


ScienceDirect

Contents lists available at **sciencedirect.com**
Journal homepage: **www.elsevier.com/locate/jval**

Notions of "Value" in Healthcare

Exploring the Internal Structure of the EQ-5D Using Non–Preference-Based Methods

You-Shan Feng, PhD,[1,*] Ruixuan Jiang, PharmD,[2] Thomas Kohlmann, PhD,[1] A. Simon Pickard, PhD[2]

[1]*Institute for Community Medicine, University of Greifswald, Greifswald, Germany;* [2]*University of Illinois at Chicago College of Pharmacy, Chicago, IL, USA.*

ABSTRACT

Background: When the EuroQol EQ-5D is applied in settings other than resource allocation, a non–preference-based score may be more appropriate than societal, preference-weighted utility. To develop a psychometric score for the EQ-5D, its structural relationship, ie, how the 5 items/dimensions interrelate, must be understood to inform appropriate methods of summarizing the instrument.

Objectives: To explore psychometrically derived approaches of elucidating the 5-level EQ-5D (EQ-5D-5L) item structure.

Methods: Three measurement models were assessed. All 5 items were modeled as reflective indicators using confirmatory factor analysis. EQ-5D-5L items were conceptualized as formative indicators, and other health scales (eg, the short form 36 health survey) were conceptualized as reflective indicators in Multiple Indicators Multiple Causes models (external MIMIC). The EQ-5D-5L items were modeled as a combination of formative and reflective indicators in internal MIMIC models. Results across 9 data sets from various countries and patient groups were examined to determine their robustness.

Results: All items loaded well (0.63-0.96) in the confirmatory factor analysis except for anxiety/depression (0.20-0.66, excluding 1 outlier). The model fit statistics of the external MIMIC models were poor, and the coefficients of the Self-Care dimension were small. The internal MIMIC model with Mobility, Pain/Discomfort, and Anxiety/Depression as formative indicators and Self-Care and Usual Activities as reflective indicators fit best. The model results of the Spanish valuation data set were outliers.

Conclusions: Although there were some variations in results across subgroups, the relationship between the items remained robust. The evidence calls for testing of formative/reflective combination approaches to summarize the EQ-5D-5L.

Keywords: EQ-5D-5L, formative measurement model, Multiple Indicators Multiple Causes (MIMIC), noneconomic scoring approaches, reflective measurement model.

VALUE HEALTH. 2019; 22(5):527–536

Introduction

The EQ-5D is a health status questionnaire that covers 5 dimensions of health—Mobility (MO), Self-Care (SC), Usual Activities (UA), Pain/Discomfort (PD), and Anxiety/Depression (AD)—using 1 item per dimension.[1] The instrument also includes a visual analogue scale (VAS) anchored by 0 (worst imaginable health) and 100 (best imaginable health). The original 3-level version (EQ-5D-3L) was expanded to 5 levels of severity (EQ-5D-5L) to minimize ceiling effects and increase sensitivity.[2,3]

Although the EQ-5D rose to prominence as a preference-based measure of health for use in cost-utility analyses, it is currently applied in an array of broader settings.[4–8] These include measurement of individual health status in clinical practice, population health surveillance, and assessment of healthcare quality, as well as aiding medical decision making and patient communication.[4,6,8–10] EQ-5D use continues to increase in prevalence and importance as healthcare stakeholders around the world find relevant applications for the EQ-5D and other measures of health-related quality of life (HRQOL).[8]

Three methods of presenting EQ-5D responses are recommended in the EuroQol Group's user guide: (1) calculation of index-based scores using preference-based utility weights, (2) summarization of responses descriptively by using counts and percentages, and (3) use of the VAS.[1,11] Nonutility methods of summarizing the EQ-5D, that is, methods 2 and 3, can be used when no utility weights are available (eg, EQ-5D-Y). Descriptive summarization evaluates each dimension

* Address correspondence to: You-Shan Feng, PhD, Methods Department, Institute for Community Medicine, Walther-Rathenau-Str. 48, Greifswald D-17475, Germany. Email: you-shan.feng@uni-greifswald.de

https://doi.org/10.1016/j.jval.2019.02.006

separately and does not leverage the full information provided by the EQ-5D, thereby limiting statistical inferences regarding the overall health state. Analysis of the 5-digit string (eg, 11111 [no problems on any dimension]) is possible but difficult to aggregate and interpret; furthermore, classification of change in overall health is challenging when some dimensions improve while others worsen.[4] Thus, neither descriptive use of individual health dimensions nor the 5-digit EQ-5D-5L string is commonly used for statistical inference. Although the VAS appears to overcome these hurdles, it does not summarize the EQ-5D items because it measures health more broadly.[12,13]

Using a value set to score the EQ-5D-5L appeals to end users for simplicity and familiarity. Nevertheless, applying preference-based scoring when the EQ-5D-5L is used as a patient-reported outcome or measure of population health may not be theoretically justified and can introduce additional statistical biases.[6,14] Preferences captured in value sets are elicited from the general population and represent choices made in resource (eg, tax revenue) allocation decisions in healthcare. Applying a societal value set to health in clinical or population settings fails to inform the respondent's level of HRQOL (latent construct of interest) because only the general population's preference for the health state on a dead to full health (0-1) scale is reflected by the utility weight. Because societal preferences lack proven relevance outside of economic evaluations of healthcare interventions, nonpreference, psychometric scoring approaches should be considered with respect to the way the EQ-5D is presented, described, analyzed, and interpreted to potentially better measure HRQOL in non–preference-based ways.

To generate psychometric scoring of the EQ-5D, an understanding of its internal structure, that is, how the 5 items/dimensions are related to each other, is essential. The interrelationship of the EQ-5D items has not been thoroughly addressed in the literature.[15,16] This fundamental knowledge would inform appropriate grouping and summarization of the EQ-5D dimensions to develop valid psychometric weights.

Two prominent approaches to describe the structure of a set of items are the "reflective indicators model" and the "formative indicators model."[17,18] The reflective model is the classic psychometric perspective of health measurement and is illustrated in path diagrams with the arrows directed from the latent trait to the indicators (HRQOL → reflective indicators). The reflective indicators are interrelated as manifestations of the latent construct, meaning changes in the underlying construct are reflected by changes in the observable indicators. HRQOL instruments such as the short form 36 health survey (SF-36) and the Patient-Reported Outcomes Measurement Information System[19–21] are constructed on the basis of the reflective model, using classical test theory or item response theory to develop scoring algorithms.

The formative measurement model estimates a latent construct using linear combinations of observable indicators to create a composite measure.[17,18] A change in the latent construct is a result of changes to formative indicators, depicted in path diagrams with the arrows directed from the indicators to the latent trait (formative indicators → HRQOL). Formative models have roots in sociology where complex ideas, such as socioeconomic status, cannot easily be captured using a single item. The EQ-5D instrument was intended as a formative measure, using 5 dimensions to create a composite measure of health status (ie, the EQ-5D health state). Variations on the classic measurement models are also possible using combinations of formative and reflective indicators.[16]

An obvious need has grown for the application of psychometric scoring methods for the EQ-5D to keep pace with expanding HRQOL research.[8] To date, non–preference-based methods have been investigated only to a limited extent,[4,22–24] and none were based on EQ-5D internal structure findings. Because future development of psychometrically based scoring methods should be based on the underlying structural relationship of the EQ-5D items, we used reflective and formative approaches to model the EQ-5D and understand the relationship of the items with each other. The aims of this study were (1) to examine the structural relationships of the EQ-5D-5L items on the basis of reflective and/or formative frameworks and (2) to examine the robustness of these models in different patient groups and data sets.

