Quality of Life Research (2019) 28:153–162
https://doi.org/10.1007/s11136-018-2027-9

Psychometric performance assessment of Malay and Malaysian English version of EQ-5D-5L in the Malaysian population

Asrul Akmal Shafie[1] · Annushiah Vasan Thakumar[1] · Ching Jou Lim[1] · Nan Luo[2]

Accepted: 11 October 2018 / Published online: 13 October 2018

Abstract

Purpose To determine the psychometric properties and performance of Malay and English versions of the EQ-5D-5L descriptive instrument in the general Malaysian population.

Methods 1137 members of the Malaysian general public were sampled in this national study. Respondents were recruited by quota sampling of urbanicity, gender, age, and ethnicity. In face-to-face interviews, respondents first answered the EQ-5D-5L questionnaire administered using the EQ-Valuation Technology software, and then completed the EQ-5D-3L questionnaire on paper. A subgroup of the respondents were given paper form of EQ-5D-5L for completion within 2 weeks for test–retest reliability. Ceiling effects, response redistribution, informativity, and convergent validity were compared between EQ-5D-5L and ED-5D-3L separately by Malay and English language versions.

Results The proportion of 'full health' responses (11111) drastically decreased by 25.55% and 15.74% in the Malay and English language versions indicating lower ceiling effects in EQ-5D-5L. Inconsistencies from response redistribution was below 6% for all dimensions across languages. The measure of relative informativity was comparatively higher in EQ-5D-5L than in EQ-5D-3L in both language versions, with the exception of dimensions mobility and pain/discomfort in the English version. Convergent validity in terms of correlation with EQ-VAS was relatively better for EQ-5D-5L dimensions, with pain/discomfort of the Malay version having the strongest correlation ($|r| = 0.37$). Also, reliability testing revealed moderate to poor agreements on all 5L dimensions.

Conclusions EQ-5D-5L fared better in terms of psychometric performance compared to EQ-5D-3L for both language versions. This encourages the application of the EQ-5D-5L in health-related research in Malaysia.

Keywords EQ-5D-5L · EQ-5D-3L · Psychometric properties · Validity · Reliability

Introduction

The EQ-5D instrument is one of the most frequently used patient-reported outcome measures (PROMs) available today and is the recommended PROM tool by many health authorities such as the National Institute for Health and Care Excellence (NICE) in the UK. As the name suggests, the EQ-5D measures health in five dimensions, namely mobility, self-care, usual activities, pain/discomfort, and anxiety/depression [1].

First introduced in the early 90s with three severity levels (no problem, some problem, extreme problem, known as EQ-5D-3L), the EQ-5D has been applied in a variety of settings for purposes of economic evaluations and health monitoring of patient groups. Experience has demonstrated that ceiling effects and low descriptive richness may be the limiting factors of the EQ-5D descriptive system. Subsequently, a newer, five-level version (EQ-5D-5L) was initiated in 2005 [2] to counter the earlier mentioned limitations.

The availability of both the EQ-5D-3L and EQ-5D-5L for use has allowed for head-to-head comparisons between the two versions in terms of psychometric properties. From the earliest published study in 2008 [3] involving Dutch panel members valuing both their own health and hypothetical disease states to the latest studies in 2017 involving Hungarian psoriasis patients [4] and Greek general population [5] describing their personal health, these studies help to

✉ Asrul Akmal Shafie
aakmal@usm.my

[1] Discipline of Social & Administrative Pharmacy, Universiti Sains Malaysia, Penang, Malaysia

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

quantify ceiling effects and information richness in the current setting and the benefits (if any) of using either instruments. Variability has been frequently reported in the previous studies, potentially due to differences in population demographics and socioeconomic background [6]. To date, there is no similar study conducted in the Malaysia setting, with studies mainly focusing on the use of EQ-5D-3L in health-related research.

In Malaysia, the EQ-5D-3L instrument has been validated [7] both in the general population and in the patient group setting. Results from these studies reveal prominent ceiling effects, especially when used in the general population [8, 9]. Despite this, the latter is still more commonly reported in scientific work [7] in the Malaysian setting. This could partially be due to the lack of EQ-5D-5L psychometric validity evidence in Malaysia for its application in both general population [8] and disease-specific setting. Importantly, the availability of a validated EQ-5D-5L descriptive system will function as an additional patient-reported outcome measure alternative, whether in terms of gauging population health or even obtaining a quick overview of a patient's health status for monitoring or evaluation purposes.

The exploration of psychometric properties of EQ-5D-5L is warranted in the Malaysian population. Therefore, the objective of this study is to compare the psychometric performance of the EQ-5D-5L descriptive system with EQ-5D-3L of the Malay and English versions in terms of convergent validity, response redistribution, informativity, ceiling effects, and also the EQ-5D-5L questionnaire's test–retest reliability.

Methods

This study is part of a larger EQ-5D-5L valuation initiative to obtain utility values of the Malaysian population. Ethics approval has been granted by the Malaysia Medical Research & Ethics Committee (ID NMRR-13-1377-18574).

