Quantification of the level descriptors for the standard EQ-5D three-level system and a five-level version according to two methods

M. F. Janssen · E. Birnie · G. J. Bonsel

Accepted: 6 February 2008 / Published online: 5 March 2008

Abstract

Objectives Our aim was to compare the quantitative position of the level descriptors of the standard EQ-5D three-level system (3L) and a newly developed, experimental five-level version (5L) using a direct and a vignette-based indirect method.

Methods Eighty-two respondents took part in the study. The direct method represented a visual analog scale (VAS) rating of the nonextreme level descriptors for each dimension and each instrument separately. The indirect method required respondents to score 15 health scenarios with 3L, 5L and a VAS scale. Investigated were: (1) equidistance (Are 3L and 5L level descriptors distributed evenly over the VAS continuum?); (2) isoformity (Do the identical level descriptors on 3L and 5L yield similar results?); and (3) consistency between dimensions (Do the positions of similar level descriptors differ across dimensions within instruments?).

Results Equidistance without transformation was rejected for all dimensions for both 3L and 5L but satisfied for 5L after transformation. Isoformity gave mixed results. Consistency between dimensions was satisfied for both instruments and both methods.

Discussion The level descriptors have similar distributions across comparable dimensions within each system, but the pattern differs between 3L and 5L. This methodological study provides evidence of increased descriptive power and a broadened measurement continuum that encourages the further development of an official five-level EQ-5D.

Keywords EQ-5D · Methodology · Health-related quality of life · Psychometrics · Health status

Introduction

The EQ-5D is a widely used instrument to describe and value generic health (status) in terms of five dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Each dimension comprises three levels, indicating no problems, some or moderate problems, and extreme problems, resulting in a total of 243 (3^5) unique health states [1].

The condensed format of the EQ-5D has undoubtedly contributed to its global dissemination, as it is easy to include in existing surveys by questionnaire designers, easy to fill out by respondents, and easy to report by analysts. However, compared with other generic preference based instruments such as the Health Utilities Index Mark 2 and Mark 3 (HUI2 and HUI3) and the Short Form 6D (SF-6D), which define respectively 24,000, 972,000, and 18,000 unique health states, the EQ-5D is lacking descriptive richness [2–5]. Although the EQ-5D descriptive system has demonstrated strong psychometric properties in general, its restricted ability to discriminate (clinically relevant) small to moderate differences in health status between individuals or within individuals over time is recognized [6–9].

M. F. Janssen (✉) · E. Birnie · G. J. Bonsel
Public Health Epidemiology, Department of Social Medicine,
Academic Medical Center, P.O. Box 22660, Amsterdam DD 1100,
The Netherlands
e-mail: m.f.janssen@amc.uva.nl

M. F. Janssen · G. J. Bonsel
EuroQol Group, Amsterdam, The Netherlands

E. Birnie · G. J. Bonsel
Institute of Health Policy and Management, Erasmus MC,
Rotterdam, The Netherlands

Moreover, several studies have reported on the ceiling effect of the EQ-5D in the general population as well as in patient populations [10–15].

A straightforward way of improving the discriminatory potential of the EQ-5D descriptive system is to increase the number of response options. In most health-status classification systems, the response options are ordered in terms of severity along a hypothetical measurement continuum. Since the exact position of the response options defines the discriminatory abilities of the descriptive system [16, 17], it is important to know where on the measurement continuum the level descriptors are quantitatively positioned.

Previous research in which a five-level (5L) version of EQ-5D was compared with the standard three-level (3L) EQ-5D demonstrated increased discriminatory power, increased reliability, and satisfactory validity [18, 19]. This paper presents a head-to-head comparison of the quantitative positioning of the level descriptors of the standard 3L EQ-5D descriptive system versus a newly developed, experimental 5L system, which covers 3,125 unique health states (5^5). Two independent methods were used. The first method directly compared the nonextreme level descriptors (for 3L: the level two midcategory; for 5L: the level two, level three, and level four categories) for each dimension separately on a visual analogue scale (VAS). The second, indirect, method required respondents to score complete health scenarios (vignettes) on dimension-specific VAS scales and subsequently to classify the same vignettes on the two EQ-5D instruments (3L and 5L).

Methods

Instruments

Three instruments were used in this study: the standard EQ-5D3L version, an adapted Dutch 5L version developed in 1993 [20], and a set of five dimension-specific VAS scales. The version of the 5L EQ-5D used in this study was an experimental version, since at the time of this study, no official five-level version had been advocated by the EuroQol Group. We chose to test a five-level EQ-5D system, even though we also could have chosen four or six levels. An increase in the number of levels is always an increase of discriminatory potential at the cost of a more complex descriptive system (which might compromise the robustness of the value function). Five levels appears to be an optimal number of response options concerning reliability [21, 22]. Furthermore, Preston et al. (2000) investigated feasibility for 11 different rating formats (ranging from 2 to 11 and a 101 point scale) and found that feasibility peaked at five levels [23]. We chose to add two in-between levels to the existing 3L descriptive system (between levels 1 and 2 and levels 2 and 3) because we considered this the most obvious option in regard to the objective of refining the EQ-5D instrument. In any preference-based instrument, level descriptors are practically required for valuation research in which generic profiles are to be valued. A small focus group was assigned to determine the wording of the level descriptors. The level descriptors presented here were translated from Dutch. The one-, three-, and five-level descriptors in 5L were the same as the one-, two-, and three–level descriptors in the standard EQ-5D3L. The grading terms that were used for the intermediate levels two and four in the 5L-system were “a little” for level 2 (5L-2) in Anxiety/Depression and “mild problems” for the remaining dimensions; and “severe” for level 4 (5L-4) in Pain/Discomfort, “very” for Anxiety/Depression, and “many problems” for the remaining dimensions. One further alteration was made to both the 3L and 5L systems: the most severe response category in Mobility was changed from “confined to bed” to “unable to walk about”, so it would be analogous to the extreme response categories of the other dimensions. Table 1 displays the exact wording of the descriptors in the 3L and 5L systems, respectively.

