ScienceDirect

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

Systematic Literature Review

Head-to-Head Comparisons of the Distributional Characteristics and Measurement Properties of the 3-Level and 5-Level Versions of the EQ-5D-Y: A Systematic Review

Ling Jie Cheng, PhD, MPH, RN, Simone Schieskow, PhD, Le Ann Chen, MPH, Jing Ying Cheng, BSN (Hons), RN, Michael Herdman, MSc, Nan Luo, PhD

ABSTRACT

Objectives: This systematic review compared the distributional characteristics and measurement properties of EuroQol's EQ-5D-Y-3L and EQ-5D-Y-5L instruments, using results from published head-to-head comparative studies.

Methods: The review was reported following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses-COnsensus-based Standards for the selection of health Measurement INstruments (PRISMA-COSMIN) guideline. We searched 8 databases up to February 14, 2024 for studies on the measurement properties and distributional characteristics (missing data, ceiling/floor effects, consistency, and informativity) of both EQ-5D-Y versions. Measurement quality was assessed in terms of the proportion of tests meeting COSMIN recommended criteria for reliability, validity, and responsiveness.

Results: Eighteen studies were included. The Y-5L reduced ceiling effects by 0.8% to 16.5% and had a slightly higher evenness index, indicating greater informativity. Both versions showed sufficient construct validity across patient and general population samples, with the Y-5L slightly outperforming the Y-3L in known-group validity. Both the Y-3L and Y-5L versions showed sufficient test-retest reliability for mobility, pain/discomfort, and worried, sad, or unhappy and inconsistent reliability for looking after myself and usual activities, indicating similar performance across these dimensions. The Y-5L showed better responsiveness (standardized effect size range = 0.03-2.05) than Y-3L (range = 0.13-0.94). In proxy-reported data, the Y-5L appears to have slightly lower test-retest reliability than the Y-3L, despite demonstrating better agreement with self-reported data.

Conclusions: Both EQ-5D-Y versions demonstrated varying psychometric performances across the evaluated populations, with the Y-5L slightly outperforming the Y-3L in responsiveness and proxy-child agreement. The availability of the Y-5L expands the options of health-related quality-of-life instruments for assessing pediatric populations. Further research is needed to assess its performance across diverse populations and administrative modes.

Keywords: EQ-5D-Y, measurement properties, patient-reported outcomes, systematic review.

VALUE HEALTH. 2025; 28(10):1574–1586

Highlights

- This review provides insights into the EQ-5D-Y-3L and Y-5L relative performance, aiding tool selection for researchers and clinicians.
- The Y-5L reduces ceiling effects, improves responsiveness, and enhances proxy-child agreement, whereas both versions show similar discriminatory capacity and test-retest reliability.
- The Y-5L may be preferable for self-reporting, but its proxy advantage over Y-3L remains unclear and requires further research.

Introduction

Preference-weighted measures of health-related quality of life (HRQoL) are instruments used to both measure and value an individual's or population's HRQoL. They generally consist of 2 components: a descriptive system that defines the aspects of HRQoL to be measured, and a preference-based scoring algorithm that assigns a (usually societal) value to each health state generated by the descriptive system.[1] Preference-weighted measures of HRQoL can be generic (ie, not specifically designed for any one population) or condition specific.[2] The EQ-5D-Y is a generic preference-weighted measure specifically developed for use in children and adolescents.[3–6] It provides a quick, reliable measure of HRQoL and is widely used in clinical trials, health technology assessments, and population health surveys.[5–9]

Since its original publication in 2010,[5] the EQ-5D-Y has been validated in various populations and has demonstrated generally satisfactory measurement properties.[9] To enhance its measurement properties, a new version of EQ-5D-Y, the EQ-5D-Y-5L, has been developed, offering 5 levels of response in each dimension compared with the 3 levels in the original EQ-5D-Y (now the EQ-5D-Y-3L). Although 3 recent reviews by Kwon et al,[6] Golicki et al,[9] and Tan et al[8] have examined the measurement properties of the EQ-5D-Y, their focus was primarily on the EQ-5D-Y-3L, without including direct comparisons between the 2 versions. Additionally, their searches were limited to studies published before

January 2022,[6,8,9] omitting more recent studies examining the psychometric properties of the Y-5L.

Notably, recent head-to-head studies have provided direct comparisons of the 2 versions under identical conditions, minimizing confounding factors, and providing robust evidence of their relative performance.[10-14] Such comparisons are essential for helping users decide which version of the EQ-5D-Y is best suited to their needs. However, no systematic review has yet synthesized the findings from these studies.

Therefore, this review aims to systematically compare the distributional characteristics and measurement properties of the Y-3L and Y-5L using results from published head-to-head studies.

Methods

The study protocol was registered in the International Prospective Register of Systematic Reviews (PROSPERO) database (CRD42024503995). The review was conducted in alignment with the latest COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) guideline[15-17] and reported following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA-COSMIN) guideline.[18]

Information Sources and Search Strategy

Preliminary scoping searches were conducted in the COSMIN database, PubMed Clinical Queries, and PROSPERO to avoid duplicating existing systematic reviews. A comprehensive 3-step search strategy was then developed based on the COSMIN methodology guide[19] and the Cochrane Handbook for systematic review,[20] with the assistance of a university librarian. The search began with an extensive investigation across 8 databases: MEDLINE (PubMed), EMBASE (Elsevier), CINAHL (EBSCO), CENTRAL (Ovid), PsycINFO (Ovid), ProQuest Theses and Dissertations (ProQuest), Scopus (Elsevier), and Web of Science (Clarivate), covering studies from inception to February 14, 2024. Multiple combinations of search terms were used (see Appendix Table 1 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2025.03.020). The reference lists of eligible studies were screened for additional relevant studies. A manual search of gray literature and target journals was also conducted (see Appendix Table 2 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2025.03.020).[21] Direct communication with the authors of selected studies was initiated to obtain supplemental information and clarifications, if necessary. References were managed and exported using EndNote version X20.0.[22]

EQ-5D-Y Instruments

The EQ-5D-Y is a child-friendly adaptation of the widely used EQ-5D, designed to assess HRQoL in younger populations across 5 dimensions: mobility (MO; walking about), looking after myself (LAM; self-care), doing usual activities (UA; such as school, hobbies, sports, or spending time with family or friends), having pain or discomfort (PD), and feeling worried, sad, or unhappy (WSU; anxiety/depression). Each dimension in the Y-3L has 3 levels of severity (no problems, some problems, and a lot of problems), whereas the Y-5L offers 5 levels (no problems, a little bit problems, some problems, a lot of problems, and cannot do/extreme problems). This results in a 5-digit health profile, with 243 possible health states for the Y-3L version and 3125 for the Y-5L version.[4,23] Additionally, the instrument includes a visual analog scale, in which respondents rate their overall health from 0 to 100, with 0 representing the worst imaginable health and 100 the best imaginable health. The EQ-5D-Y is designed for self-completion by children and adolescents aged 8 to 15 years and is available in both paper and digital format.[4] For younger children aged 4 to 7, a proxy version is available for completion by a parent or caregiver.[4]

Eligibility Criteria

Studies were included if they met the following criteria: (1) validation studies reporting on measurement properties and distributional characteristics using head-to-head comparisons between Y-3L and Y-5L, regardless of the participants' health status or medical condition, (2) study samples involving children and/or adolescents, and (3) studies written in English and published in peer-reviewed journals.

We excluded reviews, study protocols, conference proceedings, editorials, and practice guidelines (see Appendix Table 3 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2025.03.020).