Sampling and recruitment

In the valuation study, a sample size of 1000 was recommended as part of the computer-assisted EQ-Valuation Technology (EQ-VT) experimental design [10] employed. Quota-based sampling by age, gender, ethnicity, and area of living (urban/rural) was carried out based on the 2010 Malaysian National Census [11]. Malaysians aged 18 and above who could speak or write in English or Malay met the inclusion criteria. However, those with low cognitive ability or unable to understand instructions were excluded from the study. The non-formal assessment of cognitive ability of using body language cues was based on the judgement of the trained interviewers (from pharmacy and medicine backgrounds) during respondent recruitment and interview process.

Malay and English were chosen for this study, as these are the two most frequently spoken languages in Malaysia. Malay is the official and national language of Malaysia, while English is widely used in teaching many subjects at secondary and tertiary levels. Additionally, Malay and English are both compulsory subjects taught in schools from ages 7 to 17.

A convenience sample of respondents was acquired at targeted population gathering points (e.g. shopping complexes, markets, food courts, or community centres) and the computer-assisted questionnaires were administered to individual participants at rented community halls nearby.

Instrument

The EQ-5D-5L (here forth identified as 5L) descriptive system measures health in five dimensions, namely mobility, self-care, usual activities, pain/discomfort, and anxiety/depression using five levels of severity roughly corresponding to no problems, slightly, moderate, severe, and extreme. The 5L was answered by participants with the help of trained interviewers. Each 5L question and response displayed on the computer screen was read aloud to ensure respondents were aware of all the available response options. Subsequently, participants filled in the visual analogue scale known as EQ-VAS. This instrument comprised of a 20-cm thermometer (0 representing worst imaginable health, 100, best imaginable health) inquiring about one's general health for the day. Participants then completed the valuation tasks, followed by some sociodemographic questions, and concluded with paper and pencil form of the EQ-5D-3L (here forth identified as 3L) instrument. Interviewers explained to respondents that the 3L was a different version of the EQ-5D and provided guidance on filling-in the paper and pencil form of the 3L. Briefly, both the 3L and 5L measure the similar five dimensions while differing in the number of severity levels present in the descriptive system, with the 3L questionnaire having three levels and 5L questionnaire comprising five levels.

The values of the five dimensions of 3L or 5L can be combined in the order of mobility, self-care, usual activities, pain/discomfort, and anxiety/depression to form health states. A health state of 11111 would mean rating level 1 ("no problem") on all dimensions.

Respondents can choose either the Malay or Malaysian English version of the questionnaire. The language versions were developed through translations commissioned by the EuroQol Group.

The first 50 respondents from each interview day were asked to complete another set of 5L paper questionnaire within 2 weeks and to return it to the investigators by pre-paid self-addressed mail for the purpose of assessing its test–retest reliability.

Data analysis

Ceiling effects were assessed by measuring the proportion of level 1 (‘no problem’) responses on the dimensions and health state ‘11111’ (no problem on all dimensions). Percentage of absolute ceiling effect changes was also measured and compared between 3L and 5L.

Since respondents answered both the 3L and 5L versions, these responses can be tabulated to display distribution patterns when shifting from 5L to 3L. Response redistribution was described as proportions of 5L responses of each dimension redistributing into 3L responses, forming 3L-5L response pairs. A $3L_2$-$5L_1$ response would mean the respondent answered level 2 on the 3L and correspondingly, a level 1 on the 5L. The corresponding mean and median EQ-VAS scores were calculated for each subgroup of paired responses, except for inconsistent pairs. Inconsistency and its size were defined as in Janssen et al. [3]. Briefly, after projecting the 3L response scale on a 5L response scale (i.e. producing $3L_{5L}$ by recoding severity levels $1=1$, $2=3$, $3=5$), the size of inconsistency was calculated as $|3L_{5L}-5L|-1$. An inconsistency size of 1 and above denotes an inconsistency. For example, when level 1 in 3L was redistributed as level 1 or 2 in 5L, it was considered as consistent response. However, if the response was redistributed as level 3, 4, or 5 in 5L, it was considered as inconsistent response and the sizes of inconsistency were 1, 2, and 3, respectively.

Convergent validity was assessed by comparing the strength of correlation of 5L and 3L dimension values with the EQ-VAS values obtained using Spearman’s rank coefficient.

Informativity was tested by using Shannon entropy (Shannon index, H′) [3, 12] and Shannon information efficiency (Shannon evenness index, *J*′). The Shannon entropy was calculated as below:

$$H' = -\sum_{i=1}^{C} P_i \log_2 P_i,$$

where H' is the absolute amount of informativity captured; C is the number of possible categories (or levels in this study); $P_i = n_i/N$ is the proportion of observations in the ith category ($i = 1,\ldots, C$); and n_i is the observed number of scores (responses) in category i, while N is the total sample size.