Methods

Data Sets

To assess item structure for the EQ-5D-5L, we conducted secondary data analysis on 9 data sets, characterized as general population data (n = 4), patient data (n = 3), and patient and healthy population mix (n = 2). Diverse data sets were chosen to evaluate the robustness of results across respondent types, countries, study designs, and modes of administration of the EQ-5D-5L questionnaire. Subgroup analyses were conducted for each data set and 9 disease groups were determined by the information available (respiratory diseases, musculoskeletal diseases, heart disease, depression/mental health, stroke, diabetes, cancer, liver disease, and kidney disease).

Population Data Sets

In valuation studies, general population respondents self-described their health using the EQ-5D-5L descriptive system at the time of the valuation interview. For these analyses, we used valuation data sets from Indonesia,[25] Spain,[25,26] Canada,[27] and the United States.[28]

Patient Data Sets

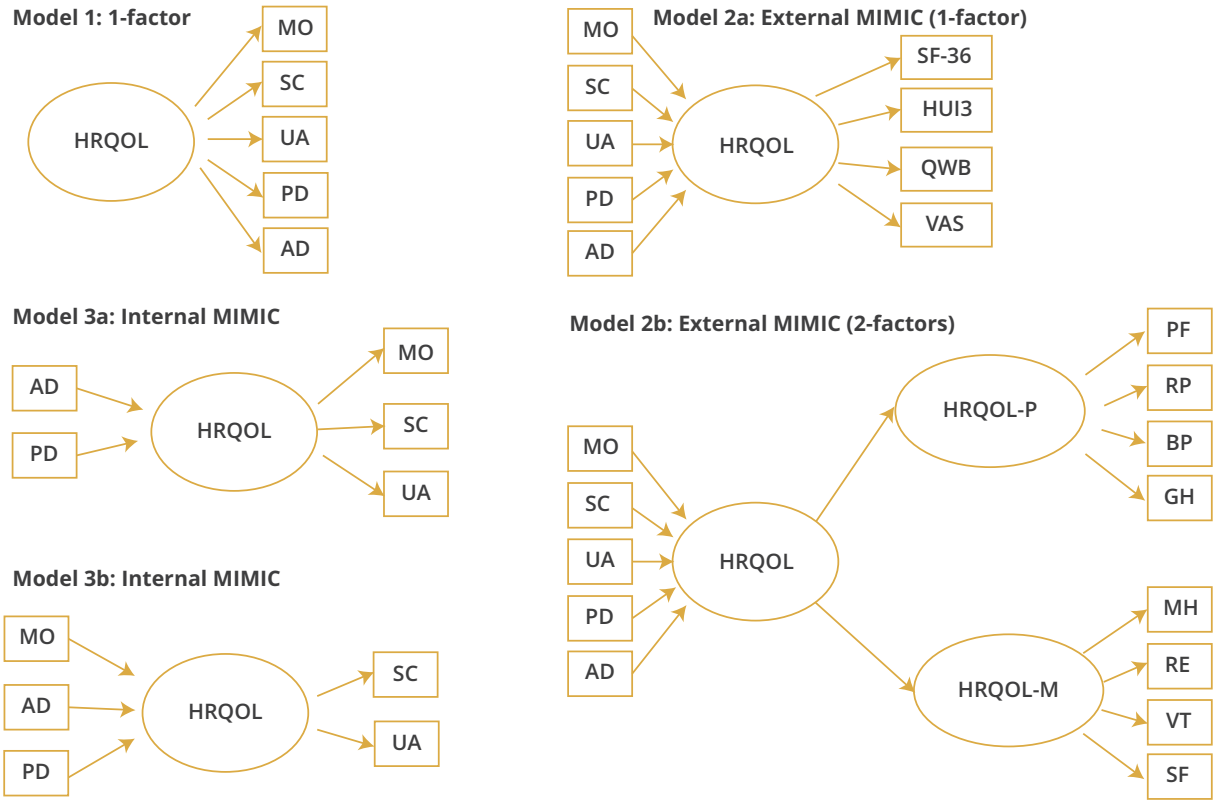
Baseline data from several patient longitudinal data sets were used. The first data set was a German study that sought to validate the Veteran's Rand 36 and 12 instruments against the SF-36 and SF-12 (VR validation study) at the beginning and the end of physical or mental rehabilitation. A separate group of German physical and psychosomatic rehabilitation patients was sampled to conduct a head-to-head comparison study of the EQ-5D-3L and the EQ-5D-5L[29] at the beginning and the end of their rehabilitation (German 3L/5L study).

The Alberta PROMs and EQ-5D Research Support Unit (APERSU)[5] used the EQ-5D-5L in the Alberta Caring for Diabetes Study.[30] More than 2000 patients with type 2 diabetes completed the EQ-5D-5L and other self-assessed health instruments, and provided information about self-reported behaviors, symptoms, and healthcare.

General Population/Patient Mix

Two data sets had a mix of general population and patient respondents: (1) a large data set that included patient samples from 7 countries, which was the basis for developing interim, preference-based scoring for the 5L (crosswalk data set),[3] and (2) the data set from the international Multi-Instrument Comparison (MIC) study that collected data encompassing a range of preference and subjective well-being instruments as well as various disease states from more than 8000 respondents representing 6 countries using online surveys.[31]

Figure 1. Examined item structures.



AD indicates Anxiety/Depression; BP, bodily pain; GH, general health perceptions; HRQOL, health-related quality of life; HRQOL-P, physical health-related quality of life; HRQOL-M, mental health-related quality of life; HUI3, Health Utility Index; MH, mental health; MIMIC, multiple indicators, mutliple causes; MO, Mobility; PD, Pain/Discomfort; PF, Physical Functioning; QWB, Quality of Well Being Scale; RE, Role Emotional; RP, Role Physical; SC, Self-Care; SF, social role functioning; UA, Usual Activities; VAS, Visual Analogue Scale; VT, vitality.

Analysis

Measurement models

To examine the underlying structure of the EQ-5D-5L, we tested (1) reflective (model 1), (2) formative (models 2a and 2b), and (3) reflective and formative combination (models 3a and 3b) approaches to conceptualizing the EQ-5D-5L measurement model, as summarized in Figure 1. Factor and regression structures were allowed to vary across data sets and respondent groups. Data cleaning and organization were conducted using SAS 9.4 (SAS Institute, Cary, NC),[32] STATA 14.0 (StataCorp, College Station, TX),[33] and SPSS 22 (IBM, Armonk, NY),[34] and analyses were conducted using M-Plus 8.0 (Muthén & Muthén, Los Angeles, CA).[35] Although many approaches can be used to understand the relationship between the EQ-5D items, we focused on 3 different approaches in this article that best facilitate understanding the extent to which the items are interrelated and the dimensionality of the EQ-5D as an operationalization of HRQOL.

Reflective approach. We tested a purely reflective model: whether 5 items of the EQ-5D-5L were reflective indicators of a single latent construct (HRQOL → EQ-5D-5L items) and therefore be summarized as a single score (model 1) using confirmatory factor analysis (CFA). Although the EQ-5D-5L may contain more than 1 latent factor, we tested, as a baseline assessment, whether its 5 items can be summarized as a single reflective score. EQ-5D items were modeled continuously (maximum likelihood) as well as categorically (robust mean- and variance-adjusted weighted least square).

