

**Psychometric performance of the Chichewa versions of the EQ-5D-Y and EQ-5D-Y-5L
among healthy and sick children and adolescents in Malawi**

Lucky G. Ngwira^{1,2}, Hendramoorthy Maheswaran³, Janine Verstraete^{4,5}, Stavros Petrou⁶,
Louis Niessen^{2, 7}, Sarah C. Smith⁸

Affiliations: [1] Malawi-Liverpool-Wellcome Trust Clinical Research Programme, Malawi; [2] Department of Clinical Sciences, Liverpool School of Tropical Medicine, UK; [3] Imperial College London, UK; [4] University of Cape Town, South Africa; [5] EuroQol Group VMC, Netherlands; [6] University of Oxford, UK; [7] John Hopkins School of Public Health, Baltimore, MD, USA; [8] London School of Hygiene & Tropical Medicine, UK.

ABSTRACT

Objectives

The EuroQol Group has developed a new EQ-5D-Y-5L version with five response levels for each of the five dimensions. The psychometric performance of the EQ-5D-Y has been reported in a number of studies involving both healthy children and those with different health conditions, but that of the EQ-5D-Y-5L has only been reported in a few studies. The aim of this study was to psychometrically evaluate the Chichewa (Malawi) versions of the EQ-5D-Y and EQ-5D-Y-5L.

Methods

The EQ-5D-Y, EQ-5D-Y-5L and PedsQL™ 4.0 self-report Chichewa versions were administered to children and adolescents aged 8-17 years in Blantyre, Malawi.

Results

A total of 298 participants (95 healthy, and 194 chronic and acute) self-completed the questionnaires. There was little problem with missing data except in children aged 8-12 years. The internal consistency reliability was acceptable (Cronbach $\alpha > 0.7$). Convergent validity tested with PedsQL™ 4.0 self-report was found to be satisfactory (correlation > 0.4). However, correlation between some of the EQ-5D-Y dimensions and PedsQL™ 4.0 sub-scales was mixed. While there was evidence of discriminant validity for gender and age (no association with EQ-5D-Y and EQ-5D-Y-5L sum scores), but not with school grade ($p < 0.05$). As regards empirical validity, the EQ-5D-Y-5L was 31%-91% less efficient than the EQ-5D-Y at detecting differences in health status using external measures.

Conclusions

The EQ-5D-Y and EQ-5D-Y-5L were found to be reliable and valid for use among children and adolescents although with some limitation. Further, psychometric testing is required for test re-test reliability and responsiveness that could not be carried out in this study due to COVID-19 restrictions.

Key words: EQ-5D-Y-5L, EQ-5D-Y, childhood, 'health-related quality of life, HRQoL, translation, adaptation, ranking exercise, sub-Saharan Africa

Introduction

The adult EQ-5D-3L, is one of the most widely used preference-based HRQoL measures in health economic evaluations. (1) The prominence of the EQ-5D-3L can be attributed to among others, the simple descriptive system which makes it relatively easy for self-completion and valuation exercises to generate utilities, and recommendation for use by the National Institute for Health and Care Excellence in health technology assessments in the UK. (2) The EQ-5D-3L has been validated in different populations but has been criticized for being too simplistic and less sensitive to small changes in health status, which led to the development of the five response level EQ-5D-5L. (3) Research suggests the EQ-5D-5L performs better, less affected by the ceiling effect and improved known-groups validity compared to the EQ-5D-3L. (4-6). The EQ-5D-Y and the EQ-5D-Y-5L, currently experimental before becoming an official version, have emerged from the adult EQ-5D versions. (7, 8) The EQ-5D-Y-5L has been developed on the same premise as the adult EQ-5D-5L version to increase its sensitivity and to reduce ceiling effects. (3) Psychometric performance of the EQ-5D-Y has been reported in different studies involving children with different health conditions, (9-11) To a large extent, the EQ-5D-Y has demonstrated good reliability, with acceptable levels of convergent, discriminant and known-group validity. (12-14) One systematic review on the use of the EQ-5D-Y between 2000-2010 reported that most studies have reported problems with missing values. (15) On the other hand, the performance of the EQ-5D-Y-5L has only been reported in a few studies owing to it being a new instrument. (16-18) In those studies, the EQ-5D-Y-5L has demonstrated to be feasible and to reduce ceiling effect, but it has not performed differently on other psychometric properties than the EQ-5D-Y. (16, 18, 19)

Neither the EQ-5D-Y nor the EQ-5D-Y-5L have been psychometrically evaluated in Malawi. As reported previously, (20) economic evaluation is becoming increasingly important in

Malawi in health programs. The application of the adult EQ-5D-3L (and EQ-5D-5L), EQ-5D-Y and EQ-5D-Y-5L could be key instruments for assessing whether these goals are achieved. The EQ-5D-Y and EQ-5D-Y-5L are preference-based HRQoL measures commonly used in cost utility (and effectiveness) analysis, as such their utilities need to be evaluated as well as their psychometric properties. (21) The EQ-5D-Y utility algorithm has only been developed recently and due to this, researchers have tended to apply adult utilities. (15, 22) This approach although not encouraged by the EuroQol Group, (23) is used in absence of youth utilities. The pros and cons of using adult or proxy utilities, as well as de novo valuation techniques have been discussed at length elsewhere and are not the subject of this paper. (9, 24-28) Generally, there is a discrepancy between children and adult/proxy utilities, (29) and this is the case even among adults. (30) The sensitivity of the EQ-5D-Y vis-à-vis that of the EQ-5D-Y-5L have not been assessed elsewhere owing to the latter being an experimental version.

The aim of this study was to psychometrically evaluate the EQ-5D-Y and EQ-5D-Y-5L Chichewa language versions in general population, chronic and acutely ill children and adolescents in Malawi.

