[95% CI]		SRM ratio	[95% CI]		SES ratio	[95% CI]		SRM ratio	[95% CI]		SES ratio	[95% CI]	
Canada	1.33	0.95	1.72	1.11	0.81	1.41	1.02	0.74	1.30	0.84	0.62	1.05	0.75	0.43	1.07	0.79	0.53	1.05
China	1.51	1.07	1.95	1.25	0.94	1.56	0.95	0.70	1.20	0.81	0.60	1.01	0.71	0.44	0.99	0.72	0.50	0.95
England/UK	1.46	1.02	1.89	1.17	0.84	1.49	0.88	0.61	1.16	0.76	0.55	0.97	0.72	0.36	1.08	0.74	0.49	1.00
Germany	1.82	1.00	2.64	1.46	0.96	1.97	0.82	0.59	1.04	0.69	0.51	0.87	0.77	0.45	1.08	0.73	0.48	0.98
Japan	1.74	1.31	2.17	1.22	0.87	1.58	1.07	0.81	1.33	0.85	0.65	1.04	0.87	0.51	1.23	0.84	0.59	1.10
Netherlands	1.60	1.11	2.09	1.28	0.95	1.62	0.90	0.60	1.19	0.78	0.56	1.01	0.70	0.40	1.00	0.72	0.49	0.95
Po land	1.43	1.02	1.84	1.28	0.94	1.62	0.88	0.68	1.07	0.80	0.64	0.96	0.77	0.49	1.05	0.83	0.60	1.05
South Korea	1.44	1.01	1.87	1.32	1.00	1.64	0.94	0.74	1.14	0.77	0.61	0.93	0.78	0.49	1.07	0.80	0.57	1.02
Spain	1.71	1.08	2.35	1.43	1.02	1.83	0.89	0.69	1.10	0.74	0.56	0.92	0.77	0.45	1.09	0.77	0.51	1.03
US[b]	1.59	1.17	2.01	1.26	0.96	1.56	1.04	0.78	1.30	0.83	0.62	1.04	0.73	0.45	1.01	0.77	0.52	1.02

SRH self-rated health, SRM standardized response mean, SES standardized effect size, mRS modified Rankin Scale, BI Barthel Index, CI confidence interval, UK United Kingdom, US United States

[a] Green cells indicate a significant SES or SRM ratio showing better responsiveness for 5L, orange cells for 3L (95% CI, 1000 bootstrap samples)

[b] US value sets were derived using an identical methodological approach within the same sample of respondents

Table 4 Descriptive LSS change scores[a] (baseline minus follow-up) and 3L–5L differences for the improved subsamples (Δ = 5L–3L)

	Mobility	Self-care	Usual activities	Pain/discomfort	Anxiety/ depression	Sum
Change scores						
Rehab—SRH-based (*n* = 84)						
3L	22	12	48	44	44	170
5L	40	20	51	54	53	218
Stroke—mRS-based (*n* = 43)						
3L	36	42	40	12	12	142
5L	28	49	42	− 1	17	135
Stroke—BI based (*n* = 37)						
3L	36	38	28	14	24	140
5L	27	39	30	5	26	127
Δ improvement (5L–3L)						
Rehab—SRH-based (*n* = 84)	18	8	3	10	9	48
Stroke—mRS-based (*n* = 43)	− 8	7	2	− 13	5	− 7
Stroke—BI based (*n* = 37)	− 9	1	2	− 9	2	− 13

LSS level sum score, *3L* EQ-5D-3L, *5L* EQ-5D-5L, *SRH* self-rated health, *BI* Barthel Index, *mRS* modified Rankin Scale, *rehab* rehabilitation

[a] Recoded: no problems = 0; 3L and 5L on the same scale. For 3L, level 2 = 2 and level 3 = 4; and for 5L, level 2 = 1, level 3 = 2, level 4 = 3 and level 5 = 4