Basically, informativity measures the richness of the descriptive system and the highest value indicates all levels are filled equally. Measuring informativity allows us to gauge the usefulness of the levels present in a system. If responses are clustered around only certain levels, the informativity is smaller than descriptive systems with more evenly distributed responses around all levels. Based on the formula, Shannon entropy increases with number of filled categories (levels) and evenness of distribution of responses between levels available. Therefore, the highest attainable H' for 3L is 1.58 and 5L is 2.32. Relative comparison of informativity between instruments of varying number of levels can be accomplished with Shannon information efficiency, J'. J' is defined as H' divided by maximum H' for an instrument.

The 2-week test–retest reliability of the 5L dimensions were measured using kappa agreement.

The frequency and percentage of respondent characteristics were also presented based on the following demographics: gender (male/female); age categorised according to quota sampling range (18–39, 40–64, > 64); ethnicity (Malay, Chinese, Indian, other); level of education (no formal, primary, lower secondary, higher secondary, pre-university, university); monthly household income categorised roughly to low, middle, and above-middle (less or equal to MYR 3000, less or equal to MYR 6000, more than MYR 6000, did not disclose), marital status (single, married/cohabiting, widowed, divorced/separated), and residential area as indicated by the Malaysian Statistics Department (urban/rural).

All data were analysed separately by language versions using statistical software SPSS version 22 and a p value < 0.05 was considered as statistically significant.

Results

A total of 1137 respondents participated in the study (Table 1) with quotas for age, gender, residential area, and ethnicity coinciding with ratios of the actual population. Slightly less than three-quarter of the sample chose to answer the questionnaires in Malay (813 respondents) and the remaining answered in English (324). There were no missing data from 5L but there is one missing information from the Malay version of 3L. Therefore, for analyses involving only 5L data, no exclusions were made, but for those analyses involving solely 3L data, or both 3L and 5L data, only 812 respondents were included in the Malay version analyses.

Frequency and ceiling effects

The top five most frequently reported health states in the Malay 3L accounted for 90.5% of the total responses, dropping to 72.8% in the 5L instrument. Similar trends, but of smaller proportions were observed in the English version. While the order of the top five health states were similar in the 3L and 5L English version, the ranks of top two and top three were reversed in the Malay 5L responses.

Table 1 Demographic characteristics of study sample ($N = 1137$)

	Total *n* (%)	Malay *n* (%)	English *n* (%)	*p* value[a]
Gender				0.991
Male	584 (51.4)	417 (51.3)	167 (51.5)	
Female	553 (48.6)	396 (48.7)	157 (48.5)	
Age (years), mean (SD)	39.1 (16.2)	39.7 (15.5)	37.5 (17.6)	0.045
18–39	635 (55.8)	431 (53.0)	204 (63.0)	0.000
40–64	411 (36.1)	326 (40.1)	85 (26.2)	
> 64	91 (8.0)	56 (6.9)	35 (10.8)	
Ethnicity				
Malay	772 (67.9)	717 (88.2)	55 (17.0)	0.000
Chinese	288 (25.3)	61 (7.5)	227 (70.1)	
Indian	67 (5.9)	31 (3.8)	36 (11.1)	
Other	10 (0.9)	4 (0.5)	6 (1.9)	
Level of education				0.000
No formal	6 (0.5)	6 (0.7)	0 (0.0)	
Primary	62 (5.5)	54 (6.6)	8 (2.5)	
Lower secondary	82 (7.2)	75 (9.2)	7 (2.2)	
Higher secondary	325 (28.6)	269 (33.1)	56 (17.3)	
Pre-University	231 (20.3)	168 (20.7)	63 (19.4)	
University	431 (37.9)	241 (29.6)	190 (58.6)	
Monthly household income[b]				0.000
Less or equal to MYR 3000	580 (51.4)	467 (57.9)	113 (35.1)	
Less or equal to MYR 6000	294 (26.1)	201 (24.9)	93 (28.9)	
More than MYR 6000	233 (20.7)	131 (16.3)	102 (31.7)	
Did not disclose	21 (1.9)	7 (0.9)	14 (4.3)	
Marital status[c]				0.000
Single	455 (40.4)	268 (33.3)	187 (58.1)	
Married/cohabiting	589 (52.3)	463 (57.6)	126 (39.1)	
Widowed	39 (3.5)	38 (4.7)	1 (0.3)	
Divorced/separated	43 (3.8)	35 (4.4)	8 (2.5)	
Residential area				0.000
Urban	799 (70.3)	527 (64.8)	272 (84.0)	
Rural	338 (29.7)	286 (35.2)	52 (16.0)	
EQ-VAS, mean (SD)	85.52 (12.3)	86.27 (12.4)	83.64 (11.69)	0.000

[a] Chi-square test was used for categorical variables and t-test for continuous age variable and non-parametric Mann–Whitney test for EQ-VAS

