PharmacoEconomics
https://doi.org/10.1007/s40273-018-0623-8

ORIGINAL RESEARCH ARTICLE

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

Mathieu F. Janssen[1] · **Gouke J. Bonsel**[2,3] · **Nan Luo**[4]

Abstract

Objective This study describes the first empirical head-to-head comparison of EQ-5D-3L (3L) and EQ-5D-5L (5L) value sets for multiple countries.

Methods A large multinational dataset, including 3L and 5L data for eight patient groups and a student cohort, was used to compare 3L versus 5L value sets for Canada, China, England/UK (5L/3L, respectively), Japan, The Netherlands, South Korea and Spain. We used distributional analyses and two methods exploring discriminatory power: relative efficiency as assessed by the *F* statistic, and an area under the curve for the receiver-operating characteristics approach. Differences in outcomes were explored by separating descriptive system effects from valuation effects, and by exploring distributional location effects.

Results In terms of distributional evenness, efficiency of scale use and the face validity of the resulting distributions, 5L was superior, leading to an increase in sensitivity and precision in health status measurement. When compared with 5L, 3L systematically overestimated health problems and consequently underestimated utilities. This led to bias, i.e. over- or underestimations of discriminatory power.

Conclusion We conclude that 5L provides more precise measurement at individual and group levels, both in terms of descriptive system data and utilities. The increased sensitivity and precision of 5L is likely to be generalisable to longitudinal studies, such as in intervention designs. Hence, we recommend the use of the 5L across applications, including economic evaluation, clinical and public health studies. The evaluative framework proved to be useful in assessing preference-based instruments and might be useful for future work in the development of descriptive systems or health classifications.

Key Points for Decision Makers

EQ-5D-5L (5L) is superior to EQ-5D-3L (3L) with respect to various measurement properties, enabling improvements in sensitivity and precision in health status measurement.

5L provides more precise measurements than 3L at individual and group levels, both in terms of responses to EQ-5D items and the resultant utilities.

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

5L is recommended for use across applications, including economic evaluation, clinical studies, quality of care and in public health studies.

Disclaimer The views expressed by the authors in this paper do not necessarily reflect the views of the EuroQol Group.

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

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

[2] Department of Public Health, Erasmus MC, Erasmus University, Rotterdam, The Netherlands

[3] Division Mother and Child, UMC Utrecht, University of Utrecht, Utrecht, The Netherlands

[4] Saw Swee Hock School of Public Health, National University of Singapore, Singapore, Singapore

Published online: 22 February 2018

1 Introduction

Since the introduction of the original EQ-5D descriptive system in 1990 [1] and the first value set in 1997 [2], the EuroQol Group has continuously furthered research aimed at enhancing the instrument [3, 4]. This entailed refining the descriptive system, developing new valuation methodology and also developing new EQ-5D instruments for specific use. Examples of the latter include the child-friendly EQ-5D version (EQ-5D-Y) as a more comprehensible instrument suitable for children and adolescents [5, 6], and the exploration of EQ-5D versions with one or two additional dimensions to the descriptive system [7–10]. Arguably, the biggest change has been in refining the 'granularity' of the five dimensions by replacing the three response options (levels) of the original EQ-5D (now 'EQ-5D-3L') with five levels. The official EQ-5D-5L descriptive system (for convenience we use the term '5L' from here) has been available since 2011 [11] and is currently available in more than 150 translations and multiple modes of administration [12]. In parallel, a new valuation protocol for the 5L was developed (EQ-VT) to establish new country-specific value sets, warranting a high level of standardisation and quality control as well as introducing new and improved valuation methods [13, 14].

Several studies have compared the descriptive systems of EQ-5D-3L (for convenience we use the term '3L' from here) and 5L in terms of their measurement properties, including distributional characteristics such as ceiling effects and evenness, reliability and various types of validity [15–22]. Most studies showed that the 5L descriptive system had better or at least similar measurement properties compared with 3L, but two remarks apply. First, we must establish whether the increased descriptive richness of 5L will increase measurement precision rather than measurement error, as this a trade-off. Further, considering that the EQ-5D is a preference-based instrument, it is essential also to investigate whether the increased descriptive richness translates into increased sensitivity of its utility-based index values (hereafter 'utility values' or 'utilities'); again, error may increase due to the increased difficulty in valuing more refined health states. The final question is whether the combined descriptive and valuation effects of 5L improve the discriminatory potential of the utility instrument in, for example, the estimation of quality-adjusted life-years (QALYs) in economic evaluation. As the measurement of health status with the descriptive system is independent from the derivation of utility values and involves different methodologies, improved sensitivity and discrimination of the descriptive system does not necessarily translate into better discriminatory power using utilities (comparing groups or comparing pre- and post-intervention health state). For economic evaluation (e.g. cost-utility analysis), improved discriminatory performance of the utility values would represent a major advantage.

To compare the performance of 3L and 5L in terms of QALYs gained, longitudinal patient-level data on both 3L and 5L in one or multiple study populations would be preferred. In the absence of such longitudinal data we compared 3L and 5L using data from a large multi-country cross-sectional survey, applying country-specific value sets for seven countries.

We first compared the distributional characteristics of the observed utility values by value set, and standard descriptive statistics by condition group and value set. Our main analysis consisted of two tests of discriminatory power. In order to further clarify and explain the results, we performed an exploratory analysis to determine the factors responsible for certain patterns in the results. In this analysis, a clear distinction was made between differences caused by descriptive system results and by the utility values applied to the descriptive data. The separation of descriptive and valuation effects has proven to be of use in an earlier study exploring differences in utilities derived from different preference-based instruments [23]. We introduce an evaluative framework consisting of a novel combination of non-parametric methods to establish increased measurement refinement (if any), with parametric methods to demonstrate improved discrimination (if any); 5L is only better than (rather than 'different from') 3L if (1) more response levels are efficiently used without a decrease of uniformity of the distribution and (2) this increased use is not offset by more measurement error, both in terms of description and valuation.

Our study had two research questions: (1) Do 5L value sets perform better than 3L value sets in terms of discriminatory power, as a direct result of the improved descriptive sensitivity? (2) What are the underlying factors affecting this performance? Our approach allowed us to make normative assessments on the performance of both instruments and to offer recommendations to users of EQ-5D instruments.

2 Methods

2.1 Paired EQ-5D-3L–EQ-5D-5L (3L–5L) Descriptive Data

A large multinational dataset that included paired descriptive 3L and 5L data for eight patient groups and a student cohort was used [15, 24]. These data were obtained with the standard 3L and 5L versions for self-report use in adults, describing health on the dimensions of mobility,

self-care, usual activities, pain/discomfort and anxiety/depression. The 3L version applied the level descriptors (or labels) 'no problems', 'some/moderate problems' and 'extreme problems/unable to', and the 5L version used 'no problems', 'slight problems', 'moderate problems', 'severe problems' and 'extreme problems/unable to'. For mobility, the most severe response option was changed from 'confined to bed' for 3L to 'unable to walk about' for 5L. The 3L classification describes 243 unique health states (or health profiles) that are often reported as vectors ranging from 11111 (full health) to 33333 (worst health), whereas the 5L defines 3125 unique health states, with 55555 as the worst health state.

Paper-and-pencil versions of the questionnaires were used in all countries except in England where data collection took place online. Since there were many condition-specific subgroups with small sample sizes, it was decided to combine related patient groups, resulting in nine main condition groups. Only respondents who completed both the 3L and 5L[1] without any missing responses were included in the analyses (a 3L–5L comparison of missing values is reported elsewhere [15]). It was assumed that within a specific condition group country differences were not important so that descriptive data could be pooled.

2.2 Paired 3L–5L Value Sets

At the time of this study there were seven countries with both 3L and 5L value sets available, namely Canada, China, England/UK (5L/3L, respectively), Japan, The Netherlands, South Korea and Spain [2, 25–37]. All EQ-5D value sets were obtained using representative samples of the general public, ensuring that they represented the societal perspective. A value set is a set of weights that can convert each health state into an index value on a scale anchored at 1 (referring to full health) and 0 (referring to a state as bad as being dead), allowing for negative values for health states considered to be worse than dead. Most 3L valuation studies followed similar protocols, although there were notable differences with regard to the sampling of respondents (affecting representation), sample size and health state design (varying from 17 to 101 valued health states) [38, 39]. All 3L valuation studies were performed with face-to-face interviews and paper-and-pencil methods except for Canada where a web survey was used. All 3L value sets were based on time trade-off (TTO) data. With the introduction of 5L a standardised valuation protocol was developed, the EQ-VT (EuroQol Valuation Technology Platform) [13]. In addition to standardisation in terms of health state design, valuation methodology and a computer-assisted personal interview mode of administration, a strict protocol of interviewer training and quality control during the entirety of the data collection process was developed and implemented [14]. Discrete choice experiment (DCE) methodology was introduced in the EQ-VT, along with composite TTO as the main valuation method. Since there is no standardised analytic protocol, some 5L value sets were based on hybrid models utilising both TTO and DCE data while others were based on TTO data only. After the initial valuation studies were performed using EQ-VT version 1.0 (Canada, China, England, The Netherlands, Spain) some data quality issues and interviewer effects were apparent and a cyclic quality control process was introduced in version 1.1, which led to a substantial improvement [14].

Usually country-specific utility values are used to conduct analyses in a population or patient sample from that particular country, reflecting the appropriate preferences. Since our research questions were of a methodological nature, aiming at making generalisations across value set characteristics, we used the pooled multi-country dataset to compare the characteristics of 14 country-specific 3L and 5L value sets.

2.3 Analyses

2.3.1 3L and 5L Value Sets for Seven Countries

Characteristics of all value sets were reported in terms of model parameters and model characteristics, such as the modelled value range, intercept, interaction parameters and histograms of all possible values (3L: 243; 5L: 3125), which may be responsible for differences in performance between 3L and 5L (see Table 1).

2.3.2 Distributional Analyses of 3L and 5L Utility Values

Country-specific 3L and 5L utility values were calculated for each value set for all condition groups combined and described numerically and graphically using histograms. We examined clusters and discontinuities ('gaps') in the histograms as such patterns theoretically diminish the sensitivity and the accuracy of the instruments and might lead to estimation problems [40].

In order to assess the frequency and efficiency of use of the utility scale we applied Shannon's indices as a means of assessing distributional evenness [17, 18, 21, 22]. While Shannon's H' captures absolute informativity and is simultaneously powered by evenness and the number of categories used, Shannon's J' index of relative informativity solely reflects the evenness of a distribution [41]. Since Shannon's J' corrects for the total number of possible categories (here: possible utility values), which could be

[1] We use the notation '3L-5L' to refer to '3L compared to 5L', '3L versus 5L' or '3L and 5L', depending on the context.

Table 1 Characteristics of EQ-5D-3L and EQ-5D-5L value sets from seven countries

3L and 5L value set models	Canada		China		England/UK		Japan		The Netherlands		South Korea		Spain	
	3L	5L	3L	5L[a]	3L	5L	3L	5L	3L	5L	3L	5L	3L	5L
Intercept	0.111	0.051	0.039		0.081		0.152	0.061	0.071	0.047	0.050	0.096	0.024	
Mobility														
Slight		0.039		0.066		0.058		0.064		0.035		0.046		0.084
Some/moderate	0.046	0.078	0.099	0.158	0.069	0.076	0.075	0.113	0.036	0.057	0.096	0.058	0.106	0.099
Severe		0.168		0.287		0.207		0.179		0.166		0.133		0.250
Confined to bed/unable to	0.322	0.207	0.246	0.345	0.314	0.274	0.418	0.243	0.161	0.203	0.418	0.251	0.430	0.337
Self-care														
Slight		0.046		0.048		0.050		0.044		0.038		0.032		0.050
Some/moderate	0.071	0.092	0.105	0.116	0.104	0.080	0.054	0.077	0.082	0.061	0.046	0.050	0.134	0.053
Severe		0.196		0.210		0.164		0.124		0.168		0.078		0.164
Unable to	0.224	0.242	0.208	0.253	0.214	0.203	0.102	0.160	0.152	0.168	0.136	0.122	0.309	0.196
Usual activities														
Slight		0.020		0.045		0.050		0.050		0.039		0.021		0.044
Some/moderate	0.072	0.039	0.074	0.107	0.036	0.063	0.044	0.091	0.032	0.087	0.051	0.051	0.071	0.048
Severe		0.169		0.194		0.162		0.148		0.192		0.100		0.135
Unable to	0.105	0.188	0.193	0.233	0.094	0.184	0.133	0.175	0.057	0.192	0.208	0.175	0.195	0.153
Pain/discomfort														
Slight		0.044		0.058		0.063		0.045		0.066		0.042		0.078
Moderate	0.045	0.089	0.092	0.138	0.123	0.084	0.080	0.068	0.086	0.092	0.037	0.053	0.089	0.101
Severe		0.274		0.252		0.276		0.131		0.360		0.166		0.245
Extreme	0.298	0.319	0.236	0.302	0.386	0.335	0.194	0.191	0.329	0.415	0.151	0.207	0.261	0.382
Anxiety/depression														
Slight		0.038		0.049		0.078		0.072		0.070		0.033		0.081
Moderate	0.063	0.075	0.086	0.118	0.071	0.104	0.063	0.110	0.124	0.145	0.043	0.046	0.062	0.128
Severe		0.241		0.215		0.285		0.168		0.357		0.102		0.270
Extreme	0.280	0.278	0.205	0.258	0.236	0.289	0.112	0.196	0.325	0.421	0.158	0.137	0.144	0.348
Interaction parameters														
N3			0.022		0.269				0.234		0.050		0.291	
Num45sq		0.0085												
C4												0.078		
Highest value (11111)	1	0.949	1	1	1	1	1	1	1	1	1	1	1	1
Second highest value	0.844	0.929	0.887	0.955	0.883	0.950	0.804	0.895	0.897	0.918	0.913	0.883	0.914	0.956
Lowest value (33333/55555)	− 0.340	–0.148	− 0.149	− 0.386	− 0.594	− 0.285	− 0.111	− 0.025	− 0.329	− 0.446	− 0.171	− 0.066	− 0.654	− 0.416