3 Results

3.1 Descriptive Cross-Sectional Results

At baseline, 37 unique 3L profiles were observed versus 99 for 5L for rehabilitation, and 37 3L profiles versus 84 5L profiles for the stroke sample (results from follow-up were comparable). Level distributions by dimension are available elsewhere [4, 6]. Overestimation of 3L was observed on most dimensions (average 14% increase of reported problems in terms of LSS by dimension when compared with 5L), although not for mobility and self-care for the rehabilitation sample, and for baseline mobility for stroke, where underestimation is observed (Table 2). Overestimation of 3L health problems compared with 5L across the two patient samples was largest in pain/discomfort, followed by anxiety/depression and usual activities. Standardized differences varied from − 0.029 (rehabilitation baseline mobility) to 0.087 (rehabilitation follow-up pain/discomfort). For all dimensions combined, the standardized differences varied from 0.022 for rehabilitation baseline to 0.040 for stroke follow-up, generally showing larger 3L–5L differences for the stroke sample.

3.2 Descriptive Responsiveness

There were very few inconsistencies between 3L and 5L regarding the change data, with an average of 1.0% over

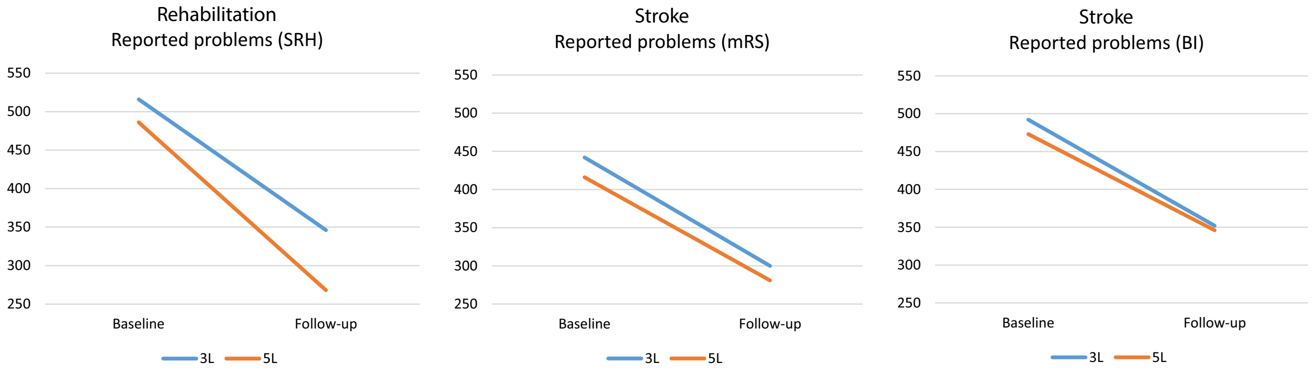

Fig. 3 Reported 3L versus 5L health problems over time (LSS*) for the improved subsamples. *Recoded: no problems = 0; 3L and 5L on the same scale. For 3L, level 2 = 2 and level 3 = 4; and for 5L, level 2 = 1, level 3 = 2, level 4 = 3 and level 5 = 4, summed across dimensions. *3L* EQ-5D-3L, *5L* EQ-5D-5L, *LSS* level sum score, *SRH* self-rated health, *BI* Barthel Index, *mRS* modified Rankin scale

dimensions for the rehabilitation sample and 1.1% for stroke. A considerable increase of reported changes (moves) was observed in 5L over 3L, with an average number of moves across dimensions of 47 (3L) versus 100 (5L) for rehabilitation, and 38 (3L) versus 65 (5L) for stroke. Average moves per respondent were 0.22 (3L) versus 0.46 (5L) in rehabilitation, and 0.34 (3L) versus 0.58 (5L) in stroke. The change data further show that 5L results in both more improvement and deterioration, which implies that more change is captured by 5L in both directions (Fig. 1). According to the PCHC, 50% (3L) versus 60% (5L) of patients improved in the rehabilitation sample, while 51% (3L) versus 44% (5L) improved in stroke. The PS index shows a clear but modest improvement of 5L over 3L in both samples, with $PS_{3L} = 0.57$ and $PS_{5L} = 0.59$ for rehabilitation, and $PS_{3L} = 0.59$ and $PS_{5L} = 0.63$ for stroke.