To obtain quantitative values for each level descriptor of 3L and 5L, the VAS was used. We used five VAS scales, one for each EQ-5D dimension. Each VAS consisted of a horizontal hashmarked line without corresponding numbers, with the extreme-level descriptors belonging to that dimension as anchors. Respondents were asked to indicate their score on the VAS by marking the line. For the most severe category of Pain/Discomfort and Anxiety/Depression, the original descriptor was labeled “extreme”. Because the study was part of a larger process of choosing the definite level descriptors for the official five-level version of the EQ-5D, we decided to use the entire continuum of disability (extreme included), and used “worst imaginable” as upper VAS anchor for these two dimensions. This is analogous to the other three dimensions, which ranged from “no problems” to “unable to”.

Study design

Data collection took place in the form of one of two panel sessions and a follow-up postal survey 2 weeks later. A convenience sample of 82 laypeople from an existing general population panel ($N = 560$) participated. All participants were familiar with the vignette presentation form used in the indirect method.

All participants completed both the direct and the indirect quantification task. For the direct method, all 3L answers were obtained during the panel sessions and all 5L answers as part of the postal survey to avoid memory effects. For the indirect method, participants scored ten health states in the panel sessions (acute pharyngitis,

Table 1 Direct quantification of three- and five- level (3L, 5L) descriptors

	Number	Mean	Median	95% CI
3L				
Mobility				
No problems in walking about[a]	–	–	–	–
Some problems in walking about	74	26.70	22	22.82−30.59
Unable to walk about[a]	–	–	–	–
Self-care				
No problems with self-care[a]	–	–	–	–
Some problems washing or dressing self	74	30.18	28	26.10−34.25
Unable to wash or dress self[a]	–	–	–	–
Usual activities				
No problems with performing usual activities[a]	–	–	–	–
Some problems with performing usual activities	77	29.74	25	25.95−33.53
Unable to perform usual activities[a]	–	–	–	–
Pain/Discomfort				
No pain or discomfort[a]	–	–	–	–
Moderate pain or discomfort	66	32.33	31	28.56−36.10
Extreme pain or discomfort	66	86.36	89	83.75−88.98
Worst imaginable pain or discomfort[a]	–	–	–	–
Anxiety/Depression				
Not anxious or depressed[a]	–	–	–	–
Moderately anxious or depressed	67	33.94	34	29.89−37.99
Extremely anxious or depressed[a]	67	88.82	90	86.88−90.77
Worst imaginable anxiety or depression[a]	–	–	–	–
5L				
Mobility				
No problems in walking about[a]	–	–	–	–
Mild problems in walking about	75	11.31	11	9.73–12.88
Some problems in walking about	75	38.39	40	35.39–41.39
Many problems in walking about	75	79.80	82	76.81–82.79
Unable to walk about[a]	–	–	–	–
Self-care				
No problems with self-care[a]	–	–	–	–
Mild problems washing or dressing self	76	11.24	10	9.72–12.76
Some problems washing or dressing self	76	37.14	38	34.14–40.15
Many problems washing or dressing self	76	80.61	81	77.81–83.40
Unable to wash or dress self[a]	–	–	–	–
Usual activities				
No problems with performing usual activities[a]	–	–	–	–
Mild problems with performing usual activities	77	11.08	10	9.29–12.87
Some problems with performing usual activities	77	39.01	40	36.12–41.90
Many problems with performing usual activities	77	80.81	83	77.70–83.91
Unable to perform usual activities[a]	–	–	–	–
Pain/Discomfort				
No pain or discomfort[a]	–	–	–	–
Mild pain or discomfort	53	8.85	8	7.43–10.26
Moderate pain or discomfort	53	32.32	31	29.58–35.06
Severe pain or discomfort	53	67.94	68	64.98–70.90
Extreme pain or discomfort	53	91.26	94	88.96–93.57

Table 1 continued

	Number	Mean	Median	95% CI
Worst imaginable pain or discomfort[a]	–	–	–	–
Anxiety/Depression				
Not anxious or depressed[a]	–	–	–	–
A little anxious or depressed	59	9.46	8	7.97–10.94
Moderately anxious or depressed	59	32.56	33	30.01–35.11
Very anxious or depressed	59	67.37	66	64.55–70.20
Extremely anxious or depressed	59	91.34	92	89.42–93.25
Worst imaginable anxiety or depression[a]	–	–	–	–

CI confidence interval

[a] Level descriptor used as anchor in visual analog scale

exacerbation of eczema, hip fracture, cerebrovascular accident/stroke with moderate impairments, moderate gastritis, low spinal cord lesion, mild depression, back and neck pain, severe dementia, and acute multiple injury) and the remaining five in the survey (otitis externa, severe stable brain injury, irritable bowel syndrome, acute large burn, and posttraumatic stress disorder), because we expected that more than ten health states within one session could lead to concentration problems. The two sets of health states were balanced according to severity and duration. Following this design, the indirect method provided 225 responses for each respondent: 15 diseases × 5 dimensions × 3 response scales.

Direct quantification of level descriptors

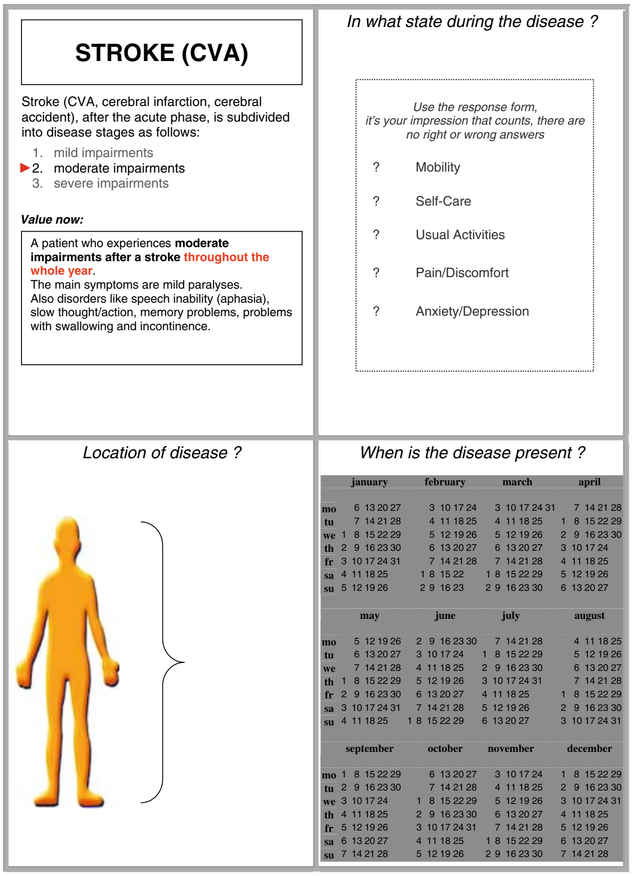
In the direct method, respondents were asked to project the 3L and the 5L descriptors on the VAS scales for each dimension separately. As the extreme levels were used as anchors of the VAS, for 3L only, the midcategory (3L-2) level descriptor needed to be scored, except for Pain/Discomfort and Anxiety/Depression, which needed additional scoring of 3L-3 (extreme). Similarly, the midcategories 5L-2, 5L-3, and 5L-4 descriptors were scored for each dimension, except for Pain/Discomfort and Anxiety/Depression, which included the scoring of 5L-5.