Selection Process

The study selection process followed the PRISMA 2020 guideline.[24] During the screening phase, 3 reviewers (C.L.J., C.J.Y., and C.L.A.) independently screened studies based on titles and abstracts, excluding irrelevant ones according to predefined criteria. Excluded studies are listed in Appendix Table 4 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2025.03.020. The same reviewers (C.L.J., C.J.Y., and C.L.A.) independently assessed full-text eligibility, with any disagreements resolved through discussion and, if needed, by the senior author (L.N.). The degree of agreement among reviewers was measured using Landis and Koch's classification, which categorizes agreement as slight (0.20-0.40), fair (0.21-0.40), moderate (0.41-0.60), substantial (0.61-0.80), or almost perfect (0.81-1.00).[25]

Data Collection Process and Data Items

A pilot-tested extraction form, based on the COSMIN guideline template, was used.[19] This form included data on child characteristics (population type, sample size, age, and gender proportion), proxy characteristics (sample size, age, and gender proportion), design factors (study design, country, survey language, EQ-5D-Y level, proxy version, administrator, follow-up period, and summary of results), statistical methods, and detailed results on the psychometric properties of Y-3L and Y-5L or data that could assess their psychometric properties. For studies that reported data on both the adapted and original versions of the Y-3L and Y-5L, we extracted only the data related to the original versions for inclusion in this review. The definitions of psychometric properties are provided in Appendix Table 5 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2025.03.020. Two independent reviewers (C.J.Y. and C.L.A.) conducted data extraction and cross-validated the findings,[26] with any disagreements resolved through discussion with a third reviewer (C.L.J.).

Risk of Bias Assessment

Two reviewers (C.L.J. and C.L.A.) independently evaluated the design quality of individual tests using the modified COSMIN Risk of Bias checklist. Tests were rated as very good (1), adequate (2), doubtful (3), or inadequate (4). A test, used in place of "study" as per COSMIN terminology to enhance clarity and minimize confusion, was defined as an independent empirical assessment of a specific psychometric property of the Y-3L or Y-5L. A study could include multiple tests or provide data for conducting multiple tests for 1 or more psychometric properties. The ratings were determined following the "worst score counts" principle. All assessments were conducted using

COSMIN criteria, except for construct validity (both convergent and known groups) and responsiveness because some items are unsuitable for appraising preference-based measures. Modified criteria (see Appendix Table 6 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2025.03.020) were used to assess the design of construct validity and responsiveness to ensure comparability with previous EQ-5D reviews.[27] All standards for responsiveness were modified because all included tests evaluated standardized effect size (SES) or standardized response mean (SRM).

Distributional Characteristics and Psychometric Property Assessment

In this review, we separately evaluated and synthesized evidence for the self-completed and proxy versions of the Y-3L and Y-5L. Distributional characteristics were compared using descriptive statistics, with feasibility assessed by calculating missing response rates and ceiling effects calculated as the percentage of respondents reporting full health (state 11111). Shannon's evenness index (Shannon's *J*'), based on information theory, is a measure of informativity. This index assesses the informational and discriminatory power of each descriptive system, in which higher values indicate more uniformity in response distribution and better discriminatory power.[28,29] Descriptive statistics (including median, mean, interquartile range, and standard deviation) were used to summarize point estimates.

Following COSMIN guidelines, 2 reviewers (C.L.J. and C.L.A.) independently evaluated the design and results of all individual psychometric tests reported in the reviewed studies. Discrepancies between reviewers were resolved through consensus-based discussions. Figure 1 illustrates the evaluation process used to assess Y-3L/5L tests for construct validity, responsiveness, test-retest reliability, and proxy-child agreement.

In line with the latest COSMIN guidelines,[19] our review team formulated 122 a priori hypotheses for known groups and convergent and divergent validity to assess the construct validity of EQ-5D-Y. These hypotheses were generated through a thorough review of existing literature and consultation with content experts (see Appendix Table 7 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2025.03.020). A positive rating was assigned when the results of a study supported the hypothesis, and a negative rating was given when they did not. Responsiveness results were graded based on SES or SRM values, with <0.20 considered negligible (assigned a negative rating).[30,31] A consensus within the team led to assigning a positive rating to tests with SES or SRM ≥ 0.20. Tests reporting other measures received an indeterminate rating.

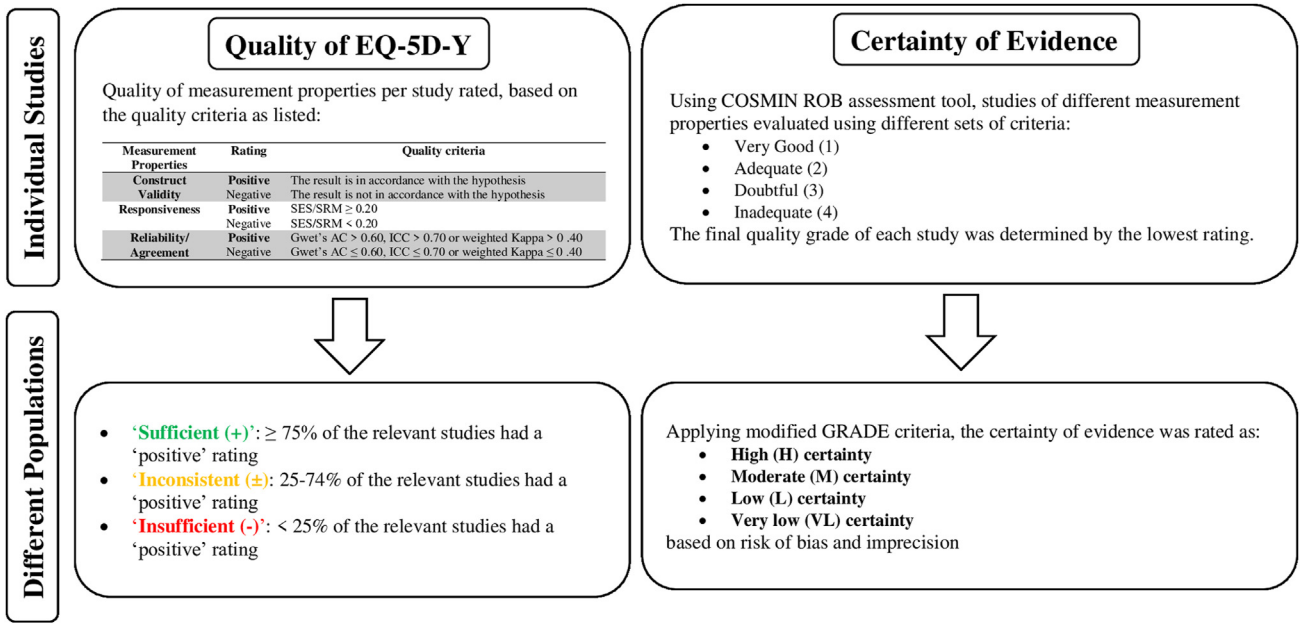
Reliability and proxy-child agreement for Y-3L or Y-5L were rated positive if the intraclass correlation coefficient was >0.70, weighted kappa was >0.40, or Gwet's AC was >0.60.[32,33] A negative rating was assigned if intraclass correlation coefficient was ≤0.70, weighted kappa was ≤0.40, or Gwet's AC was ≤0.60. Tests were rated indeterminate if only weighted kappa was reported for the EQ-5D-Y index/level sum score (LSS) or not reported at all for dimensions.[19]

Synthesis Methods

For each psychometric property of the Y-3L/Y-5L, we aggregated the ratings from all tests. The COSMIN guideline recommends classifying quality as "sufficient (+)" for ≥75% positive ratings (green), "inconsistent (±)" for 25% to 74% (amber), and "insufficient (−)" for <25% (red), with traffic light colors introduced by our team to aid interpretation.[19]

These measures were calculated separately for Y-3L and Y-5L to support comparison. We also analyzed the evidence for self-completed and proxy versions separately to better understand their validation properties.

Figure 1. Flowchart of the methodological quality and the instrument quality rating based on COSMIN. In accordance with COSMIN's taxonomy, one study is defined as a single test of a measurement property. Each included article in this review can consist of more than one study of different measurement properties.



ICC indicates intraclass correlation coefficient; SES, standardized effect size; SRM, standardized response mean.