Formative approach. To explore the EQ-5D-5L using a purely formative approach, we examined Multiple Indicators Multiple Causes (MIMIC) models—a specific type of structural equation model comprising (1) a measurement model that accounts for the reflective factors addressing 1 or more continuous latent factor(s) and (2) a regression model that estimates the association between the formative variables and the latent factor(s).[36] For this approach, external, established HRQOL measures were used to assess how the EQ-5D-5L items may be composites of HRQOL. The inclusion of other health scales in this MIMIC model allowed for better exploration and triangulation of EQ-5D item structure; future scoring algorithms need not depend on non–EQ-5D measures. The MIC data, because of the richness of included measures, were well suited for this purpose. All EQ-5D-5L items were conceptualized as composite indicators (ie, regression covariates) and self-assessed health scales as reflective indicators (EQ-5D-5L items → HRQOL → self-assessed health scales). Because these MIMIC models used HRQOL measures other than the 5 items of the EQ-5D-5L, we defined them as "external" MIMIC models.

Two external MIMIC models were fitted: (1) HRQOL as a single latent construct (model 2a) and (2) HRQOL as a second-order factor made up of physical and mental health factors (model 2b). We first fitted external MIMIC using the 8 SF-36 subscales because it is

Table 1A. Data set characteristics.

			Indonesian 5L valuation, N (%)	Spanish 5L valuation, N (%)	Canadian 5L valuation, N (%)	US 5L valuation, N (%)	MIC data set, N (%)
		Total N	1055	1000	1210	1134	8022
		Women	526 (49.91)	525 (52.50)	664 (54.88)	565 (49.78)	4174 (52.03)
		Men	528 (50.09)	475 (47.50)	546 (45.12)	565 (49.78)	3848 (47.97)
		Age (y), mean ± SD	35.81 ± 13.65	43.78 ± 17.27	47.59 ± 17.38	46.93 ± 18.11	44.34 ± 21.74
		EQ-5D VAS, mean ± SD	79.39 ± 14.01	82.33 ± 14.54	82.3 ± 14.22	80.08 ± 16.22	67.13 ± 21.67
		EQ-5D-5L full health (11111)	464 (44.02)	580 (58.00)	401 (33.14)	346 (30.48)	1531 (19.09)
		Worst health (55555)	0 (0.00)	0 (0.00)	1 (0.08)	2 (0.18)	0 (0.00)
EQ-5D-5L dimension	Mobility	No problems	970 (92.03)	886 (88.60)	919 (75.95)	795 (70.04)	5337 (66.53)
		Slight problems	71 (6.74)	73 (7.30)	204 (16.86)	209 (18.41)	1491 (18.59)
		Moderate problems	11 (1.04)	33 (3.30)	66 (5.45)	85 (7.49)	824 (10.27)
		Severe problems	2 (0.19)	7 (0.70)	17 (1.40)	33 (2.91)	340 (4.24)
		Unable to	0 (0.00)	1 (0.10)	4 (0.33)	13 (1.15)	30 (0.37)
	Self-Care	No problems	1034 (98.10)	957 (95.70)	1118 (92.40)	1042 (91.81)	7033 (87.67)
		Slight problems	18 (1.71)	32 (3.20)	77 (6.36)	45 (3.96)	646 (8.05)
		Moderate problems	1 (0.09)	10 (1.00)	9 (0.74)	30 (2.64)	273 (3.40)
		Severe problems	1 (0.09)	1 (0.10)	4 (0.33)	10 (0.88)	62 (0.77)
		Unable to	0 (0.00)	0 (0.00)	2 (0.17)	8 (0.70)	8 (0.10)
	Usual Activities	No problems	940 (89.18)	688 (68.80)	903 (74.63)	840 (74.01)	5182 (64.60)
		Slight problems	102 (9.68)	222 (22.20)	208 (17.19)	179 (15.77)	1739 (21.68)
		Moderate problems	12 (1.14)	73 (7.30)	87 (7.19)	85 (7.49)	794 (9.90)
		Severe problems	0 (0.00)	16 (1.60)	11 (0.91)	18 (1.59)	256 (3.19)
		Unable to	0 (0.00)	1 (0.10)	1 (0.08)	48 (4.23)	51 (0.64)
	Pain/ Discomfort	No problems	636 (60.34)	792 (79.20)	569 (47.02)	550 (48.46)	2340 (29.17)
		Slight problems	385 (36.53)	154 (15.40)	465 (38.43)	370 (32.60)	3251 (40.53)
		Moderate problems	27 (2.56)	38 (3.80)	131 (10.83)	153 (13.48)	1619 (20.18)
		Severe problems	6 (0.57)	10 (1.00)	40 (3.31)	45 (3.96)	697 (8.69)
		Extreme problems	0 (0.00)	6 (0.60)	5 (0.41)	17 (1.50)	115 (1.43)
	Anxiety/ Depression	No problems	693 (65.75)	913 (91.30)	751 (62.07)	689 (60.70)	4012 (50.01)
		Slight problems	297 (28.18)	60 (6.00)	322 (26.61)	267 (23.52)	2348 (29.27)
		Moderate problems	58 (5.50)	22 (2.20)	112 (9.26)	135 (11.89)	1107 (13.80)
		Severe problems	4 (0.38)	4 (0.40)	19 (1.57)	26 (2.29)	393 (4.90)
		Extreme problems	2 (0.19)	1 (0.10)	6 (0.50)	18 (1.59)	162 (2.02)

ABCD indicates Alberta Caring for Diabetes; APERSU, Alberta PROMs and EQ-5D Research Support Unit; MIC, Multi-Instrument Comparison; SD, standard deviation; VAS, visual analogue scale; VR, Veteran's Rand.

considered a criterion standard of self-assessed HRQOL (models 2a-1 and 2b-1).[21] Well-regarded subjective health summary scores (Health Utility Index 3 [HUI3], Quality of Well-Being Scale [QWB], and VAS) were then added to model 2a-1 to assess any substantive changes in the model. HUI3 subscales were added to model 2b-1 to assess the robustness of results.[37] SF-36 subscales were used in external MIMIC models; orthogonal scorings of the mental component and physical components were not used. Standardized coefficients of the EQ-5D-5L items were examined to learn about items' relationship with the latent factor(s).

Reflective and formative combination approach. The internal structure of the EQ-5D-5L was further investigated using MIMIC models. Because these models used only the 5 items of the EQ-5D-5L, they were defined as "internal" MIMIC models (EQ-5D-5L item subgroup → HRQOL → EQ-5D-5L item subgroup). Unlike the CFA and the external MIMIC models, internal MIMIC models cannot be used to generate usable scoring weights; they can, however, clarify the internal structure of EQ-5D-5L items, which can then be harnessed using other structural equation modeling methods to generate weights for the EQ-5D.

The evaluated internal structures were based on previous work by Gamst-Klaussen et al[16]; the authors hypothesized certain dimensions of the EQ-5D-5L to be reflective indicators and others to be formative indicators on the basis of previous investigations of the EQ-5D in existing models of HRQOL, such as the Wilson and Cleary model and the International Classification of Functioning, Disability and Health.[38–41] Two structures from Gamst-Klaussen et al[16] were investigated as internal MIMIC models: (1) model 3a: causal indicators, AD, PD; reflective indicators, MO, SC, UA, and (2) model 3b: causal indicators, MO, AD, PD; reflective indicators, SC, UA.