3.3 Responsiveness of Values

Mean values, SDs and differences for all value sets at baseline and follow-up are available in Appendix 1. Due to small sample sizes, SRM and SES statistics for the deteriorated subsamples were not statistically significant (Appendix 2); we will only report on the improved subsamples from here. All SRM and SES statistics for the improved subsamples were statistically significant and, with a few exceptions, were considered to be of medium to large magnitude. SRM values ranged from 0.51 (3L Germany) to 1.30 (3L China), while SES statistics ranged from 0.46 (3L Germany) to 1.01 (5L China). Both SRM and SES results were generally highest for the BI-based stroke subsample. Overall, 5L SRM and SES statistics were higher for rehabilitation and lower for the stroke subsamples. Overall, there was little variation in SRM and SES results between value sets in the three improved subsamples.

Differences in responsiveness demonstrated a clear and strong trend in favor of 5L for the rehabilitation subsample for both SRM and SES (Fig. 2). For the BI-based stroke sample, results are in favor of 3L, while for mRS there is a clear trend in favor of 3L for SES, while for SRM results are mixed and close to 0.

Relative efficiency results demonstrated the strongest results in favor of 5L for the rehabilitation subsample for SRM, with significant results for all value sets except Canada (Table 3). SES results for rehabilitation were significant for South Korea and Spain. The mRS-based subsample does not result in significant relative efficiency for SRM, while SES shows significant results for five value sets in favor of 3L. For the BI subsample, significant SRM results in favor of 3L were obtained for two value sets (China and The Netherlands), while SES results indicated significance for five value sets. Note that all non-significant SES results for the mRS stroke subsample were borderline significant, while many BI subsample results (both SRM and SES) were also close to significance. Although there were differences between value sets in terms of significance, all statistics were in the same direction (except for mRS stroke), and differences in CIs were relatively small across value sets, including the US value sets.

LSS change and difference scores between 3L and 5L by dimension are shown in Table 4. For rehabilitation, the largest effect between 3L and 5L was in mobility (an LSS increase of reported problems of 18 for 5L over 3L), which was due to the lack of level 3 scores for 3L, as only one respondent scored ‘confined to bed’. For both stroke subsamples, only mobility and pain/discomfort showed increased differences for 3L over 5L, while the other dimensions showed similar or increased responsiveness for 5L. Many stroke patients were evidently confined to bed at baseline but improved during follow-up, contributing to larger differences in LSS for 3L over 5L. Overestimation of 3L reported problems for the improved subsamples increased over time for rehabilitation, while 3L overestimation slightly decreased for stroke (Fig. 3).

4 Discussion

Although responsiveness results between 3L and 5L varied across patient samples, general conclusions could be formulated across study results. Cross-sectional descriptive results confirmed findings from our earlier study [10], concluding that the increased sensitivity of 5L over 3L leads to a more accurate measurement of health at individual and group levels. Again, we found that descriptive system results are the largest driver in outcomes, which confirms findings from Richardson et al. [43], showing that differences between utilities of different preference-accompanied instruments are mainly attributable to the descriptive data. Results for the US value sets further confirmed this finding, where US results were largely comparable with results of the other value sets.

Encouragingly, descriptive responsiveness results showed very few inconsistencies between 3L and 5L patient-level data (confirming the robustness of the data), but 5L showed a large increase of moves across dimensions. As previously demonstrated, 3L systematically overestimated health problems when compared with 5L, leading to biased (underestimated) utility values. The extent to which overestimation changed over time influenced responsiveness differences between 3L and 5L in both patient samples. On average, rehabilitation patients moved from moderately to mildly impaired health, and overestimation of 3L increased over time, therefore leading to worse responsiveness results for 3L. Apparently the tipping point in moving to milder levels is easier reached for 5L than for 3L. For stroke, opposite results were found, with 3L overestimation decreasing over time, although the effect was small for the mRS subsample.