Certainty Assessment

The overall certainty of evidence for each psychometric property was evaluated using modified Grading of Recommendations Assessment, Development and Evaluation (GRADE) criteria (see Appendix Table 8 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2025.03.020), including ratings of high (H), moderate (M), low (L), and very low (VL).[34]

Results

Study Selection

A total of 469 records were identified from the databases, with 1 additional document retrieved from other sources. After removing 228 duplicates, 3 reviewers independently screened titles and abstracts, excluding 212 records. We assessed 29 full-text reports, excluding 12 for various reasons (Figure 2, Appendix Table 6 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2025.03.020). One relevant record was found through a website search. Ultimately, 18 studies were included in the review.[10-14,35-47] The interrater agreement among the reviewers was substantial (κ = 0.76, $P < .001$).

Characteristics of Included Articles

Tables 1 and 2 summarize the characteristics of the included articles. The 18 validation studies analyzed a total of 9906 children and adolescents, mainly from South-east/East Asia (N = 9) and Africa (N = 5). The median age was 11.3 years, ranging from 2.6 to 14.0 years. Study populations included individuals with hematological and oncological conditions,[13,14,37,45] orthopedic and musculoskeletal issues,[36,39,42,46,47] or the general pediatric population.[10,12,14,35,38,40,41,43,44] Most studies used a cross-sectional design (N = 13) and consecutive sampling (N = 17).

Most studies (9 out of 14) used the Y-5L, Y-3L, and visual analog scale sequence, whereas one study randomized the questionnaire order.[37] Five studies reversed the sequence at follow-up time points.[13,42,45-47] Most studies reported self-administered questionnaires (N = 13) on-site (N = 14). The interval between baseline and follow-up varied, with 11 studies reporting intervals of less than a month[10-13,35,38,39,42,43,45,46] and others conducting follow-ups after 3 months or more.[36,47] The questionnaires were mainly in Bahasa Indonesia,[35,43] Chinese,[13,36,42,44-47] English,[10-12,14,38-41] and Spanish.[37] There were 15 studies using the self-reported version[10,11,13,14,35,37-42,44-47] and 7 studies using the proxy version.[10-13,36,43,45]

Self-Reported Version

Distributional characteristics

Table 3 shows the missing data, ceiling effects, floor effect, and informativity of both self- and proxy-reported data. Five studies[35,37,39,40,44] reported missing values at the profile level, whereas 3[,37,38] reported them at the dimension level for both Y-3L (dimension range: 0%-10.4%; profile range: 0%-2.0%) and Y-5L (dimension range: 0%-10.4%; profile range: 0%-2.0%).

Almost all studies (N = 12) reported information on the number or proportion reporting no problems in any dimension or for the profile (11111).[14,35-37,39-43,45-47] Ceiling effects for full health state profiles ranged from 6.0% to 64.0% for Y-3L and from 5.0% to 62.0% for Y-5L. Using Y-5L could reduce ceiling effects by up to 14.0 percentage points (MO) to 34.2 percentage points (UA). The greatest reduction in ceiling effects was observed for UA (−4.0 to 34.2 percentage points), followed by PD (−1.6 to 28.8 percentage points) and WSU (0.0 to 16.4 percentage points). In most studies, reduced ceiling effects were observed across all dimensions for Y-5L compared with Y-3L.

Figure 2. PRISMA 2020 flow diagram for systematic review.

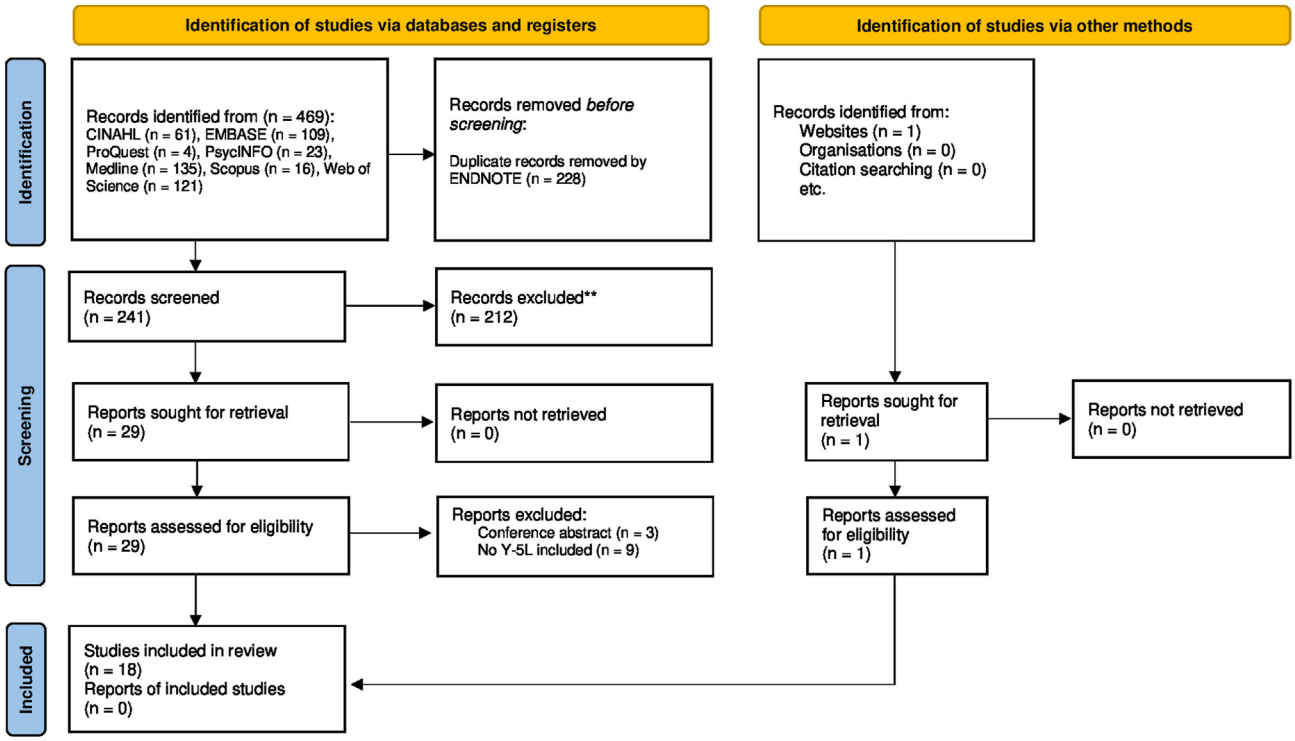


Table 1. Characteristics of the 18 included studies.