Indirect quantification of level descriptors

As an alternative to the direct method, we developed an indirect method that we believe lies closer to the actual use of the EQ-5D instrument, as it uses a (hypothetical) health state as a calibrator or medium to derive a VAS score. In contrast to the direct method, the object of measurement in the indirect method is not a 3L or 5L descriptor but a complete health scenario (vignette). Each vignette was scored with the 3L and 5L descriptors and on a VAS, one for each separate dimension, independently. Consequently, an indirect head-to-head comparison of 3L and 5L scores could be made, calibrated via the common VAS score.

Figure 1 shows one of the vignettes. Each vignette was designed to present a disease as close to clinical reality as possible, therefore also including information on disease duration. All 15 diseases were presented on a standardized sheet (vignette) that contained (1) a disease label with a naturalistic description of the disease; (2) the course of the disease over a 1-year period using a calendar (the grey scales represent the duration of the disease); (3) the location of the disease with, if relevant, a visual representation; and (4) the EQ-5D dimensions, of which the levels were left unspecified, as the respondents were invited to select the appropriate EQ-5D level (according to his or her own view) for each dimension. Respondents were asked to read each vignette carefully and to select the level of each dimension of the EQ-5D descriptive system that best described the presented health state in their view using three response scales: the standard 3L response scale, the new 5L scale, and the VAS scale (similar to the VAS used in the direct method).

The 5L and 3L response scales were presented on the left and the right side of one page (per dimension), respectively. The respondents were first invited to score the 5L descriptors for all dimensions and all vignettes while covering the right side of the page that showed the 3L descriptors. Next, they were instructed to return to the first vignette, asked to cover the left side with the 5L scores, and provide the 3L response for all vignettes. Pilot testing revealed that when respondents scored 3L first, there was a tendency to avoid the in-between levels 2 and 4 of 5L, and for this reason, all respondents were asked to score 5L first. Adequate instruction was critical, stressing that 3L and 5L were two independent ways of scoring (in the postal survey, these instructions were repeated in writing). Subsequently, VAS scores were obtained on a separate form without respondents having access to the 3L and 5L scores. The demanding task of first providing 5L



STROKE (CVA)

Stroke (CVA, cerebral infarction, cerebral accident), after the acute phase, is subdivided into disease stages as follows:

1. mild impairments
2. moderate impairments
3. severe impairments

Value now:

A patient who experiences **moderate impairments after a stroke throughout the whole year.**
The main symptoms are mild paralyses.
Also disorders like speech inability (aphasia), slow thought/action, memory problems, problems with swallowing and incontinence.

In what state during the disease ?

Use the response form, it's your impression that counts, there are no right or wrong answers

? Mobility

? Self-Care

? Usual Activities

? Pain/Discomfort

? Anxiety/Depression

Location of disease ?

When is the disease present ?

	january	february	march	april
mo	6 13 20 27	3 10 17 24	3 10 17 24 31	7 14 21 28
tu	7 14 21 28	4 11 18 25	4 11 18 25	1 8 15 22 29
we	1 8 15 22 29	5 12 19 26	5 12 19 26	2 9 16 23 30
th	2 9 16 23 30	6 13 20 27	6 13 20 27	3 10 17 24
fr	3 10 17 24 31	7 14 21 28	7 14 21 28	4 11 18 25
sa	4 11 18 25	1 8 15 22	1 8 15 22 29	5 12 19 26
su	5 12 19 26	2 9 16 23	2 9 16 23 30	6 13 20 27

	may	june	july	august
mo	5 12 19 26	2 9 16 23 30	7 14 21 28	4 11 18 25
tu	6 13 20 27	3 10 17 24	1 8 15 22 29	5 12 19 26
we	7 14 21 28	4 11 18 25	2 9 16 23 30	6 13 20 27
th	1 8 15 22 29	5 12 19 26	3 10 17 24 31	7 14 21 28
fr	2 9 16 23 30	6 13 20 27	4 11 18 25	1 8 15 22 29
sa	3 10 17 24 31	7 14 21 28	5 12 19 26	2 9 16 23 30
su	4 11 18 25	1 8 15 22 29	6 13 20 27	3 10 17 24 31

	september	october	november	december
mo	1 8 15 22 29	6 13 20 27	3 10 17 24	1 8 15 22 29
tu	2 9 16 23 30	7 14 21 28	4 11 18 25	2 9 16 23 30
we	3 10 17 24	1 8 15 22 29	5 12 19 26	3 10 17 24 31
th	4 11 18 25	2 9 16 23 30	6 13 20 27	4 11 18 25
fr	5 12 19 26	3 10 17 24 31	7 14 21 28	5 12 19 26
sa	6 13 20 27	4 11 18 25	1 8 15 22 29	6 13 20 27
su	7 14 21 28	5 12 19 26	2 9 16 23 30	7 14 21 28

Fig. 1 Disease vignette with empty EQ-5D descriptive system

classifications on all five dimensions of all 15 vignettes minimized possible memory effects when the participants were instructed to return to the first vignette to score the 3L classifications while covering the 5L responses.