The 3L ‘confined to bed’ level descriptor played a crucial role in differences between 3L and 5L in both patient samples. For rehabilitation, this response option was only endorsed once, leading to a less efficient use of 3L mobility compared with 5L, and, consequently, to lower 3L responsiveness. In stroke however, 21% of respondents scored ‘confined to bed’ at baseline, versus 7% at follow up, ultimately leading to an increase in value responsiveness. We did not expect there to be an underestimation of reported problems for 3L on self-care for the rehabilitation sample, although the effect is small. This could partly be explained by inconsistencies between 3L and 5L of the cross-sectional data [6]. For the German 3L value set, it must be noted that several utility decrements are lacking (usual activities levels 2 and 3, and anxiety/depression level 2), which might have affected the comparative results with the 5L value set.

It was striking that relative efficiency results for rehabilitation only showed two significant value set comparisons using SES, despite relatively large 3L–5L absolute differences in SES values (Fig. 2), which was due to large baseline SDs within the subsample.

SRM and SES statistics led to different results in relative efficiency, especially for the mRS stroke sample. As the denominator in SRM is the SD of the difference score, and in SES the baseline SD, apparently there were significant differences between SDs of the difference scores and baseline SDs, with the former being larger for 5L when compared with 3L, leading to similar SRM results. It seems that 5L shows more dispersion for assessing change scores for this particular subsample. Arguably, SRM could be considered to be the preferred method, making optimal use of the change data.

Contrary to our earlier study that assessed discriminatory power between patient groups, we judge the differences in responsiveness between values sets to be rather small, which is likely due to the nature of the data. When comparing the same group of patients over time, individual differences are cancelled out to a large extent, leading to less heterogeneity in values. This allowed for general conclusions with regard to 3L and 5L across value sets.

Some limitations must be acknowledged. There may have been an order effect for stroke, as 5L was always tested first. For the rehabilitation sample, only one anchor was used. Ideally, responsiveness would be tested with multiple anchors, preferably including a clinical anchor. Small sample sizes for the stroke improvement subsamples could have limited the number of significant results, especially for SES where many results were borderline significant. Proxy assessments were used in 18.7% of the stroke sample, which could have affected responsiveness results. However, Pickard et al. showed that although patient versus proxy agreement in a sample of 124 stroke patients was suboptimal, no systematic bias was observed [44]. Moreover, proxy assessments were only used in a relatively small proportion of our study sample. Responsiveness was assessed for only two patient samples, which, although very different in nature and disease severity, do not allow for generalization to other disease areas, especially about how 3L bias will manifest itself, e.g., dimension-specific problems frequently occurring in certain conditions might play a deciding role, like confined to bed did in our study samples. Furthermore, because patient samples from only two different countries were used, there could be country-specific effects related to translations of EQ-5D or to response behavior. Bias may also be present in 5L when compared with more refined classifications, although these effects will be substantially smaller. Moreover, an extended classification system beyond 5L will increasingly complicate valuation studies in terms of health state selection, valuation and modeling.

Based on the results from these two patient samples, it is expected that for moderate conditions where an improvement towards mild health states might be expected (such as in our rehabilitation sample), 3L underestimation of QALY gains may be observed. In severe conditions where overall improvements might be expected towards mild to moderate health (such as in our stroke sample), 3L might overestimate QALY gains, although this will be dependent on which health aspects are affected in any given condition.

5 Conclusion

This study shows that 5L leads to more accurate estimations of self-reported health and utility values over time, which will result in more accurate QALY estimations. Future 3L–5L comparative studies will have to show how the 3L bias will affect responsiveness when compared with 5L in other patient groups.

Appendix 1

See Table 5.