Author, year	Nature of population	Country/ region	Age range	Mean age (SD)	% female	Sample size	Proxy mean age (SD)	% female (proxy)	Sample size (proxy)
Fitriana et al.[35]	Children with major beta-thalassemia, severe-to-moderate hemophilia, AcLL (Hematological)	Indonesia/ South-east Asia	Between 8 and 16 years old	11.2 (2.4)	38.1%	286 (Base) 44 (Retest) 219 (FU)	-	-	-
Fitriana et al[43]	Children with major beta-thalassemia, severe-to-moderate hemophilia, AcLL (Hematological)	Indonesia/ South-east Asia	Between 8 and 16 years old	11.2 (2.6)	38.1%	286	40.6 (8.0)	-	286
Jones et al[10]	Children and adolescents (General)	Australia/ Oceania	Between 5 and 18 years old	10.9 (3.9)	46.0%	5945	-	-	-
Lin et al[36]	JIS or AIS (Orthopedic)	Hong Kong/ East Asia	Between 10 and 18 years old	13.8 (1.4) (Base) 11.8 (1.5) (FU)	90.8% (Base) 88.2% (FU)	130 (Base) 68 (FU)	45.0 (8.0)	71.10%	128 (Base) 68 (FU)
Ngwira et al[14]	Healthy and sick children and adolescents (General)	Malawi/Africa	Between 8 and 17 years old	13.6 (NR)	56.0%	289	-	-	-
O'Loughlin et al[11]	Children and Adolescents with ADHD, Anxiety and/or Depression (Mental health)	Australia/ Oceania	Between 4 and 18 years old	11.4 (4.1) (Base) 11.7 (3.8) (FU)	41.5% (Base) 45.0% (FU)	1013 (Base) 284 (FU)	39.6 (8.9) (Base) 41.5 (8.5) (FU)	82.5 (Base) 79.9 (FU)	1013
Perez-Sousa et al[37]	Children with cancer (Oncological)	Spain/Europe	Between 6 and 18 years old	9.7 (2.3)	52.1%	73	-	-	-
van Heusden et al[12]	Children (General)	Australia/ Oceania	Between 2 and 4 years old	2.6 (0.6)	46.0%	842	-	-	842
Verstraete et al, 2022[38]	Acute or chronic health condition and from the general population (General)	South Africa/ Africa	Between 6 and 15 years old	11.3 (1.6)	50.0%	550 (Base) 173 (FU)	-	-	-
Verstraete et al, 2022[39]	Children/adolescents receiving acute orthopedic management (Orthopedic)	South Africa/ Africa	Between 8 and 15 years old	11.5 (1.9)	53.0%	83 (Base) 78 (24 hours) 71 (48 hours)	-	-	-
Verstraete et al, 2022[41]	Children and adolescents (General)	South Africa/ Africa	Between 8 and 15 years old	11.3 (1.6)	50.0%	550	-	-	-
Verstraete et al, 2022[40]	Children and adolescents with cerebral palsy (Neurological)	South Africa/ Africa	Between 8 and 15 years old	11.8 [IQR 9.9, 13.35]	49.0%	51	-	-	-
Wang et al[44]	Children and adolescents attending schools (General)	China/East Asia	Between 8 and 18 years old	12.7 (3.1)	58.4%	262			
Wong et al, 2019[46]	Pediatric patients with idiopathic scoliosis (Orthopedic)	Hong Kong/ East Asia	Between 8 and 17 years old	14.0 (1.9) (Base) 14.3 (1.9) (Retest)	80.6% (Base) 75.7% (Retest)	129 (Base) 70 (Retest)	-	-	-
Wong et al, 2019[47]	Pediatric patients with idiopathic scoliosis (Orthopedic)	Hong Kong/ East Asia	Between 8 and 17 years old	14.0 (1.9) (Base) 14.6 (1.7) (FU)	80.6% (Base) 80.2% (FU)	129 (Base) 111 (FU)	-	-	-
Xu et al[42]	Children and adolescents with osteogenesis imperfecta (Orthopedic)	China/East Asia	Between 8 and 18 years old	11.3 (5.0)	65.00%	157	-	-	-

continued on next page

Table 1. Continued

Author, year	Nature of population	Country/ region	Age range	Mean age (SD)	% female	Sample size	Proxy mean age (SD)	% female (proxy)	Sample size (proxy)
Zhou et al[45]	Pediatric patients with hematological malignancies (Hematological)	China/East Asia	Between 8 and 17 years old	10.5 (2.2) (Base) 10.7 (2.2) (FU)	35.4% (Base) 39.3% (FU)	96 (Base) 84 (FU)	40.1 (9.3) (Base) 39.0 (8.0) (FU)	67.7% (Base) 61.0% (FU)	96 (Base) 84 (FU)
Zhou et al[13]	Children with hematological malignancies and their caregivers (Hematological)	China/East Asia	Between 8 and 17 years old	10.5 (2.2)	35.4%	96	40.1 (9.3)	67.7%	96

AcLL indicates acute lymphoblastic leukemia; AIS, adolescent idiopathic scoliosis; Base, baseline; FU, follow-up; JIS, juvenile; NR, not reported; SD, standard deviation; %, percentage.

Floor effects by dimension were reported in 2 studies (Y-3L: 4.0%-58.0%; Y-5L: 0.0%-60.0%).[39,40] The absolute reduction in floor effects ranged from 0.0 percentage points (LAM) to 6.0 percentage points (UA, PD, and WSU). For the profile, floor effects ranged from 0% to 2.0% for Y-3L and 0% to 1.0% for Y-5L (N = 3).[14,39,40]

Six studies reported on Shannon's H' and Shannon's J'. In general, Shannon's H' was always higher for Y-5L than Y-3L, but Shannon's J' was higher for Y-3L than Y-5L. Across all dimensions, Shannon's H' ranged from 0.22 to 1.39 for Y-5L and from 0.18 to 1.05 for Y-3L, indicating that Y-5L descriptive classification system is more informative than Y-3L.

Measurement Properties

Table 4 compares the measurement properties of the Y-3L and Y-5L at baseline and follow-up, covering construct validity, test-retest reliability, and responsiveness for self-reported data. "Sufficient" quality is indicated by green cells, "inconsistent" quality by amber cells, and "insufficient" quality by red cells.

Construct validity

Eight studies involving 8289 self-reported cases[10,11,13,14,38-40,44] assessed construct validity for Y-3L and Y-5L, showing high certainty of evidence supporting sufficient validity. The Y-5L (100%) performed slightly better in known-group validity than the Y-3L (90%).

Test-retest reliability

Eight studies, including 599 individuals for the index score and 663 for dimensional levels, assessed test-retest reliability.[10,11,35,38,42,45-47] Both Y-3L and Y-5L showed inconsistent to sufficient reliability across most dimensions (N = 6) and index/level sum scores (N = 4), with inconsistent reliability observed for LAM and UA dimensions. The certainty of evidence was consistently downgraded because of inconsistent test conditions during the initial and retest periods. Specifically, there were changes in the mode of administration (shifting from face-to-face in the initial test to telephone for the retest) and changes in the order of the Y-3L and Y-5L versions.

Responsiveness

Five studies with 589 individuals assessed responsiveness, showing inconsistent responsiveness for both versions, though with low certainty of evidence.[10,11,13,40,47] Y-5L demonstrated a larger effect size (SES: 0.03-2.05) than Y-3L (SES: 0.13-0.94). One additional study[35] reported that Y-5L was more responsive but could not be aggregated because it only reported the proportion of patients showing improvement. For proxy versions, 4 studies, including 318 individuals, also showed inconsistent responsiveness with low certainty.[10-13]

Proxy-Reported Version

Distributional characteristics

For the proxy version, only one study reported data,[43] and it was comparable for both versions (Y-3L dimension range: 0%-1.4%; Y-5L dimension range: 0%-1.0%). As for the ceiling effects, the full health state profiles ranged from 17.7% to 62.8% for Y-3L and from 15.5% to 62.0% for Y-5L.[10-12,36,43,45] Similar to the self-reported version, most studies showed that reduced ceiling effects were observed across all dimensions for Y-5L compared with Y-3L.

Measurement Properties

Construct validity

Four studies using 2079 proxy-reports[11-13,36] assessed construct validity for Y-3L and Y-5L, showing high certainty of evidence supporting sufficient validity. The Y-5L (100%) performed equally well for both convergent/divergent and known-group validity than the Y-3L (100%).

Test-retest reliability

For the proxy versions, 946 proxies from 6 studies (n = 453 for the index score and n = 493 for the dimensions) were included.[10-12,36,43,45] Both versions showed similar performance, with overall inconsistent test-retest reliability supported by moderate certainty. However, the dimension-level results differed from the self-reported version, with sufficient test-retest reliability observed for the LAM and UA dimensions. Notably, the Y-5L version demonstrated insufficient test-retest reliability for the index or level sum score level, whereas the Y-3L version showed inconsistent quality.

Responsiveness

Four studies including 318 individuals also showed inconsistent responsiveness with low certainty.[10-13] Y-5L demonstrated a larger effect size (SES: 0.16-1.00) than Y-3L (SES: 0.02-0.89).