Analysis

Results of the direct and indirect methods are presented with conventional descriptive statistics. Results of the indirect method were derived by grouping 3L-VAS pairs and 5L-VAS pairs for each respondent per vignette and subsequently by calculating level means over all vignettes and all respondents combined. For each respondent, scorings were removed for the combined 3L, 5L, and VAS scores if at least one of the 3L, 5L, or VAS scores was missing, equalizing the number of VAS observations between 3L and 5L.

Characteristics

For both the direct and indirect methods, the 3L–5L extension of EQ-5D was investigated in terms of three characteristics. First, equidistance addresses the degree to which 3L and 5L level descriptors are distributed evenly over the VAS continuum, either without or with transformation. Equidistance is determined for each dimension and each instrument (3L and 5L) separately. Untransformed equidistance implies that level descriptors are distributed according to VAS ratings of 0–50–100 for 3L and 0–25–50–75–100 for 5L. There is evidence that the precision of the VAS might be illusory, as respondents mentally divide the VAS continuum in a smaller number of segments, which is nine or ten at maximum [23, 24]. Therefore, we defined a deviation of 5 VAS points as the maximum acceptable deviation (which makes a segment of 10 VAS points, as the deviation can be either way). Furthermore, a deviation of 5 VAS points has been used before [16]. If untransformed equidistance is rejected, equidistance using power [$y = (ax)^b$] transformation is considered. A power relation of, e.g., $y = (5.38*x)^{1.5}$ for 5L would result in a VAS rating distribution of 0–12–35–65–100. Note that transformation is only possible for 5L, as there is only one 3L observation apart from the anchors.

Part of the evaluation of equidistance is analysis of the position of the extreme levels according to the indirect method: are the VAS ratings for the extreme level descriptors close to the supposed anchor values for the indirect method? Ideally, 3L-1 and 5L-1 scores would equal 0 and 3L-3 and 5L-5 scores would equal 100, except for Pain/Discomfort and Anxiety/Depression in which the 3L and 5L extreme level descriptors were not identical to the VAS anchors.

Second, isoformity is the degree to which the positions of 3L-2 and 5L-3 level descriptors (and also 3L-3 versus 5L-5 for Pain/Discomfort and Anxiety/Depression) are similar. Isoformity directly compares the 3L and 5L descriptive systems for each separate dimension between instruments. For the indirect method, all 3L level means, including 3L-1 and 3L-3, can be compared with 5L. Analysis of isoformity is based on paired 3L–5L response means for each dimension separately. For the direct method, isoformity was tested with a paired *t* test between the 3L and 5L scorings. For the indirect method, a deviation of 5 VAS points was defined as the maximum acceptable deviation.

Finally, consistency between dimensions is the degree to which the positions of the same level descriptors differ across dimensions. Consistency, between dimensions was tested for each instrument (3L, 5L) separately. The first three dimensions (Mobility, Self-Care, and Usual Activities) were distinguished from the last two (Pain/Discomfort and Anxiety/Depression), as these—in Dutch—share identical level descriptors, e.g., some problems for Mobility, Self-Care, and Usual Activities. For the direct method, analysis of variance (ANOVA) was used for each identical level descriptor for the first three dimensions combined (one comparison for 3L and three for 5L) and

Pain/Discomfort and Anxiety/Depression combined (two comparisons for 3L and four for 5L), resulting in a total of ten comparisons . For the indirect method, consistency is tested with a generalizability study (G-study). In a G-study, one is able to separate multiple sources of error variance [25]. Generalizability coefficients (G-coefficients) can be constructed as functions of the estimated variance components, expressing consistency on a 0–1 scale, with 1 expressing perfect consistency [26, 27]. We used a variance components analysis based on the restricted maximum likelihood method and identified four possible sources of variance: label, vignette, dimension, and respondent. Four separate G-studies were conducted, one on the first three dimensions and one on the remaining two dimensions, for each instrument (3L, 5L) separately. A G-coefficient expressing consistency between dimensions was calculated on the basis of these variance components ("Appendix A").

We regarded transformed or untransformed equidistance to be a desirable characteristic for the new 5L system as opposed to no systematic relation between the quantitative position of the level descriptors at all. Consistency between identical-level descriptors across dimensions was also regarded as a desirable property because this expresses that respondents have a consistent conceptualization of the grading terms used over different dimensions of health. When consistency is achieved, this does not imply that utility values would also be expected to be consistent over dimensions, because utility values are an expression of an entire EQ-5D profile, whereas we investigated VAS scores within each dimension separately. Furthermore, a choice-based method presumably leads to different results than the dimension-specific VAS scales we used. We investigated isoformity to see whether the new 5L system was a refinement or a new system, and whether isoformity was achieved or not does not tell us anything about the 5L system in itself.

Results

The mean age of the participants was 53.6 years, with 42.7% being men. Of the 82 respondents who attended in the panel sessions, 81 returned the survey. Three respondents (4%) were of Turkish nationality, two (2%) were of Moroccan nationality, and the remaining 75 (94%) were of Dutch origin. In the Pain/Discomfort and Anxiety/Depression dimensions, respondents often failed to score the extreme-level descriptor when using the direct method (8 and 9 for 3L, respectively, and 22 and 16 for 5L, respectively). For these respondents, the remaining scorings were deleted for that dimension because of possible context effects (i.e., spreading out the VAS scores of the remaining 3L descriptors over the VAS scale). For the direct method, missing responses for 3L ranged from 6.1% (Usual Activities) to 19.5% (Pain/Discomfort) and for 5L from 4.9% (Usual Activities) to 34.6% (Pain/Discomfort). For the indirect method, missing responses ranged from 1.1% (Usual Activities) to 2.5% (Pain/Discomfort) for the three response scales (3L, 5L, and VAS) combined.