[b] Missing data in 9 respondents

[c] Missing data in 11 respondents

The most common response of the 3L and 5L in all dimensions (Table 2) for both Malay and English versions was level 1 or ‘no problem’. Comparing the ceiling effects between the descriptive systems of 3L and 5L (Table 3) revealed substantial reductions in the frequency of reported “11111” (no problem on all dimensions) health state of both the 5L Malay (25.55%) and English versions (15.74%). In terms of dimension-specific trends, pain/discomfort and self-care had consistently the least and highest amount of no problem responses across versions and languages, with all of the self-care responses in the English 3L being categorised as no problem. Ceiling effects

Table 2 Frequency of top 5 health states with the most responses

	n (%)			
	Malay ($n = 813$)		English ($n = 324$)	
	3L	5L	3L	5L
11111	559 (68.8)	352 (43.3)	215 (66.4)	164 (50.6)
11121	72 (8.9)	88 (10.8)	33 (10.2)	40 (12.3)
11112	68 (8.4)	92 (11.3)	25 (7.7)	34 (10.5)
11122	24 (3.0)	35 (4.3)	17 (5.2)	20 (6.2)
21121	12 (1.5)	25 (3.1)	8 (2.5)	9 (2.8)
	735 (90.5)	592 (72.8)	298 (92.0)	267 (82.4)

Table 3 Proportion of 'no problem' responses by dimensions and ceiling effect change

	n (%)				Absolute ceiling effect change (%)	
	Malay		English		Malay	English
	3L	5L	3L	5L		
Mobility	757 (93.2)	665 (81.8)	299 (92.3)	291 (89.8)	− 11.43	− 2.47
Self-care	802 (98.8)	777 (95.6)	324 (100.0)	317 (97.8)	− 3.20	− 2.16
Usual activities	761 (93.7)	684 (84.1)	309 (95.4)	286 (88.3)	− 9.59	− 7.10
Pain/discomfort	659 (81.2)	517 (63.6)	255 (78.7)	219 (67.6)	− 17.57	− 11.11
Anxiety/depression	689 (84.9)	576 (70.8)	270 (83.3)	236 (72.8)	− 14.00	− 10.49
Health state '11111'	559 (68.8)	352 (43.3)	215 (66.4)	164 (50.6)	− 25.55	− 15.74

decreased in the 5L with larger reductions observed in the Malay version.

Response redistribution and inconsistency

Most of the consistent $3L_1$ (level one responses from 3L) tended to redistribute into $5L_1$ (level one responses from 5L) in both the Malay and English versions (Table 4). Subsequently, $3L_1$–$5L_2$ transitions accounted for 2.9% (1.6%) to 17.5% (15.3%) of responses on the Malay (English) versions, respectively. Comparing $3L_2$ distributions to the 5L, a large proportion of responses clustered around $5L_2$ and $3L_2$–$5L_4$ transitions were non-existent in the English responses. There were only few $3L_3$ responses recorded and the one consistent response recorded was from pain/discomfort dimension in English. Mean and median of EQ-VAS

Table 4 Consistent responses redistributed from 5L into 3L values

Dimension	*n* (%)				EQ-VAS			
	3L	5L	Malay	English	Malay		English	
					Mean	Median	Mean	Median
Mobility	1	1	658 (90.1)	284 (95.9)	88.3	90.0	84.6	85.0
		2	72 (9.9)	12 (4.1)	82.0	83.0	73.3	72.5
	2	2	32 (66.7)	11 (61.1)	74.2	74.0	81.4	80.0
		3	12 (25.0)	7 (38.9)	69.6	72.5	72.9	80.0
		4	4 (8.3)	0 (0.0)	58.0	55.0	–	–
Self-care	1	1	770 (97.1)	317 (98.4)	87.1	90.0	83.9	85.0
		2	23 (2.9)	5 (1.6)	74.9	75.0	77.0	75.0
	2	2	3 (75.0)	0 (0.0)	63.3	50.0	–	–
		4	1 (25.0)	0 (0.0)	50.0	50.0	–	–
Usual activities	1	1	672 (89.8)	284 (93.1)	88.4	90.0	84.7	88.5
		2	76 (10.2)	21 (6.9)	79.3	80.0	79.5	80.0
	2	2	23 (60.5)	11 (84.6)	75.7	75.0	73.6	75.0
		3	12 (31.6)	2 (15.4)	76.3	75.0	70.0	70.0
		4	3 (7.9)	0 (0.0)	43.7	50.0	–	–
Pain/discomfort	1	1	493 (77.4)	213 (84.5)	89.7	90.0	86.1	90.0
		2	144 (22.6)	39 (15.5)	84.1	90.0	79.7	85.0
	2	2	104 (81.3)	54 (87.1)	80.1	80.0	79.9	80.0
		3	22 (17.2)	8 (12.9)	69.2	70.0	73.9	75.0
		4	2 (1.6)	0 (0.0)	70.0	70.0	–	–
	3	5	0 (0.0)	1 (100.0)	–	–	80.0	80.0
Anxiety/depression	1	1	553 (82.5)	227 (84.7)	88.5	90.0	86.0	90.0
		2	117 (17.5)	41 (15.3)	85.5	90.0	81.0	80.0
	2	2	79 (82.3)	33 (78.6)	82.7	80.0	78.2	80.0
		3	16 (16.7)	9 (21.4)	70.1	71.5	73.3	70.0
		4	1 (1.0)	0 (0.0)	70.0	70.0	–	–

tended to decrease as severity increases in both language versions.