Proxy-child agreement

Two studies involving 382 proxy-child dyads[43,45] reported varied agreement for both Y-3L and Y-5L at baseline and follow-up. At baseline, MO and UA showed sufficient agreement, whereas PD showed inconsistent agreement. Y-3L displayed inconsistent agreement for LAM and WSU, whereas Y-5L showed sufficient agreement. Follow-up assessments showed improved agreement across most dimensions, except for PD in Y-3L, which remained inconsistent.

Table 2. EQ-5D-Y and study design characteristics of the included studies (*N* = 18).

Author, year	Study design/ sampling	Proxy version/ sequence	Administrator/ mode	Interval period (for test-retest reliability or responsiveness)	Survey language
Fitriana et al[35]	Cross-sectional/ Consecutive	Version 1/Y-5L-Y-3L-VAS	Self/On-site	Baseline, retest, follow-up (after treatment)	Bahasa Indonesia
Fitriana et al[43]	Cross-sectional/ Consecutive	Version 1/Y-5L-Y-3L-VAS	Self/On-site	Baseline, retest, follow-up (after treatment)	Bahasa Indonesia
Jones et al[10]	Cross-sectional/ Consecutive	Version 1/Randomized	Self/Online	2 days	English
Lin et al[36]	Cross-sectional/ Consecutive	Version 1/Y-3L-Y-5L-VAS	Self/On-site	3 months	Chinese
Ngwira et al[14]	Cross-sectional/ Convenience	NA/Y-5L-Y-3L-VAS	Self/On-site	-	English
O'Loughlin et al[11]	Cross-sectional/ Consecutive	Version 1/Randomized	Self/Online	4 weeks	English
Perez-Sousa et al[37]	Cross-sectional/ Consecutive	NA/50%: Y-3L-Y-5L-VAS 50%: Y-5L-Y-3L-VAS	Interviewer/On-site	-	Spanish
van Heusden et al[12]	Cross-sectional/ Consecutive	Version 1/Randomized	Self/Online	4 weeks	English
Verstraete et al, 2022[38]	Cohort study/ Consecutive	NA/Y-5L-Y-3L-VAS	Self/On-site	7 to 14 days	English
Verstraete et al, 2022[39]	Cross-sectional/ Consecutive	NA/Y-5L-Y-3L-VAS	Self/On-site	1 to 2 days	English
Verstraete et al, 2022[41]	Cohort study/ Consecutive	NA/Y-5L-Y-3L-VAS	Self/On-site	-	English
Verstraete et al, 2022[40]	Cross-sectional/ Consecutive	NA/Y-5L-Y-3L-VAS	Self/On-site	-	English
Wang et al[44]	Cross-sectional/ Consecutive	NA/Y-5L-Y-3L-VAS	Self/On-site	-	Chinese
Wong et al, 2019[46]	Cross-sectional/ Consecutive	NA/Y-3L-Y-5L-VAS (Baseline) Y-5L-Y-3L-VAS (Retest)	Interviewer/On-site (Baseline) Telephone (Retest)	2-3 weeks	Chinese
Wong et al, 2019[47]	Cross-sectional/ Consecutive	NA/Y-3L-Y-5L-VAS (Baseline) Y-5L-Y-3L-VAS (Retest)	Interviewer/On-site (Baseline) Electronic platform (Retest)	3 months	Chinese
Xu et al[42]	Cross-sectional/ Consecutive	NA/Y-3L-Y-5L-VAS (Baseline) Y-5L-Y-3L-VAS (Retest)	Self/Online	2 weeks	Chinese
Zhou et al[45]	Cross-sectional/ Consecutive	Version 1/Proxy: Y-5L-Y-3L-VAS Self: Y-5L-Y-3L-VAS (Baseline) Y-3L-Y-5L-VAS (Follow-up)	Interviewer/On-site	2-13 days	Chinese
Zhou et al[13]	Cross-sectional/ Consecutive	Version 1/Proxy: Y-5L-Y-3L-VAS Self: Y-5L-Y-3L-VAS (Baseline) Y-3L-Y-5L-VAS (Follow-up)	Interviewer/On-site	2-13 days	Chinese

Discussion

To our knowledge, this is the first systematic review of head-to-head studies comparing the 2 available versions of EQ-5D-Y, Y-3L and Y-5L. The Y-5L, recently introduced by the EuroQol Group, offers children and adolescents more options for assessing their HRQoL. It aims to reduce ceiling effects, improve discriminatory capacity and responsiveness, and ensure better comparability with the adult version, the EQ-5D-5L.[48]

Distributional Characteristics

One of the key findings of this review is the reduction in ceiling effects observed when using the Y-5L compared with the Y-3L. The

Table 3. Feasibility, ceiling effects, floor effect, and informativity of EQ-5D-Y.

Distributional characteristics	MO [No.]	LAM [No.]	UA [No.]	PD [No.]	WSU [No.]	Overall [No.]
Self-reported Version						
Feasibility						
Range of missing data rate for the Y-3L (%)	1.0-6.9 [3]	0.0-10.4 [3]	0.0-8.3 [3]	0.0-7.6 [3]	0.0-10.4 [3]	<2.0 [5]
Range of missing data rate for the Y-5L (%)	0.0-8.3 [3]	0.0-9.0 [3]	0.0-9.0 [3]	0.0-10.4 [3]	0.8-9.7 [3]	<2.0 [5]
Proportion of participants reporting "somewhat" or "very" easy for Y-3L (%)	-	-	-	-	-	69.5 [1]
Proportion of participants reporting "somewhat" or "very" easy for Y-5L (%)	-	-	-	-	-	72.1 [1]
Ceiling (11111)						
Range of ceiling effects for the Y-3L (%)	17.0-93.5 [12]	41.0-97.7 [12]	14.0-94.6 [12]	39.2-80.4 [12]	35.0-81.5 [12]	6.0-64.0 [13]
Follow-up	73.8-95.7 [2]	74.1-100.0 [2]	66.1-92.4 [2]	62.9-84.8 [2]	68.2-87.0 [2]	53.5 [1]
Range of ceiling effects for the Y-5L (%)	11.0-95.4 [12]	43.0-97.8 [12]	14.0-92.3 [12]	36.0-81.3 [12]	28.0-77.5 [12]	5.0-62.0 [13]
Follow-up	74.5-96.7 [2]	74.8-100.0 [2]	66.8-95.6 [2]	61.0-78.0 [2]	62.9-84.8 [2]	50.7 [1]
Range of absolute reduction in ceiling effects (% points)	0.8-14.0 [5]	0.0-16.5 [5]	-4.0-34.2 [5]	-1.6-28.8 [5]	0.0-16.4 [5]	0.8-16.5 [6]
Range of relative reduction in ceiling effects (% points)	0.8-44.0 [5]	0.0-23.6 [5]	-10.0-42.4 [5]	0.0-36.9 [5]	0.0-63.0 [5]	1.2-44.6 [6]
Number of studies reporting on ceiling effects	12	12	12	12	12	13
Number of studies reporting lower ceiling effects for the Y-5L than for the Y-3L	11	12	10	11	11	12
Floor (33333/55555)						
Range of floor effects for the Y-3L (%)	22.0-58.0 [2]	14.0-13.0 [2]	12.0-58.0 [2]	4.0-10.0 [2]	4.0-8.0 [2]	0.0-2.0 [3]
Range of floor effects for the Y-5L (%)	18.0-60.0 [2]	13.0-18.0 [2]	6.0-58.0 [2]	0.0-4.0 [2]	0.0-2.0 [2]	0.0-1.0 [3]
Range of absolute reduction in floor effects (% points)	-2.0-4.0 [2]	-4.0-0.0 [2]	0.0-6.0 [2]	4.0-6.0 [2]	4.0-6.0 [2]	0.0-1.0 [2]
Range of relative reduction in floor effects (% points)	-4.0-18.0 [2]	-29.0-0.0 [2]	0.0-50.0 [2]	63.0-100.0 [2]	71.0-100.0 [2]	0.0-50.0 [2]
Number of studies reporting on floor effects	2	2	2	2	2	3
Number of studies reporting lower floor effects for the Y-5L than for the Y-3L	1	1	2	2	2	2
Informativity*						
Range of Shannon index (*H'*) for the Y-3L	0.43-1.05 [6]	0.18-1.00 [6]	0.30-0.98 [6]	0.65-0.94 [6]	0.67-0.89 [6]	-
Range of Shannon evenness index (*J'*) for the Y-3L	0.16-0.41 [6]	0.11-0.36 [6]	0.17-0.36 [6]	0.20-0.54 [6]	0.08-0.53 [6]	-
Range of Shannon index (*H'*) for the Y-5L	0.36-1.39 [6]	0.22-1.28 [6]	0.43-1.34 [6]	0.75-1.32 [6]	0.75-1.14 [6]	-
Range of Shannon evenness index (*J'*) for the Y-5L	0.14-0.31 [6]	0.10-0.29 [6]	0.16-0.31 [6]	0.14-0.37 [6]	0.10-0.40 [6]	-