Characteristics: direct method

Results for the direct method are shown in Table 1 and Fig. 2. Untransformed equidistance was rejected for all level descriptors except 5L-4 in Mobility (80), although Self-Care and Usual Activities were only 1 VAS point away for Mobility. Regardless of dimension, level descriptors were positioned systematically lower than the expected value for equidistance for 3L-2 (16–23 VAS points lower), 5L-2 (14–16 points lower), and 5L-3 level (11–18 points lower), whereas 5L-4 was sometimes higher (4–5 points) and sometimes lower (7–8 points). Transformed equidistance (power function) provided an excellent fit for all dimensions of 5L ($R^2 \geq 0.99$).

Isoformity could not be established except for the middle-level descriptors (3L-2 vs. 5L-3) for Pain/Discomfort and Anxiety/Depression (Table 2). Relatively large gaps appeared between 3L-2 and 5L-3 for Mobility (11), Self-Care (8), and Usual Activities (9), with 5L-3 showing systematically higher values. Although there was a statistically significant difference between the extreme level descriptors (3L-3 vs. 5L-5) for Anxiety/Depression, the absolute difference was 3 VAS points.

Consistency between dimensions gives supportive results for both 3L and 5L, as none of the ten comparisons (ANOVA) showed significant differences (see Fig. 2). Generally, VAS means are similar among the first three dimensions as well as among Pain/Discomfort and Anxiety/Depression.

Characteristics: indirect method

Results of the indirect method are shown in Table 3 and Fig. 3. Untransformed equidistance of 3L-2 was rejected for all dimensions (systematically 7–14 VAS points too low) as well as for 5L-2 (systematically 8–13 points lower) and 5L-3 (systematically 8–17 points lower). Untransformed equidistance was achieved only for the 5L-4 level for all dimensions (systematically 1–5 points lower), with VAS scores ranging from 70 (Mobility and Usual Activities) to 74 (Anxiety/Depression). Transformed equidistance (power function) provided an excellent fit for all dimensions of 5L ($R^2 \geq 0.99$).

VAS results for the extreme-level descriptors show that the lower extreme is close to 0, except for Pain/Discomfort

Fig. 2 Direct quantification of the three- and five-level (3L, 5L) descriptors. Visual analog scale (VAS) means by dimension

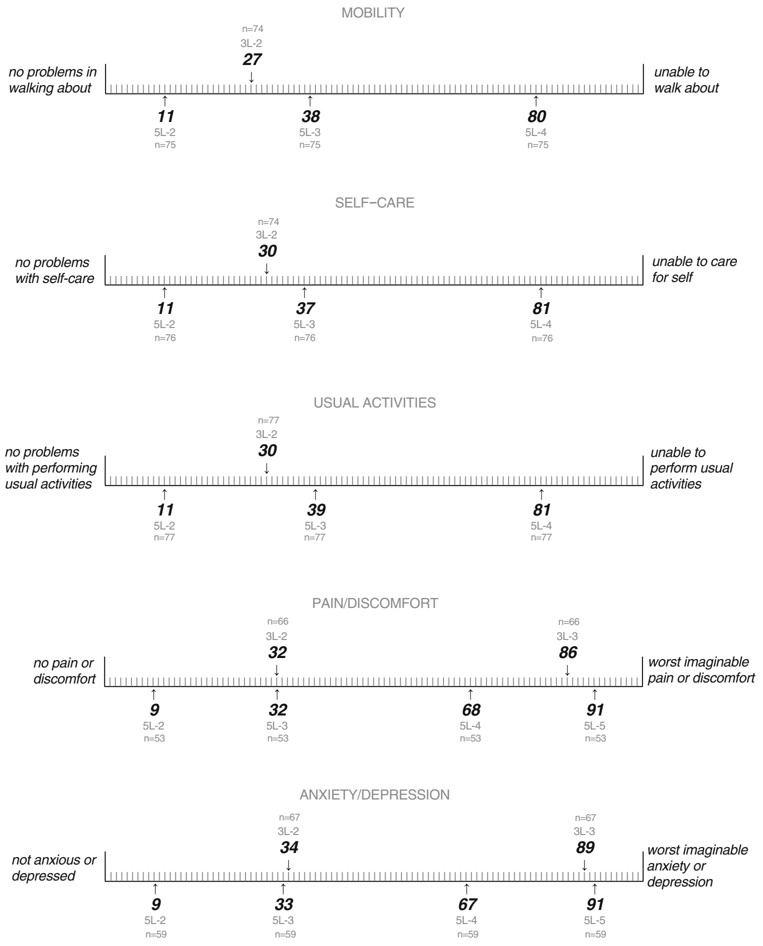


Table 2 Isoformity of identical three-and five-level (3L, 5L) descriptors for the direct quantification method

Dimension	Comparison		Mean difference	*P* value
Mobility	3L-2	5L-3	−11.4	<0.001
Self-care	3L-2	5L-3	−8.0	0.002
Usual activities	3L-2	5L-3	−9.4	<0.001
Pain/Discomfort	3L-2	5L-3	−1.4	0.501
Pain/Discomfort	3L-3	5L-5	−4.9	0.012
Anxiety/Depression	3L-2	5L-3	2.8	0.276
Anxiety/Depression	3L-3	5L-5	−3.0	0.025

(3L-1 = 13; 5L-1 = 8). VAS results for the upper extreme values are systematically higher for 5L than for 3L (range of difference: 6–10). Noticeable are large deviations in Self-Care (3L-3 = 85; 5L-5 = 91) and Usual Activities (3L-3 = 89). Isoformity was accepted for 3L-1 versus 5L-1 for all dimensions and for 3L-2 vs. 5L-3 for all dimensions except Mobility (showing a gap of 7 points). Isoformity was rejected for the upper extreme comparison 3L-3 versus 5L-5 for all dimensions. Consistency between dimensions gave supportive results for both 3L and 5L. Table 4 shows the G-study results. Most variance is attributed to the label component, whereas less than 2% of variance is attributed to the components including dimension, which is reflected in high G-coefficients for all comparisons. Consistency for 5L is somewhat higher (0.87; 0.86) than for 3L (0.86; 0.81).

Discussion

In this study, we compared the quantitative position of the level descriptors of the standard EQ-5D3L and a new five-level version using two independent methods. The study showed that the extension of the EQ-5D3L to a five-level version by inserting two extra levels, leaving the existing descriptors unaltered, is not a simple refinement but a redesign. The inserted levels pushed the extreme levels closer to the anchors, which indicates that 5L makes better use of the measurement continuum, contributing to superior descriptive power of the 5L version. In both the 3L and 5L versions, the position of the 3L or 5L descriptors, reassuringly, was independent of dimension.