Inconsistencies from response redistribution (Table 5) peaked in the pain/discomfort dimension in the Malay version and anxiety/depression for the English version. There were slightly less counts and smaller sizes of inconsistencies in the English version in comparison to the Malay version.

Informativity

Among dimensions, Shannon entropy (*H*′) for pain/discomfort was the highest and lowest for self-care for the 3L and 5L of the two language versions. In terms of information efficiency (*J*′), 5L was comparatively more informative/descriptive in all dimensions than 3L of the Malay version. Mobility and pain/discomfort of the English version had slightly stronger values in 3L in contrast to the 5L.

Convergent validity

Comparatively, the dimensions of 5L showed stronger correlation to EQ-VAS than those of 3L for both language versions. The self-care dimension of the English 3L had only no problem responses so the strength of correlation to EQ-VAS could not be measured. While the correlation coefficients of the Malay and English versions had comparable values, the 3L and 5L dimensions of the Malay version generally had stronger values with the exception of anxiety/depression.

Reliability

Of the 717 respondents who agreed to the test–retest survey, 528 questionnaires (73.8%; 358 Malay and 170 English copies) were returned and answered fully. Kappa agreement (Table 6) revealed results ranging from 0.208 (self-care) to 0.382 (anxiety/depression) on the Malay version and − 0.015 (self-care) to 0.553 (mobility) on the English version. Based on Landis and Koch's [13] standards for kappa strength of agreement [< 0.0 = poor, 0.0–0.20 = slight, 0.21–0.40 = fair, 0.41–0.60 = moderate, 0.61–0.80 = substantial, 0.81–1.00 = almost], the dimensional test–retest reliability of the Malay version had slight to fair agreement, while the responses on the English version had poor to moderate agreement.

Table 5 Inconsistencies from response redistribution

Dimension	Dimension *n* (%)		Average size of inconsistencies	
	Malay	English	Malay	English
Mobility	34 (4.19)	10 (3.09)	1.03	1.00
Self-care	15 (1.85)	2 (0.62)	1.20	1.00
Usual activities	26 (3.20)	6 (1.85)	1.12	1.00
Pain/discomfort	47 (5.79)	9 (2.78)	1.09	1.00
Anxiety/depression	46 (5.67)	11 (3.40)	1.13	1.00

Discussion

The aim of this study was to compare the psychometric performance of the EQ-5D-5L descriptive system with EQ-5D-3L of the Malay and English and measure test–retest reliability of the EQ-5D-5L questionnaire. In the three psychometric properties comparing performance of 3L and 5L, 5L fared better, with evidently lower ceiling effects, higher absolute descriptive informativity, and stronger correlation of dimensions with EQ-VAS for both the Malay and English version of responses. The 5L's wider selection of severity levels is better adapted to detect small health changes especially common in general population studies, which may have been missed by the less sensitive levels of the 3L.

The top five frequently responded health states for both the 3L and 5L instruments (Table 2) appear to be similar, although the strength of the levels represented by the numbers of the health states in the 3L and 5L differ. Coincidently, two previous Malaysian 3L validation studies, one focusing on dialysis patients [14] and another involving the general population [8], also recorded similar top four and five health states trends, respectively. This similarity in 3L health state trends exhibits dimensions pain/discomfort and anxiety/depression dimensions as the top two with the most common problems. However, applying the 5L revealed that health problems faced by respondents are actually milder than when 3L is answered as demonstrated by Craig et al. [15].

These five health states accounted fewer responses in 5L for both language versions, indicating there was a wider distribution of health states recorded in the 5L version compared to the 3L version as demonstrated. Having two filler levels in between the original 3 levels, 'slight problems' and 'severe problems' resulted in the 5L version more sensitive in capturing the health status of the population. This is also reflected in reduction of ceiling effects in the 5L, most evidently in the health state of perfect health or '11111'. Similar reductions can be observed in recent validation study involving the Greek population [5]. The trends of ceiling effects in our study is comparable to that of a Korean general population-based study [16] with self-care dimension exhibiting the highest ceiling effects and pain/discomfort dimension the least. Reduction of ceiling effects signifies that the 5L is more sensitive in capturing mildly problematic health states, which the 3L may miss due to its limited availability of levels.