continued on next page

Table 3. Continued

Distributional characteristics	MO [No.]	LAM [No.]	UA [No.]	PD [No.]	WSU [No.]	Overall [No.]
Proxy-reported version						
Feasibility						
Range of missing data rate for the Y-3L proxy (%)	0.0 [1]	0.0 [1]	0.0 [1]	1.4 [1]	0.0 [1]	
Range of missing data rate for the Y-5L proxy (%)	1.0 [1]	1.0 [1]	1.0 [1]	1.0 [1]	1.0 [1]	
Proportion of participants reporting "somewhat" or "very" easy for Y-3L proxy (%)	-	-	-	-	-	71.0-84.0 [2]
Proportion of participants reporting "somewhat" or "very" easy for Y-5L proxy (%)	-	-	-	-	-	73.4-80.0 [2]
Ceiling ('11111')						
Range of ceiling effects for the Y-3L Proxy (%)	71.0-85.3 [3]	58.9-97.7 [3]	53.5-94.6 [3]	33.0-77.5 [3]	51.0-79.8 [3]	17.7-62.8 [6]
Follow-up	93.2 [1]	92.7 [1]	80.5 [1]	75.0 [1]	87.3 [1]	63.2 [1]
Range of ceiling effects for the Y-5L Proxy (%)	65.0-84.5 [3]	52.0-96.9 [3]	51.1-92.3 [3]	29.2-79.1 [3]	49.0-77.5 [3]	15.5-62.0 [6]
Follow-up	93.6 [1]	92.3 [1]	79.5 [1]	72.7 [1]	90.5 [1]	61.8 [1]

LAM indicates Looking after myself; MO, Mobility; No., Number of tests; PD, Having pain/Discomfort; UA, Doing usual activities; WSU, Feeling worried, sad, or unhappy.
*The study conducted by Perez-Sousa et al[37] was excluded because of inaccuracies in the reporting of *H'* and *J'*.

Y-5L, with its 5 response levels, allows for more granular reporting, particularly in higher-functioning respondents, thereby reducing the proportion of respondents reporting no problems in any health dimension. This aligns with prior reviews in both younger populations[9] and adult populations,[49] in which the increased response options in the Y-5L were shown to reduce ceiling effects and were able to capture finer distinctions in patient health states. Consistent with findings from the adult EQ-5D version,[28] the distributional characteristics observed in the Y-5L also make it a more informative tool, particularly in detecting subtle differences in health states, as evidenced by the higher Shannon's *H'* values. Additionally, both versions show similar missing response rates, as supported by previous reviews,[7,9] demonstrating their feasibility in various patient and general population samples. However, floor effects remained largely similar between the 2 versions, indicating that although Y-5L improves discrimination at the higher end of the scale, it offers limited advantages at the lower end.

Self-Reported Measurement Properties

In self-reported data, the Y-5L also exhibited superior measurement properties in several domains. In terms of construct validity, the Y-5L slightly outperformed the Y-3L, particularly in known-group validity, in which it showed better discriminatory capacity in distinguishing between clinically relevant groups. This finding is consistent with those reported for the adult EQ-5D-5L,[28,50] suggesting that the 5-level version offers better accuracy and sensitivity in measuring health states. The Y-5L's superior ability to differentiate health states may be attributed to its detailed response options, which better capture the range of patient health experiences, enabling a more nuanced assessment of health outcomes.[23,51] The Y-5L also demonstrated marginally better responsiveness, with larger effect sizes observed across various studies.[7] This improvement in responsiveness likely stems from its finer gradation of health states, which allows for more precise detection of changes over time. However, the overall certainty of evidence for responsiveness remains low, because of the methodological variations between studies, such as differences in administration modes and questionnaire order, thus limiting the generalizability of these results. Future studies should focus on standardizing these conditions to strengthen the evidence base before strong recommendations can be made, especially for clinical trials and economic analyses.

In terms of test-retest reliability, both the Y-3L and Y-5L exhibited varied results across dimensions but demonstrated consistent performance in most of them. The Y-5L generally showed sufficient test-retest reliability for the MO, PD, and WSU dimensions in both self-reported versions, but inconsistencies were noted in the LAM and UA dimensions.[9] This may be because these dimensions are more familiar and clear to respondents, making them easier to recall and evaluate consistently.[52] These variations highlight the need for further research to refine the Y-5L's dimensional structure and ensure that it captures the health states of younger individuals more accurately.

Proxy-Reported Measurement Properties

The proxy-reported data highlighted similar patterns, with sufficient construct validity and inconsistent responsiveness

Table 4. Construct validity, test-retest reliability, and responsiveness of EQ-5D-Y-3L and Y-5L quality at baseline and follow-up.