Table 1 continued

3L and 5L value set models	Canada		China		England/UK		Japan		The Netherlands		South Korea		Spain	
	3L	5L	3L	5L[a]	3L	5L	3L	5L	3L	5L	3L	5L	3L	5L
Value range	1.340	1.097	1.149	1.386	1.594	1.285	1.111	1.025	1.329	1.446	1.171	1.066	1.654	1.416

3L EQ-5D-3L, *5L* EQ-5D-5L, *C4* number of level 4s or 5s beyond the first one, *N3* any level 3, *Num45sq* number of level 4s or 5s beyond the first one squared

[a]The 5L model for China is based on an underlying multiplicative eight-parameter model

potentially close (or equal) to 243 for 3L and 3125 for 5L, it was not considered to be a fair comparison (we expected that J' would result in higher values for 3L for this reason). Hence, we also calculated both indices by subdividing the scale range in categories ('bins') with a width of 0.05, making the number of categories between 3L and 5L more comparable.

Subsequently, we presented mean utility values (and standard deviations [SDs]) by condition group for all 14 value sets, with the addition of an equal weighting score (Level Sum Score [LSS] transformed to a 0–1 scale) in order to assess the impact of the descriptive data without the effect of utility weights. The transformed LSS (tLSS) was calculated by summing the level scores for the five dimensions and performing a linear transformation on this sum score to a 0–1 scale so that the value for 11111 is equal to 1.0 and 33333 (for 3L) or 55555 (for 5L) is equal to 0.

2.3.3 Discriminatory Performance of 5L Versus 3L

Two tests of discriminatory power were conducted, accommodating different distributional assumptions with respect to utility values: one based on the *F* statistic (parametric), the second on receiver-operating characteristics (non-parametric).

Discriminatory power was assessed using the *F* statistic derived from analysis of variance (ANOVA) to test the equality of means. The *F* statistic is widely used to assess the relative efficiency of patient-reported outcome measures [21, 42, 43] and is based on differences in group means divided by the standard error of the difference. A higher *F* statistic means a higher likelihood for a measure to show statistical significance when used to compare groups. Hence, higher *F* statistic values indicate higher discriminatory power. To express the discriminatory power of 5L relative to 3L we computed the ratio of their *F* statistics resulting from comparisons of different condition groups, in such a way that a ratio higher than 1.0 indicated that 5L was more discriminative than 3L: relative efficiency $= F$ statistic$_{5L}$/F statistic$_{3L}$.

Comparisons were made between (1) the eight disease groups and the student cohort, assuming the students were a valid proxy for a healthy population sample; and (2) patients with a mild condition versus those with a moderate or severe condition. Using the observed mean EQ-5D visual analogue scale (EQ VAS) ratings as reference, we defined diabetes and liver disease as mild conditions (relative to the other conditions), and the remaining six as moderate to severe conditions. Since our main aim was to compare measurement properties of 3L and 5L, we considered this method to be suitable for assessing their ability

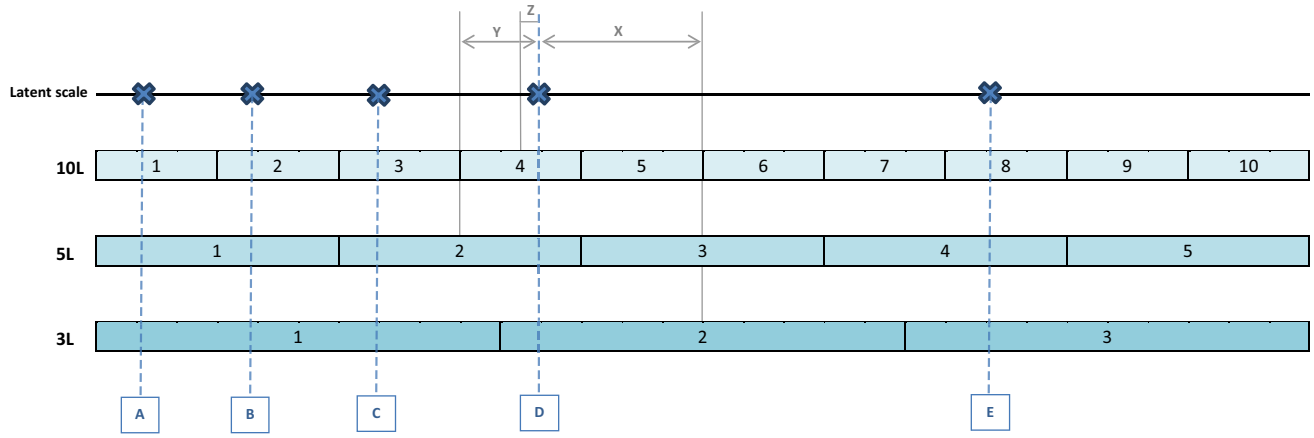

Fig. 1 Illustration of location effects when five hypothetical latent health states (A through E) are measured on three scales with varying levels of granularity (3L, 5L, 10L). *3L* 3 levels, *5L* 5 levels, *10L* 10 levels

to distinguish between mild and moderate/severe condition groups.

As a second analysis, we calculated the area under the receiver-operating characteristics curve (AUROC) as a non-parametric method of assessing discriminatory power. AUROC analyses were performed for each pair of condition group comparisons using pooled data on the groups, with group membership being the outcome and the 3L/5L utility score being the exposure. AUROCs for 3L and 5L were calculated and the ratio (5L/3L) was used as the measure of discriminatory power. The AUROC value can range from 0.5 (no prediction) to 1.0 (perfect prediction). Consequently, a 5L/3L AUROC ratio >1.0 indicates 5L to be more discriminative than 3L. While the *F* statistic is directly based on means and dispersion, the AUROC employs the full distribution.

For all comparisons 95% confidence intervals (CIs) of the *F* statistic and AUROC ratios were calculated using 3000 bootstrap samples, enabling us to test whether the ratio was statistically different from 1.0.

2.3.4 Exploration of Factors Affecting Discriminatory Power

At least three separate factors determine discriminatory power results:

1. The effects of the descriptive system, involving choice of dimensions, number of levels and corresponding labels, translation effects and reporting heterogeneity.
2. Valuation effects, relating to the valuation protocol, the valuation study (interviewer effects, quality control, etc.) but also to the modelling of the valuation data. Valuation effects also encompass true country-specific variation in preferences, which may be caused by many underlying factors, e.g. cultural, geographical or related to demographics, language or health system.
3. A third factor is related to the ability of any scale to capture the location of a respondent on the true latent scale. The precision of measuring this location will have an impact on the descriptive data and consequently the utility distribution of any study sample. As it appears this important factor is often ignored, we discuss this in some detail.

A graphical example can illustrate potential misclassification effects due to distributional descriptive 3L–5L effects (Fig. 1). The general methodology has been widely discussed in research on reporting heterogeneity [44–48]. Imagine a health dimension scaled with three levels of granularity: 3L, 5L and 10L (3, 5 and 10 levels respectively). In this example we do not take specific labels into account (although '1' refers to no problems). There is an underlying unobservable latent scale which is assumed to be continuous: all three measurement systems (3L, 5L, 10L) will only be approximations of the true latent value. The transition area of two adjacent categories is called the cut-off point (or 'cut-point'), and in the development of measurement scales one strives for clearly defined cut-points with little overlap (as defined by the labels), to avoid error. The distribution of observed scores of the 3L, 5L and 10L ordinal scales depends on the cut-points. Random error may occur at the cut-points when overlap exists, and this overlap may differ between 3L, 5L and 10L. Note that random error may cause a shift of average values for the extreme categories of the scale, as misclassification can only be towards the middle level of the scale due to the censored nature of the EQ-5D dimensions. Also note that when applying labels, the middle category of 3L does not necessarily coincide with the middle level of 10L, or would have the same latent midpoint, i.e. the middle point of the

category, equidistant from both cut-points. Various types of misclassification may occur between the three systems. Imagine five different locations on the latent scale (A through to E), which we here refer to as respondents, although these also might indicate group averages. For respondent A there is no discrepancy between 3L, 5L and 10L: no problems are scored in all three systems. For respondent B both 3L and 5L lack refinement (no problems) as evidently there are reported problems on 10L. Respondent C illustrates the reduced ceiling effect with the introduction of 5L over 3L: no problems are reported in 3L whereas problems are reported on 5L. Respondent D might contribute to an overestimation of reported health problems in 3L when compared to 5L: the middle 3L category is chosen whereas a milder category is chosen for 5L. The distance from the 3L midpoint to the true latent value (X) is larger than the distance from the 5L midpoint to the latent value (Y) and smallest with 10L (Z). The same goes for respondent E: the most extreme category is chosen for 3L whereas a less severe category is scored on 5L. As mentioned, these location effects may also apply to group means, potentially leading to misclassification, especially when the group is rather homogeneous. Random error will increase if the mass of observations of a group is close to a cut-point of the scale such as location D, and may then have a strong impact on a crude scale such as 3L, but may only have a small effect on a more refined scale such as 5L, and even less on 10L. Generally, we assume that more levels theoretically will lead to less measurement bias.

With regard to factor 2, specific modelling outcomes on the intercept and dimension coefficients and the use of interaction terms such as the N3 term (representing whether any dimension is at level 3) will affect the resulting utility distributions and may subsequently affect discriminatory power. To explore the role of these modelling effects we studied the impact of altering the models (based on the original valuation data) by performing a sensitivity analysis in which we excluded the N3 term for two 3L value sets (The Netherlands, UK).

We explored the role of factors 1–3 both numerically and graphically. The point of departure was the LSS of the descriptive data, both by dimension and summed over all dimensions. From the LSS, difference scores between 3L and 5L were calculated by condition. We investigated how various value set characteristics contributed to discriminatory power results using tLSS (LSS transformed to a 0–1 scale) as a reference.

As a way of disentangling the intertwined effects of various factors affecting discriminatory power, we performed a multiple regression analysis with the F statistic and AUROC as dependent variables and the following variables representing value set or descriptive system characteristics as independent variables: intercept (continuous), modelled range (continuous), N3 (continuous, we included only N3 since this was the most prominent interaction term), version (with 3L as reference) and country (with Canada as reference).

3 Results

3.1 3L and 5L Value Sets for Seven Countries

There were substantial differences in the models across value sets (Table 1). For most countries the modelled 5L value range was smaller than that for the 3L, with the exception of China and The Netherlands. If 5L value sets included an intercept, its size was much smaller than 3L (except for South Korea where the intercept was 0.050 for 3L and 0.096 for 5L). The 'upper gap' between the value for 11111 and the second best health state was reduced quite substantially in 5L, ranging from a 0.02 reduction for The Netherlands and 0.04 for Spain to 0.09 for Japan and 0.14 for Canada, with South Korea as the exception (0.09 for 3L and 0.12 for 5L). Note that for Canada the upper gap was only 0.02 for 5L, because the value for 11111 was set at 0.949 (1 minus the intercept). Five countries included the N3 term in their 3L model, while for 5L only two countries used a similar interaction term (Canada and South Korea). Considerable variation was apparent in the model coefficients indicating the utility value decrement ('disutility') of dimensions, with mobility showing the highest decrements for level 3 (3L) for Canada, China, Japan, South Korea and Spain and for level 5 (5L) for China, Japan and South Korea. Pain/discomfort had the highest decrement for level 3 (3L) for the UK and The Netherlands and for level 5 (5L) for Canada, England and Spain. Anxiety/depression showed the second largest disutility in 5L for Canada, England, Japan and Spain and the largest for The Netherlands. For The Netherlands, both 3L and 5L value sets include large disutility values for anxiety/depression.