Methods

Participants, recruitment and procedure

The study recruited participants from a convenience sample of healthy and ill, children (≤ 12 years) and adolescents (> 12 years) in Blantyre, Malawi's commercial capital. Children and adolescents attending schools (primary and secondary) and those visiting the out-patient department at the Queen Elizabeth Central Hospital made up healthy and sick participants respectively. Written assent and consent were obtained from children and parents/guardian, respectively. The invitation to participate for sick participants came at the end of clinical care, and through the schoolteacher for healthy participants. For both sets of participants,

questionnaires were distributed by the research team.

As it has been observed that patients avoid the intermediate responses in the equivalent adult EQ-5D-5L questionnaire if the EQ-5D-3L is administered first, (4) the EQ-5D-Y-5L was administered before the EQ-5D-Y. The PedsQL™4.0 child self-reports and PedsQL™ 4.0 teen self-reports were then administered to those aged 8-12 years and 13-17 years respectively.

i. The instruments

The EQ-5D-Y

The EQ-5D-Y consists of five dimensions: ‘mobility’, ‘looking after oneself’, ‘doing usual activities’, ‘having pain or discomfort’, and ‘feeling worried, sad or unhappy’, (and visual analogue scale (EQ VAS)), each with three severity (response) levels: 1 “no problems”, 2 “some problems /a bit”, and 3 “a lot of problems /very”. Self-rated health is also assessed with an EQ VAS on a thermometer like vertical scale with scores ranging between 0 (representing worst imaginable health) and 100 (representing best imaginable health). The EQ-5D-Y has a recall period of “today”. (143)

The EQ-5D-Y-5L

The EQ-5D-Y-5L instrument consists of the same five dimensions as the EQ-5D-Y version but with five responses: 1 ‘no problems/not’, 2 ‘a little bit of a problem’, 3 ‘some problems /quite’, 4 ‘a lot of problems/really’ and 5 ‘extreme problems/extremely/cannot’. The Chichewa version of both the EQ-5D-Y and EQ-5D-Y-5L has been reported elsewhere. (20) Additionally, the EQ-5D-Y-5L included a page for collecting demographic and medical condition.

The EQ-5D-Y and EQ-5D-Y-5L were firstly scored using the sum scores by summing the responses. A health state (represented by responses) '11111' (denoting a one for each of the five dimensions) had a minimum sum score of 5. Similarly, '33333' and '55555' health states had maximum sum scores of 15 and 25 for the EQ-5D-Y and EQ-5D-Y-5L respectively. The other health states had a sum score ranging between minimum and maximum scores. The sum score is a crude measure with some limitations but gives a better indication of the dimension performance. (31) Secondly, utility scores were generated from adult utilities as the youth sets were not available at the time of doing this study. Further, few countries have adult utilities for both the EQ-5D-3L and EQ-5D-5L, and none of these are in Africa. (32) The utility scores for the EQ-5D-Y and EQ-5D-Y-5L were generated from the USA utilities developed by Shaw et al (33) and Pickard et al (34), respectively on recommendation from the EuroQol Group.

Self-rated general health

A self-rated general health was also included through the question: How would you rate your health today? Excellent, very good, good, fair, or poor? Self-rated health question is considered a reliable and valid, although limited, measure of HRQoL. (35) Nevertheless, it has been used in various studies involving children and adolescents as a crude measure for comparison. (18)

The PedsQL™ 4.0 self-report

The Chichewa versions of the PedsQL™ 4.0 child self-report (for ages 8-12 years) and the PedsQL™ 4.0 teen self-report (for ages 13-18 years) were also administered. The translation process and approval were provided by the Mapi Trust. (36) Both the PedsQL™ 4.0 child self-report and the PedsQL™ 4.0 teen self-report (herein referred to as PedsQL™ 4.0 if referring

to both for brevity) have 23 items in five dimensions. The only difference between the PedsQL[™] 4.0 child self-report and the PedsQL[™] 4.0 teen self-report version is with reference to use of ‘kids’ and ‘teens’ respectively in some items.

The PedsQL[™] 4.0 total scale score is obtained by scoring across all 23 items (higher = better) and the Physical Functioning subscale (sum of eight Physical Health items), whereas the last three subscales (15 items) are combined to form Psychosocial Health scale score. The subscale scores are obtained through the summation of all the items that are answered. This is done in order to account for the missing items if present. (37, 38)

Psychometric analyses

Data analysis were performed using IBM SPSS 26.0.0 for Mac (IBM Corp. Armonk, New York, USA). (39) The sample was divided into two groups: children and adolescents. This was to reflect the age ranges for the PedsQL[™] 4.0 child self-reports (for age 8-12 years) and PedsQL[™] 4.0 teen self-reports (for age 13-18 years). Psychometric analyses were conducted using gold standard techniques based on four psychometric guidelines. (40-43)

ii. Item performance

The proportion of missing responses and distribution of responses in each EQ-5D-Y and EQ-5D-Y-5L dimension was calculated. The following four aspects of item analysis were evaluated.

Missing data was defined as EQ-5D-Y and EQ-5D-Y-5L dimensions that had $\geq 5\%$ missing data were flagged as potentially problematic. For maximum endorsement frequency,

dimensions were considered problematic if they had >80% endorsement at either end of the scale. Dimensions were considered problematic if any two or more adjacent response options summed to <10% for adjacent aggregate endorsement frequency. The criteria for item redundancy was not met if pairs of items had correlations > 0.75.

iii. Reliability

This was assessed using Cronbach's alpha ≥ 0.70 to indicate acceptable internal consistency.

iv. Construct validity

Convergent validity

It was hypothesized that the EQ-5D-Y and EQ-5D-Y-5L sum scores would be correlated with PedsQL™4.0 total scale scores. Further, the EQ-5D-Y mobility and feeling worried, sad or unhappy would be correlated with PedsQL™ 4.0 physical and psychosocial health summary scores respectively. The correlation should be negative because lower sum scores for EQ-5D-Y and EQ-5D-Y-5L means better HRQoL whereas higher PedsQL™ 4.0 score means better HRQoL. Generally, a correlation ≥ 0.4 is considered moderate to strong.