Equidistance was not achieved for both systems, in most cases showing values lower than the equidistant values.

Table 3 Indirect quantification of three- and five-level (3L, 5L) descriptors

	Number	Mean	Median	CI
3L				
Mobility				
No problems in walking about	599	1.69	0	1.31–2.07
Some problems in walking about	403	42.94	40	40.24–45.64
Unable to walk about	180	91.70	99	89.28–94.12
Self-care				
No problems with self-care	482	3.24	0	2.40–4.08
Some problems washing or dressing self	435	39.18	34	36.58–41.78
Unable to wash or dress self	273	85.47	95	82.77–88.16
Usual activities				
No problems with performing usual activities	235	4.50	2	3.49–5.51
Some problems with performing usual activities	582	36.55	30	34.40–38.71
Unable to perform usual activities	378	88.54	95	86.87–90.22
Pain/Discomfort				
No pain or discomfort	246	12.64	4	9.94–15.34
Moderate pain or discomfort	643	35.76	31	33.92–37.60
Extreme pain or discomfort	275	83.21	89	80.82–85.61
Anxiety/Depression				
Not anxious or depressed	433	6.29	1	5.01–7.57
Moderately anxious or depressed	478	42.45	40	40.26–44.63
Extremely anxious or depressed	270	84.80	90	82.73–86.86
5L				
Mobility				
No problems in walking about	547	1.30	0	0.92–1.69
Mild problems in walking about	147	15.33	11	12.64–18.02
Some problems in walking about	159	36.48	31	33.00–39.97
Many problems in walking about	217	69.82	76	66.72–72.92
Unable to walk about	112	97.36	100	95.24–99.48
Self-care				
No problems with self-care	398	2.45	0	1.43–3.48
Mild problems washing or dressing self	204	12.70	9	10.76–14.64
Some problems washing or dressing self	184	36.09	33	33.00–39.17
Many problems washing or dressing self	257	71.20	78	68.33–74.06
Unable to wash or dress self	147	91.37	99	87.80–94.94
Usual activities				
No problems with performing usual activities	136	3.22	0	1.49–4.95
Mild problems with performing usual activities	268	12.39	9	10.68–14.10
Some problems with performing usual activities	228	32.53	30	29.97–35.09
Many problems with performing usual activities	351	69.54	75	67.18–71.90
Unable to perform usual activities	212	95.35	100	93.74–96.96
Pain/Discomfort				
No pain or discomfort	145	8.34	0	5.32–11.37
Mild pain or discomfort	274	17.27	12	15.13–19.41
Moderate pain or discomfort	367	36.83	35	34.91–38.76
Severe pain or discomfort	263	71.72	79	69.05–74.39
Extreme pain or discomfort	115	92.76	98	89.86–95.65
Anxiety/Depression				
Not anxious or depressed	305	4.75	0	3.07–6.43

Table 3 continued

	Number	Mean	Median	CI
A little anxious or depressed	241	16.48	10	14.33–18.63
Moderately anxious or depressed	271	41.98	41	39.72–44.25
Very anxious or depressed	248	74.19	80	71.69–76.70
Extremely anxious or depressed	116	92.33	97	89.61–95.04

CI confidence interval

Both methods revealed a large gap between the 5L-3 and 5L-4 levels, regardless of dimension. This could be caused by the wording of 5L-3 [some and moderate(ly)] being interpreted as fairly mild.

In Pain/Discomfort, respondents tended to avoid the lower anchor of the scale, indicating some pain or discomfort on VAS while scoring no problems on 3L and 5L. This indicates that respondents preferred a more refined response scale for scoring pain or discomfort, maybe a scale with even more than five response options (as is the case of, e.g., the HUI3 or SF-36). Also noticeable were the gaps observed for the upper extreme in Self-Care, for which we cannot provide an explanation.

Isoformity between 3L and 5L showed mixed results. The 3L-1 vs. 5L-1 descriptors showed isoformity (indirect method only), as expected, as these both indicated the upper ceiling (no problems). Isoformity was also established for the middle level descriptors of Pain/Discomfort and Anxiety/Depression for both methods. This could be due to the wording of the middle level descriptors, as the descriptor some problems represented a wider range and hence more potential variation, than moderate(ly), as used in Pain/Discomfort and Anxiety/Depression. Assuming that the descriptor some problems was a well-considered choice in the development of the original EQ-5D3L system in order to cover the entire range between the two extremes, it is questionable whether that descriptor is still suitable in a 5L version.

Direct quantification is a well-known method of estimating the magnitude of level descriptors or response