Table 6 Psychometric properties of the EQ-5D instrument

	Mobility		Self-care		Usual activities		Pain/discomfort		Anxiety/depression	
	3L	5L	3L	5L	3L	5L	3L	5L	3L	5L
Informativity										
Malay										
H′	0.36	0.87	0.10	0.31	0.35	0.78	0.71	1.21	0.64	1.09
J′	0.23	0.38	0.06	0.13	0.22	0.34	0.45	0.52	0.40	0.47
English										
H′	0.39	0.56	0.00	0.17	0.27	0.60	0.77	1.09	0.65	1.05
J′	0.25	0.24	0.00	0.07	0.17	0.26	0.49	0.40	0.41	0.45
Convergent validity										
Malay	−0.24	−0.32	−0.09	−0.21	−0.22	−0.30	−0.28	−0.37	−0.23	−0.26
English	−0.08	−0.21	X	−0.13	−0.18	−0.21	−0.24	−0.26	−0.28	−0.29
	Kappa (SE)		Kappa (SE)		Kappa (SE)		Kappa (SE)		Kappa (SE)	
Test–retest reliability on 5L										
Malay	0.34 (0.05)		0.21 (0.11)		0.33 (0.06)		0.33 (0.05)		0.38 (0.05)	
English	0.55 (0.11)		−0.02 (0.01)		0.40 (0.10)		0.50 (0.07)		0.43 (0.07)	

H′ Shannon entropy, *J′* Shannon information efficiency, *SE* standard error

Additionally, response redistribution patterns revealed the tendencies of the general population respondents to choose lower severity levels if the option is available as evidenced by $3L_2$–$5L_2$ transitions making up the mass of $3L_2$ redistributions. Most of the no problem responses from all dimensions of the 5L tended to redistribute back into the same level on the 3L signifying that these are true responses of 'no problem'. This is to be expected, seeing that the respondents were made up of the general population with no specific health conditions.

The dimensions of anxiety/depression and pain/discomfort benefited the most from the presence of level 2 of the 5L with a minimum of 17.5% (15.3%) $3L_1$–$5L_2$ transitions observed. In terms of $3L_2$ responses, majority of 3L responses shifted from 'moderate' to 'slight' on the 5L. This trend is common in most studies both using disease-specific [17, 18] and general population respondents [5, 16]. However, in our study, the most severe level responses are lacking both in 3L (level 3) and 5L (level 5) so $3L_3$ transition trends were not observable in this study. The EQ-VAS values tended to decrease as severity of the dimensions increased, signifying respondents were able to consistently shift between 3L and 5L versions and relate higher severity levels to poorer general health as reflected by EQ-VAS values.

The 3L instrument was presented after 5L and was not randomised throughout the study. Previously, a pilot study in the pioneering 3L–5L comparison paper [3] showed that respondents were less likely to fill levels 2 and 4 of the 5L if the respondents answered the 3L first. Interestingly, two other studies [19, 20] that randomised the order of 3L–5L presentation showed contradictory findings whereby the sequence of instrument presentation did not have any priming effect on response trends. Having said that, it would be interesting to have further studies measuring sequence effects as most previous 3L–5L comparative studies tended to adopt the 5L first, 3L second approach [6].

Generally, although slightly higher than previous studies, inconsistencies in response redistribution were low. This is a reflection that respondents were able comprehend and express their health states well in both the 3L and 5L. Inconsistencies present tended to focus around $3L_1$–$5L_3$ and $3L_2$–$5L_1$ responses, especially in the dimensions of pain/discomfort and anxiety/depression.

In terms of absolute higher information richness in 5L, the additional level served its purpose in being useful alternative responses compared to the existing 3L levels for the respondents. Comparing relative information richness using J', 5L generally fared better with responses more fairly distributed among the levels compared to 3L. The only two exceptions were mobility and pain/discomfort dimension in the English version where levels of the 3L had slightly better relative efficiency. Both absolute and relative information efficiency recorded in this study were lower than other studies, especially disease-specific diseases. This was partly contributed by the lower frequency responses from level 3 on the 3L and level 4 and 5 on the 5L. General population respondents tend to record better health status, resulting in better display of usefulness of an additional mild level between level 1 and 2, then a severe level between level 2 and 3 on the 5L.

Convergent validity revealed 5L dimensions to be more strongly correlated to EQ-VAS than 3L on all dimensions. Strength of correlations of the EQ-5D dimensions with the EQ-VAS were comparable to some studies [21, 22] except self-care dimension. This is due to the large proportion of respondents in this study tended to face no problems with self-care, even if they exhibit lower EQ-VAS values.

Test–retest reliability results revealed consistency in values within the range of fair to moderate apart from self-care for the English version of the 5L which had poor agreement. Comparatively, the results obtained in this study was lower than commonly reported for EQ-5D-5L studies [6]. A possible reason for the occurrence would be an actual change severity of the respondents in this particular dimension which was not captured as a separate question in the retest questionnaire, counting as a limitation of the study. It should be noted that while respondents completed the 5L on an electronic questionnaire at the baseline survey, they completed the 5L as paper and pencil format in the follow-up survey. The mode of instrument administration adopted might have confounded the test–retest results to some extent. However, measurement equivalence between paper and electronic forms has been established for the EQ-5D-5L [23, 24], as well as a variety of instruments [25–27].