Discriminant validity

It was anticipated that there would be correlation <0.2 between EQ-5D-Y, EQ-5D-Y-5L sum scores and age. No association at the 5% significance level was also anticipated between EQ-5D-Y and EQ-5D-Y-5L sum scores, with gender and grade. Grade was dichotomised based on general distribution and in line with the former scaling for primary school education in Malawi, grades 1-5 made a lower primary grade (group 1), grades 6-8 made upper primary grade (group 2), and Secondary/High school grades (grades 9-12) made secondary school grade (group 3). A one-way ANOVA test was performed to assess association between sum scores and grade.

v. Known-groups validity

It was hypothesised that the EQ-5D-Y and EQ-5D-Y-5L sum scores would be worse for the sick children compared to that of their healthy counterpart. The relationship was investigated using a t-test evaluated using an effect size according to Cohen's criterion: <0.2 poor, 0.3-0.49 small, 0.5-0.8 moderate, and >0.8 large. (44, 45)

vi. Utility performance (Empirical validity)

It was hypothesized that the utility scores generated by the EQ-5D-Y and EQ-5D-Y-5L would detect differences in external indicators of health status with the latter possessing greater ability. The EQ-5D-Y-5L was developed on the premise to address ceiling effect and sensitivity and therefore expected to have better characteristics compared to the EQ-5D-Y. (3)

The relative ability to assess external indicators of health status was investigated by comparing the utility scores with the self-reported general health and the PedsQL™4.0 total scale scores. This relative ability was estimated using the relative efficiency statistic, defined as 'the ratio of the square of the t-statistic of the comparator instrument over the square of the t-statistic of the reference instrument'. (46) The EQ-5D-Y-5L acted as the comparator instrument in this study and the EQ-5D-Y as the reference since the latter has been widely used and psychometrically validated. (8) Relative efficiency equal to 1.0 indicates that the EQ-5D-Y-5L has the same efficiency as the EQ-5D-Y at detecting differences in health status using the external indicator; Relative efficiency >1.0 indicates that the EQ-5D-Y-5L is more efficient than the EQ-5D-Y at detecting differences in health status; and the converse is true.(46)

Self-reported general health and PedsQL™4.0 total scale scores were dichotomised. A frequency distribution determined the cut-off (46) for dichotomising self-reported general health status in two: i) excellent or very good versus good or fair or poor, and ii) excellent versus very good or good or fair or poor. The mean for the total scale scores provided a cut-off for the PedsQL™4.0 such that below mean, and mean and above formed two categories.

All empirical validity analyses were based on participants who completed both the EQ-5D-Y-5L and EQ-5D-Y. All missing responses for the EQ-5D-Y-5L and EQ-5D-Y measures were not included in this analysis but not for the PedsQL™4.0 as per respective scoring algorithm. There is possibility that lower scores below the utility score of 0 (which could lead to under predicting poorest health states) would be different for the two instruments since utility scores are based on two different valuation models. (34) To overcome this, the relative efficiency statistic was re-calculated by restricting utility scores between 0 and 1 only.

Results

Participant characteristics

In total, 289 participants completed the EQ-5D-Y, EQ-5D-Y-5L, and PedsQL™ 4.0, aged 8-17 years (mean 13.6, median 14) as presented in last chapter and re-presented in Table 1.

Table 1 Participant characteristics

Characteristic		N (%)	Age group 1 (8-12yrs)	Age group 2 (13-17yrs)
Participants		289	98	191
Gender*	Male	121 (44%)	39	82
	Female	153 (56%)	51	102
Health condition	healthy	95 (33%)	12	83
	acute	155 (54%)	85	70
	chronic	39 (13%)	1	38
Grade [#]	1-5	71 (25%)	53	18
	6-8	97 (35%)	40	57
	9-12	111 (40%)	0	111

*missing data: 15 (age group 1 =8, age group 2=7); [#]missing data: 10 (5 in each age groups)

Age group 1 completed EQ-5D-Y, EQ-5D-Y-5L and PedsQL 4.0 child self-report

Age group 2 completed EQ-5D-Y, EQ-5D-Y-5L and PedsQL 4.0 teen self-report

Item performance

Table 2 shows that all items failed criterion for missing data for the EQ-5D-Y among children compared to none among adolescents. While no items failed on maximum endorsement frequency among children, three items failed criterion among adolescents. All items passed aggregate endorsement frequency criterion among and only one item failed among adolescents. No item failed on the item redundancy criterion in both age groups.

For the EQ-5D-Y-5L, three items in children compared to one among adolescents failed criterion for missing data. None of the items failed maximum endorsement frequency criterion in children compared to two that failed in adolescents. While all items failed maximum endorsement frequency criterion, all items passed item redundancy in both age groups.

Table 2 Item performance for EQ-5D-Y and EQ-5D-Y-5L by age

Measure	Dimension	Item level performance *							
		MD		MEF		AEF		IR	
		8- 12yrs	13- 18yrs	8- 12yrs	13- 18yrs	8- 12yrs	13- 18yrs	8- 12yrs	13- 18yrs
EQ-5D-Y	Mobility	x			x				
	Looking after oneself	x			x		x		
	Doing usual activities	x			x				
	Having pain or discomfort	x							
	Feeling worried, sad or unhappy	x							
EQ-5D-Y-5L	Mobility				x	x	x		
	Looking after oneself				x	x	x		
	Doing usual activities	x				x	x		
	Having pain or discomfort	x	x			x	x		
	Feeling worried, sad or unhappy	x				x	x		

* x represent dimension failed on the criteria; MD: Missing data; MEF: Maximum endorsement frequency; AEF: Adjacent aggregate endorsement frequency; IR: Item redundancy

The item level analysis based on health status of the participants for both the EQ-5D-Y and EQ-5D-Y-5L are presented in Table 3. For the EQ-5D-Y, missing data was not observed in healthy and chronic children, but two items failed criterion among the acutely ill. While all items failed maximum endorsement frequency among the healthy and chronically ill, none failed among the acutely ill. No items failed aggregate endorsement frequency and item redundancy among healthy and acutely ill, but in the chronically ill children.