EQ-5D-Y	Y-3L					Y-5L				
	N (Articles)	'+' rating (%)	No.	Range of coefficient/ SES	Quality (COE)	*N* (Articles)	'+' rating (%)	No.	Range of coefficient/ SES	Quality (COE)
Self-reported Version										
Construct validity										
Hypothesis testing	33 (8)	30 (90.9)	8289	-	+ (H)	33 (8)	30 (90.9)	8289	-	+ (H)
Convergent/ Divergent	23 (5)	21 (91.3)	2248	-	+ (H)	23 (5)	21 (91.3)	2248	-	+ (H)
Known groups	10 (6)	9 (90.0)	8155	-	+ (H)	10 (6)	10 (100.0)	8155	-	+ (H)
Test-retest reliability										
Overall	34 (8)	24 (70.6)	1068	-	± (L ‡)	34 (8)	25 (73.5)	1068	-	± (L ‡)
Index score or LSS	4 (4)	3 (75.0)	599	GAC = 0.780-0.820; ICC = 0.61-0.84	+ (M †)	4 (4)	3 (75.0)	599	GAC = 0.790-0.840; ICC = 0.56-0.94	+ (M †)
MO	7 (6)	6 (85.7)	663	GAC = 0.420-0.920; κ = 0.47-0.55	+ (L ‡)	7 (6)	6 (85.7)	663	GAC = 0.500-0.916; κ = 0.54-0.67	+ (L ‡)
LAM	7 (6)	5 (71.4)	663	GAC = 0.480-0.970; κ = 0.38-0.43	± (L ‡)	7 (6)	5 (71.4)	663	GAC = 0.570-0.978; κ = 0.32-0.73	± (L ‡)
UA	7 (6)	5 (71.4)	663	GAC = 0.310-0.937; κ = 0.21-0.42	± (L ‡)	7 (6)	4 (57.1)	663	GAC = 0.390-0.92; κ = 0.29-0.45	± (L ‡)
PD	7 (6)	6 (85.7)	663	GAC = 0.560-0.853; κ = 0.42-0.59	+ (L ‡)	7 (6)	6 (85.7)	663	GAC = 0.562-0.808; κ = 0.43-0.54	+ (L ‡)
WSU	7 (6)	6 (85.7)	663	GAC = 0.370-0.768; κ = 0.41-0.48	+ (L ‡)	7 (6)	6 (85.7)	663	GAC = 0.540-0.824; κ = 0.40-0.53	+ (L ‡)
Responsiveness										
Index score/LSS*	27 (5)	19 (70.3)	589	Range = 0.03-2.05	± (L ‡)	27 (5)	21 (77.8)	589	Range = 0.13-0.94	+ (L ‡)
	11 (5)			Improved = 0.03-0.67		11 (5)			Improved = 0.13-0.94	
	5 (2)			Unchanged = 0.11-0.42		5 (2)			Unchanged = 0.15-0.53	
	11 (5)			Deteriorated = 0.06-2.05		11 (5)			Deteriorated = 0.16-0.76	
Proxy-reported Version										
Construct validity										
Hypothesis testing	28 (4)	28 (100.0)	2079	-	+ (H)	28 (4)	28 (100.0)	2079	-	+ (H)
Convergent/ Divergent	7 (1)	7 (100.0)	1013		+ (H)	1 (7)	7 (100.0)	1013		+ (H)
Known groups	21 (4)	21 (100.0)	2079		+ (H)	21 (4)	21 (100.0)	2079		+ (H)
Test-retest reliability										
Overall	23 (6)	14 (60.9)	946		± (M †)	23 (6)	12 (52.2)	946	-	± (M †)
Index score or LSS	3 (3)	1 (33.3)	453	ICC = 0.49-0.79	± (M †)	3 (3)	0 (0.0)	453	ICC = 0.44-0.54	− (M †)
MO	4 (4)	2 (50.0)	493	GAC = 0.590-0.910; κ = −0.06	± (M †)	4 (4)	2 (50.0)	493	GAC = 0.332-0.940; κ = 0.66	± (M †)
LAM	4 (4)	3 (75.0)	493	GAC = 0.746-0.920; κ = 0.34	+ (M †)	4 (4)	3 (75.0)	493	GAC = 0.624-0.940; κ = 0.29	+ (M †)
UA	4 (4)	4 (100.0)	493	GAC = 0.753-0.860; κ = 0.71	+ (M †)	4 (4)	3 (75.0)	493	GAC = 0.688-0.840; κ = 0.23	+ (M †)
PD	4 (4)	2 (50.0)	493	GAC = 0.267-0.710; κ = 0.17	± (M †)	4 (4)	2 (50.0)	493	GAC = 0.343-0.770; κ = 0.38	± (M †)
WSU	4 (4)	2 (50.0)	493	GAC = 0.421-0.910; κ = 0.14	± (M †)	4 (4)	2 (50.0)	493	GAC = 0.428-0.840; κ = 0.21	± (M †)
Responsiveness										
Index score/LSS*	15 (4)	7 (46.7)	318	Range = 0.02-0.89	± (L †)	15 (4)	11 (73.3)	318	Range = 0.16-1.00	± (L †)
	7 (4)			Improved = 0.02-0.34		7 (4)			Improved = 0.16-1.00	
	1 (1)			Unchanged = 0.22		1 (1)			Unchanged = 0.19	
	7 (4)			Deteriorated = 0.15-0.89		7 (4)			Deteriorated = 0.31-0.96	
Proxy-child agreement										
Baseline										
Overall	25 (2)	17 (68.0)	382	-	± (H)	25 (2)	20 (80.0)	382	-	+ (H)
MO	5 (2)	5 (100.0)	382	GAC = 0.630-0.860	+ (H)	5 (2)	5 (100.0)	382	GAC = 0.701-0.910	+ (H)
LAM	5 (2)	3 (60.0)	382	GAC = 0.587-0.870	± (H)	5 (2)	4 (80.0)	382	GAC = 0.290-0.950	+ (H)
UA	5 (2)	4 (80.0)	382	GAC = 0.030-0.720	+ (H)	5 (2)	4 (80.0)	382	GAC = 0.070-0.810	+ (H)
PD	5 (2)	3 (60.0)	382	GAC = 0.550-0.720	± (H)	5 (2)	3 (60.0)	382	GAC = 0.530-0.810	± (H)
WSU	5 (2)	2 (40.0)	382	GAC = 0.470-0.800	± (H)	5 (2)	4 (80.0)	382	GAC = 0.561-0.770	+ (H)
Follow-up										
Overall	25 (2)	23 (92.0)	382	-	+ (H)	25 (2)	22 (88.0)	382	-	+ (H)
MO	5 (2)	5 (100.0)	382	GAC = 0.675-0.960	+ (H)	5 (2)	4 (80.0)	382	GAC = 0.582-0.970	+ (H)
LAM	5 (2)	5 (100.0)	382	GAC = 0.769-0.980	+ (H)	5 (2)	5 (100.0)	382	GAC = 0.629-0.980	+ (H)

continued on next page

Table 4. Continued

EQ-5D-Y	Y-3L					Y-5L				
	N (Articles)	'+' rating (%)	No.	Range of coefficient/ SES	Quality (COE)	*N* (Articles)	'+' rating (%)	No.	Range of coefficient/ SES	Quality (COE)
UA	5 (2)	5 (100.0)	382	GAC = 0.644-0.930	+ (H)	5 (2)	4 (80.0)	382	GAC = 0.503-0.970	+ (H)
PD	5 (2)	3 (60.0)	382	GAC = 0.460-0.900	± (H)	5 (2)	5 (100.0)	382	GAC = 0.603-0.930	+ (H)
WSU	5 (2)	5 (100.0)	382	GAC = 0.680-0.900	+ (H)	5 (2)	4 (80.0)	382	GAC = 0.565-0.960	+ (H)

Note. Certainty of evidence: H indicates "high"; M indicates "moderate"; L indicates "low"; VL indicates "very low." Quality of EQ-5D, + indicates "sufficient" results; ± indicates "inconsistent" results; − indicates "insufficient" results.
COE indicates certainty of evidence; LAM, looking after myself; LSS, level sum score; MO, mobility; *N*, number of tests; No., number of samples; PD, having pain/ discomfort; UA, doing usual activities; WSU, feeling worried, sad, or unhappy.
*The studies conducted by Fitriana et al[35] and Fitriana et al[43] were excluded because they reported the synthesis of responsiveness as a proportion of improvement instead of SRM/SES. However, the review team included their results in a narrative format.
†Quality downgraded by 1 level due to ROB.
‡Quality downgraded by 2 level due to ROB.

observed for both Y-3L and Y-5L versions. However, the Y-5L demonstrated slightly lower test-retest reliability compared with the Y-3L, a difference predominantly driven by 3 studies conducted in Australia[11,12,53] out of the 6 included. Theoretically, the reliability of the Y-5L should be comparable to that of the Y-3L under stable health conditions. Several plausible factors may explain this discrepancy. First, the Australian studies included younger age groups (eg, 2-4 years[12] and 4-18 years[10,11]), in which proxy responses are inherently more variable because of the challenges faced in consistently assessing very young children; yet, the Y-5L's higher sensitivity to subtle shifts—such as from "no problems" to "slight problems" in mobility—may amplify this variability, particularly for nonclinical changes during the test-retest period. Although this sensitivity allows the Y-5L to capture nuanced health changes that the Y-3L might miss, it may also increase inconsistency in proxy ratings. Furthermore, it is worth noting that the EQ-5D-Y is not intended for use with children younger than 4 years.[3] Second, two-thirds of the Australian studies used retest intervals of 4 weeks,[11,12] whereas the remaining studies varied between 2 days and 2 weeks, potentially violating the assumption of stable health conditions over time.[54] These findings underscore the importance of tailoring study designs to the target population and adhering to best practices in test-retest reliability assessments for proxy-reported outcomes.