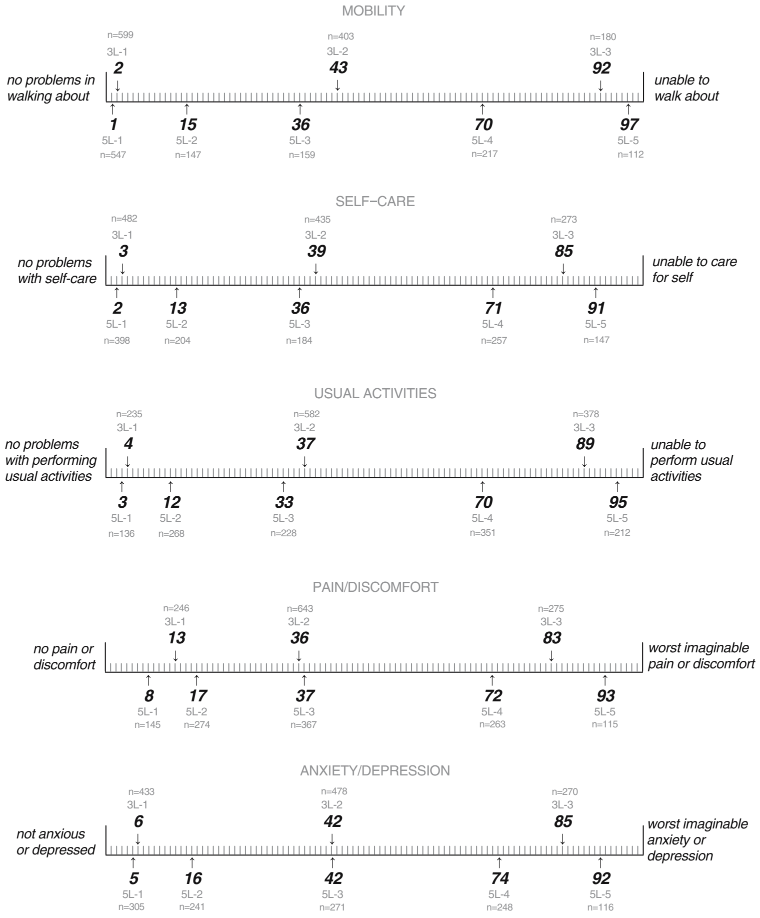


Fig. 3 Indirect quantification of the three- and five-level (3L, 5L) descriptors. Visual analog scale (VAS) means by dimension

Table 4 Consistency between dimensions for the indirect quantification method. Variance components estimates (percentages) and generalizability coefficients (G-coefficients) for comparable dimensions of three- and five-level (3L, 5L) instruments

3L		5L	
Mobility/Self-care/Usual activities			
Label	66.12	Label	71.52
Vignette	8.05	Vignette	6.35
Dimension	0.26	Dimension	0.04
Respondent	0.33	Respondent	0.79
Label × vignette	5.60	Label × vignette	2.91
Label × dimension	0.22	Label × dimension	0.12
Label × respondent	2.20	Label × respondent	2.59
Vignette × dimension	0.60	Vignette × dimension	0.17
Vignette × respondent	3.77	Vignette × respondent	2.57
Dimension × respondent	0.76	Dimension × respondent	0.60
Residual	12.09	Residual	12.34
G-coefficient	0.86	G-coefficient	0.87
Pain/Discomfort; Anxiety/Depression			
Label	65.25	Label	73.58
Vignette	4.95	Vignette	2.73
Dimension	0.00	Dimension	0.00
Respondent	0.65	Respondent	0.77
Label × vignette	1.91	Label × vignette	1.02
Label × dimension	0.04	Label × dimension	0.00
Label × respondent	2.96	Label × respondent	3.36
Vignette × dimension	1.06	Vignette × dimension	0.17
Vignette × respondent	5.52	Vignette × respondent	4.50
Dimension × respondent	0.88	Dimension × respondent	0.45
Residual	16.78	Residual	13.42
G-coefficient	0.81	G-coefficient	0.86

labels [16, 17, 28, 29]. This approach, however, ignores the fact that the VAS values expressed for the level descriptors did not necessarily reflect the self-report use of such descriptors (and the use in subsequent valuation studies) in a similar way, because the valuation of an abstract level descriptor might lead to different results than self-reported health. The indirect method is novel: to our knowledge, this is the first time a quantification of level descriptors is estimated with this method. The indirect method has several advantages. First, we believe it is a better representation of the hypothesized measurement continuum of EQ-5D, as the medium of the vignette (disease) was used to calibrate 3L and 5L descriptors on a VAS scale. Second, it is closer to the general use of the EQ-5D instrument as a self-report health status assessment measure and is therefore likely to be more valid. Classifying a vignette can be regarded similarly to a health status classification by proxy assessment. Other advantages of the indirect method are analytical: values can be calculated for all level descriptors, including the anchors, and it is possible to investigate explained variance for various components (G-study). Furthermore, the indirect method proved to be much more feasible than the direct method, considering the lower number of missing responses. Disadvantages are that no direct comparison (e.g., paired *t* test) between 3L and 5L is possible, as there is only one VAS value for each 3L–5L response pair, and that the indirect method is more time consuming.

A potential weakness of the study procedure is that 3L and 5L were presented on one sheet, and panelists were asked to score 5L dimensions first while covering 3L and vice versa. We cannot be sure that respondents actually complied to the blinding procedure in the follow-up measurement. Also, there might have been an order effect, as 5L always preceded 3L.

The 5L instrument presented here obviously improves the discriminatory potential of the EQ-5D descriptive system, as the level descriptors generally capture a larger part of the measurement continuum and broaden the measurement space. Furthermore, 5L showed slightly better consistency between levels. In a previous study, we demonstrated increased discriminatory power of the same 5L version of EQ-5D, as well as superior reliability (interobserver and test–retest) and face validity when compared with the standard EQ-5D3L [18]. Awaiting a valuation study for an official version of 5L, a set of preference weights was developed for this 5L version of EQ-5D using item response theory (IRT) methodology [30]. An officially sanctioned five-level descriptive system will become available within a short period [31] and is expected to be in use alongside the standard three-level EQ-5D.

The experimental five-level EQ-5D version presented here is likely to demonstrate a less severe ceiling effect. Assuming that milder states are more common in the general population, we expect increased benefit in the detection of mild problems and in measuring and monitoring general population health, although the extra 5L-4 level is expected to also lead to better differentiation and detection of more severe health states. The methodology presented here can be of use in the development of generic or disease-specific health status measures.

Acknowledgments This research was funded by the EuroQol Group. An earlier draft of this paper was presented and discussed at the 2006 Annual EuroQol Meeting in Barcelona.