For the EQ-5D-Y-5L, missing data was similarly not observed in healthy and chronic children, but one item failed criterion among the acutely ill. Three and five items failed maximum endorsement frequency among the healthy and chronically ill respectively, but none failed among the acutely ill. Whereas all items failed aggregate endorsement frequency among healthy and acutely ill, this was not observed in the chronically ill children. No item failed on item redundancy among the healthy and acutely ill, but this was observed in all items among the chronically ill.

Table 3 Item performance for EQ-5D-Y and EQ-5D-Y-5L by health status

Measure	Dimension	Item level performance*											
		Missing data			Maximum endorsement frequency			Adjacent endorsement frequency			Item redundancy		
		healthy	chronic	acute	healthy	chronic	acute	healthy	chronic	acute	healthy	chronic	acute
EQ-5D-Y	Mobility				x	x			x				
	Looking after oneself			x	x	x							
	Doing usual activities				x	x							
	Having pain or discomfort				x	x			x			x	
	Feeling worried, sad or unhappy			x	x	x							
EQ-5D-Y-5L	Mobility				x	x		x		x		x	
	Looking after oneself				x	x		x		x		x	
	Doing usual activities			x	x	x		x		x		x	
	Having pain or discomfort					x		x		x		x	
	Feeling worried, sad or unhappy					x		x		x		x	

* x represent dimension failed on the criteria

Reliability

The overall Cronbach's alpha was 0.766 and 0.838 for the EQ-5D-Y and EQ-5D-Y-5L respectively.

Construct validity

Convergent validity

Results of convergent validity are summarised in Table 4. For the EQ-5D-Y, there was correlation above the hypothesized threshold between EQ-5D-Y sum scores and PedsQL™ 4.0 total scale score: -0.392 with PedsQL™ 4.0 child self-report and -0.399 with PedsQL™ 4.0 teen self-report. The correlation between EQ-5D-Y dimensions and PedsQL™ 4.0 health summary scores was rather mixed, e.g., the EQ-5D-Y mobility/PedsQL™ 4.0 Physical health correlation was: -0.254 and -0.343 with PedsQL™ 4.0 child and teen self-report respectively.

The EQ-5D-Y-5L sum score and PedsQL™ 4.0 total scale score correlation was similarly as hypothesized -0.397 with PedsQL™ 4.0 child self-report and -0.375 with PedsQL™ 4.0 teen self-report. The EQ-5D-Y-5L mobility dimension and PedsQL™ 4.0 Physical health summary score correlation was similarly mixed -0.265 and -0.391 with PedsQL™ 4.0 child and PedsQL™ 4.0 teen self-report respectively.

Table 4 Convergent validity of the EQ-5D-Y and EQ-5D-Y-5L with PedsQL™ 4.0 self-report sub-scale.

Measure	Dimension / Sub-scale	PedsQL 4.0 child self-report			PedsQL 4.0 teen self-report		
		Psychosocial Health Summary	Physical Health Summary	Total Scale score	Psychosocial Health Summary	Physical Health Summary	Total Scale score
EQ-5D-Y	Mobility	-.301**	-.254*	-.312**	-.371**	-.343**	-.380**
	Looking after oneself	-.240*	-.250*	-.276*	-.212**	-0.117	-.163*
	Doing usual activities	-.344**	-.316**	-.376**	-.266**	-.213**	-.252**
	Having pain or discomfort	-0.122	-0.138	-0.146	-.190**	-.155*	-.172*
	Worried, sad or unhappy	-.449**	-.303**	-.432**	-.259**	-.283**	-.291**
	EQ-5D-Y sum score	-.347**	-.351**	-.392**	-.406**	-.353**	-.399**
EQ-5D-Y-5L	Mobility	-0.148	-.265*	-0.202	-.455**	-.391**	-.440**
	Looking after oneself	-.330**	-.325**	-.370**	-.156*	-0.056	-0.092
	Doing usual activities	-.221*	-0.177	-.229*	-.309**	-.211**	-.260**
	Having pain or discomfort	-.328**	-.365**	-.377**	-.357**	-.267**	-.314**
	Worried, sad or unhappy	-.359**	-.276**	-.353**	-.341**	-.365**	-.379**
	EQ-5D-Y-5L sum score	-.362**	-.366**	-.397**	-.405**	-.324**	-.375**

** . Correlation is significant at the 0.01 level (2-tailed); * . Correlation is significant at the 0.05 level (2-tailed).

Discriminant validity

There was no statistically significant difference between EQ-5D-Y or EQ-5D-Y-5L sum scores and gender: EQ-5D-Y (mean = 0.036, $t = 1.799$, $p < 0.074$) and the EQ-5D-Y-5L (mean = 0.05, $t = 0.143$, $p < 0.886$) among adolescents. There was similar no significance between sum scores and gender in children.

There was low correlation between age and the sum scores of both the EQ-5D -Y, and EQ-5D-Y-5L (range 0.113 – 0.173) showing age was not associated with scores. The correlation between age and EQ-5D-Y in adolescents was borderline. (-0.201)

While there was no evidence of difference between the EQ-5D-Y and EQ-5D-Y-5L sum scores and grade categories in the younger age group, there was a statistically significant association of sum scores and grade categories in adolescents ($p < 0.05$). When age was combined, the association was found to be statistically significant between EQ-5D-Y and EQ-5D-Y-5L sum scores with grade categories ($p < 0.001$). The lack of association between grade and EQ-5D-Y and EQ-5D-Y-5L sum score in children would indicate that grade had no effect on scores but not among adolescents and when age was combined. (Table 5)