Another study limitation would be the quota sampling strategy employed that may lead to selection bias. Noting the scarcity of very severe responses on both the 3L and 5L, ideally a random sampling strategy of households would possibly get more frail or physically impaired respondents to participate in the study. Additionally, no formal assessment of cognitive ability was made and no data on the share of exclusions were collected. Respondents who were approached in the market and rural areas tended to immediately decline but many approached the interviewers at a more convenient time. Due to the nature of the recruitment process, we did not keep track of the rate of true responses. More data are needed to assess the instrument's applicability in the wider Malay population.

Both the Malay and English versions generally demonstrated similar improvement trends in the psychometric properties measured with the pain/discomfort and anxiety/depression dimensions demonstrating the most improvement and self-care dimension the least when switching to the 5L version. The availability of an additional level between 'no problem' and 'moderate' was shown to be especially useful in this general population sample to better express one's health-related quality of life. It should be noted that our analysis cannot answer the question whether the English and Malay versions of the 5L have measurement equivalence. Since the main purpose of using EQ-5D is to assist the calculation of utility score for resource allocation, whether the choice of different language versions of EQ-5D questionnaire within a single country can potentially lead to different health profiles been recorded (and consequently different EQ-5D utility scores been calculated) is an important question to be answered. A proper measurement equivalence study [28] aiming to compare language differences in the Malay and English versions is recommended before conclusive stands on language differences in EQ-5D responses can be made. This is crucial as the extend of equivalence between languages will determine whether language versions can be interchangeably used without contributing to differences in health state responses.

Conclusion

The Malay and English versions of 5L fared better in terms of absolute informativity and convergent validity, with significantly lower ceiling effects when compared to 3L, further supporting the future application of 5L version in the Malaysian population.

Funding This study was co-funded by EuroQol Research Foundation and Universiti Sains Malaysia's Research University Grant.

Compliance with ethical standards

Conflict of interest Nan Luo is a member of EuroQol Research Foundation. There is no other conflict of interest.

Ethics approval The study received ethical approval from the Malaysia Medical Research & Ethics Committee (ID NMRR-13-1377-18574) and was conducted in accordance with the Declaration of Helsinki.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

1. Gudex, C. (2006). The descriptive system of the EuroQol instrument. In P. Kind, R. Brooks & R. Rabin (Eds.), *EQ-5D concepts and methods: A developmental history* (pp. 19–27). Dordrecht: Springer.
2. Herdman, M., Gudex, C., Lloyd, A., Janssen, M., Kind, P., Parkin, D., et al. (2011). Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Quality of Life Research, 20*(10), 1727–1736. https://doi.org/10.1007/s11136-011-9903-x.
3. Janssen, M. F., Birnie, E., Haagsma, J. A., & Bonsel, G. J. (2008). Comparing the standard EQ-5D three-level system with a five-level version. *Value in Health, 11*(2), 275–284. https://doi.org/10.1111/j.1524-4733.2007.00230.x.
4. Poór, A. K., Rencz, F., Brodszky, V., Gulácsi, L., Beretzky, Z., Hidvégi, B., et al. (2017). Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L in psoriasis patients. *Quality of Life Research*. https://doi.org/10.1007/s11136-017-1699-x.
5. Yfantopoulos, J. N., & Chantzaras, A. E. (2017). Validation and comparison of the psychometric properties of the EQ-5D-3L and EQ-5D-5L instruments in Greece. *The European Journal of Health Economics, 18*(4), 519–531. https://doi.org/10.1007/s10198-016-0807-0.
6. Buchholz, I., Janssen, M. F., Kohlmann, T., & Feng, Y.-S. (2018). A systematic review of studies comparing the measurement properties of the three-level and five-level versions of the EQ-5D. *Pharmacoeconomics*. https://doi.org/10.1007/s40273-018-0642-5.
7. Shafie, A. A. (2014). EuroQol 5-Dimension measures in Malaysia. In A. C. Michalos (Ed.), *Encyclopedia of quality of life and well-being research* (pp. 2041–2044). Dordrecht: Springer.
8. Shafie, A., Hassali, M., & Liau, S. (2011). A cross-sectional validation study of EQ-5D among the Malaysian adult population. *Quality of Life Research, 20*(4), 593–600. https://doi.org/10.1007/s11136-010-9774-6.
9. Varatharajan, S., & Chen, W.-S. (2011). Reliability and validity of EQ-5D in Malaysian population. *Applied Research in Quality of Life*. https://doi.org/10.1007/s11482-011-9156-4.
10. Oppe, M., & Van Hout, B. (2017). The ''power'' of eliciting EQ-5D-5L values: the experimental design of the EQ-VT. *EuroQol Working Paper Series, 17003*.
11. Department of Statistics Malaysia Population distribution and basic demographic characteristic report 2010. https://www.statistics.gov.my/index.php?r=column/ctheme&menu_id=L0pheU43NWJwRWVSZklWdzQ4TlhUUT09&bul_id=MDMxdHZjWTk1SjFzTzNkRXYzcVZjdz09.
12. Shannon, C. E. (1949). Communication theory of secrecy systems. *Bell System Technical Journal, 28*(4), 656–715.
13. Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159–174.
14. Faridah, A., Jamaiyah, H., Goh, A., & Soraya, A. (2010). The validation of the EQ-5D in Malaysian dialysis patients. *Medical Journal of Malaysia, 65*(Suppl *A*), 114–119.
15. Craig, B. M., Pickard, A. S., & Lubetkin, E. I. (2014). Health problems are more common, but less severe when measured using newer EQ-5D versions. *Journal of Clinical Epidemiology, 67*(1), 93–99. https://doi.org/10.1016/j.jclinepi.2013.07.011.
16. Kim, T. H., Jo, M.-W., Lee, S., Kim, S. H., & Chung, S. M. (2013). Psychometric properties of the EQ-5D-5L in the general population of South Korea. *Quality of Life Research, 22*(8), 2245–2253.
17. Kim, S. H., Kim, H. J., Lee, S., & Jo, M.-W. (2012). Comparing the psychometric properties of the EQ-5D-3L and EQ-5D-5L in cancer patients in Korea. *Quality of Life Research, 21*(6), 1065–1073.
18. Janssen, M., Pickard, A. S., Golicki, D., Gudex, C., Niewada, M., Scalone, L., et al. (2013). Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: A multi-country study. *Quality of Life Research, 22*(7), 1717–1727.
19. Buchholz, I., Thielker, K., Feng, Y.-S., Kupatz, P., & Kohlmann, T. (2015). 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. *Quality of Life Research, 24*(4), 829–835. https://doi.org/10.1007/s11136-014-0838-x.
20. Greene, M. E., Rader, K. A., Garellick, G., Malchau, H., Freiberg, A. A., & Rolfson, O. (2015). The EQ-5D-5L improves on the EQ-5D-3L for health-related quality-of-life assessment in patients undergoing total hip arthroplasty. *Clinical Orthopaedics and*