Figure 9 (Appendix) depicts the distribution plots for all possible values for the 3L and 5L value sets. Note that these plots are ranked by utility value for 3L and 5L separately, implying that 'comparable' health states such as 21111 for 3L and 31111 for 5L can be at different positions on the common utility space (X-axis). For England/UK and Spain, most 3L index values were concentrated at a much lower segment of the utility scale when compared to 5L, while for China it was vice versa, although to a lesser extent.

3.2 Distributional Analyses of 3L and 5L Utility Values

The descriptive final dataset consisted of 3L and 5L health profile data for 3467 respondents, with the smallest and

Table 2 Characteristics of descriptive EQ-5D data for nine condition groups

Condition groups	*N*	Countries	Mean age (years)	% female	Mean EQ VAS (SD)	Ceiling 3L (% 11111)	Ceiling 5L (% 11111)	Floor 3L (% 33333)	Floor 5L (% 55555)
Healthy population									
Students	443	Poland	22	79	79 (16)	47.0	34.3	0	0
Mild disease									
Diabetes mellitus	271	Denmark, England	52	48	74 (20)	33.6	28.8	0	0
Liver disease	588	Italy	56	36	70 (21)	38.6	35.7	0	0
Moderate/severe disease									
Cardiovascular disease	251	England, Scotland	67	46	61 (21)	13.2	8.0	0	0.4
Stroke	582	England, Poland	68	47	52 (26)	7.0	6.2	2.8	1.9
Asthma/COPD	342	England, Scotland	67	52	58 (21)	8.5	7.0	0	0
RA/arthritis	367	Denmark, England, Scotland	61	52	63 (21)	6.5	1.9	0	0
Depression	250	England	42	56	62 (21)	12.0	6.4	0	0
Personality disorder	373	The Netherlands	32	67	59 (18)	15.8	13.3	0	0
Total	3467	6 countries	52	53	64 (23)	20.5	16.1	0.5	0.3

3L EQ-5D-3L, *5L* EQ-5D-5L, *COPD* chronic obstructive pulmonary disease, *EQ VAS* EQ-5D visual analogue scale, *RA* rheumatoid arthritis, *SD* standard deviation

largest condition groups being depression ($n = 250$) and liver disease ($n = 588$), respectively (Table 2). The ceiling was always lower in 5L, ranging from a difference between 3L and 5L of 0.8% (stroke) to 12.7% (students). Floor effects were negligible.

Figure 2 depicts the empirically observed utility values for all countries. The 5L distributions are smoother and more evenly distributed than those for 3L. The 3L value distributions often show clusters and discontinuities across the entire range of the scale. Due to the intercept for 3L there is a large upper gap for Japan and Canada, and to a slightly lesser extent for The Netherlands and the UK. The 5L country-specific distributions look rather similar despite the model heterogeneity, although for South Korea and Japan the effect of the intercept is also clearly visible. While for England and Spain most possible 3L utility values (Fig. 9, Appendix) were concentrated at a much lower segment of the scale than 5L, the observed values did not show this pattern.

The non-parametric Shannon's H' and J' indices numerically reflected the graphical results (Table 3). For all comparisons, Shannon's H' was much higher for 5L and Shannon's Evenness J' index also was consistently higher for 5L. After subdivision into 0.05 utility space categories 5L clearly showed substantially higher values than 3L for both indices in all countries, establishing better distributional evenness for 5L overall.

Figure 3 shows the observed country-specific mean utility values for each condition group (means and SDs are listed in Table 6, Appendix). The presentation as a line graph was chosen to facilitate pattern comparison between 3L and 5L. Overall, the same ranking of average utilities per condition group across countries is visible in the figure and also a strong similarity of utilities with tLSS (showing only descriptive 3L–5L differences). Two patterns are visible: between-country valuation effects appeared larger than 3L versus 5L effects (judging from the scale differences between countries), and 3L–5L utility differences did not seem to add very much to the difference based on tLSS between 3L and 5L. For mild conditions 5L SDs were generally smaller, except for England/UK and Spain where SDs in 5L were smaller overall. Two countries displayed close to identical 3L and 5L condition group means (Canada and Japan). The other countries and tLSS values generally indicated an upward or downward shift. The UK showed a universal upward shift of 5L, South Korea a downward shift, the remaining countries (China, The Netherlands and Spain) showed a general shift plus a modifying effect in four conditions: CVD, stroke, asthma/chronic obstructive pulmonary disease (COPD) and

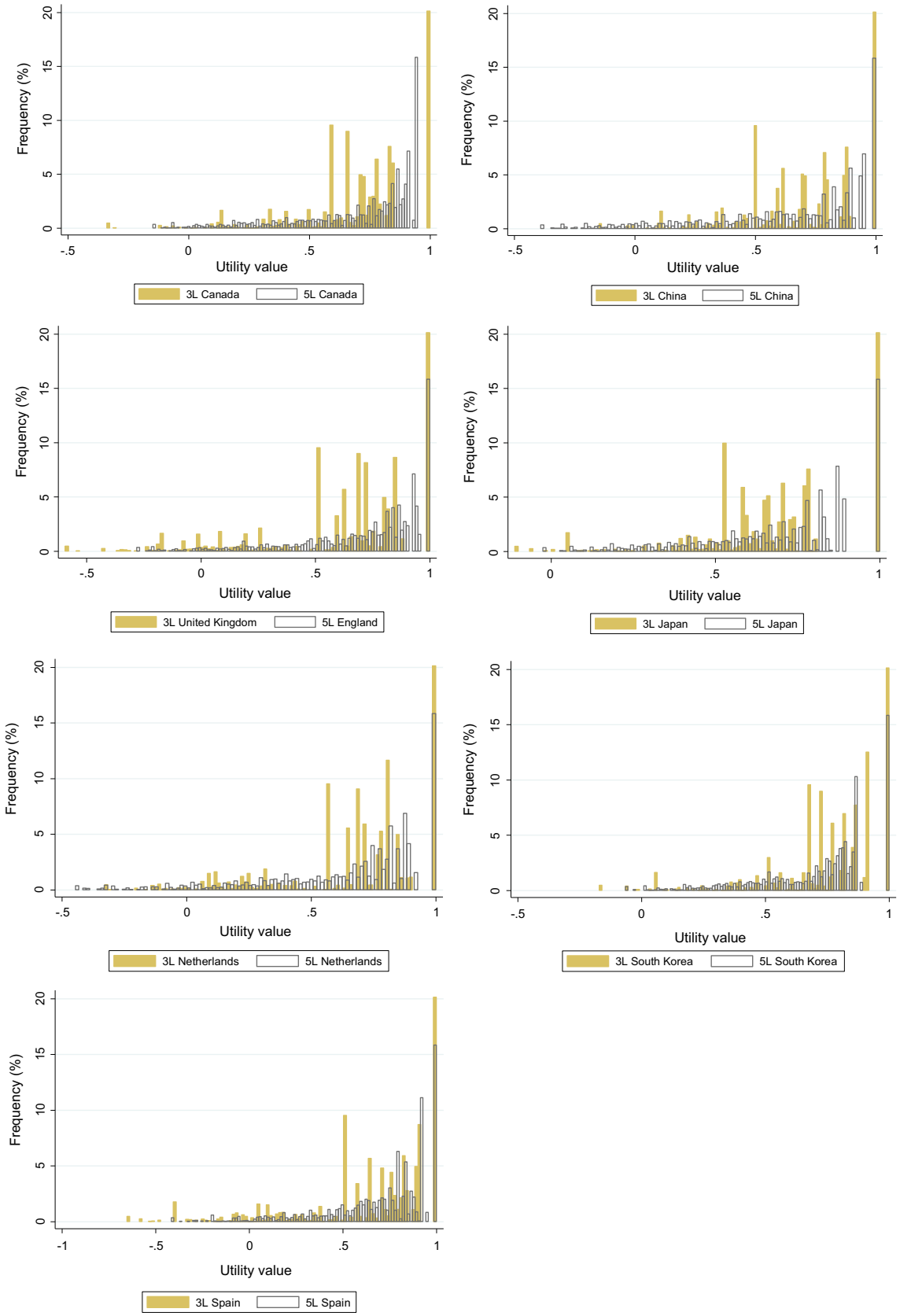

Fig. 2 Histograms of observed utility values based on value sets from seven countries, all condition groups combined. *3L* EQ-5D-3L, *5L* EQ-5D-5L

Table 3 Distributional evenness (Shannon's indices) of EQ-5D-3L and EQ-5D-5L utility values from seven countries: all condition groups combined

	Canada				China				England/UK				Japan			
	No. categories	Categories used	H'	J'	No. of categories	Categories used	H'	J'	No. of categories	Categories used	H'	J'	No. of categories	Categories used	H'	J'
All																
3L	236	123	4.79	0.61	223	118	4.76	0.61	237	121	4.78	0.61	239	123	4.78	0.61
5L	2978	668	7.21	0.62	2139	628	7.17	0.65	1569	562	6.99	0.66	3125	675	7.22	0.62
Bins 0.05[a]																
3L	27	22	3.38	0.71	23	20	3.59	0.79	32	30	3.79	0.76	23	21	3.39	0.75
5L	23	22	3.61	0.80	28	28	4.07	0.85	26	26	3.91	0.83	21	20	3.77	0.86

	The Netherlands				South Korea				Spain			
	No. of categories	Categories used	H'	J'	No. of categories	Categories used	H'	J'	No. of categories	Categories used	H'	J'
All												
3L	228	118	4.77	0.61	229	119	4.77	0.61	229	122	4.78	0.61
5L	2000	582	7.11	0.65	1224	504	6.88	0.67	3125	675	7.22	0.62
Bins 0.05[a]												
3L	27	25	3.55	0.75	24	23	3.50	0.76	33	33	3.95	0.78
5L	29	29	4.07	0.84	22	21	3.64	0.82	29	29	4.00	0.82

3L EQ-5D-3L, *5L* EQ-5D-5L, *H'* absolute informativity, *J'* relative informativity

[a]The utility scale was subdivided in categories ('bins') with 0.05 of utility space each

rheumatoid arthritis (RA)/arthritis, which may have been caused by location effects.

3.3 Discriminatory Performance of 5L Versus 3L

Both 3L and 5L distinguished well between the healthy and the disease groups as well as between mild and moderate/severe condition groups for all country-specific value sets. All comparisons resulted in statistically significant results. However, performance in terms of relative efficiency varied noticeably across version (3L/5L), value set (country and model effects) and the condition groups compared. Generally, 3L performed better in the healthy–disease comparisons while 5L performed better comparing mild and moderate/severe conditions (Fig. 4). Japanese and Dutch 5L value sets performed better overall while Canadian and Chinese 3L value sets performed better overall. The bootstrap analysis showed that although most significant results were quite robust, some were borderline significant while others were borderline non-significant.

The results for the AUROC analysis generally supported the relative efficiency results (Fig. 5), with 3L showing a better performance in the healthy–disease comparison, and 5L in the mild versus moderate/severe comparisons. However, overall results showed a significantly better performance for 5L over 3L when compared to the relative efficiency results, except for Japan.

3.4 Exploration of Factors Affecting Discriminatory Power

For the exploratory analysis we initially focused on the descriptive data, comparing LSS by dimension. Table 4 shows a pronounced shift effect between 3L and 5L (LSS by dimension recoded to no problems = 0; 3L on the same scale as 5L). A standardised difference score (Δ) was calculated, adjusting for sample size. For almost all condition groups and all dimensions, a shift to less reported health problems on 5L when compared with 3L occurred, except for mobility, where 5L represents more health problems for five condition groups due to 'confined to bed' barely being endorsed in 3L. The sum of the standardised differences scores shows that over all five dimensions the 3L–5L difference (shift) was smallest for the healthy population (28.4) and largest for liver disease (75.0). Level distributions by dimension for the pooled dataset graphically depict this main trend (Fig. 6). The shift was mainly caused by the very large proportion of respondents scoring level 2 on 3L who scored a level 2 or level 3 on 5L (average 85% over dimensions), leaving a very small proportion scoring level 4 on 5L. For pain/discomfort and anxiety/depression this also occurred at the extreme end of the scale, with a larger proportion scoring level 3 on 3L who scored level 4 on 5L rather than level 5. These observations translate into the conclusion that 3L as a scale tended to overestimate health problems when compared with 5L.