Table 5 EQ-5D-Y and EQ-5D-Y-5L sum score discriminant validity by grade

Measure		Age 8-12 years (grade* grp1=53, grade grp 2=40)				Age 13-18yrs years (grade* grp1=18, grade grp 2=57, grade grp3=111)				Age 7-18yrs years (grade* grp1=71, grade grp 2=97, grade grp3=111)			
		df	Mean Square	F	p-value	df	Mean Square	F	p-value.	Df	Mean Square	F	p-value.
EQ-5D-Y sum score	Between Groups	1	0.389	0.114	0.737	2	5.122	3.074	0.049	2	23.110	10.335	0.000
	Within Groups	79	3.420			171	1.667			252	2.236		
	Total	80				173				254			
EQ-5D-Y-5L sum score	Between Groups	1	1.539	0.146	0.703	2	28.585	4.741	0.010	2	56.435	7.491	0.001
	Within Groups	89	10.530			175	6.030			266	7.534		
	Total	90				177				268			

*grade group1 = grade 1-5; group2 = grade 6-8; group3 = grade 9-12

Known-group validity

In children, effect size was low (0.23) for EQ-5D-Y-5L compared to high (-1.15) for the EQ-5D-Y. In adolescents, effect sizes were generally much higher (>0.5) suggesting reasonably good known-groups validity. (Table 6)

Table 6 EQ-5D-Y and EQ-5D-Y-5L sum score known group validity

Measure	Age 8-12 years (N=98, healthy=12, 81= sick) *					Age 13-18 years (N=191, healthy=83, 106= sick) *				
	t-statistic #		Mean Diff	SD	Effect size ¶	t-statistic #		Mean Diff	SD	Effect size ¶
	t	p-value				t	p-value			
EQ-5D-Y	-0.850	0.398	-0.957	4.196	-0.23	-3.480	0.001	-1.262	1.220	-1.03
5L sum score	-0.751	0.466	-0.957	3.556		-3.715	0.000	-1.262	3.088	
EQ-5D-Y	-2.154	0.034	-1.391	1.206	-1.15	-2.757	0.006	-0.536	0.884	-0.61
sum score	-3.172	0.005	-1.391	2.083		-2.856	0.005	-0.536	1.566	

*2 and 5 did not complete EQ-5D-Y and EQ-5D-Y-5L respectively

assuming equal variance

¶ effect size designated as <0.2 poor, 0.3-0.49 small, 0.5-0.8 moderate, and >0.8 large

Empirical validity

Table 7 presents the relative efficiency statistics for the EQ-5D-Y and EQ-5D-Y-5L over the dichotomous self-reported general health status and PedsQL™ 4.0 HRQoL measures, respectively. When the EQ-5D-Y was referenced at 1.0, the EQ-5D-Y-5L was 31% to 91% and 5% to 44% less efficient than the EQ-5D-Y at detecting differences in self-reported general health and the PedsQL™ 4.0 total scale score respectively.

Restricting the analyses to utility scores between 0 and 1 reduced the sample size by four, but the EQ-5D-Y-5L continued not to perform as hypothesized against the EQ-5D-Y with just one exception.

Table 7 Efficiency of the EQ-5D to detect differences in self-reported health status

Measure	Age	Categorisation of self-reported health status and PedsQL 4.0 scale score	Utility score [#]		t-test*		Relative efficiency
			mean	(SD)	t-statistic	p-value	
EQ-5D-Y	age 8-12yrs (n=81)	excellent or v. good	0.838	0.222	2.075	0.041	1.000
		good or fair	0.746	0.175			
EQ-5D-Y		excellent or v. good	0.812	0.328	0.510	0.612	0.060
5L		good or fair	0.780	0.243			
EQ-5D-Y		excellent	0.872	0.200	2.197	0.033	1.000
		v. good, good, fair or poor	0.766	0.205			
EQ-5D-Y	age 13-17yrs (n=172)	excellent	0.832	0.321	0.660	0.513	0.090
5L		v. good, good, fair or poor	0.783	0.284			
EQ-5D-Y		excellent or v. good	0.903	0.137	0.148	0.883	1.000
		good or fair	0.899	0.150			
EQ-5D-Y		excellent or v. good	0.911	0.175	0.123	0.902	0.693
5L		good or fair	0.907	0.160			
EQ-5D-Y	combined ages 7-17yrs (n=253)	excellent	0.933	0.116	2.205	0.029	1.000
		v. good, good, fair or poor	0.887	0.150			
EQ-5D-Y		excellent	0.924	0.210	0.704	0.483	0.102
5L		v. good, good, fair or poor	0.902	0.147			
EQ-5D-Y		excellent or v. good	0.883	0.169	1.733	0.085	1.000
		good or fair	0.844	0.175			
EQ-5D-Y	age 8-12yrs (n=81)	excellent or v. good	0.881	0.235	0.702	0.484	0.164
5L		good or fair	0.862	0.202			
EQ-5D-Y		excellent	0.913	0.150	3.027	0.003	1.000
		v. good, good, fair or poor	0.848	0.178			
EQ-5D-Y		excellent	0.895	0.253	0.945	0.346	0.098
5L		v. good, good, fair or poor	0.864	0.208			
EQ-5D-Y	age 8-12yrs (n=81)	≥ 72.79	0.840	0.159	2.298	0.025	1.000
		< 72.79	0.727	0.249			
EQ-5D-Y	age 13-17yrs (n=172)	≥ 72.79	0.865	0.223	2.237	0.030	0.948
5L		< 72.79	0.705	0.363			
EQ-5D-Y	combined ages 7-17yrs (n=253)	≥ 78.68	0.946	0.087	3.837	0.000	1.000
		< 78.68	0.864	0.167			
EQ-5D-Y	age 8-12yrs (n=81)	≥ 78.68	0.947	0.151	2.863	0.005	0.557
5L		< 78.68	0.872	0.177			
EQ-5D-Y	age 13-17yrs (n=172)	≥ 76.81	0.918	0.114	4.716	0.000	1.000
		< 76.81	0.812	0.210			
EQ-5D-Y	combined ages 7-17yrs (n=253)	≥ 76.81	0.929	0.170	4.102	0.000	0.756
5L		< 76.81	0.808	0.265			

[#] US utilities

*assuming equal variance

Discussion

In this setting, the Chichewa versions demonstrated good internal consistency, and some mixed evidence of item performance and validity. Overall, both the EQ-5D-Y and EQ-5D-Y-5L showed good evidence of reliability but with mixed validity. Although reliability results are acceptable, there are some issues with item performance especially endorsement and redundancy for EQ-5D-Y-5L. This suggests that the response options are possibly not working as intended and also dimensions may not be making unique contributions to the scale. In general, both Chichewa versions demonstrated that they can be used with some limitations in children and adolescents in this setting. Other psychometric properties like test-retest and responsiveness also need to be evaluated.