Table 5 Mean 3L and 5L values, standard deviations and differences (Δ = follow-up minus baseline) by patient sample at baseline and follow-up

	Rehabilitation						Stroke					
	Baseline		Follow-up		Δ		Baseline		Follow-up		Δ	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Canada												
3L	0.69	0.17	0.75	0.15	0.07	0.13	0.53	0.30	0.63	0.24	0.11	0.25
5L	0.70	0.19	0.77	0.17	0.07	0.15	0.52	0.31	0.62	0.27	0.10	0.21
China												
3L	0.68	0.17	0.75	0.16	0.07	0.13	0.49	0.28	0.60	0.25	0.10	0.21
5L	0.63	0.23	0.73	0.23	0.10	0.17	0.40	0.39	0.54	0.36	0.14	0.25
England/UK												
3L	0.60	0.27	0.69	0.23	0.09	0.23	0.40	0.40	0.53	0.34	0.13	0.33
5L	0.71	0.19	0.77	0.18	0.06	0.16	0.51	0.34	0.61	0.30	0.10	0.24
Germany												
3L	0.75	0.23	0.82	0.19	0.07	0.21	0.58	0.34	0.69	0.29	0.11	0.29
5L	0.73	0.21	0.80	0.19	0.08	0.18	0.52	0.40	0.64	0.33	0.12	0.29
Japan												
3L	0.65	0.13	0.71	0.14	0.06	0.11	0.49	0.27	0.59	0.22	0.10	0.21
5L	0.66	0.16	0.74	0.16	0.07	0.12	0.51	0.26	0.60	0.24	0.09	0.17
The Netherlands												
3L	0.65	0.24	0.73	0.20	0.09	0.21	0.50	0.32	0.59	0.28	0.10	0.27
5L	0.63	0.24	0.72	0.22	0.09	0.19	0.44	0.37	0.56	0.33	0.11	0.27
Poland												
3L	0.74	0.22	0.82	0.15	0.08	0.18	0.58	0.35	0.69	0.28	0.11	0.30
5L	0.83	0.15	0.88	0.13	0.05	0.13	0.63	0.37	0.74	0.29	0.11	0.28
South Korea												
3L	0.76	0.15	0.82	0.13	0.06	0.11	0.58	0.31	0.69	0.24	0.11	0.25
5L	0.70	0.15	0.76	0.15	0.06	0.12	0.55	0.27	0.64	0.24	0.09	0.19
Spain												
3L	0.65	0.27	0.74	0.23	0.09	0.22	0.37	0.47	0.53	0.39	0.16	0.38
5L	0.68	0.19	0.75	0.18	0.08	0.15	0.48	0.34	0.59	0.30	0.12	0.24
US[a]												
3L	0.68	0.20	0.76	0.16	0.08	0.16	0.45	0.40	0.59	0.31	0.14	0.33
5L	0.62	0.23	0.71	0.22	0.10	0.18	0.41	0.35	0.53	0.33	0.12	0.25

SD standard deviation, *3L* EQ-5D-3L, *5L* EQ-5D-5L, *UK* United Kingdom, *US* United States

[a] US value sets were derived using an identical methodological approach within the same sample of respondents

Appendix 2

See Table 6.

Table 6 Responsiveness statistics for 3L and 5L values by patient sample based on external anchors[a]