Overall, 3L resulted in higher relative efficiency ratios for the healthy–disease comparison whereas 5L performed better for the mild versus moderate/severe comparisons. Figure 7 provides an example explaining this trend. Using tLSS as reference, 3L has overall lower average values than 5L, but as mentioned earlier the differences for the healthy population were smallest, while for the other condition groups they were larger, resulting in a larger difference in means between the healthy and disease groups for 3L (X) than for 5L (Y), reflected in higher F statistics for 3L. For the mild disease showing the most pronounced results on relative efficiency (liver disease), the descriptive difference between 3L and 5L was largest (Table 4). Here the difference pattern was reversed, as indicated at the foot of Fig. 7. The difference in means between the mild and moderate/severe diseases was larger for 5L (Y) than for 3L (X), resulting in higher discriminatory power for 5L.

When exploring 3L–5L differences of the country-specific utilities, various model characteristics emerged as important underlying factors. A large intercept generally results in a lower mean and increased variance around the mean. The net effect on the F statistic is difficult to predict since both the difference of means and the standard error of the difference are affected. Overall, we detected a negative effect on discriminatory power, exemplified by the very large 3L intercept of Japan, leading to inferior performance when compared to 5L. Second, an effect of the use of model interaction terms was visible. The large N3 terms for the UK, Spain (and to a lesser extent for The Netherlands) appeared to negatively influence discriminatory power, caused by a substantial increase in variance. Note particularly that the Canadian 3L set did not contain an N3 term, but the 5L did include an 'N4 or N5' term which might have contributed to poorer discriminatory performance of 5L. Partly caused by the N3 term, but also due to other characteristics of 3L value sets, clusters and gaps occurred in the utility distributions, especially in the moderate to severe region (0–0.5), whereas 5L employed the utility scale more efficiently, resulting in smoother distributions. The histograms for the separate condition group comparisons demonstrate that the modelled range of a given value set bore no relation to the F statistic results. Instead, the use of the scale was decisive (as also shown in Fig. 2). One example is liver disease: while the modelled range for 5L in Canada and Japan was much smaller than for 3L, the available value range was being used much more frequently and efficiently in 5L, contributing to higher discriminatory power in 5L.

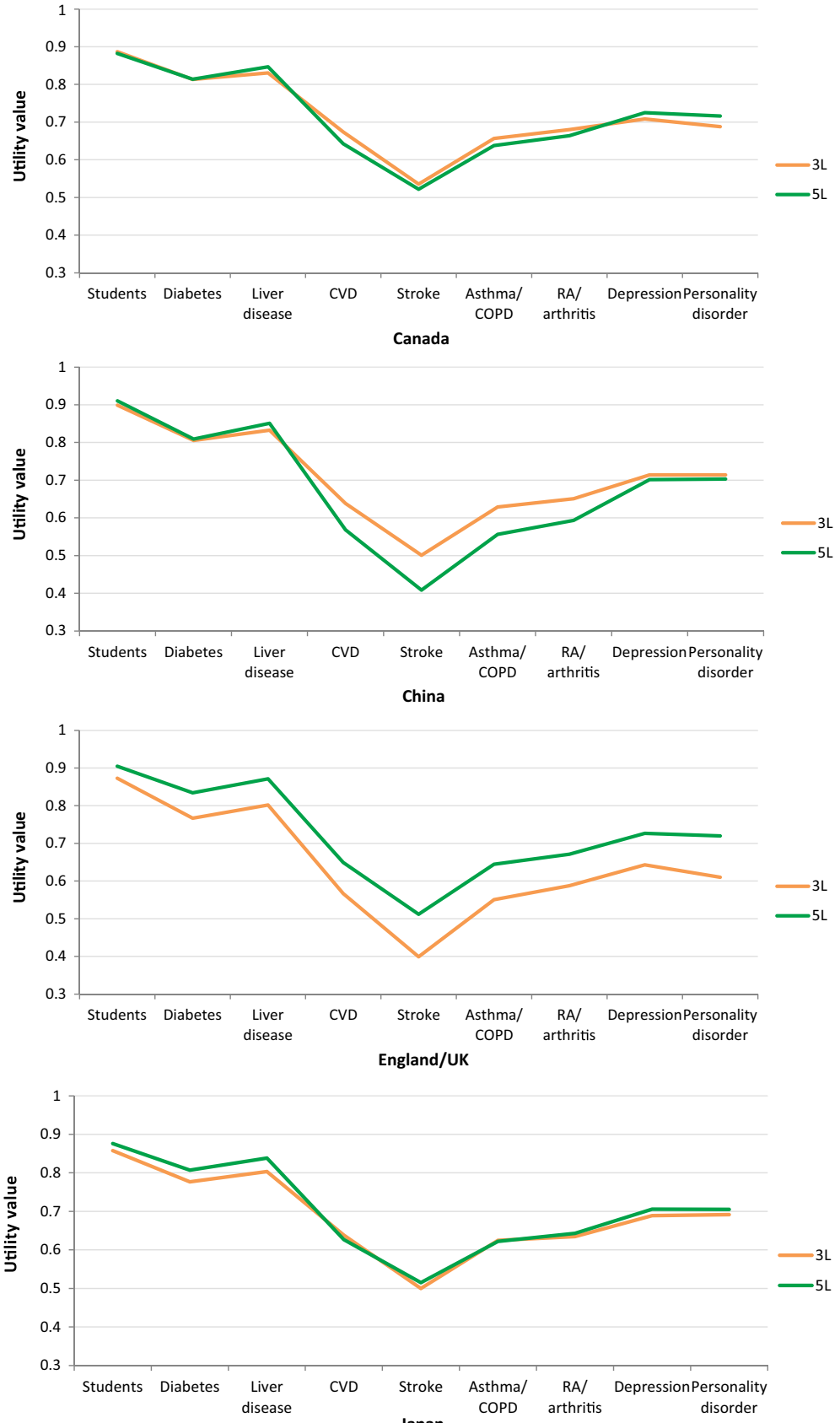

Fig. 3 Mean 3L and 5L utility value per condition group for seven countries and the transformed Level Sum Score. *3L* EQ-5D-3L, *5L* EQ-5D-5L, *COPD* chronic obstructive pulmonary disease, *CVD* cardiovascular disease, *RA* rheumatoid arthritis

Fig. 3 continued

	Canada			China			England/UK			Japan			Netherlands			South Korea			Spain			tLSS		
	F ratio	95% CI		F ratio	95% CI		F ratio	95% CI		F ratio	95% CI		F ratio	95% CI		F ratio	95% CI		F ratio	95% CI		F ratio	95% CI	
Healthy vs disease																								
Healthy vs diabetes	0.94	0.55	1.34	0.87	0.63	1.10	0.69	0.43	0.95	0.89	0.57	1.22	0.91	0.46	1.35	0.51	0.34	0.67	0.52	0.35	0.69	0.77	0.58	0.97
Healthy vs liver disease	0.53	0.25	0.82	0.64	0.42	0.87	0.40	0.16	0.65	0.59	0.29	0.90	0.39	0.08	0.70	0.21	0.06	0.36	0.30	0.14	0.46	0.56	0.37	0.75
Healthy vs cardiovascular dis.	1.10	0.91	1.29	1.03	0.89	1.16	1.02	0.86	1.18	1.31	1.11	1.51	1.20	1.00	1.40	0.85	0.73	0.98	0.88	0.76	1.00	0.99	0.87	1.12
Healthy vs stroke	1.00	0.90	1.09	0.84	0.78	0.91	0.96	0.88	1.03	1.14	1.05	1.23	0.94	0.86	1.03	1.04	0.96	1.12	1.00	0.93	1.07	0.81	0.75	0.87
Healthy vs asthma/COPD	0.89	0.75	1.02	0.88	0.78	0.99	0.91	0.79	1.03	1.00	0.87	1.13	1.11	0.94	1.27	0.77	0.67	0.87	0.84	0.74	0.94	0.86	0.77	0.95
Healthy vs RA/arthritis	0.82	0.69	0.96	0.83	0.73	0.93	0.87	0.74	0.99	0.96	0.81	1.10	1.18	1.00	1.36	0.82	0.71	0.92	0.87	0.77	0.98	0.85	0.75	0.95
Healthy vs depression	0.74	0.58	0.90	0.82	0.68	0.96	0.96	0.78	1.15	1.11	0.89	1.33	0.95	0.76	1.15	0.84	0.69	0.99	0.98	0.82	1.14	0.87	0.74	1.01
Healthy vs personality dis.	0.76	0.62	0.90	0.99	0.83	1.14	1.06	0.87	1.24	1.31	1.09	1.53	1.08	0.90	1.26	0.91	0.78	1.05	0.95	0.80	1.10	1.04	0.89	1.20
Mild vs moderate/severe																								
Diabetes vs cardiovascular dis.	1.14	0.82	1.45	1.17	0.92	1.42	1.19	0.86	1.52	1.44	1.11	1.78	1.25	0.88	1.61	1.04	0.83	1.25	1.11	0.85	1.38	1.21	0.96	1.45
Diabetes vs stroke	0.99	0.86	1.12	0.91	0.81	1.00	1.05	0.93	1.17	1.16	1.04	1.28	0.96	0.83	1.08	1.04	0.93	1.15	1.07	0.95	1.18	0.91	0.83	1.00
Diabetes vs asthma/COPD	0.90	0.69	1.11	1.03	0.84	1.22	1.03	0.79	1.27	1.08	0.89	1.28	1.12	0.85	1.38	0.94	0.76	1.11	1.06	0.84	1.28	1.05	0.88	1.22
Diabetes vs RA/arthritis	0.84	0.63	1.06	0.97	0.76	1.17	1.04	0.77	1.31	0.99	0.79	1.20	1.31	0.94	1.68	1.08	0.86	1.30	1.25	0.94	1.55	1.05	0.85	1.24
Diabetes vs depression	0.58	0.31	0.84	0.82	0.49	1.14	1.13	0.59	1.66	1.18	0.74	1.62	0.88	0.57	1.18	1.13	0.71	1.56	1.75	0.70	2.79	1.05	0.68	1.43
Diabetes vs personality dis.	0.54	0.34	0.75	0.91	0.57	1.25	1.00	0.65	1.35	1.38	0.94	1.83	0.92	0.66	1.18	1.15	0.80	1.49	1.23	0.77	1.68	1.29	0.88	1.69
Liver dis. vs cardiovascular dis.	1.61	1.23	2.00	1.35	1.10	1.60	1.42	1.09	1.75	1.57	1.28	1.86	1.68	1.25	2.10	1.32	1.06	1.58	1.23	0.97	1.49	1.35	1.12	1.57
Liver dis. vs stroke	1.18	1.05	1.30	1.00	0.90	1.09	1.16	1.04	1.28	1.21	1.11	1.32	1.16	1.03	1.29	1.18	1.07	1.28	1.13	1.02	1.24	1.00	0.91	1.08
Liver dis. vs asthma/COPD	1.27	1.03	1.51	1.20	1.01	1.39	1.23	1.00	1.47	1.24	1.05	1.43	1.48	1.17	1.79	1.20	1.00	1.40	1.18	0.96	1.39	1.19	1.03	1.36
Liver dis. vs RA/arthritis	1.25	1.01	1.50	1.16	0.96	1.35	1.27	1.01	1.53	1.17	0.97	1.37	1.77	1.35	2.19	1.39	1.12	1.66	1.36	1.09	1.64	1.20	1.02	1.38
Liver dis. vs depression	1.00	0.71	1.30	1.07	0.78	1.35	1.41	0.97	1.85	1.33	1.00	1.66	1.30	0.93	1.66	1.58	1.10	2.07	1.80	1.19	2.40	1.23	0.93	1.52
Liver dis. vs personality dis.	0.92	0.69	1.15	1.16	0.88	1.45	1.27	0.95	1.58	1.50	1.16	1.84	1.25	1.01	1.49	1.56	1.18	1.94	1.38	1.01	1.75	1.43	1.11	1.75

Fig. 4 Observed relative efficiency of 5L over 3L using the *F* statistic ratio. Green cells indicate a significant *F* ratio showing better discriminatory power for 5L, orange cells for 3L (95% CI, 3000 bootstrap samples). *3L* EQ-5D-3L, *5L* EQ-5D-5L, *CI* confidence interval, *COPD* chronic obstructive pulmonary disease, *dis.* disease/disorder, *RA* rheumatoid arthritis, *tLSS* transformed Level Sum Score

The sensitivity analysis, exploring the effect of excluding the N3 term for the 3L value sets for the UK and The Netherlands, confirmed this pattern. Discriminatory power clearly increased for 3L as the number of significant results in favour of 3L increased from 3 versus 4 (3L vs. 5L) to 9 versus 2 for the UK and from 1 versus 10 (3L vs. 5L) to 4 versus 5 for The Netherlands (Figs. 4, 8). Descriptive statistics showed that this was mainly due to lower levels of dispersion for the models without N3.