The item performance shows that both the EQ-5D-Y and EQ-5D-Y-5L had no item that is redundant. Adolescents in the combined sample ably self-completed the questionnaires with no missing values, and this confirms what has been observed in several EQ-5D-Y studies. (10, 11, 47) The adult EQ-5D-3L and EQ-5D-5L versions have similarly demonstrated negligible missing values. (48) However, missing data found in children and in acute participants would indicate difficulty in providing good quality HRQoL data among the younger population when using self-completion questionnaires. The tendency to pick responses from the top end of an item (ceiling effect) has been less evident in adolescents compared to children in most studies, (49) although this study found the opposite. Like the adult EQ-5D-5L counterpart, the EQ-5D-Y-5L edged the EQ-5D-Y in reducing answering from top end of a scale. (48, 50) The EQ-5D-Y-5L failing the adjacent aggregate endorsement frequency might indicate that some responses were not relevant to this population. This finding is comparable to what has been reported elsewhere. (47, 49) One previous study suggested that the use of a 3-point response scale is

better to achieving reliability of as it gets rid of irrelevant responses. (51) However, doing so would be in direct contrast to the intention of the EQ-5D-Y-5L which was developed precisely to increase sensitivity in health status reporting. (8) It could be reasoned that perhaps the EQ-5D-Y-5L is doing just that.

The evidence for construct validity shows that criteria for convergent validity were met at scale but not dimension level. This might imply that the EQ-5D-Y and EQ-5D-Y-5L are best suited to assess physical functioning as opposed to other aspects of HRQoL. While the adult EQ-5D-5L has been found to be associated with high correlation with other health measure compared to the EQ-5D-3L, (52-54) this was not the case between the two youth versions. The EQ-5D-Y and EQ-5D-Y-5L equally demonstrated convergent validity with PedsQL™4.0.

There were no differences between the discriminant ability of the EQ-5D-Y and EQ-5D-Y-5L as regards gender and age. The adult EQ-5D-3L and EQ-5D-5L have equally demonstrated discriminant validity with gender in young adults in a previous study, (55) with little difference between the two measures. (50) This study confirms these findings although there was a negative correlation observed between scores and age among adolescents and a positive one in children. This could not be established in this study and might benefit from further research. While age has been associated with different scores depending on the EQ-5D-3L or EQ-5D-5L, (50) this study did not find such a difference between the EQ-5D-Y and EQ-5D-Y-5L.

Although discriminant validity criteria between both the EQ-5D-Y and EQ-5D-Y-5L and grade was met in children, this was not the case among adolescents and when age was combined. This may indicate that older children or education exposure contributes to better

comprehension of health possibly due to confidence and experience with completing questionnaires for older children.

There were differences between the scores of the healthy and sick children across different health conditions for both the EQ-5D-Y and EQ-5D-Y-5L. While the EQ-5D-Y had the largest effect size in children, the EQ-5D-Y-5L was able to discriminate better in adolescents. The adult EQ-5D-5L ability to distinguish between groups based on age. (56) The findings from this study confirms that the EQ-5D-Y-5L may be best suited for older population since they have better distinguishing ability of the responses.

Finally, unlike what was hypothesized, the EQ-5D-Y outperformed the EQ-5D-Y-5L in establishing empirical validity which was rather surprising. Elsewhere, the SF-6D has performed better than the EQ-5D-3L because of the better sensitivity with the former. (46) The fact that the participants were relatively healthy (even among those sick) could possibly have resulted in better response for the EQ-5D-Y compared to EQ-5D-Y-5L. The EQ-5D-5L has also been found to overestimate health problems leading to underestimation of utilities compared to the EQ-5D-3L. (5) This overestimation can lead to underestimation of utilities resulting in the EQ-5D-3Y outperforming the EQ-5-Y-5L. In general, the EQ-5D-5L has been found to demonstrate better relative efficiency compared to the EQ-5D-3L. (57-59) The full understanding of why the EQ-5D-Y performed better than the EQ-5D-Y-5L in predicting health scores is beyond this study. Future studies could test the hypothesis stated above.

The psychometric evaluation of the Chichewa EQ-5D-Y and EQ-5D-Y-5L versions has some limitations. Firstly, data was only collected from one time point and consequently test-retest reliability as well as responsiveness could not be evaluated due to COVID-19 restrictions. This

might have an effect in ascertaining the reliability of the Chichewa versions of the EQ-5D-Y and EQ-5D-Y-5L in a clinical setting.

Another limitation is to do with the utilities used for empirical validity. While none are available for the EQ-5D-Y-5L, these have only recently been developed in two countries for the EQ-5D-Y. (60, 61) This is an area that need further research globally.

Conclusion

The EQ-5D-Y established internal consistency among children and adolescents but was not able to assess test-retest reliability. The EQ-5D-Y had problems with item performance and known-groups validity. The EQ-5D-Y-5L similarly demonstrated internal consistency and validity except discriminant validity with grade. The study found both the EQ-5D-Y and the EQ-5D-Y-5L have problems with missing data in younger children. In addition, both the EQ-5D-Y and the EQ-5D-Y-5L have problems with endorsement suggesting that the response options are not working as intended in this setting. These issues are likely to be not specific to Malawi though as shown by evidence from elsewhere. However, some psychometric problems seen in Malawi are not caused by the translation/cultural but may be specific to the context.