	Rehabilitation				Stroke							
	SRH-based				mRS-based				Barthel index-based			
	Deteriorated		Improved		Deteriorated		Improved		Deteriorated		Improved	
	SRM	SES	SRM	SES	SRM	SES	SRM	SES	SRM	SES	SRM	SES
Canada												
3L	– 0.08	– 0.09	0.75	0.61	– 0.34	– 0.49	0.76	0.61	– 0.54	– 0.53	1.12	0.86
5L	– 0.14	– 0.20	1.00	0.68	– 0.37	– 0.27	0.78	0.51	– 0.25	– 0.14	0.84	0.68
China												
3L	0.02	0.02	0.79	0.59	– 0.37	– 0.37	0.89	0.64	– 0.58	– 0.44	1.30	0.97
5L	– 0.07	– 0.07	1.19	0.74	– 0.28	– 0.18	0.84	0.52	– 0.19	– 0.09	0.93	0.71
England/UK												
3L	– 0.13	– 0.15	0.63	0.55	– 0.38	– 0.49	0.82	0.65	– 0.59	– 0.53	1.12	0.90
5L	– 0.22	– 0.33	0.92	0.64	– 0.52	– 0.38	0.72	0.49	– 0.39	– 0.23	0.81	0.67
Germany												
3L	– 0.14	– 0.17	0.51	0.46	– 0.38	– 0.53	0.81	0.64	– 0.54	– 0.49	1.03	0.89
5L	– 0.24	– 0.39	0.92	0.68	– 0.36	– 0.32	0.66	0.47	– 0.44	– 0.30	0.79	0.65
Japan												
3L	– 0.09	– 0.09	0.77	0.67	– 0.22	– 0.23	0.80	0.63	– 0.48	– 0.31	1.18	0.94
5L	– 0.12	– 0.12	1.33	0.81	– 0.30	– 0.19	0.85	0.53	– 0.24	– 0.11	1.03	0.79
The Netherlands												
3L	– 0.12	– 0.15	0.66	0.57	– 0.43	– 0.62	0.77	0.60	– 0.62	– 0.59	1.10	0.90
5L	– 0.25	– 0.37	1.05	0.73	– 0.54	– 0.42	0.69	0.47	– 0.39	– 0.22	0.77	0.65
Poland												
3L	– 0.10	– 0.11	0.57	0.49	– 0.38	– 0.57	0.77	0.63	– 0.58	– 0.69	1.06	0.82
5L	– 0.24	– 0.40	0.82	0.62	– 0.24	– 0.25	0.68	0.50	– 0.51	– 0.41	0.81	0.68
South Korea												
3L	0.03	0.03	0.69	0.54	– 0.30	– 0.33	0.85	0.67	– 0.54	– 0.44	1.21	0.92
5L	– 0.18	– 0.23	1.00	0.72	– 0.23	– 0.17	0.79	0.52	– 0.29	– 0.16	0.94	0.73
Spain												
3L	– 0.03	– 0.03	0.62	0.51	– 0.33	– 0.36	0.87	0.69	– 0.61	– 0.47	1.14	0.93
5L	– 0.17	– 0.22	1.06	0.73	– 0.42	– 0.31	0.77	0.51	– 0.36	– 0.22	0.88	0.71
US[b]												
3L	– 0.05	– 0.05	0.72	0.58	– 0.32	– 0.41	0.77	0.62	– 0.56	– 0.52	1.17	0.90
5L	– 0.17	– 0.22	1.14	0.73	– 0.47	– 0.33	0.79	0.51	– 0.26	– 0.15	0.86	0.70

SRH self-rated health, *mRS* modified Rankin Scale, *SRM* standardized response mean, *SES* standardized effect size, *3L* EQ-5D-3L, *5L* EQ-5D-5L, *UK* United Kingdom, *US* United States

[a] All SRM and SES statistics for the three deteriorated subsamples were not statistically significant; all SRM and SES statistics for the three improvement subsamples were statistically significant ($p < 0.001$)

[b] US value sets were derived using an identical methodological approach within the same sample of respondents

Acknowledgements The authors would like to thank Aki Tsuchiya for a constructive review of an earlier version of this manuscript.

Declarations

Funding This research was supported by the EuroQol Research Foundation under Grant number 20170600.

Conflict of interest Mathieu F. Janssen, Ines Buchholz, Dominik Golicki, and Gouke J. Bonsel are active members of the EuroQol Group, license holder of the EQ-5D family of instruments. The views expressed by the authors in this paper do not necessarily reflect the views of the EuroQol Group.

Compliance with ethical standards Not applicable since this study describes secondary data analyses.

Data availability As this study was based on a secondary data analysis, we refer to the original publications as mentioned in the manuscript for data access requests.

Code availability The code however is available upon request.

Consent to participate Not applicable.