The results from the regression on the *F* statistic were a way to validate our interpretation of the relative impact of various factors. Our findings were confirmed (Table 5), demonstrating a significant negative coefficient for 5L for the healthy–disease comparison and a positive coefficient for the mild versus moderate/severe comparisons. The modelled range was not significant for both types of comparison, confirming that the modelled range did not significantly impact upon the *F* statistic. The intercept showed a significant negative value for the healthy–disease comparison, implying that the use of an intercept decreases discriminatory power. The N3 term did not show a significant impact. It is of interest to note that the value sets for the Asian countries resulted in higher discriminatory power than for the non-Asian countries. Using the AUROC as the independent variable showed similar patterns, where the intercept consistently showed a negative effect, as did the N3 term for the mild versus moderate/severe comparison. The modelled range, however, appeared to contribute to discriminatory power.

4 Discussion

Our study showed that the 5L version of the EQ-5D instrument was in many respects superior to the original 3L version. By separating the performance of description and valuation, it became clear that these benefits mainly arise from the improved descriptive system: 5L was superior in terms of the distributional evenness, efficiency of scale use and the face validity of the resulting distributions, leading to an increase in sensitivity and precision in health status measurement. Refinement of 5L was not offset by more error, neither in terms of description nor in valuation.

The fewer cut-points of 3L (two instead of four in 5L) and the position of the cut-points relative to the true latent scale position could be the main drivers of the larger error component in 3L. The net effect was that 3L overestimated self-reported health problems by displaying 'moderate problems' where the true latent score most often was more likely to be in between 'no problems' and 'moderate', i.e. 3L suffered from a rather high cut-point between levels 1 and 2 (and for pain/discomfort and anxiety/depression also between levels 2 and 3). The impact of this artefact of the descriptive system decreased when the number of levels increased. The fact that 3L systematically overestimated reported health problems was unexpected, as for certain condition groups (e.g. in severe patients) the level of reported health problems between 3L and 5L could have been similar, or 3L could have led to the reverse finding. i.e. an underestimation of health problems. The overestimation of 3L was not trivial and affected any difference score when making comparisons: differences may be underestimated or overestimated, such as the overestimation of the difference between a healthy population and

	Canada			China			England/UK			Japan			Netherlands			South Korea			Spain			tLSS		
	auc ratio	95% CI		auc ratio	95% CI		auc ratio	95% CI		auc ratio	95% CI		auc ratio	95% CI		auc ratio	95% CI		auc ratio	95% CI		auc ratio	95% CI	
Healthy vs disease																								
Healthy vs diabetes	0.98	0.94	1.01	0.96	0.93	0.99	0.93	0.90	0.97	0.93	0.89	0.96	0.99	0.95	1.03	0.96	0.93	0.99	0.92	0.89	0.95	0.96	0.93	0.99
Healthy vs liver disease	0.92	0.89	0.95	0.93	0.90	0.96	0.92	0.88	0.95	0.92	0.89	0.95	0.94	0.90	0.97	0.91	0.88	0.94	0.90	0.87	0.94	0.93	0.90	0.96
Healthy vs cardiovascular dis.	1.01	0.99	1.03	1.01	0.99	1.03	1.00	0.98	1.03	1.00	0.98	1.02	1.04	1.01	1.07	0.99	0.97	1.01	0.99	0.96	1.01	1.01	0.99	1.03
Healthy vs stroke	0.98	0.97	0.99	0.98	0.97	0.99	0.98	0.97	1.00	0.99	0.98	1.00	0.99	0.98	1.00	0.98	0.96	0.99	0.97	0.96	0.98	0.99	0.98	1.00
Healthy vs asthma/COPD	0.99	0.97	1.01	0.99	0.97	1.01	0.97	0.95	0.99	0.97	0.95	0.99	1.03	1.00	1.05	0.97	0.95	0.99	0.95	0.93	0.97	0.99	0.97	1.00
Healthy vs RA/arthritis	1.02	1.00	1.04	1.00	0.98	1.02	0.98	0.96	1.00	0.98	0.96	1.00	1.06	1.03	1.08	1.00	0.98	1.02	0.97	0.95	0.98	1.01	0.99	1.02
Healthy vs depression	0.98	0.95	1.01	1.02	0.99	1.05	1.03	0.99	1.06	1.03	1.00	1.06	1.01	0.98	1.05	0.98	0.95	1.01	1.01	0.98	1.04	1.01	0.98	1.04
Healthy vs personality dis.	0.98	0.96	1.00	1.03	1.01	1.06	1.03	1.00	1.06	1.04	1.02	1.06	1.03	1.01	1.06	0.98	0.96	1.00	1.00	0.97	1.02	1.02	1.00	1.04
Mild vs moderate/severe																								
Diabetes vs cardiovascular dis.	1.03	1.00	1.06	1.04	1.01	1.06	1.04	1.02	1.07	1.04	1.02	1.07	1.05	1.02	1.07	1.02	0.99	1.04	1.03	1.00	1.05	1.04	1.01	1.06
Diabetes vs stroke	0.99	0.97	1.00	1.00	0.98	1.01	1.01	0.99	1.02	1.01	1.00	1.02	1.00	0.98	1.01	0.99	0.98	1.01	1.00	0.98	1.01	1.00	0.99	1.01
Diabetes vs asthma/COPD	1.01	0.99	1.04	1.02	1.00	1.04	1.02	0.99	1.04	1.01	0.99	1.04	1.03	1.00	1.06	1.00	0.98	1.02	1.00	0.98	1.02	1.02	1.00	1.04
Diabetes vs RA/arthritis	1.05	1.02	1.07	1.04	1.01	1.06	1.03	1.00	1.05	1.02	1.00	1.05	1.06	1.03	1.09	1.03	1.01	1.06	1.02	1.00	1.05	1.04	1.02	1.06
Diabetes vs depression	1.00	0.96	1.03	1.05	1.01	1.08	1.08	1.04	1.11	1.08	1.05	1.11	1.00	0.97	1.04	1.02	0.98	1.05	1.08	1.05	1.12	1.03	1.00	1.06
Diabetes vs personality dis.	1.00	0.97	1.03	1.07	1.04	1.10	1.07	1.04	1.10	1.11	1.08	1.15	1.02	0.99	1.05	1.03	1.00	1.06	1.05	1.02	1.09	1.07	1.04	1.09
Liver dis. vs cardiovascular dis.	1.06	1.04	1.08	1.05	1.02	1.07	1.05	1.03	1.07	1.04	1.02	1.06	1.07	1.05	1.10	1.05	1.02	1.07	1.04	1.02	1.06	1.05	1.03	1.07
Liver dis. vs stroke	1.01	1.00	1.02	1.00	0.99	1.02	1.02	1.01	1.03	1.01	1.00	1.02	1.02	1.01	1.03	1.01	1.00	1.02	1.01	1.00	1.02	1.01	1.00	1.02
Liver dis. vs asthma/COPD	1.04	1.03	1.06	1.03	1.01	1.05	1.02	1.00	1.04	1.01	1.00	1.03	1.06	1.04	1.08	1.02	1.01	1.04	1.00	0.99	1.02	1.03	1.01	1.04
Liver dis. vs RA/arthritis	1.08	1.06	1.10	1.04	1.03	1.06	1.03	1.01	1.05	1.02	1.01	1.04	1.09	1.07	1.12	1.06	1.04	1.08	1.03	1.01	1.05	1.05	1.03	1.07
Liver dis. vs depression	1.04	1.01	1.07	1.06	1.03	1.09	1.08	1.05	1.11	1.07	1.04	1.10	1.05	1.02	1.07	1.06	1.03	1.09	1.09	1.06	1.12	1.05	1.03	1.08
Liver dis. vs personality dis.	1.04	1.02	1.06	1.08	1.05	1.10	1.07	1.05	1.10	1.10	1.08	1.13	1.06	1.04	1.08	1.06	1.04	1.08	1.05	1.03	1.08	1.07	1.05	1.10

Fig. 5 Observed relative efficiency of 5L over 3L using the AUROC. Green cells indicate a significant AUROC comparison showing better discriminatory power for 5L, orange cells for 3L (95% CI, 3000 bootstrap samples). *3L* EQ-5D-3L, *5L* EQ-5D-5L, *auc* area under the curve, *AUROC* area under the receiver-operating characteristics curve, *CI* confidence interval, *COPD* chronic obstructive pulmonary disease, *dis.* disease/disorder, *RA* rheumatoid arthritis, *tLSS* transformed Level Sum Score

most patient groups in our study. This disadvantage of 3L has further consequences in the valuation procedure: if respondents were to value a 3L health profile with moderate problems, and no information was available to inform them that this would actually (empirically) refer to a mix of moderate and predominantly milder health problems, then the disutility would also be overestimated.

When adding utility values to the descriptive data, it was apparent that although absolute utility means varied substantially, 3L–5L differences were not very large, as usually a constant upward or downward shift was observed. Nevertheless, this study showed that seemingly small differences do affect results in discriminating between groups, and are likely to also affect responsiveness. A more precise

Table 4 EQ-5D-3L versus EQ-5D-5L Level Sum Score by dimension[a] and condition group, including a standardized level shift indicator ($\Delta = 3L - 5L$ adjusted for sample size)[b]

Condition groups	Mobility			Self-care			Usual activities			Pain/discomfort			Anxiety/depression			Sum (Δ)
	3L	5L	Δ	3L	5L	Δ	3L	5L	Δ	3L	5L	Δ	3L	5L	Δ	
Healthy population																
Students	18	19	− 0.2	2	2	0.0	92	91	0.2	294	210	19.0	404	362	9.5	28.4
Mild disease																
Diabetes mellitus	172	185	− 4.8	92	59	12.2	230	181	18.1	314	255	21.8	180	145	12.9	60.1
Liver disease	298	236	10.5	140	87	9.0	376	305	12.1	480	368	19.0	552	409	24.3	75.0
Moderate/severe disease																
Cardiovascular disease	366	396	− 12.0	236	195	16.3	434	401	13.1	406	368	15.1	278	238	15.9	48.6
Stroke	1140	1128	2.1	1042	970	12.4	1280	1191	15.3	1030	950	13.7	978	847	22.5	66.0
Asthma/COPD	518	562	− 12.9	298	267	9.1	586	551	10.2	624	530	27.5	374	305	20.2	54.1
RA/arthritis	524	526	–0.5	270	225	12.3	588	522	18.0	730	657	19.9	322	287	9.5	59.1
Depression	172	157	6.0	92	75	6.8	288	233	22.0	330	288	16.8	466	409	22.8	74.4
Personality disorder	116	86	8.0	44	27	4.6	576	555	5.6	448	378	18.8	816	715	27.1	64.1

3L EQ-5D-3L, *5L* EQ-5D-5L, *COPD* chronic obstructive pulmonary disease, *LSS* Level Sum Score, *RA* rheumatoid arthritis, *Δ* difference

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

[b]The difference between LSS by dimension (3L − 5L), adjusted for sample size: '28.4' means that the average level shift per respondent was 0.284