Model Fit

For the CFA and the external MIMIC models, standardized factor loadings and coefficients, respectively, were examined to learn about items' relationship with the latent trait. For all model types, χ^2, root mean square error of approximation (RMSEA), standardized root mean square residual (SRMR), comparative fix index (CFI), and Tucker-Lewis index (TLI) were examined to assess model fit. As guided by the evaluation recommendations reported by Hu et al, RMSEA and SRMR less than 0.08 were considered to indicate good

Table 1B. Data set characteristics.

		APERSU ABCD, N (%)	Crosswalk study, N (%)	German 3L/5L study (baseline), N (%)	VR validation (baseline), N (%)
		2040	3691	230	559
		917 (44.95)	1936 (52.45)	160 (69.57)	182 (32.56)
		1110 (54.41)	1740 (47.14)	70 (30.43)	361 (64.58)
		64.40 ± 10.72	51.46 ± 20.08	56.46 ± 11.45	53.63 ± 10.67
		Not available	63.82 ± 22.74	57.66 ± 16.80	57.80 ± 19.15
		338 (16.57)	537 (14.55)	6 (2.61)	9 (1.61)
		1 (5.00)	13 (0.35)	9 (3.91)	0 (0.00)
EQ-5D-5L dimension	Mobility	965 (47.73)	1828 (49.53)	83 (36.09)	216 (38.64)
		508 (25.12)	676 (18.31)	55 (23.91)	127 (22.72)
		401 (19.83)	612 (16.58)	72 (31.30)	133 (23.79)
		141 (6.97)	418 (11.32)	18 (7.83)	69 (12.34)
		7 (0.35)	140 (3.79)	1 (0.43)	4 (0.72)
	Self-Care	1798 (88.92)	2533 (68.63)	169 (73.48)	409 (73.17)
		151 (7.47)	499 (13.52)	39 (16.96)	85 (15.21)
		61 (3.02)	337 (9.13)	16 (6.96)	41 (7.33)
		9 (0.45)	150 (4.06)	5 (2.17)	12 (2.15)
		3 (0.15)	146 (3.96)	0 (0.00)	(0.00)
	Usual Activities	1074 (53.14)	1443 (39.10)	48 (20.87)	115 (20.57)
		541 (26.77)	836 (22.65)	69 (30.00)	177 (31.66)
		301 (14.89)	703 (19.05)	80 (34.78)	145 (25.94)
		84 (4.16)	420 (11.38)	28 (12.17)	100 (17.89)
		21 (1.04)	255 (6.91)	4 (1.74)	11 (1.97)
	Pain/Discomfort	535 (26.47)	1205 (32.65)	19 (8.26)	37 (6.62)
		830 (41.07)	1074 (29.10)	65 (28.26)	146 (26.12)
		490 (24.25)	880 (23.84)	109 (47.39)	256 (45.80)
		140 (6.93)	409 (11.08)	36 (15.65)	107 (19.14)
		26 (1.29)	92 (2.49)	(0.00)	5 (0.89)
	Anxiety/Depression	1093 (54.08)	1416 (38.36)	82 (35.65)	237 (42.40)
		599 (29.64)	1068 (28.94)	77 (33.48)	147 (26.30)
		275 (13.61)	742 (20.10)	52 (22.61)	98 (17.53)
		46 (2.28)	334 (9.05)	14 (6.09)	57 (10.20)
		8 (0.40)	103 (2.79)	4 (1.74)	9 (1.61)

ABCD indicates Alberta Caring for Diabetes; APERSU, Alberta PROMs and EQ-5D Research Support Unit; MIC, Multi-Instrument Comparison; SD, standard deviation; VAS, visual analogue scale; VR, Veteran's Rand.

fit, whereas CFI and TLI more than 0.9 indicated good fit.[42–44] Evaluation recommendations are important but can be relatively arbitrary.[45] Relative model performance was more important than absolute model fit in this study because the primary aim was to compare measurement models. Because model fit metrics determine fit from different aspects of the model, a comprehensive evaluation of model differences and/or these metrics may be necessary if 2 indices report conflicting information.[45]

Results

The individual data sets had between 230 and 8022 respondents. Across all data sets, 26 521 respondents (Table 1) and 25 types of diseases/conditions from 13 countries were captured. Relatively healthier populations (valuation studies) were sometimes missing endorsements of level 4 (severe problems) and/or level 5 (extreme problems) of the level 5-EQ-5D.

We tested model 1 using CFA; all EQ-5D-5L items were treated as reflective indicators of a single latent trait. Modeling the EQ-5D item responses as categorical variables (RMSEA 0.049-0.225, good to mediocre fit; Table 2) demonstrated a better fit than modeling them as continuous variables (RMSEA 0.080-0.225, mostly poor fitting; data not shown). CFI and TLI were mostly higher than 0.9, indicating acceptable fit (Table 2).

Model 1 was robust across nearly all subgroups and data sets, with dimension factor loadings ranging from 0.63 to 0.96 except for AD (standard loadings 0.20-0.66). AD loading was overall moderate (0.527), weak in 1 data set (0.203; German 3L/5L study), and strong in 3 data sets (>0.6; stroke, Spanish valuation, VR validation). AD had smaller loadings than did other dimensions even for patients with depression and mental health concerns. In contrast to other data sets, the Spanish valuation data showed the highest factor loading on AD and low loadings for UA and PD. Furthermore, neither a 1-factor CFA without AD nor a second-order CFA with AD on one factor and non-AD dimensions on a second factor improved model fit (data not shown).

Only the external MIMIC results in the MIC data set were presented because it was the only data set that could support the entire range of analyses (Table 3). The external MIMIC models had poor absolute model fit, but they can be compared with each other in relative terms to glean information about the EQ-5D item structure. When only the SF-36 subscales were modeled as reflective indicators, the 2-factor external MIMIC (model 2b-1) performed marginally better than the 1-factor external MIMIC (model 2a-1) (Table 3).

The addition of more reflective indicators (HUI3, VAS, and QWB scores) (model 2a-2) did not substantially improve model fit over model 2a-1 (Table 3). Similarly, the combination of HUI3 and

Table 2A. Model 1 CFA (1-factor solution).

Group	Mobility		Self-Care		Usual Activities	
	Factor loading	SE	Factor loading	SE	Factor loading	SE
Overall (all data)	0.906	0.003	0.899	0.004	0.909	0.003
Sex						
Women	0.903	0.004	0.899	0.006	0.903	0.005
Men	0.912	0.005	0.900	0.006	0.917	0.004
Illness status						
With 1+ self-reported illness	0.888	0.004	0.888	0.005	0.904	0.004
Healthy/general population	0.882	0.011	0.902	0.019	0.808	0.014
Disease group						
Respiratory diseases	0.932	0.010	0.913	0.010	0.944	0.010
Musculoskeletal disease	0.859	0.010	0.818	0.014	0.913	0.009
Heart disease	0.909	0.012	0.869	0.015	0.923	0.011
Depression/mental health	0.852	0.017	0.829	0.021	0.826	0.015
Stroke	0.903	0.012	0.957	0.009	0.947	0.010
Diabetes	0.881	0.009	0.829	0.015	0.903	0.009
Cancer	0.909	0.015	0.827	0.029	0.913	0.015
Liver disease	0.867	0.030	0.870	0.034	0.915	0.019
Kidney disease	0.716	0.073	0.921	0.092	0.674	0.100
Data set						
Indonesian valuation	0.712	0.043	0.791	0.063	0.632	0.044
Spanish valuation	0.888	0.019	0.915	0.024	0.771	0.021
Canadian valuation	0.888	0.014	0.879	0.022	0.921	0.012
US Valuation	0.845	0.016	0.895	0.021	0.872	0.015
MIC data set	0.908	0.004	0.856	0.007	0.923	0.004
APERSU ABCD	0.876	0.009	0.813	0.019	0.891	0.009
Crosswalk study	0.908	0.005	0.940	0.005	0.896	0.005
German 3L/5L study	0.815	0.034	0.717	0.057	0.815	0.036
VR validation	0.865	0.016	0.852	0.022	0.884	0.015