Consent for publication Not applicable.

Author contributions Concept and design: All authors. Analysis: MJ. Interpretation of data: All authors. Drafting of the manuscript: MJ. Critical revision of the paper for important intellectual content: All authors. Obtaining funding: All authors.

References

1. Devlin NJ, Brooks R. EQ-5D and the EuroQol group: past, present and future. Appl Health Econ Health Policy. 2017;15:127–37.
2. 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;20:1727–36.
3. Buchholz I, Janssen MF, Kohlmann T, et al. A systematic review of studies comparing the measurement properties of the three-level and five-level versions of the EQ-5D. Pharmacoecon. 2018;36:645–61.
4. Golicki D, Niewada M, Karlinska A, et al. Comparing responsiveness of the EQ-5D-5L, EQ-5D-3L and EQ VAS in stroke patients. Qual Life Res. 2015;24:1555–63.
5. Jia YX, Cui FQ, Li L, et al. Comparison between the EQ-5D-5L and the EQ-5D-3L in patients with hepatitis B. Qual Life Res. 2014;23:2355–63.
6. Buchholz I, Thielker K, Feng YS, et al. Measuring changes in health over time using the EQ-5D 3L and 5L: a head-to-head comparison of measurement properties and sensitivity to change in a German inpatient rehabilitation sample. Qual Life Res. 2015;24:829–35.
7. Gandhi M, Ang M, Teo K, et al. EQ-5D-5L is more responsive than EQ-5D-3L to treatment benefit of cataract surgery. Patient. 2019;12:383–92.
8. Golicki D, Niewada M, van Hout B, et al. Interim EQ-5D-5L value set for Poland: first crosswalk value set in Central and Eastern Europe. Value Health Reg Issues. 2014;4:19–23.
9. Key EuroQol references: value sets. Available at: https://euroqol.org/publications/key-euroqol-references/value-sets/. Accessed 24 Dec 2021.
10. Janssen MF, Bonsel GJ, Luo N. Is EQ-5D-5L better than EQ-5D-3L? A head-to-head comparison of descriptive systems and value sets from seven countries. Pharmacoeconomics. 2018;36:675–97.
11. Xie F, Pullenayegum E, Gaebel K, et al. A time trade-off-derived value set of the EQ-5D-5L for Canada. Med Care. 2016;54:98–105.
12. Bansback N, Tsuchiya A, Brazier J, et al. Canadian valuation of EQ-5D health states: preliminary value set and considerations for future valuation studies. PLoS ONE. 2012;7: e31115.
13. Luo N, Liu G, Li M, et al. Estimating an EQ-5D-5L value set for China. Value Health. 2017;20:662–9.
14. Liu GG, Wu H, Li M, et al. Chinese time trade-off values for EQ-5D health states. Value Health. 2014;17:597–604.
15. Devlin N, Shah K, Feng Y, et al. Valuing health-related quality of life: an EQ-5D-5L value set for England. Health Econ. 2018;27:7–22.
16. Dolan P. Modeling valuations for EuroQol health states. Med Care. 1997;35:1095–108.
17. Ludwig K, Graf von der Schulenburg JM, Greiner W. German value set for the EQ-5D-5L. Pharmacoeconomics. 2018;36:663–74.
18. Greiner W, Claes C, Busschbach JJ, et al. Validating the EQ-5D with time trade off for the German population. Eur J Health Econ. 2005;6:124–30.
19. Shiroiwa T, Ikeda S, Noto S, et al. Comparison of value set based on DCE and/or TTO data: scoring for EQ-5D-5L health states in Japan. Value Health. 2016;19:648–54.
20. Tsuchiya A, Ikeda S, Ikegami N, et al. Estimating an EQ-5D population value set: the case of Japan. Health Econ. 2002;11:341–53.
21. Versteegh MM, Vermeulen KM, Evers SM, et al. Dutch tariff for the five-level version of EQ-5D. Value Health. 2016;19:343–52.
22. Lamers LM, McDonnell J, Stalmeier PF, et al. The Dutch tariff: results and arguments for an effective design for national EQ-5D valuation studies. Health Econ. 2006;15:1121–32.
23. Golicki D, Jakubczyk M, Niewada M, et al. Valuation of EQ-5D health states in Poland: first TTO-based social value set in Central and Eastern Europe. Value Health. 2010;13:289–97.
24. Golicki D, Jakubczyk M, Graczyk K, et al. Valuation of EQ-5D-5L Health States in Poland: the First EQ-VT-Based Study in Central and Eastern Europe. Pharmacoeconomics. 2019;37:1165–76.
25. Kim SH, Ahn J, Ock M, et al. The EQ-5D-5L valuation study in Korea. Qual Life Res. 2016;25:1845–52.
26. Lee YK, Nam HS, Chuang LH, et al. South Korean time trade-off values for EQ-5D health states: modeling with observed values for 101 health states. Value Health. 2009;12:1187–93.
27. Ramos-Goñi JM, Craig BM, Oppe M, et al. Handling data quality issues to estimate the Spanish EQ-5D-5L value set using a hybrid interval regression approach. Value Health. 2018;21:596–604.