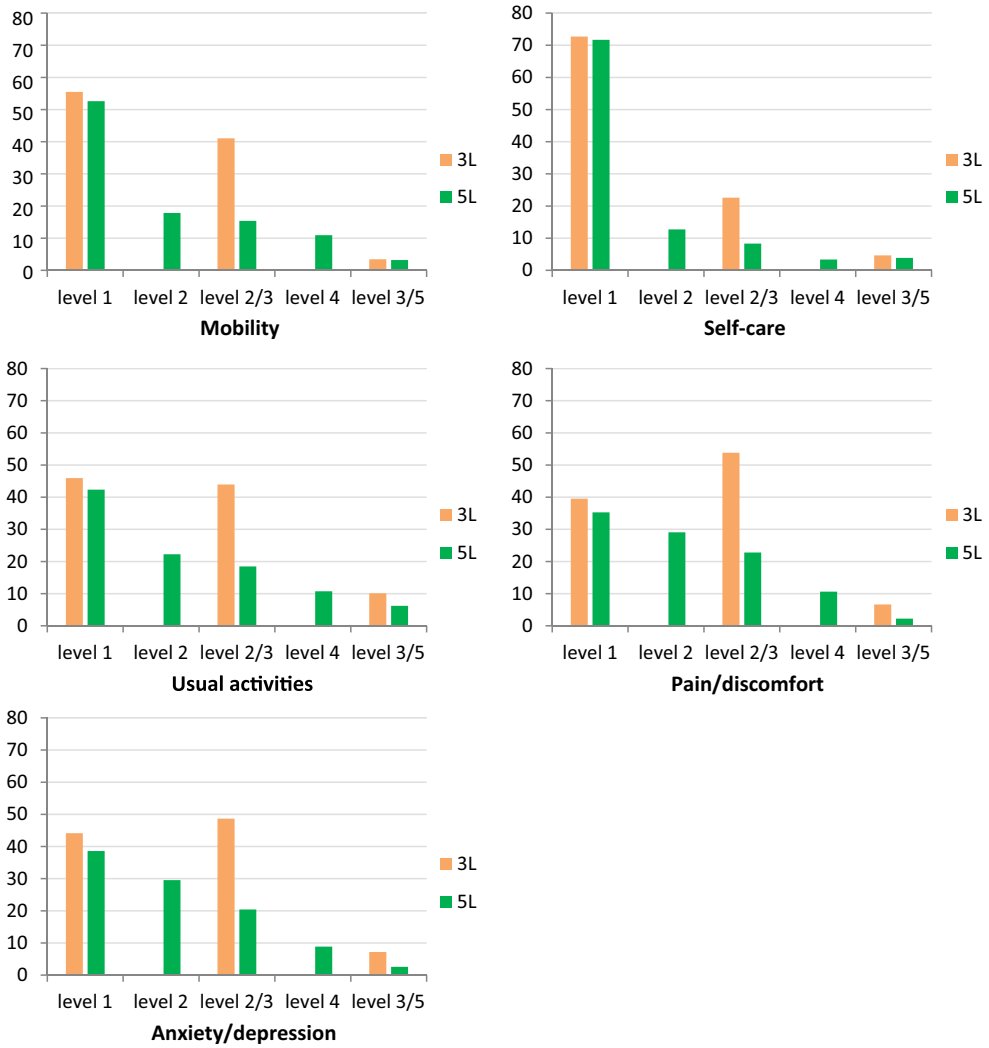

Fig. 6 Percentage of reported problems to 3L and 5L descriptive systems: all condition groups combined. *3L* EQ-5D-3L, *5L* EQ-5D-5L

discrimination between subgroups is achieved with 5L. The effect on QALY comparisons might be smaller since here it would mainly be the difference of mean utilities that would determine the outcome, with the exception being heterogeneous diseases and/or populations where the redistribution effects were non-linear (in our study CVD, stroke, asthma/COPD and RA/arthritis), where larger differences might be expected.

On the assumption that the increased number of levels in 5L led to less bias in the resulting utilities, we concluded that 3L overestimated health problems and consequently underestimated utilities when compared with 5L. This was generally observed across condition groups, but was most pronounced in liver disease (caused by a large misclassification at location D, as depicted in Fig. 1). Against our expectation, health problems in this group were apparently very mild [49], as confirmed by the high mean EQ VAS rating. A result of 3L misclassification is a biased assessment of discriminatory power that could lead to an overestimation of discriminatory power of 3L in the healthy versus disease comparisons in our study, or an underestimation of discriminatory power in the mild versus moderate/severe comparisons.

For mild conditions SDs were lower in 5L, which may be a consequence of 3L overestimation being larger in these conditions, as 5L was better equipped to capture the (very) mild skewed distribution, resulting in lower SDs. For moderate and severe condition groups, 5L SD rates were higher. Graphical and numerical (Shannon's indices) evidence clearly showed that 5L covered a much wider range of the utility scale in these condition groups and was more evenly distributed, which in our view resulted in a much better reflection of the true underlying distribution. Note also that for the UK and Spain, 3L levels of dispersion were higher overall, which was in part due to the inclusion of the N3 term.

The analysis additionally proved useful in detecting inter-country differences. The relatively poor performance of 5L in some countries may relate to the use of the initial EQ-VT version 1.0. For instance, in Canada and England

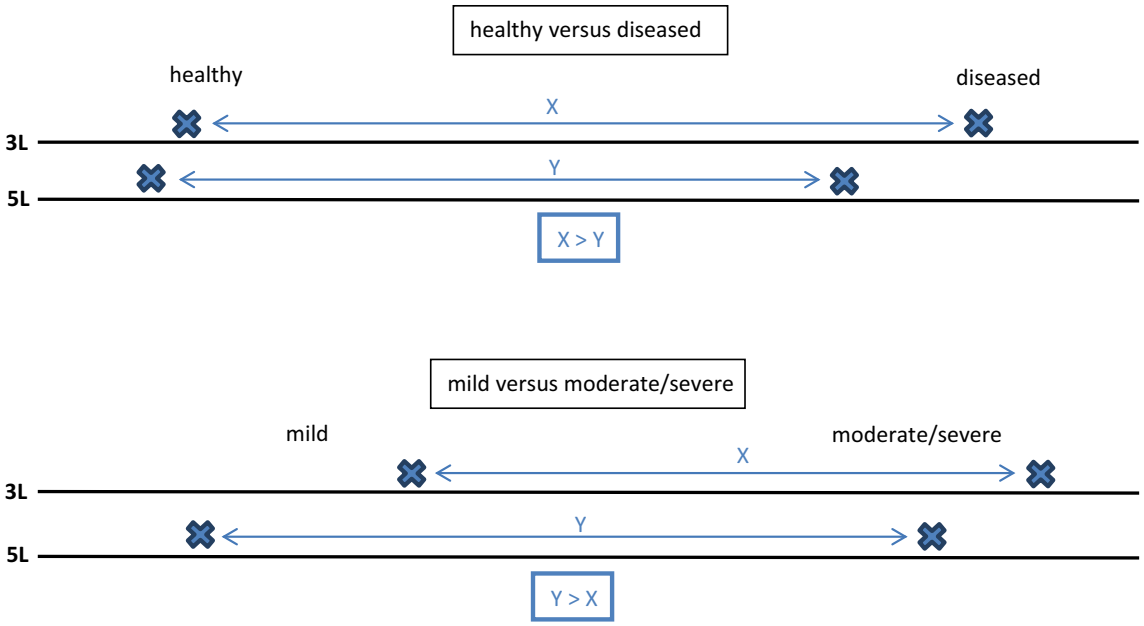

Fig. 7 Observed redistribution of latent health states from 3L to 5L if descriptive refinement increases. *3L* EQ-5D-3L, *5L* EQ-5D-5L

Fig. 8 Sensitivity analysis main effects without N3 for UK and The Netherlands: observed relative efficiency of 5L over 3L using the *F* statistic ratio. Green cells indicate a significant F ratio showing better discriminatory power for 5L, orange cells for 3L (95% CI, 3000 bootstrap samples). *3L* EQ-5D-3L, *5L* EQ-5D-5L, *CI* confidence interval, *COPD* chronic obstructive pulmonary disease, *dis.* disease/disorder, *N3* any level 3, *RA* rheumatoid arthritis

	UK withouth N3 (3L)			Netherlands without N3 (3L)		
	F ratio	95% CI		F ratio	95% CI	
Healthy vs disease						
Healthy vs diabetes	0.64	0.39	0.88	1.06	0.46	1.66
Healthy vs liver disease	0.34	0.14	0.55	0.40	0.08	0.71
Healthy vs cardiovascular dis.	0.90	0.76	1.03	1.31	1.07	1.55
Healthy vs stroke	0.81	0.74	0.88	0.93	0.83	1.03
Healthy vs asthma/COPD	0.72	0.62	0.82	1.08	0.92	1.25
Healthy vs RA/arthritis	0.67	0.57	0.76	1.14	0.95	1.33
Healthy vs depression	0.78	0.64	0.93	0.83	0.66	1.00
Healthy vs personality dis.	0.81	0.67	0.95	0.88	0.72	1.04
Mild vs moderate/severe						
Diabetes vs cardiovascular dis.	1.07	0.78	1.36	1.30	0.90	1.70
Diabetes vs stroke	0.93	0.82	1.04	0.92	0.79	1.06
Diabetes vs asthma/COPD	0.83	0.66	1.00	1.03	0.78	1.28
Diabetes vs RA/arthritis	0.77	0.58	0.96	1.14	0.82	1.46
Diabetes vs depression	0.88	0.52	1.23	0.66	0.42	0.89
Diabetes vs personality dis.	0.80	0.52	1.08	0.68	0.48	0.88
Liver dis. vs cardiovascular dis.	1.40	1.09	1.71	1.94	1.42	2.46
Liver dis. vs stroke	1.07	0.96	1.18	1.20	1.05	1.35
Liver dis. vs asthma/COPD	1.09	0.90	1.28	1.51	1.22	1.80
Liver dis. vs RA/arthritis	1.05	0.85	1.25	1.73	1.34	2.12
Liver dis. vs depression	1.24	0.90	1.59	1.11	0.82	1.41
Liver dis. vs personality dis.	1.15	0.88	1.41	1.07	0.85	1.29

very few negative values were derived, which could be caused by poor protocol compliance of the interviewers and/or a poor explanation of the worse than dead task in the composite TTO exercise. In general, the value sets for the Asian countries showed better discriminatory power than non-Asian countries. We must also accept that structural components influence preferences, with many possible underlying factors involved (e.g. culture, demographics, language, geography), which was also noted by Olsen et al. [50].

Our study rested on two unique features:

1. The development of an innovative framework to assess the performance of preference-based measures of health with varying levels of sensitivity. Note that a framework such as the COSMIN (COnsensus-based Standards for the selection of health Measurement Instruments) taxonomy only partially applies to instruments with separate descriptive and valuation components [51, 52].

Table 5 Effect of value set characteristics on discriminatory performance in terms of *F* statistic and area under the receiver-operating characteristics curve

	F statistic				AUROC			
	Coefficient	*t* value	*p* value	95% CI	Coefficient	*t* value	*p* value	95% CI
Healthy vs. disease								
Version 5L[a]	− 37.3	− 4.29	0.00	− 56.1 to − 18.5	− 0.015	− 7.20	0.00	− 0.020 to − 0.011
Country[b]								
China	60.0	3.91	0.00	26.8 to 93.1	0.000	0.00	1.00	− 0.007 to 0.007
England/UK	− 27.3	− 2.00	0.07	− 56.7 to 2.1	− 0.023	− 6.42	0.00	− 0.031 to − 0.015
Japan	73.4	3.57	0.00	29.0 to 117.8	0.013	3.75	0.00	0.005 to 0.020
The Netherlands	− 50.2	− 3.67	0.00	− 79.8 to − 20.7	− 0.037	− 7.75	0.00	− 0.047 to − 0.027
South Korea	53.3	5.54	0.00	32.5 to 74.1	0.018	6.66	0.00	0.012 to 0.024
Spain	− 8.4	− 0.47	0.65	− 47.4 to 30.6	− 0.015	− 2.39	0.03	− 0.028 to − 0.001
Intercept	− 342.5	− 3.67	0.00	− 543.9 to − 141.0	− 0.109	− 4.43	0.00	− 0.162 to − 0.056
Modelled range	50.4	1.25	0.23	− 36.8 to 137.6	0.051	3.92	0.00	0.023 to 0.079
N3	− 57.5	− 1.17	0.26	− 163.9 to 49.0	− 0.029	− 1.68	0.12	− 0.066 to 0.008
Constant	293.5	5.24	0.00	172.4 to 414.6	0.746	45.08	0.00	0.710 to 0.782
Mild vs. moderate/severe								
Version 5L[a]	27.2	8.81	0.00	20.5 to 33.8	0.023	25.02	0.00	0.021 to 0.025
Country[b]								
China	20.9	3.79	0.00	9.0 to 32.8	− 0.004	− 2.61	0.02	− 0.008 to − 0.001
England/UK	0.0	0.00	1.00	− 11.0 to 11.1	− 0.008	− 3.84	0.00	− 0.012 to − 0.003
Japan	21.1	2.66	0.02	3.9 to 38.2	0.002	0.94	0.37	− 0.003 to 0.007
The Netherlands	− 5.0	− 1.04	0.32	− 15.5 to 5.4	− 0.006	− 3.62	0.00	− 0.010 to − 0.003
South Korea	10.4	2.72	0.02	2.1 to 18.7	0.002	1.49	0.16	− 0.001 to 0.006
Spain	8.5	1.33	0.21	− 5.3 to 22.3	− 0.008	− 3.31	0.01	− 0.013 to − 0.003
Intercept	60.5	1.70	0.11	− 16.4 to 137.4	− 0.057	− 4.44	0.00	− 0.085 to − 0.029
Modelled range	− 11.1	− 0.86	0.40	− 39.0 to 16.7	0.015	4.80	0.00	0.008 to 0.022
N3	− 11.8	− 0.67	0.52	− 50.1 to 26.5	− 0.017	− 2.71	0.02	− 0.031 to − 0.004
Constant	126.6	6.98	0.00	87.4 to 165.8	0.695	159.70	0.00	0.686 to 0.705

3L EQ-5D-3L, *5L* EQ-5D-5L, *AUROC* area under the receiver-operating characteristics curve, *CI* confidence interval, *N3* any level 3

[a]With 3L as reference

[b]With Canada as reference

2. The use of a large number of published value sets 'as is' in a large multinational parallel 3L–5L dataset across nine condition groups.