Note. Internal MIMIC only used the EQ-5D-5L items, so all data was collapsed for this analysis; reflective EQ-5D-5L items modeled as categorical variables.
ABCD indicates Alberta Caring for Diabetes; APERSU, Alberta PROMs and EQ-5D Research Support Unit; CFA, confirmatory factor analysis; CFI, comparative fix index; MIC, Multi-Instrument Comparison; RMSEA, root mean square error of approximation; SE, standard error; TLI, Tucker-Lewis index; VR, Veteran's Rand.

SF-36 subscales in model 2b-2 improved model fit minimally over SF-36 subscales alone in model 2b-1. Standardized coefficients were larger for MO and PD in the 2-factor external MIMIC than in the 1-factor model, whereas AD coefficients were larger in the 1-factor model. SC coefficients were much smaller than the other dimensions in all external MIMIC models. Model 2b-1 fit statistics were better for healthy subjects than for those with at least 1 health condition. Models 2a-1 and 2b-1 demonstrated similar coefficients and model fit in other data sets (results not shown), but models 2a-2 and 2b-2 could be fit only in the MIC data set.

Results of internal MIMIC models using categorical reflective indicators were reported because of improved model fit over the same models using continuous reflective indicators (Table 4). Model 3b (formative indicators: MO, AD, PD; reflective indicators: SC, UA) demonstrated superior χ^2 and RMSEA (<0.001-0.147) over all other explored models across all subgroups and data sets, except the Spanish valuation data set and liver disease. Furthermore, the relative χ^2 improvements were substantial based on percent change. The model 3b CFI and TLI statistics were slightly smaller than those of CFA models but typically remained higher than 0.95, demonstrating good model fit. On the basis of the totality of model fit evidence, model 3b was determined to fit the data best. An internal MIMIC model with only AD as a formative indicator and all non-AD dimensions as reflective indicators had poorer model fit than models 3a and 3b (data not shown).

In sensitivity analyses for internal MIMIC models, subgroup analyses showed better fit for healthy respondents than for patients. Nevertheless, model fit was not consistently better for population data sets (ie, valuation studies) than for patient samples (German 3L/5L, VR validation, and Alberta Caring for Diabetes) (Table 4). Patient/health respondent mix data sets (crosswalk and MIC) had the poorest fit.

Discussion

When developing scoring for an instrument, we consider (1) how items should be combined and (2) the weights that should be applied. In this article, we explored the first consideration by examining the structure of EQ-5D-5L items, applying reflective, formative, and combination models in 9 data sets covering various languages, countries, study designs, and patient types. Although standardized loadings and coefficients varied across different data sets and subgroups, 2 findings regarding the item structure of the EQ-5D instrument were robust: (1) separate reflective and formative indicator subscores may be appropriate given that the internal MIMIC models fit best and (2) weights developed from purely reflective methods and purely formative methods would differ drastically.

The single-factor CFA had acceptable absolute model fit statistics and found that MO, SC, UA, and PD were highly relevant to the latent trait, whereas AD demonstrated a weaker relationship, that is, had smaller factor loadings. The general poorer loading of AD may indicate that it does not address physical health like the other items, matching reporting elsewhere.[46–48] Furthermore, CFA may be suboptimal when both

Table 2B. Model 1 CFA (1-factor solution).

Group	Pain/Discomfort		Anxiety/Depression		Model fit statistics			
	Factor loading	SE	Factor loading	SE	χ^2	RMSEA	CFI	TLI
Overall (all data)	0.773	0.005	0.527	0.007	963.964	0.101	0.993	0.986
Sex								
Women	0.773	0.007	0.496	0.010	498.513	0.100	0.993	0.985
Men	0.773	0.007	0.561	0.010	452.314	0.099	0.994	0.987
Illness status								
With 1+ self-reported illness	0.739	0.006	0.458	0.009	905.538	0.115	0.990	0.980
Healthy/general population	0.681	0.016	0.484	0.020	151.618	0.070	0.984	0.967
Disease group								
Respiratory diseases	0.715	0.022	0.482	0.027	65.021	0.096	0.996	0.993
Musculoskeletal disease	0.738	0.016	0.457	0.023	87.665	0.090	0.994	0.988
Heart disease	0.704	0.027	0.514	0.030	50.945	0.088	0.995	0.991
Depression/mental health	0.737	0.020	0.459	0.026	107.372	0.119	0.982	0.964
Stroke	0.661	0.031	0.659	0.031	166.314	0.225	0.989	0.978
Diabetes	0.798	0.012	0.524	0.019	93.874	0.074	0.996	0.992
Cancer	0.770	0.022	0.471	0.041	20.698	0.063	0.997	0.994
Liver disease	0.803	0.038	0.541	0.047	17.671	0.077	0.995	0.990
Kidney disease	0.755	0.081	0.480	0.144	6.270	0.071	0.990	0.981
Data set								
Indonesian valuation	0.752	0.044	0.467	0.041	26.884	0.064	0.963	0.926
Spanish valuation	0.514	0.045	0.956	0.018	16.787	0.049	0.997	0.994
Canadian valuation	0.810	0.017	0.471	0.033	29.254	0.063	0.996	0.992
US Valuation	0.785	0.017	0.538	0.030	40.643	0.079	0.992	0.984
MIC data set	0.804	0.005	0.481	0.011	256.583	0.079	0.996	0.991
APERSU ABCD	0.809	0.009	0.542	0.020	54.996	0.070	0.992	0.897
Crosswalk study	0.732	0.009	0.476	0.015	391.063	0.145	0.990	0.979
German 3L/5L study	0.815	0.037	0.203	0.064	18.542	0.109	0.984	0.968
VR validation	0.802	0.019	0.619	0.029	14.233	0.060	0.998	0.996

Note. Internal MIMIC only used the EQ-5D-5L items, so all data was collapsed for this analysis; reflective EQ-5D-5L items modeled as categorical variables.
ABCD indicates Alberta Caring for Diabetes; APERSU, Alberta PROMs and EQ-5D Research Support Unit; CFA, confirmatory factor analysis; CFI, comparative fix index; MIC, Multi-Instrument Comparison; RMSEA, root mean square error of approximation; SE, standard error; TLI, Tucker-Lewis index; VR, Veteran's Rand.

causal and effect indicators are present without a well-understood conceptual model.[16,49] Thus, either the latent construct of health-related quality of life has 2 domains as measured by the EQ-5D-5L or AD is a formative indicator of HRQOL.[49,50]

The external MIMIC model was especially powerful because it allowed "regression" of EQ-5D-5L items onto 1 or more latent traits as captured by other HRQOL metrics. The poor fit of external MIMIC models may be due to the additional variance introduced by HRQOL measures external to the EQ-5D, for example, the SF-36 and the QWB, because these measures assess different dimensions of health or well-being. The improved fit statistics of the 2-factor model over the 1-factor model may have been induced by the mental and physical domains contained within the SF-36, which were used to fit all external MIMIC models. Despite generally poor model fit statistics of these models, the finding that SC had small standardized coefficients converged with other regression-based analyses of the EQ-5D items; the same phenomenon was observed in linear regressions of the EQ-5D items onto the VAS in data sets explored in this study and others.[51–54] Information overlap between SC and UA, also hypothesized by Gamst-Klaussen et al,[16] may minimize the contribution of SC in regression-derived EQ-5D-5L weights.