In conclusion, there was not much difference between the performance of the EQ-5D-Y and the EQ-5D-Y-5L. Whereas the EQ-5D-Y-5L performed better than the EQ-5D-Y in discriminant validity, the latter outperformed the EQ-5D-Y-5L in the utilization of response options as well as empirical validity. Both instruments can be used among children and adolescents in Malawi although with some limitation. Further psychometric testing is required for both test re-test reliability and responsiveness which could not be carried out in this study.

Reference

1. Chen G, Ratcliffe J. A Review of the Development and Application of Generic Multi-Attribute Utility Instruments for Paediatric Populations. *Pharmacoeconomics*. 2015; 33: 1013-28.
2. NICE. Guide to the methods of technology appraisal 2013. Proces and methods. 2013.
3. Herdman M, Gudex C, Lloyd A, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res*. 2011; 20: 1727-36.
4. Janssen M, Birnie E, Haagsma JA, et al. Comparing the Standard EQ-5D Three-Level System with a Five-Level Version. *Value Health*. 2008; 11: 275-84.
5. Janssen MF, Bonsel GJ, Luo N. Is EQ-5D-5L Better Than EQ-5D-3L? A Head-to-Head Comparison of Descriptive Systems and Value Sets from Seven Countries. *Pharmacoeconomics*. 2018; 36: 675-97.
6. Janssen MF, Szende A, Cabases J, et al. Population norms for the EQ-5D-3L: a cross-country analysis of population surveys for 20 countries. *Eur J Health Econ*. 2019; 20: 205-16.
7. Wille N, Badia X, Bonsel G, et al. Development of the EQ-5D-Y: a child-friendly version of the EQ-5D. *Qual Life Res*. 2010; 19: 875-86.
8. Kreimeier S, Astrom M, Burstrom K, et al. EQ-5D-Y-5L: developing a revised EQ-5D-Y with increased response categories. *Qual Life Res*. 2019.
9. Kreimeier S, Greiner W. EQ-5D-Y as a Health-Related Quality of Life Instrument for Children and Adolescents: The Instrument's Characteristics, Development, Current Use, and Challenges of Developing Its Value Set. *Value Health*. 2019; 22: 31-37.
10. Ravens-Sieberer U, Wille N, Badia X, et al. Feasibility, reliability, and validity of the EQ-5D-Y: results from a multinational study. *Qual Life Res*. 2010; 19: 887-97.
11. Burstrom K, Bartonek A, Brostrom E, et al. EQ-5D-Y as a health-related quality of life measure in children and adolescents with functional disability in Sweden: testing feasibility and validity. *Acta Paediatrica*. 2014; 103: 426-35.
12. Shirowa T, Fukuda T, Shimozuma K. Psychometric properties of the Japanese version of the EQ-5D-Y by self-report and proxy-version: reliability and construct validity. *Quality of Life Research*. 2019; Epub ahead of print.
13. Scott D, Ferguson GD, Jelsma J. The use of the EQ-5D-Y health related quality of life outcome measure in children in the Western Cape, South Africa: psychometric properties, feasibility and usefulness - a longitudinal, analytical study. *Health Qual Life Outcomes*. 2017; 15: 12.
14. Scalone L, Ciampichini R, Fagioli S, 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: 1707-16.
15. Noyes J, Edwards RT. EQ-5D for the assessment of health-related quality of life and resource allocation in children: a systematic methodological review. *Value Health*. 2011; 14: 1117-29.
16. Wong CKH, Cheung PWH, Luo N, et al. A head-to-head comparison of five-level (EQ-5D-5L-Y) and three-level EQ-5D-Y questionnaires in paediatric patients. *Eur J Health Econ*. 2019.
17. Wong CKH, Cheung PWH, Luo N, et al. Responsiveness of the EQ-5D youth version 5-level (EQ-5D-5L-Y) and 3-level (EQ-5D-3L-Y) in

patients with idiopathic scoliosis. *Spine*. 2019; 44: 1507-14.

18. Astrom M, Krig S, Ryding S, et al. EQ-5D-Y-5L as a patient-reported outcome measure in psychiatric inpatient care for children and adolescents - a cross-sectional study. *Health Qual Life Outcomes*. 2020; 18: 164.

19. Zhou W, Shen A, Yang Z, et al. Patient-caregiver agreement and test-retest reliability of the EQ-5D-Y-3L and EQ-5D-Y-5L in paediatric patients with haematological malignancies. *Eur J Health Econ*. 2021.

20. Ngwira LG, Jelsma J, Maheswaran H, et al. Cross-cultural adaptation of the beta EQ-5D-Y-5L into Chichewa (Malawi) Accepted VIHRI. 2021.

21. Brazier J, Deverill M. A checklist for judging preference-based measures of health related quality of life: learning from psychometrics. *Health Econ*. 1999; 8: 41-51.

22. Thorrington D, Eames K. Measuring Health Utilities in Children and Adolescents: A Systematic Review of the Literature. *PLoS One*. 2015; 10: e0135672.

23. The EuroQol G. EQ-5D-Y User Guide Basic information on how to use the EQ-5D-Y instrument. 2014.

24. Petrou S. Methodological issues raised by preference-based approaches to measuring the health status of children. *Health Econ*. 2003; 12: 697-702.

25. Ravens-Sieberer U, Erhart M, Wille N, et al. Generic Health-Related Quality-of-Life Assessment in Children and Adolescents: Methodological Considerations. *Pharmacoeconomics*. 2006; 24: 1199-220.

26. Petrou S. Should health gains by children be give the same value as health gains by adults in an economic evaluation framework. In: Ungar W, ed., *Economic Evaluation in Child Health*. Oxford: Oxford Unviersity Press, 2010.

27. Powell PA, Rowen D, Rivero-Arias O, et al. Valuing child and adolescent health: a qualitative study on different perspectives and priorities taken by the adult general public. *Health Qual Life Outcomes*. 2021; 19: 222.

28. Rowen D, Mukuria C, Powell P, et al. Valuing child health: exploring the use of a mixed sample of adolescents and adults to value child and adolescent health states. NICE DSU Report, 2021.