Our innovative framework started with the separation of potential systematic effects in description and valuation. This enabled us to clarify hitherto poorly understood mechanisms underlying differences with a 3L versus a 5L system [19, 53]. Our study confirms some of the findings from an earlier study by Richardson et al. [23], showing that differences between utility results of different preference-based instruments are mainly attributable to the descriptive data, although a different methodological approach was followed in their study, based on parametric techniques. Our framework incorporated ceiling and floor effects, and Shannon's indices as expressions of the evenness of a distribution. Distributional characteristics were based on the straightforward assumption that we should expect normal or lognormal distributed outcomes, as commonly observed in many naturally occurring phenomena, including self-reported health [54–56]. We improved on the use of the F ratio to quantify discriminatory power, differentiating between the various underlying sources, e.g. random error, cut-point-related bias and dispersion in heterogeneous samples. The successful use of the AUROC is an example of the wide applicability of this method beyond diagnostics. This study shows only part of its potential, as described elsewhere [57, 58]. A main advantage of our framework lies in the combined strength of the distributional approaches and different methods to assess discriminatory power, enabling us to make claims of the superiority of one measure over another. Our methods make clear that 5L is better than 3L, but they could also demonstrate that a hypothetical 10L might be a poor choice.

There were some limitations that must be acknowledged for the current study. First, the condition samples were not optimal for all groups. We used a student cohort to represent a healthy population, whereas a better matched general population sample, especially in terms of age and education, would have been more suitable. Second, we cannot exclude the possibility that inter-country differences in the descriptive data existed. The condition groups were from various countries, e.g. the liver disease sample was derived from an Italian cohort, the student cohort was entirely Polish and the personality disorder sample was Dutch. The *F* statistic was a key component of our study, assuming a normal distribution. The 3L and 5L utility scores used in our study were often not normally distributed due to ceiling effects or clusters, although in the context of health measurement the key factors are similarity of the distributions rather than normality, and approximately equal-sized samples [42]. Our conclusion that 3L overestimated health problems might be challenged for the first three dimensions where level 2 of 3L (some problems) was not identical to level 3 of 5L (moderate problems), although we felt justified generalising over all five dimensions since for pain/discomfort and anxiety/depression, where all labels are identical, overestimation was largest. Finally, as our study was based on cross-sectional data, we cannot make firm conclusions about the 3L versus 5L impact on QALYs. However, in the main pharmacoeconomic application of EQ-5D (cost-utility analysis), the utilities for different health states that are modelled are typically based on cross-sectional data, often derived from different patients subgroups.

5 Conclusions

Our study has several implications. Although the 3L can be considered to be a valid measure in itself, we demonstrated that its lack of refinement did lead to more reported health problems on average when compared to a more sensitive and precise measure. We are aware that an even more refined system might reveal misclassification in 5L, but these effects will on average be much smaller. We conclude that 5L results in more precise and valid outcomes, both descriptive and in terms of valuation. The increased sensitivity and precision of 5L is likely to be generalisable to longitudinal designs, such as intervention studies. Hence, we recommend the use of 5L across applications, including economic evaluation, clinical studies and burden of disease or public health studies (e.g. for establishing population norms). Our results indicate that in situations where patient groups would experience a uniform recovery to nearly full health, 3L might artificially show a large effect. This might have led to the overestimation of QALY gains in past economic evaluations, especially in assessing the impact of drugs for mild diseases.

With regard to modelling of the utility data, it was apparent that the inclusion of an interaction term (such as N3) and an intercept would lead to undesirable distributional characteristics such as discontinuities and clusters in the utility scale and would be likely to reduce discriminatory power. It is notable that for the two countries that included an interaction term in their 5L model (Canada and South Korea), discriminatory power was not outstanding. Note that a large intercept might have been caused by misspecification of mild health states in the valuation procedure (by assigning low utility values), which could be due to interviewer effects (especially apparent in EQ-VT version 1.0) or cognitive overload in respondents. Our finding that the use of the scale was an important determinant of discriminatory performance (as opposed to the modelled range) shows that the previous preoccupation with the modelled range is not really justified [29, 50], which was also reflected in our regression results (Table 5). The use of 3L in conditions with problems with mobility could lead to severe underreporting of mobility problems. In our study COPD or CVD patients showed many reported problems in walking about on 5L, but since these respondents were not confined to bed they were restricted to score level 2 on 3L, thereby reducing its sensitivity and discriminatory power substantially. This is corroborated by results from a study among patients to receive hip replacement surgery in the UK. Not a single patient reported a level 3 problem on mobility on the 3L, whereas there were many reported problems with mobility in the Oxford Hip Score, a condition-specific measure [59]. Changing the most severe level descriptor of 3L 'confined to bed' to 'unable to walk about' in 5L appeared to be a huge improvement.

A final implication of our study includes the introduction of a powerful evaluative framework, allowing for further extension by using evidence resulting from longitudinal 3L–5L data. Our framework combines parametric (*F* statistic) with non-parametric (AUROC) methods, and may be more broadly applied than assessing granularity of the system (the number of response options), such as to investigate the impact of adding dimensions to the EQ-5D, or assessing translation effects.

The current 5L system would profit from more knowledge on the random error of descriptive data (reliability) and cut-point effects, which would also be useful in the development of any new measure. This includes investigating whether the latent scale people use when responding to the EQ-5D for self-classification is the same as when valuing hypothetical health states.

Acknowledgements The authors would like to thank two anonymous reviewers for valuable comments and suggestions.

Author Contributions MFJ led the data analysis and interpretation and was primarily responsible for drafting the manuscript. GJB devised the AUROC approach, and GJB and NL supported data analysis and interpretation and commented on and amended the draft manuscript.

Compliance with Ethical Standards

Funding This work was funded by the EuroQol Research Foundation (Grant number EQ Project 2016620).

Conflict of interest All authors (Mathieu F. Janssen, Gouke J. Bonsel and Nan Luo) are members of the EuroQol Group.

Data availability statement All data analysed in this study are stored at the central data archive of the EuroQol Research Foundation. The data are available from the EuroQol Research Foundation upon reasonable request.

Appendix

See Table 6 and Fig. 9.

Table 6 Mean EQ-5D-3L and EQ-5D-5L utility values and standard deviations by condition group for seven countries and the transformed Level Sum Score

Condition groups	*N*	Canada		China		England/UK		Japan		The Netherlands		South Korea		Spain		tLSS	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Healthy population																	
Students																	
3L	443	0.89	0.12	0.90	0.11	0.87	0.14	0.86	0.14	0.87	0.14	0.93	0.07	0.91	0.11	0.91	0.10
5L	443	0.88	0.09	0.91	0.11	0.90	0.11	0.88	0.11	0.86	0.15	0.88	0.10	0.89	0.12	0.92	0.09
Mild disease																	
Diabetes mellitus																	
3L	271	0.81	0.17	0.81	0.19	0.77	0.24	0.78	0.18	0.80	0.21	0.86	0.14	0.80	0.24	0.82	0.18
5L	271	0.81	0.18	0.81	0.23	0.83	0.20	0.81	0.18	0.79	0.24	0.82	0.17	0.81	0.21	0.85	0.16
Liver disease																	
3L	588	0.83	0.17	0.83	0.18	0.80	0.23	0.80	0.18	0.81	0.21	0.88	0.14	0.83	0.22	0.84	0.18
5L	588	0.85	0.15	0.85	0.21	0.87	0.17	0.84	0.17	0.83	0.21	0.85	0.15	0.85	0.18	0.88	0.15
Moderate/severe disease																	
CVD																	
3L	251	0.67	0.21	0.64	0.23	0.57	0.32	0.64	0.19	0.63	0.28	0.72	0.19	0.57	0.34	0.66	0.21
5L	251	0.64	0.25	0.57	0.32	0.65	0.27	0.63	0.21	0.58	0.31	0.65	0.21	0.62	0.27	0.68	0.21
Stroke																	
3L	582	0.54	0.29	0.50	0.28	0.40	0.40	0.50	0.27	0.49	0.32	0.58	0.31	0.37	0.47	0.53	0.25
5L	582	0.52	0.31	0.41	0.39	0.51	0.34	0.51	0.26	0.44	0.37	0.55	0.28	0.48	0.35	0.56	0.27
Asthma/COPD																	
3L	342	0.66	0.20	0.63	0.22	0.55	0.32	0.62	0.17	0.61	0.29	0.72	0.18	0.57	0.32	0.65	0.20
5L	342	0.64	0.25	0.56	0.32	0.64	0.28	0.62	0.22	0.57	0.32	0.65	0.21	0.61	0.28	0.68	0.22
RA/arthritis																	
3L	367	0.68	0.18	0.65	0.19	0.59	0.28	0.63	0.15	0.65	0.25	0.74	0.16	0.62	0.28	0.67	0.18
5L	367	0.66	0.23	0.59	0.29	0.67	0.26	0.64	0.19	0.60	0.29	0.67	0.19	0.64	0.25	0.70	0.19
Depression																	
3L	250	0.71	0.20	0.71	0.21	0.64	0.30	0.69	0.17	0.65	0.27	0.79	0.17	0.69	0.30	0.73	0.20
5L	250	0.73	0.23	0.70	0.28	0.73	0.24	0.71	0.19	0.64	0.29	0.73	0.18	0.70	0.25	0.77	0.19
Personality disorder																	
3L	373	0.69	0.17	0.71	0.16	0.61	0.27	0.69	0.13	0.61	0.26	0.78	0.14	0.67	0.25	0.73	0.15
5L	373	0.72	0.17	0.70	0.19	0.72	0.18	0.70	0.13	0.61	0.23	0.72	0.14	0.71	0.18	0.76	0.13

3L EQ-5D-3L, *5L* EQ-5D-5L, *CVD* cardiovascular disease, *COPD* chronic obstructive pulmonary disease, *RA* rheumatoid arthritis, *SD* standard deviation, *tLSS* transformed Level Sum Score

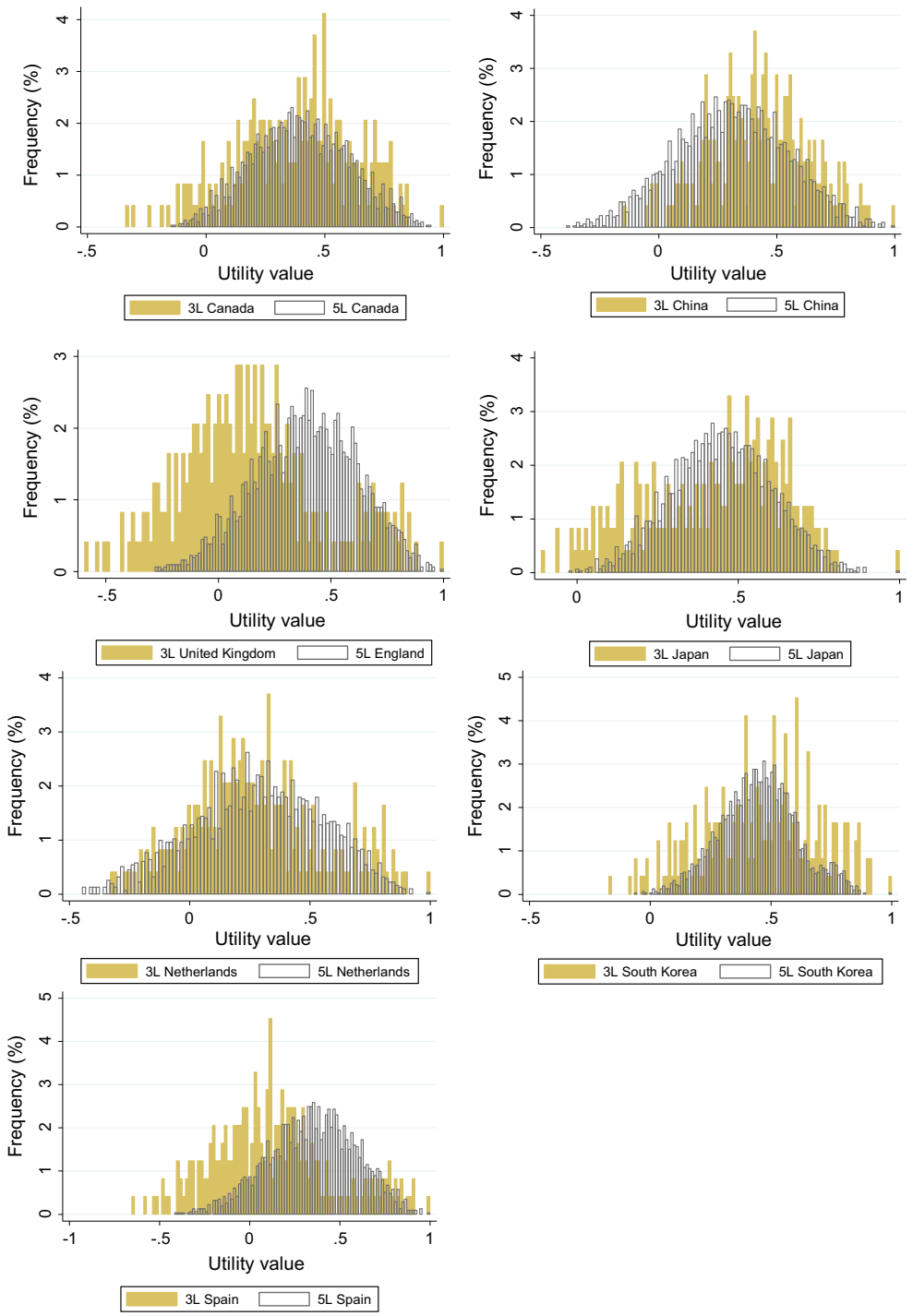

Fig. 9 Histograms of all possible 3L ($n = 243$) and 5L ($n = 3125$) utility values based on value sets from seven countries. *3L* EQ-5D-3L, *5L* EQ-5D-5L

References

1. The EuroQol Group. EuroQol–a new facility for the measurement of health-related quality of life. Health Policy. 1990;16(3):199–208.
2. Dolan P. Modeling valuations for EuroQol health states. Med Care. 1997;35(11):1095–108.
3. Brooks R. The EuroQol Group after 25 years. Dordrecht: Springer; 2013.
4. Devlin NJ, Brooks R. EQ-5D and the EuroQol Group: past, present and future. Appl Health Econ Health Policy. 2017;15(2):127–37.
5. Wille N, Badia X, Bonsel G, Burstrom K, Cavrini G, Devlin N, et al. Development of the EQ-5D-Y: a child-friendly version of the EQ-5D. Qual Life Res. 2010;19(6):875–86.