Taken together, the CFA and external MIMIC results show that EQ-5D items are unlikely to be all reflective or formative indicators and neither method alone adequately captures the EQ-5D item structure. The importance and contributions of AD and SC to the overall HRQOL construct, as evidenced by item loading/coefficient magnitude, differed by model and indicator type. AD was the least important dimension in the CFA as a reflective indicator but similarly weighted to other EQ-5D items as a formative indicator in the external MIMIC. SC had high factor loadings in the CFA but nearly negligible coefficients in the external MIMIC. Use of the CFA or the external MIMIC alone for scoring would incorrectly downweight AD or SC, respectively.

The internal MIMIC results reaffirmed that the use of the same method, that is, formative or reflective, to summarize all items may not be appropriate because they function differently. Model fit for internal MIMIC model 3b (formative indicators: MO, PD, AD; reflective indicators: UA and SC) was better than for both models 1 and 2 in most data sets and subpopulations. These findings confirm the earlier work by Gamst-Klaussen et al[16] in a broad range of respondents.

Models fit better for healthy and general population respondents than those with at least 1 illness across all 3 model types. Magnitude of factor loadings was similar across subpopulations within a narrow range; stroke patients, Spanish valuation study, and German 3L/5L study were, however, significant outliers. The Spanish valuation study data set was the most striking outlier. For CFA, AD loaded the highest of all dimensions within the data set, unlike other data sets, and its factor loading of 0.96 was far larger than the AD loading in other data sets. A CFA in Spanish patients with osteoarthritis had similar factor loadings to our overall and musculoskeletal subgroup CFA results.[55] Another CFA in Spanish patients with depression does not mirror our CFA results in patients with mental health concerns nor the CFA in the Spanish valuation data set.[56] Thus, the dissimilarity in CFA results between Spanish valuation and other data sets cannot be attributed to differences in language and culture. Model 3b also did not fit best in the Spanish valuation data set, in contrast to other data sets.

Table 3. External MIMIC model results using MIC data set.

1-Factor models				2-Factor models					
Model 2a-1 SF-36 subscales				Model 2b-1 SF-36 subscales			Model 2b-1 SF-36 subscales in healthy subjects		
	Standardized coefficient	SE	*P* value	Standardized coefficient	SE	*P* value	Standardized coefficient	SE	*P* value
MO	−0.198	0.010	<.001	−0.250	0.010	<.001	−0.158	0.020	<.001
SC	−0.014	0.008	.069	−0.005	0.008	.490	0.030	0.016	.063
UA	−0.271	0.010	<.001	−0.254	0.010	<.001	−0.157	0.020	<.001
PD	−0.341	0.009	<.001	−0.436	0.009	<.001	−0.467	0.024	<.001
AD	−0.341	0.008	<.001	−0.208	0.011	<.001	−0.001	0.021	.947
χ^2		17466.512			13838.965			2225.965	
RMSEA		0.199			0.179			0.153	
SRMR		0.094			0.090			0.550	
CFI		0.68			0.748			0.649	
TLI		0.604			0.677			0.649	
Model 2a-2 SF-36, HUI3, and VAS				Model 2b-2 SF-36 and HUI3			Model 2b-1 SF-36 subscales in respondents with any self-reported illness		
	Standardized coefficient	SE	*P* value	Standardized coefficient	SE	*P* value	Standardized coefficient	SE	*P* value
MO	−0.225	0.009	<.001	−0.290	0.010	<.001	−0.263	0.011	<.001
SC	−0.063	0.007	<.001	−0.051	0.012	<.001	−0.017	0.009	.055
UA	−0.247	0.010	<.001	−0.228	0.010	<.001	−0.258	0.011	<.001
PD	−0.326	0.008	<.001	−0.580	0.008	<.001	−0.426	0.011	<.001
AD	−0.359	0.007	<.001	−0.148	0.010	<.001	−0.217	0.013	<.001
χ^2		10489.7			24450.78			11303.342	
RMSEA		0.228			0.147			0.184	
SRMR		0.092			0.104			0.096	
CFI		0.711			0.729			0.730	
TLI		0.596			0.688			0.654	

Note. Results only presented for MIC; due to the number of other measures included, only the MIC dataset could support the most specifications of the external MIMIC; similar model results were noted in other datasets for Models 2a-1 and 2b-1.
CFI indicates comparative fix index; Dimensions: AD, Anxiety/Depression; MO, Mobility; PD, Pain/Discomfort; SC, Self-Care; UA, Usual Activities; HUI3, Health Utility Index 3; MIC, Multi-Instrument Comparison; MIMIC, Multiple Indicators Multiple Causes; RMSEA, root mean square error of approximation; SE, standard error; SF-36, short form 36 health survey; SRMR, standardized root mean square residual; TLI, Tucker-Lewis index; VAS, visual analogue scale.

The next phase of non–preference-based score development will need to explore whether measurement properties of items differ across subgroups, that is, differential item functioning, and the implications for scoring if variation exists. Despite variations across subgroups, the item structure of the EQ-5D was elucidated. Neither a reflective nor a formative model is ideal because some dimensions are formative while others are reflective, suggesting that future research should draw from the structure of model 3b. Item groupings that fit with the analyses presented can be (1) development of AD and non-AD subscores and (2) development of 2 subscores: 1 for PD, MO, and AD, and 1 for SC and UA. Candidate scoring algorithms based on the results of this study and unweighted summary scores should be compared on the basis of measurement properties such as reliability and responsiveness.[23] Other important factors to be considered in a scoring algorithm include (1) accurate description of HRQOL, (2) practicality of implementation for the end user, and (3) ease of score interpretation.

A major strength of the study is the use of a range of analyses to enrich our understanding of the EQ-5D item structure across different population and patient samples. Well-established, external HRQOL metrics were also modeled with EQ-5D items to account for HRQOL as a multidimensional latent construct. Such a broad scope of modeling approaches applied to diverse data to triangulate the item structure of the EQ-5D has not been previously reported.

The study results were subject to several limitations. The data sets were mostly from developed Western countries and not all diseases were represented; thus, despite similar item structure conclusions, the EQ-5D items may be interrelated differently in other populations. Furthermore, many data sets were from country-specific valuation studies; respondents in these studies tended to be healthier, leading to less variation in the health states captured. Modeling based on this narrow range of health may fail to fully inform the structure for the broad range of applications of the EQ-5D-5L as a generic instrument.

Conclusions

Although there were variations in results across subgroups and data sets, the elucidated item structure was generally consistent

Table 4. Fit indices of internal MIMIC models.