29. Mott DJ, Shah KK, Ramos-Goni JM, et al. Valuing EQ-5D-Y-3L Health States Using a Discrete Choice Experiment: Do Adult and Adolescent Preferences Differ? *Med Decis Making*. 2021: 272989X21999607.

30. Saigal S, Stoskopf LB, Feeny D, et al. Differences in preferences for neonatal outcomes among health care professionals, parents, and adolescents. *JAMA*. 1999; 281: 1991-7.

31. Parkin D, Rice N, Devlin N. Statistical analysis of EQ-5D profiles: does the use of value sets bias inference? *Med Decis Making*. 2010; 30: 556-65.

32. EuroQol Group. <https://euroqol.org/eq-5d-instruments/eq-5d-3l-about/valuation/choosing-a-value-set/>. Last assessed 15th September 2021.

33. Shaw JW, Johnson JA, Coon SJ. US valuation of the EQ-5D health states: development and testing of the D1 valuation model. *Med Care*. 2005; 43: 203-20.

34. Pickard AS, Law EH, Jiang R, et al. United States Valuation of EQ-5D-5L Health States Using an International Protocol. *Value in Health*. 2019; 22: 931-41.

35. Bowling A. Just one question: If one question works, why ask several? *J Epidemiol Community Health*. 2005; 59: 342-45.

36. Institute MR. Linguistic valiation of the PedsQL™ - a Quality of Life Questionnaire. 2002.

37. Trust MR, Varni JW. Scaling and Scoring of the Pediatric Quality of Life Inventory™ PedsQL™. In: Trust MR, ed., 2017.
38. Varni JW, Burwinkle TM, Seid M, et al. The PedsQLy 4.0 as a Pediatric Population Health Measure: Feasibility, Reliability, and Validity. *Ambulatory Pediatrics*. 2003; 3: 329-41.
39. IBM Corp. IBM SPSS statistics for mac. Version 26.0. In: Corp I, ed. Armonk, NY, 2018.
40. (FDA) UDoHaHSFaDA. Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims. In: FDA U, ed., Federal Register. Rockville, MD: FDA, 2009.
41. Reeve BB, Wyrwich KW, Wu AW, et al. ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. *Qual Life Res*. 2013; 22: 1889-905.
42. Mokkink LB, Prinsen CA, Bouter LM, et al. The CONsensus-based Standards for the selection of health Measurement INSTRUMENTs (COSMIN) and how to select an outcome measurement instrument. *Braz J Phys Ther*. 2016; 20: 105-13.
43. Trust MO. Assessing health status and quality-of-life instruments: attributes and review criteria. *Qual Life Res*. 2002; 11: 193-205.
44. Cohen J. Statistical power analysis for the behavioural sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.
45. Smith SC, Lamping DL, Banerjee S, et al. Measurement of health-related quality of life for people with dementia: development of a new instrument (DEMQOL) and an evaluation of current methodology. *Health Technol Assess*, 2005.
46. Petrou S, Morrell J, Spiiby H. Assessing the empirical validity of alternative multi-attribute utility measures in the maternity context. *Health Qual Life Outcomes*. 2009; 7: 40.
47. Scalone L, Tomasetto C, Matteucci M, et al. Assessing quality of life in children and adolescents: development and validation of the Italian version of the EQ-5D-Y. *Italian Journal of Public Health*. 2011; 8: 331-41.
48. Buchholz I, Janssen MF, Kohlmann T, et al. A Systematic Review of Studies Comparing the Measurement Properties of the Three-Level and Five-Level Versions of the EQ-5D. *Pharmacoeconomics*. 2018; 36: 645-61.
49. Pan CW, Zhong H, Li J, et al. Measuring health-related quality of life in elementary and secondary school students using the Chinese version of the EQ-5D-Y in rural China. *BMC Public Health*. 2020; 20: 982.
50. Feng Y, Devlin N, Herdman M. Assessing the health of the general population in England: how do the three- and five-level versions of EQ-5D compare? *Health Qual Life Outcomes*. 2015; 13: 171.
51. Weng L-J. Impact of the Number of Response Categories and Anchor Labels on Coefficient Alpha and Test-Retest Reliability. *Educational and Psychological Measurement*. 2016; 64: 956-72.
52. Conner-Spady BL, Marshall DA, Bohm E, 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: 1775-84.
53. Kim SH, Kim HJ, Lee SI, et al. Comparing the psychometric properties of the EQ-5D-3L and EQ-5D-5L in cancer patients in Korea. *Qual Life Res*. 2012; 21: 1065-73.
54. Janssen MF, Pickard AS, Golicki D, 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: 1717-27.

55. Ferreira LN, Ferreira PL, Ribeiro FP, et al. Comparing the performance of the EQ-5D-3L and the EQ-5D-5L in young Portuguese adults. *Health Qual Life Outcomes*. 2016; 14: 89.
56. Rencz F, Lakatos PL, Gulacsi L, et al. Validity of the EQ-5D-5L and EQ-5D-3L in patients with Crohn's disease. *Qual Life Res*. 2019; 28: 141-52.
57. Wang P, Luo N, Tai ES, et al. The EQ-5D-5L is More Discriminative Than the EQ-5D-3L in Patients with Diabetes in Singapore. *Value Health Reg Issues*. 2016; 9: 57-62.
58. Pickard AS, De Leon MC, Kohlmann T, et al. Psychometric comparison of the standard EQ-5D to a 5 level version in cancer patients. *Med Care*. 2007; 45: 259-63.
59. Pan CW, Sun HP, Wang X, et al. The EQ-5D-5L index score is more discriminative than the EQ-5D-3L index score in diabetes patients. *Qual Life Res*. 2015; 24: 1767-74.
60. Shiroyiwa T, Ikeda S, Noto S, et al. Valuation Survey of EQ-5D-Y Based on the International Common Protocol: Development of a Value Set in Japan. *Medical Decision Making*. 2021.
61. Prevolnik Rupel V, Ogorevc M. EQ-5D-Y Value Set for Slovenia. *Pharmacoeconomics*. 2021.