Appendix A: Formula used to estimate G-coefficients

Consistency between dimensions

$$= \frac{\sigma^2_{label} + \sigma^2_{vignette} + \sigma^2_{respondent} + \sigma^2_{label\times vignette} + \sigma^2_{label\times respondent} + \sigma^2_{vignette\times respondent}}{\sigma^2_{total}}$$

References

1. Brooks, R., Rabin, R. E., & de Charro, F. Th. (2003). *The measurement and valuation of health status using EQ-5D: A European perspective*. Dordrecht: Kluwer academic publishers.
2. Feeny, D., Furlong, W., & Torrance, G. (1999). The health utilities index: An update. *Quality of Life Newsletter, 22*, 8–9.
3. Brazier, J., Deverill, M., Green, C., Harper, R., & Booth, A. (1999). A review of the use of health status measures in economic evaluation. *Health Technology Assessment, 3*, 1–164.
4. Brazier, J., Roberts, J., & Deverill, M. (2002). The estimation of a preference-based measure of health from the SF-36. *Journal of Health Economics, 21*, 271–292.
5. Kopec, J. A., & Willison, K. D. (2003). A comparative review of four preference-weighted measures of health-related quality of life. *Journal of Clinical Epidemiology, 56*, 317–325.
6. Macran, S., Weatherly, H., & Kind, P. (2003). Measuring population health—a comparison of three generic health status measures. *Medical Care, 41*, 218–231.
7. Wu, A. W., Jacobson, K. L., Frick, K. D., Clark, R., Revicki, D. A., Freedberg, K. A., Scott-Lennox, J., & Feinberg, J. (2002). Validity and responsiveness of the EuroQol as a measure of health-related quality of life in people enrolled in an AIDS clinical trial. *Quality of Life Research, 11*, 273–282.
8. Myers, C., & Wilks, D. (1999). Comparison of Euroqol EQ-5D and SF-36 in patients with chronic fatigue syndrome. *Quality of Life Research, 8*, 9–16.
9. Willige van de, G., Wiersma, D., Nienhuis, F. J., & Jenner, J. A. (2005). Changes in quality of life in chronic psychiatric patients: a comparison between EuroQol (EQ-5D) and WHOQoL. *Quality of Life Research, 14*, 441–451.
10. Sullivan, P. W., Lawrence, W. F., & Ghushchyan, V. (2005). A national catalog of preference-based scores for chronic conditions in the United States. *Medical Care, 43*, 736–749.
11. Houle, C., Bertheloth, C. M., & Health Analysis, Modeling Group (2000). Head-to-head comparison of the health utilities index mark 3 and the EQ-5D for the population living in private households in Canada. *Quality of Life Newsletter, 24*, 5–6.
12. Badia, X., Schiaffino, A., Alonso, J., & Herdman, M. (1998). Using the EuroQol 5-D in the Catalan general population: Feasibility and construct validity. *Quality of Life Research, 7*, 311–322.
13. Wang, H., Kindig, D. A., & Mullahy, J. (2005). Variation in Chinese population health related quality of life: Results from a EuroQol study in Beijing, China. *Quality of Life Research, 14*, 119–132.
14. Brazier, J., Roberts, J., Tsuchiya, A., & Busschbach, J. (2004). A comparison of the EQ-5D and SF-6D across seven patient groups. *Health Economics, 13*, 873–884.
15. Kaarlola, A., Pettila, V., & Kekki, P. (2004). Performance of two measures of general health-related quality of life, the EQ-5D and the RAND-36 among critically ill patients. *Intensive Care Medicine, 30*, 2245–2252.
16. Szabo, S. (1996). World Health Organization Quality of Life (WHOQOL) assessment instrument. In B. Spilker (Ed.), *Quality of life and of life and pharmaeconomics in clinical trials* (pp. 355–362). Philadelphia: Lippincott-Raven.
17. Keller, S. D., Ware, J. E., Jr., Gandek, B., Aaronson, N. K., Alonso, J., Apolone, G., Bjorner, J. B., Brazier, J., Bullinger, M., Fukuhara, S., Kaasa, S., Leplege, A., Sanson-Fisher, R. W., Sullivan, M., & Wood-Dauphinee, S. (1998). Testing the equivalence of translations of widely used response choice labels: results from the IQOLA Project. International Quality of Life Assessment. *Journal of Clinical Epidemiology, 51*, 933–944.
18. Janssen, M. F., Birnie, E., Haagsma, J. A., & Bonsel, G. J. (2008). Comparing the standard EQ–5D three level system with a five level version. *Value in Health*. doi:10.1111/j.1524-4733.2007.00230.x
19. Pickard, A. S., De Leon, M. C., Kohlmann, T., Cella, D., & Rosenbloom, S. (2007). Psychometric comparison of the standard EQ-5D to a 5 level version in cancer patients. *Medical Care, 45*, 259–263.
20. Bonsel, G. J., & van Agt, H. M. E. (1993). The number of levels in the descriptive system. In J. J. van Busschbach, G. J. Bonsel, & F. Th. de Charro (Eds.), *Book of EuroQol meeting proceedings* (pp. 115–120). Rotterdam: Erasmus University Rotterdam.
21. Lissitz, R. W., & Green, S. B. (1975). Effect of the number of scale points on reliability: A Monte Carlo approach. *The Journal of Applied Psychology, 60*, 10–13.
22. Nishisato, S., & Torii, Y. (1970). Effects of categorizing continuous normal variables on the product-moment correlation. *The Japanese Psychological Research, 13*, 45–49.
23. Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica, 104*, 1–15.
24. Miller, G. A. (1956). The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63*, 81–97.
25. Streiner, D. L., & Norman, G. W. (2003). *Health measurement scales: A practical guide to their development and use*. Oxford: Oxford University Press.
26. Cardinet, J., Tourneur, Y., & Allal, L. (1981). Extension of generalizability theory and its application in educational measurement. *Journal of Educational Measurement, 4*, 183–204.
27. Krabbe, P. F., Essink-Bot, M. L., & Bonsel, G. J. (1996). On the equivalence of collectively and individually collected responses: standard-gamble and time-tradeoff judgments of health states. *Medical Decision Making, 16*, 120–132.
28. Bass, B. M., Cascio, W. F., & O'Connor, E. J. (1974). Magnitude estimations of expressions of frequency and amount. *The Journal of Applied Psychology, 59*, 313–320.
29. Spector, P. E. (1976). Choosing response categories for summated rating scales. *The Journal of Applied Psychology, 61*, 374–375.
30. Pickard, A. S., Kohlmann, T., Cella, D., Rosenbloom, S., Bonsel, G. J., & Janssen, M. F. (2007). Come together: Use of IRT models to derive preference-based algorithms for a 5 level version of the EQ-5D. *Medical Care, 45*, 259–263.
31. Kind, P. (2007). Size matters: EQ-5D in transition. *Medical Care, 45*, 809–811.