	Model 3a (formative: AD, PD; reflective: MO, SC, UA)				Model 3b (formative: MO, AD, PD; reflective: SC, UA)			
	χ^2	RMSEA	CFI	TLI	χ^2	RMSEA	CFI	TLI
Overall (all data)	1158.929	0.124	0.972	0.936	216.831	0.075	0.984	0.945
Women	574.436	0.120	0.973	0.938	135.919	0.083	0.981	0.932
Men	579.404	0.126	0.972	0.937	65.448	0.059	0.991	0.967
Illness status								
With 1+ self-reported illness	1121.730	0.144	0.967	0.925	206.311	0.087	0.982	0.937
Healthy/general population	50.367	0.044	0.988	0.973	19.129	0.038	0.989	0.961
Disease group								
Respiratory	30.725	0.072	0.996	0.992	4.588	0.032	0.998	0.994
Musculoskeletal	64.816	0.086	0.987	0.972	13.163	0.052	0.994	0.978
Heart disease	28.250	0.071	0.992	0.983	9.360	0.055	0.993	0.975
Depression/mental health	110.791	0.136	0.944	0.875	6.299	0.039	0.994	0.978
Stroke	22.211	0.085	0.997	0.993	***4.324***	0.043	0.998	0.994
Diabetes	91.508	0.082	0.986	0.968	6.990	0.028	0.997	0.991
Cancer	19.582	0.071	0.990	0.977	***0.690***	<0.001	1.000	1.010
Liver disease	19.589	0.097	0.981	0.957	17.381	0.136	0.935	0.773
Kidney disease	***2.215***	<0.001	1.000	1.130	***1.579***	<0.001	1.000	1.099
Data set								
Indonesian valuation	10.734	0.040	0.970	0.933	***0.407***	<0.001	1.000	1.063
Spanish valuation	41.223	0.096	0.937	0.859	21.183	0.098	0.899	0.646
Canadian valuation	30.084	0.073	0.983	0.962	9.178	0.054	0.988	0.957
US valuation	62.062	0.113	0.953	0.895	10.994	0.063	0.981	0.935
MIC data set	331.688	0.101	0.980	0.954	37.143	0.047	0.993	0.976
APERSU ABCD	66.676	0.088	0.982	0.959	6.944	0.035	0.995	0.983
Crosswalk study	391.676	0.163	0.974	0.941	160.271	0.147	0.963	0.870
German 3L/5L study	18.216	0.125	0.949	0.886	***3.823***	0.063	0.980	0.929
VR validation study	17.646	0.081	0.988	0.972	6.732	0.068	0.986	0.949

Note. Bolded italics: chi-squared *P*-value > .05, indicating good model fit. Internal MIMIC only used the EQ-5D-5L items, so all data was collapsed for this analysis; reflective EQ-5D-5L items modeled as categorical variables.
ABCD indicates Alberta Caring for Diabetes; APERSU, Alberta PROMs and EQ-5D Research Support Unit; CFI, comparative fix index; Dimensions: AD, Anxiety/Depression; MO, Mobility; PD, Pain/Discomfort; SC, Self-Care; UA, Usual Activities; MIC, Multi-Instrument Comparison; MIMIC, Multiple Indicators Multiple Causes; RMSEA, root mean square error of approximation; TLI, Tucker-Lewis index; VR, Veteran's Rand.

across subanalyses. In the CFA, AD does not seem to measure the same latent trait as the other items, whereas the external MIMIC demonstrated that SC contributes less to the variation in HRQOL than do the other 4 dimensions. Internal MIMIC model 3b clarified these results by showing that the EQ-5D is best modeled with both reflective (UA and SC) and formative indicators (MO, PD, and AD). The summation of evidence calls for testing of formative/reflective combination approaches of summarizing the EQ-5D-5L. These findings on the structure and relationships between the EQ-5D-5L items illustrate the challenge in identifying a broadly applicable non–preference-based approach to scoring the EQ-5D.

Acknowledgments

We are grateful to the following individuals for sharing their data graciously: (1) Alberta Caring for Diabetes data: Fatima Al Sayah and Jeff Johnson (Alberta PROMs and EQ-5D Research Support Unit); (2) Multi-Instrument Comparison Study: Gang Chen (Monash University) and Jan Abel Olsen (University of Tromsø); (3) Spanish EQ-5D-5L Valuation Study: Juan Manuel Ramos-Goñi (Axentiva Solutions) and Pedro Serrano (*Jefe de Servicio de Evaluación de la Dirección del Servicio Canario de Salud*); (4) Indonesian EQ-5D-5L Valuation Study: Frederick Purba (*Universitas Padjadjaran*) and Jan van Busschbach (Erasmus); and (5) Canadian EQ-5D-5L Valuation Study: Feng Xie (McMaster University).

Source of Financial Support

Financial support for this study was provided by a grant by the EuroQol Research Foundation (grant number 20170130). The funding agreement ensured the authors' independence in designing the study, interpreting the data, and writing and publishing the report.

REFERENCES

1. van Reenen M, Janssen B. *EQ-5D-5L User Guide: Basic information on how to use the EQ-5D-5L instrument*, 2.1 ed.. Rotterdam, The Netherlands: EuroQol Research Foundation; 2015. https://euroqol.org/wp-content/uploads/2016/09/EQ-5D-5L_UserGuide_2015.pdf. Accessed February 23, 2017
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(10):1727–1736.
3. van Hout B, Janssen MF, Feng YS, et al. Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. *Value Health*. 2012;15(5):708–715.
4. Devlin NJ, Parkin D, Browne J. Patient-reported outcome measures in the NHS: new methods for analysing and reporting EQ-5D data. *Health Econ*. 2010;19(8):886–905.
5. APERSU - Alberta PROMS and EQ-5D Research and Support Unit. http://apersu.ca/.
6. Parkin D, Rice N, Devlin N. Statistical analysis of EQ-5D profiles: does the use of value sets bias inference? *Med Decis Making*. 2010;30(5):556–565.
7. Brooks R. *EuroQol Group after 25 Years*. Berlin, Germany: Springer Science + Business; 2013.
8. Devlin NJ, Brooks R. EQ-5D and the EuroQol group: past, present and future. *Appl Health Econ Health Policy*. 2017;15(2):127–137.