6. Ravens-Sieberer U, Wille N, Badia X, Bonsel G, Burstrom K, Cavrini G, et al. Feasibility, reliability, and validity of the EQ-5D-Y: results from a multinational study. Qual Life Res. 2010;19(6):887–97.
7. Krabbe PF, Stouthard ME, Essink-Bot ML, Bonsel GJ. The effect of adding a cognitive dimension to the EuroQol multiattribute health-status classification system. J Clin Epidemiol. 1999;52(4):293–301.
8. Swinburn P, Lloyd A, Boye KS, Edson-Heredia E, Bowman L, Janssen B. Development of a disease-specific version of the EQ-5D-5L for use in patients suffering from psoriasis: lessons learned from a feasibility study in the UK. Value Health. 2013;16(8):1156–62.
9. Yang Y, Brazier J, Tsuchiya A. Effect of adding a sleep dimension to the EQ- 5D descriptive system: a "bolt-on" experiment. Med Decis Making. 2014;34(1):42–53.
10. Yang Y, Rowen D, Brazier J, Tsuchiya A, Young T, Longworth L. An exploratory study to test the impact on three "bolt-on" items to the EQ-5D. Value Health. 2015;18(1):52–60.
11. Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res. 2011;20(10):1727–36.
12. EQ-5D. http://www.euroqol.org. Accessed 24 Oct 2017.
13. Oppe M, Devlin NJ, van Hout B, Krabbe PF, de Charro F. A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. Value Health. 2014;17(4):445–53.
14. Ramos-Goñi JM, Oppe M, Slaap B, Busschbach JJ, Stolk E. Quality control process for EQ-5D-5L valuation studies. Value Health. 2017;20(3):466–73.
15. Janssen MF, Pickard AS, Golicki D, Gudex C, Niewada M, Scalone L, et al. Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: a multi-country study. Qual Life Res. 2013;22(7):1717–27.
16. Jia YX, Cui FQ, Li L, Zhang DL, Zhang GM, Wang FZ, et al. Comparison between the EQ-5D-5L and the EQ-5D-3L in patients with hepatitis B. Qual Life Res. 2014;23(8):2355–63.
17. Agborsangaya CB, Lahtinen M, Cooke T, Johnson JA. Comparing the EQ-5D 3L and 5L: measurement properties and association with chronic conditions and multimorbidity in the general population. Health Qual Life Outcomes. 2014;12:74.
18. Conner-Spady BL, Marshall DA, Bohm E, Dunbar MJ, Loucks L, Al KA, et al. Reliability and validity of the EQ-5D-5L compared to the EQ-5D-3L in patients with osteoarthritis referred for hip and knee replacement. Qual Life Res. 2015;24(7):1775–84.
19. Golicki D, Niewada M, Karlinska A, Buczek J, Kobayashi A, Janssen MF, et al. Comparing responsiveness of the EQ-5D-5L, EQ-5D-3L and EQ VAS in stroke patients. Qual Life Res. 2015;24(6):1555–63.
20. Greene ME, Rader KA, Garellick G, Malchau H, Freiberg AA, Rolfson O. The EQ-5D-5L improves on the EQ-5D-3L for health-related quality-of-life assessment in patients undergoing total hip arthroplasty. Clin Orthop Relat Res. 2015;473(11):3383–90.
21. Pan CW, Sun HP, Wang X, Ma Q, Xu Y, Luo N, Wang P. The EQ-5D-5L index score is more discriminative than the EQ-5D-3L index score in diabetes patients. Qual Life Res. 2014;24(7):1767–74.
22. Pattanaphesaj J, Thavorncharoensap M. Measurement properties of the EQ-5D-5L compared to EQ-5D-3L in the Thai diabetes patients. Health Qual Life Outcomes. 2015;13:14.
23. Richardson J, Iezzi A, Khan MA. Why do multi-attribute utility instruments produce different utilities: the relative importance of the descriptive systems, scale and 'micro-utility' effects. Qual Life Res. 2015;24(8):2045–53.
24. van Hout B, Janssen MF, Feng YS, Kohlmann T, Busschbach J, Golicki D, 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):708–15.
25. Xie F, Pullenayegum E, Gaebel K, Bansback N, Bryan S, Ohinmaa A, et al. A time trade-off-derived value set of the EQ-5D-5L for Canada. Med Care. 2016;54(1):98–105.
26. Bansback N, Tsuchiya A, Brazier J, Anis A. Canadian valuation of EQ-5D health states: preliminary value set and considerations for future valuation studies. PLoS One. 2012;7(2):e31115.
27. Luo N, Liu G, Li M, Guan H, Jin X, Rand-Hendriksen K. Estimating an EQ-5D-5L value set for China. Value Health. 2017;20(4):662–9.
28. Liu GG, Wu H, Li M, Gao C, Luo N. Chinese time trade-off values for EQ-5D health states. Value Health. 2014;17(5):597–604.
29. Devlin N, Shah K, Feng Y, Mulhern B, van Hout B. Valuing health-related quality of life: an EQ-5D-5L value set for England. Health Econ. 2017. https://doi.org/10.1002/hec.3564 **(Epub 2017 Aug 22)**.
30. Shiroiwa T, Ikeda S, Noto S, Igarashi A, Fukuda T, Saito S, Shimozuma K. Comparison of value set based on DCE and/or TTO data: scoring for EQ-5D-5L health states in Japan. Value Health. 2016;19(5):648–54.
31. Tsuchiya A, Ikeda S, Ikegami N, et al. Estimating an EQ-5D population value set: the case of Japan. Health Econ. 2002;11(4):341–53.
32. Versteegh MM, Vermeulen KM, Evers SM, de Wit GA, Prenger R, Stolk EA. Dutch tariff for the five-level version of EQ-5D. Value Health. 2016;19(4):343–52.
33. Lamers LM, McDonnell J, Stalmeier PF, Krabbe PF, Busschbach JJ. The Dutch tariff: results and arguments for an effective design for national EQ-5D valuation studies. Health Econ. 2006;15(10):1121–32.
34. Kim SH, Ahn J, Ock M, Shin S, Park J, Luo N, et al. The EQ-5D-5L valuation study in Korea. Qual Life Res. 2016;25(7):1845–52.
35. Lee YK, Nam HS, Chuang LH, Kim KY, Yang HK, Kwon IS, et al. South Korean time trade-off values for EQ-5D health states: modeling with observed values for 101 health states. Value Health. 2009;12(8):1187–93.
36. Ramos-Goñi JM, Craig BM, Oppe M, Ramallo-Fariña Y, Pinto-Prades JL, Luo L, et al. Handling data quality issues to estimate the Spanish EQ-5D-5L value set using a hybrid interval regression approach. Value Health **(In press)**.
37. Badia X, Roset R, Herdman M, Kind P. A comparison of United Kingdom and Spanish general population time trade-off values for EQ-5D health states. Med Decis Making. 2001;21(1):7–16.
38. Xie F, Gaebel K, Perampaladas K, Doble B, Pullenayegum E. Comparing EQ-5D valuation studies: a systematic review and methodological reporting checklist. Med Decis Making. 2014;34(1):8–20.
39. Oppe M, Rand-Hendriksen K, Shah K, Ramos-Goñi JM, Luo N. EuroQol protocols for time trade-off valuation of health outcomes. Pharmacoeconomics. 2016;34(10):993–1004.
40. Parkin D, Devlin N, Feng Y. What determines the shape of an EQ-5D index distribution? Med Decis Making. 2016;36(8):941–51.
41. Janssen MF, Birnie E, Bonsel GJ. Evaluating the discriminatory power of EQ-5D, HUI2 and HUI3 in a US general population survey using Shannon's indices. Qual Life Res. 2007;16(5):895–904.
42. Vickrey BG, Hays RD, Genovese BJ, et al. Comparison of a generic to disease-targeted health-related quality-of-life measures for multiple sclerosis. J Clin Epidemiol. 1997;50:557–69.
43. Luo N, Johnson JA, Shaw JW, et al. Relative efficiency of the EQ-5D, HUI2, and HUI3 index scores in measuring health

burden of chronic medical conditions in a population health survey in the United States. Med Care. 2009;47:53–60.
44. Murray CJL, Özaltin E, Tandon A, Salomon JA, Sadana R, Chatterji SA. Empirical evaluation of the anchoring vignette approach in health surveys. In: Murray CJL, Evans DB, editors. Health system performance assessment: debates, methods and empiricism. Geneva: World Health Organization; 2003. p. 369–99.
45. Lindeboom M, van Doorslaer E. Cut-point shift and index shift in self-reported health. J Health Econ. 2004;23(6):1083–99.
46. Rice N, Robone S, Smith P. Analysis of the validity of the vignette approach to correct for heterogeneity in reporting health system responsiveness. Eur J Health Econ. 2011;12(2):141–62.
47. Hirve S, Gómez-Olivé X, Oti S, Debpuur C, Juvekar S, Tollman S, et al. Use of anchoring vignettes to evaluate health reporting behavior amongst adults aged 50 years and above in Africa and Asia–testing assumptions. Glob Health Action. 2013;6(1):21064.
48. Valentine N, Verdes-Tennant E, Bonsel G. Health systems' responsiveness and reporting behaviour: multilevel analysis of the influence of individual-level factors in 64 countries. Soc Sci Med. 2015;138:152–60.
49. Scalone L, Ciampichini R, Fagiuoli S, Gardini I, Fusco F, Gaeta L, et al. Comparing the performance of the standard EQ-5D 3L with the new version EQ-5D 5L in patients with chronic hepatic diseases. Qual Life Res. 2013;22(7):1707–16.
50. Olsen JA, Lamu AN, Cairns J. In search of a common currency: a comparison of seven EQ-5D-5L value sets. Health Econ. 2017. https://doi.org/10.1002/hec.3606 **(Epub 2017 Oct 24)**.
51. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, et al. Protocol of the COSMIN study: COnsensus-based Standards for the selection of health Measurement INstruments. BMC Med Res Methodol. 2006;6:2.
52. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. International consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes: results of the COSMIN study. J Clin Epidemol. 2010;63:737–45.
53. Hernandez Alava M, Wailoo A, Grimm S, Pudney S, Gomes M, Sadique Z, et al. EQ-5D-5L versus EQ-5D-3L: the impact on cost-effectiveness. Value Health. 2018;21(1):49–56.
54. Huxley JS. Problems of relative growth. London: Methuen and Company Limited; 1932.
55. Gaddum JH. Lognormal distributions. Nature. 1945;156:463–6.
56. Fairclough DL. Design and analysis of quality of life studies in clinical trials. 2nd ed. New York: CRC Press; 2010.
57. Hilden J. The area under the ROC curve and its competitors. Med Decis Making. 1991;11(2):95–101.
58. Hilden J. Prevalence-free utility-respecting summary indices of diagnostic power do not exist. Stat Med. 2000;19(4):431–40.
59. Oppe M, Devlin N, Black N. Comparison of the underlying constructs of EQ-5D and Oxford Hip Score: implications for mapping. Value Health. 2011;14:884–91.