28. Badia X, Roset R, Herdman M, et al. A comparison of United Kingdom and Spanish general population time trade-off values for EQ-5D health states. Med Decis Mak. 2001;21:7–16.
29. Xie F, Gaebel K, Perampaladas K, et al. Comparing EQ-5D valuation studies: a systematic review and methodological reporting checklist. Med Decis Mak. 2014;34:8–20.
30. Oppe M, Rand-Hendriksen K, Shah K, et al. EuroQol protocols for time trade-off valuation of health outcomes. Pharmacoeconomics. 2016;34:993–1004.
31. Oppe M, Devlin NJ, van Hout B, et al. A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. Value Health. 2014;17:445–53.
32. Ramos-Goñi JM, Oppe M, Slaap B, et al. Quality control process for EQ-5D-5L valuation studies. Value Health. 2017;20:466–73.
33. Law EH, Pickard AS, Xie F, et al. Parallel valuation: a direct comparison of EQ-5D-3L and EQ-5D-5L societal value sets. Med Decis Mak. 2018;38:968–82.
34. Devlin NJ, Parkin D, Browne J. Patient-reported outcome measures in the NHS: new methods for analyzing and reporting EQ-5D data. Health Econ. 2010;19:886–905.
35. Grissom RJ, Kim JJ. Effect sizes for research: univariate and multivariate applications. 2nd ed. New York: Taylor & Francis; 2012.
36. Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol. 2008;61:102–9.
37. Feng YS, Kohlmann T, Janssen MF, Buchholz I. Psychometric properties of the EQ-5D-5L: a systematic review of the literature. Qual Life Res. 2021;30:647–73.
38. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Med Care. 1989;27:178–89.
39. Quinn TJ, Dawson J, Walters MR, et al. Functional outcome measures in contemporary stroke trials. Int J Stroke. 2009;4:200–5.
40. Hsieh YW, Wang CH, Wu SC, et al. Establishing the minimal clinically important difference of the Barthel Index in stroke patients. Neurorehabil Neural Repair. 2009;21:233–8.
41. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale: Lawrence Erlbaum Associates; 1988.
42. Luo N, Johnson JA, Shaw JW, et al. Relative efficiency of the EQ-5D, HUI2, and HUI3 index scores in measuring health burden of chronic medical conditions in a population health survey in the United States. Med Care. 2009;47:53–60.
43. Richardson J, Iezzi A, Khan MA. Why do multi-attribute utility instruments produce different utilities: the relative importance of the descriptive systems, scale and 'micro-utility' effects. Qual Life Res. 2015;24:2045–53.
44. Pickard AS, Johnson JA, Feeny DH, et al. Agreement between patient and proxy assessments of health-related quality of life after stroke using the EQ-5D and Health Utilities Index. Stroke. 2004;35:607–12.