9. Devlin N, Appleby J. *Getting the Most out of PROMS - Putting health outcomes at the heart of NHS decision-making*. Vol 2017. London, UK: The King's Fund; 2010. https://www.kingsfund.org.uk/sites/files/kf/Getting-the-most-out-of-PROMs-Nancy-Devlin-John-Appleby-Kings-Fund-March-2010.pdf.
10. Hostetter M, Klein S. *Using patient-reported outcomes to improve health care quality*. New York, NY: The Commonwealth Fund; 2012. http://www.commonwealthfund.org/publications/newsletters/quality-matters/2011/december-january-2012/in-focus. Accessed January 5, 2017.
11. van Reenen M, Janssen B, Oppe M, Kreimeier S, Greiner W, EQ-5D-Y User Guide. *Basic information on how to use the EQ-5D-Y instrument*, 1.0 ed.. Rotterdam, The Netherlands: EuroQol Research Foundation; 2014. https://euroqol.org/wp-content/uploads/2016/09/EQ-5D-Y_User_Guide_v1.0_2014.pdf. Accessed February 23, 2017
12. Zamora BP D, Feng Y, Bateman A, Herdman M, Devlin N. *New methods for analysing the distribution of EQ-5D observations*. London, UK: Office of Health Economics; 2018. https://www.ohe.org/publications/new-methods-analysing-distribution-eq-5d-observations. Accessed February 23, 2017.
13. Feng Y, Parkin D, Devlin NJ. Assessing the performance of the EQ-VAS in the NHS PROMs programme. *Qual Life Res*. 2014;23(3):977–989.
14. Gutacker N, Bojke C, Daidone S, Devlin N, Street A. Hospital variation in patient-reported outcomes at the level of EQ-5D dimensions: evidence from England. *Med Decis Making*. 2013;33(6):804–818.
15. Clouth J, Schmidt P, Moser G, Greiner W. *Estimating VAS values through EQ-5D by structural equation modeling for the German population*, 25th Annual EuroQol Plenary Meeting; September. Italy: Baveno Lake Maggiore; 2008.
16. Gamst-Klaussen T, Gudex C, Olsen JA. Exploring the causal and effect nature of EQ-5D dimensions: an application of confirmatory tetrad analysis and confirmatory factor analysis. *Health Qual Life Outcomes*. 2018;16(1):153.
17. Costa DS. Reflective, causal, and composite indicators of quality of life: A conceptual or an empirical distinction? *Qual Life Res*. 2015;24(9):2057–2065.
18. Hair JF, Sarstedt M, Ringle C, Gudergan S. *Advanced issues in partial least squares structural equation modeling*. Thousand Oaks, California: SAGE Publishing; 2017.
19. Hays RD, Bjorner JB, Revicki DA, Spritzer KL, Cella D. Development of physical and mental health summary scores from the patient-reported outcomes measurement information system (PROMIS) global items. *Qual Life Res*. 2009;18(7):873–880.
20. Lai JS, Stucky BD, Thissen D, et al. Development and psychometric properties of the PROMIS((R)) pediatric fatigue item banks. *Qual Life Res*. 2013;22(9):2417–2427.
21. Ware Jr JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care*. 1992;30(6):473–483.
22. Rand-Hendriksen K, Augestad LA, Kristiansen IS, Stavem K. Comparison of hypothetical and experienced EQ-5D valuations: relative weights of the five dimensions. *Qual Life Res*. 2012;21(6):1005–1012.
23. Wilke CT, Pickard AS, Walton SM, Moock J, Kohlmann T, Lee TA. Statistical implications of utility weighted and equally weighted HRQL measures: an empirical study. *Health Econ*. 2010;19(1):101–110.
24. Prieto L, Sacristan JA. What is the value of social values? The uselessness of assessing health-related quality of life through preference measures. *BMC Med Res Methodol*. 2004;4:10.
25. Purba FD, Hunfeld JAM, Iskandarsyah A, et al. The Indonesian EQ-5D-5L value set. *Pharmacoeconomics*. 2017;35(11):1153–1165.
26. Hernandez G, Garin O, Pardo Y, et al. Validity of the EQ-5D-5L and reference norms for the Spanish population. *Qual Life Res*. 2018.
27. 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(1):98–105.
28. Pickard AS, Law EH, Jiang R, et al. United states valuation of EQ-5D-5L health states: An initial model using a standardized protocol. *Value Health*. 2018;21:S4–S5.
29. Buchholz I, Thielker K, Feng YS, Kupatz P, Kohlmann T. 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(4):829–835.
30. Al Sayah F, Majumdar SR, Soprovich A, et al. The Alberta's Caring for Diabetes (ABCD) Study: Rationale, design and baseline characteristics of a prospective cohort of adults with type 2 diabetes. *Can J Diabetes*. 2015;39(Suppl 3):S113–S119.
31. Oppe M, van Hout B. *The "power" of eliciting EQ-5D-5L values: the experimental design of the EQ-VT*. EuroQol Research Foundation; 2017.
32. *The SAS System for Windows [computer program]. Version 9.4*. Cary, NC, USA: SAS Institute Inc.; 2016.
33. Stata Statistical Software. *Release 13* [computer program]. College Station, TX: StataCorp LP; 2013.
34. *SPSS Statistics for Windows [computer program]*. Armonk, NY: IBM; 2013.
35. *Mplus User's Guide [computer program]. Los Angeles, CA*. 2010.
36. Jöreskog KG, Goldberger AS. Estimation of a model with multiple indicators and multiple causes of a single latent variable. *J Am Stat Assoc*. 1975;70(351a):631–639.
37. Horsman J, Furlong W, Feeny D, Torrance G. The Health Utilities Index (HUI): concepts, measurement properties and applications. *Health Qual Life Outcomes*. 2003;1:54.
38. *The ICF: An Overview*; 2001. https://www.cdc.gov/nchs/data/icd/icfoverview_finalforwho10sept.pdf. Accessed March 28, 2017.
39. Cieza A, Brockow T, Ewert T, et al. Linking health-status measurements to the international classification of functioning, disability and health. *J Rehabil Med*. 2002;34(5):205–210.
40. Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. *JAMA*. 1995;273(1):59–65.
41. Valderas JM, Alonso J. Patient reported outcome measures: a model-based classification system for research and clinical practice. *Qual Life Res*. 2008;17(9):1125–1135.
42. Hu L-t, Bentler PM. Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychol Methods*. 1998;3(4):424–453.
43. Hu L-t, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct Equ Modeling*. 1999;6(1):1–55.
44. Marsh HW, Hau K-T, Wen Z. In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Struct Equ Modeling*. 2004;11(3):320–341.
45. Lai K, Green SB. The problem with having two watches: Assessment of fit when RMSEA and CFI disagree. *Multivariate Behav Res*. 2016;51(2-3):220–239.
46. Keeley T, Coast J, Nicholls E, Foster NE, Jowett S, Al-Janabi H. An analysis of the complementarity of ICECAP-A and EQ-5D-3 L in an adult population of patients with knee pain. *Health Qual Life Outcomes*. 2016;14:36.
47. Davis JC, Liu-Ambrose T, Richardson CG, Bryan S. A comparison of the ICECAP-O with EQ-5D in a falls prevention clinical setting: are they complements or substitutes? *Qual Life Res*. 2013;22(5):969–977.
48. Wittrup-Jensenm KL, Jørgen. *An assessment of two generic health-related quality of life (HRQoL) instruments in patients suffering from low back pain*. 2008.
49. Fayers PM, Hand DJ. Factor analysis, causal indicators and quality of life. *Qual Life Res*. 1997;6(2):139–150.
50. Fayers PM, Hand DJ, Bjordal K, Groenvold M. Causal indicators in quality of life research. *Qual Life Res*. 1997;6(5):393–406.
51. Sun S, Chen J, Kind P, Xu L, Zhang Y, Burstrom K. Experience-based VAS values for EQ-5D-3L health states in a national general population health survey in China. *Qual Life Res*. 2015;24(3):693–703.
52. Burstrom K, Sun S, Gerdtham UG, et al. Swedish experience-based value sets for EQ-5D health states. *Qual Life Res*. 2014;23(2):431–442.
53. Leidl R, Reitmeir P. A value set for the EQ-5D based on experienced health states: development and testing for the German population. *Pharmacoeconomics*. 2011;29(6):521–534.
54. Feng Y, Jiang R, Kohlmann T, AS. P. Possible redundancy in the self-care item of the EQ-5D-5L instrument. Paper presented at: 35th Annual EuroQol Plenary; 19-22 September, 2018; Lisbon, Portugal.
55. Bilbao A, Martín-Fernández J, Arenaza JC, et al. Validation of the EQ-5D-5L in patients with hip or knee osteoarthritis. *Value Health*. 2017;20(9):A760.
56. Bilbao A, García-Perez L, Retolaza-Balsategui A, et al. Psychometric properties of the EQ-5D-5L in patients with major depression disorder. *Value Health*. 2017;20(9):A758.