Related Research®, 473(11), 3383–3390. https://doi.org/10.1007/s11999-014-4091-y.

21. Yfantopoulos, J., Chantzaras, A., & Kontodimas, S. (2017). Assessment of the psychometric properties of the EQ-5D-3L and EQ-5D-5L instruments in psoriasis. *Archives of Dermatological Research*. https://doi.org/10.1007/s00403-017-1743-2.
22. Scalone, L., Ciampichini, R., Fagiuoli, S., Gardini, I., Fusco, F., Gaeta, L., et al. (2013). Comparing the performance of the standard EQ-5D 3L with the new version EQ-5D 5L in patients with chronic hepatic diseases. *Quality of Life Research, 22*(7), 1707–1716. https://doi.org/10.1007/s11136-012-0318-0.
23. Ramachandran, S., Lundy, J. J., & Coons, S. J. (2008). Testing the measurement equivalence of paper and touch-screen versions of the EQ-5D visual analog scale (EQ VAS). [Article]. *Quality of Life Research, 17*(8), 1117–1120. https://doi.org/10.1007/s11136-008-9384-8.
24. Bagattini, ÂM., Camey, S. A., Miguel, S. R., Andrade, M. V., de Souza Noronha, K. V. M., de M. A. D. C. Teixeira, et al (2018). Electronic version of the EQ-5D quality-of-life questionnaire: Adaptation to a Brazilian population sample. *Value in Health Regional Issues, 17*, 88–93. https://doi.org/10.1016/j.vhri.2017.11.002.
25. Gwaltney, C. J., Shields, A. L., & Shiffman, S. (2008). Equivalence of electronic and paper-and-pencil administration of patient-reported outcome measures: A meta-analytic review. *Value in Health, 11*(2), 322–333. https://doi.org/10.1111/j.1524-4733.2007.00231.x. doi.
26. Rutherford, C., Costa, D., Mercieca-Bebber, R., Rice, H., Gabb, L., & King, M. (2016). Mode of administration does not cause bias in patient-reported outcome results: A meta-analysis. *Quality Of Life Research: An International Journal Of Quality Of Life Aspects Of Treatment, Care And Rehabilitation, 25*(3), 559–574. https://doi.org/10.1007/s11136-015-1110-8.
27. Campbell, N., Ali, F., Finlay, A., & Salek, S. (2015). Equivalence of electronic and paper-based patient-reported outcome measures. *Quality of Life Research, 24*(8), 1949–1961. https://doi.org/10.1007/s11136-015-0937-3.
28. Luo, N., Wang, Y., How, C. H., Wong, K. Y., Shen, L., Tay, E. G., et al. (2015). Cross-cultural measurement equivalence of the EQ-5D-5L items for English-speaking Asians in Singapore. *Quality of Life Research, 24*(6), 1565–1574. https://doi.org/10.1007/s11136-014-0864-8.