Interestingly, the proxy version showed inconsistent reliability for PD and WSU and sufficient reliability for LAM and UA—results opposite to those observed in the self-reported data. This difference may occur because aspects such as PD and WSU may not be as easily observed by proxies.[45,46] The Y-5L performed less effectively than the Y-3L, possibly because of proxies interpreting and responding to the questionnaire based on their perceptions, especially with the higher number of response options in the Y-5L.[55] This highlights the importance of the careful choice between EQ-5D-Y versions, particularly in proxy settings. Future studies should explore the use of EQ-5D-Y versions in different contexts and improve proxy reporting through better training or clearer guidelines.

As expected, proxy-child agreement ranged from "inconsistent" to "sufficient" in both baseline and follow-up assessments, with the Y-5L demonstrating better performance, supported by high-quality evidence. This finding aligns with previous reviews,[6,7] and the improved agreement in the Y-5L may result from its detailed response options, which help caregivers more accurately interpret and report children's health states.[23] These findings suggest that in settings in which caregiver input is crucial, such as decisions involving incapacitated patients or children, the Y-5L may offer advantages. Future studies could focus on enhancing caregiver understanding and agreement through targeted educational programs, further improving the utility and accuracy of patient-reported outcomes in clinical settings.

This review has several strengths, including the use of rigorous methodology, collaboration with a senior librarian, transparent synthesis, and independent reviewers, all of which increase its robustness. However, several limitations should be noted. First, variations in study designs, language versions, and survey methodologies pose challenges to generalizing results and could introduce confounding factors in psychometric testing outcomes. Second, although the review focuses on English-language studies, many of the included studies involved populations from non-English-speaking countries, such as Indonesia, China, and South Africa. This diversity mitigates potential language bias and enhances the generalizability of our findings. Third, most studies compared the Y-3L and Y-5L using level-sum scores because the Y-5L currently lacks a value set. Although level-sum scores allow direct comparison between the 2 measures, they do not incorporate societal utility preferences and therefore may not perform identically to their preference-based counterparts. Lastly, the inability to assess age-specific performance is a limitation. The psychometric performance of the instruments may differ across different age groups, which may be important for users when selecting between them in practice. Future research should focus on comparing the 2 instruments in specific age groups, such as 7 to 8 years and 9 to 10 years.

Conclusions

The introduction of the Y-5L broadens the range of HRQoL instruments available for pediatric assessments. Our review suggests that both EQ-5D-Y versions demonstrated comparable psychometric performance across the evaluated populations, with the Y-5L showing slight advantages in reducing ceiling effects, improving responsiveness, and enhancing proxy-child agreement. These small, nonsignificant differences suggest that the Y-5L may offer a more nuanced assessment for self-reported health. However, careful consideration is warranted when selecting between the Y-5L and Y-3L for proxy-reported outcomes, particularly in contexts involving very young children or diverse settings. Further research is needed to evaluate the performance of both versions across a broader range of populations and administrative modes. Future studies should also investigate the underlying reasons for these observed differences, with a focus on standardizing test conditions and enhancing proxy training to improve the consistency and reliability of assessments across varied contexts.

Author Disclosures

Author disclosure forms can be accessed below in the Supplemental Material section.

The views of the authors expressed in this article do not necessarily reflect the views of the EuroQol Group. Drs Cheng, Luo, Schieskow, and Mr Herdman are members of the EuroQol Group. Dr Luo is an editor for *Value in Health* and had no role in the peer-review process of this article.

Supplemental Material

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.jval.2025.03.020.

Article and Author Information

Accepted for Publication: March 12, 2025

Published Online: May 12, 2025

doi: https://doi.org/10.1016/j.jval.2025.03.020

Author Affiliations: National Perinatal Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, Oxford, England, UK (Cheng); Alice Lee Centre for Nursing Studies, Yong Loo Lin School of Medicine, National University of Singapore, Singapore (Cheng); Saw Swee Hock School of Public Health, National University of Singapore, Singapore (Cheng, Chen, Herdman, Luo); Department of Health Economics and Health Care Management, School of Public Health, Bielefeld University, Bielefeld, Germany (Schieskow); Khoo Teck Puat Hospital, Yishun Health, National Healthcare Group, Singapore (Cheng).

Correspondence: Nan Luo, PhD, Saw Swee Hock School of Public Health, National University of Singapore, 12 Science Drive 2, #10-01, 117549, Singapore. Email: ephln@nus.edu.sg

Authorship Confirmation: All authors certify that they meet the ICMJE criteria for authorship.

Funding/Support: The authors received no financial support for this research.

Acknowledgment: The authors of this article thank the authors for sending further information on their studies for the purpose of our systematic review.

Disclaimer: Parts of the contents of this article were presented at the 30th ISOQOL Annual Conference, Calgary, Alberta, Canada and 40th EuroQol Plenary Meeting in Rome, Italy.

REFERENCES

1. Rowen D. Preference-based measures of health-related quality of life. In: Michalos AC, ed. *Encyclopedia of Quality of Life and Well-Being Research*. Dordrecht. The Netherlands: Springer Netherlands; 2014:5026–5028.
2. Rowen D. Preference-weighted health states. In: Maggino F, ed. *Encyclopedia of Quality of Life and Well-Being Research*. Cham, The Netherlands: Springer International Publishing; 2020:1–2.
3. EuroQol Research Foundation. EQ-5D-Y-5L user guide: how to apply, score, and present results from the EQ-5D-Y-5L. version 1.0. https://euroqol.org/wp-content/uploads/2024/09/EuroQol-User-Guide-EQ5D-Y5L-v1.0-0924-11.pdf. Accessed October 1, 2024.
4. EuroQol Research Foundation. EQ-5D-Y-3L user guide: how to apply, score, and present results from the EQ-5D-Y-3L (formerly EQ-5D-Y). version 2.1. https://euroqol.org/wp-content/uploads/2024/06/Userguide-EQ5D-Y3L-0424-07.pdf. Accessed October 1, 2024.
5. 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(1):31–37.
6. Kwon J, Smith S, Raghunandan R, et al. Systematic review of the psychometric performance of generic childhood multi-attribute utility instruments. *Appl Health Econ Health Policy*. 2023;21(4):559–584.
7. Rowen D, Keetharuth AD, Poku E, Wong R, Pennington B, Wailoo A. A review of the psychometric performance of selected child and adolescent preference-based measures used to produce utilities for child and adolescent health. *Value Health*. 2021;24(3):443–460.
8. Tan RL-Y, Soh SZY, Chen LA, Herdman M, Luo N. Psychometric properties of generic preference-weighted measures for children and adolescents: a systematic review. *Pharmacoeconomics*. 2023;41(2):155–174.
9. Golicki D, Młyńczak K. Measurement properties of the EQ-5D-Y: a systematic review. *Value Health*. 2022;25(11):1910–1921.
10. Jones R, O'Loughlin R, Xiong X, et al. Comparative psychometric performance of common generic paediatric health-related quality of life instrument descriptive systems: results from the Australian paediatric multi-instrument comparison study. *Pharmacoeconomics*. 2024;42(Suppl 1):39–55.
11. O'Loughlin R, Jones R, Chen G, et al. Comparing the Psychometric Performance of Generic Paediatric Health-Related Quality of Life Instruments in Children and Adolescents with ADHD, Anxiety and/or Depression. *Pharmacoeconomics*. 2024;42(Suppl 1):57–77.
12. van Heusden A, Rivero-Arias O, Herdman M, et al. Psychometric performance comparison of the adapted versus original versions of the EQ-5D-Y-3L and -Y-5L in proxy respondents for 2- to 4-year-olds. *Pharmacoeconomics*. 2024;42:129–145.
13. Zhou W, Shen A, Yang Z, et al. Validity and responsiveness of EQ-5D-Y in children with haematological malignancies and their caregivers. *Eur J Health Econ*. 2024;25(8):1361–1370.
14. Ngwira LG, Maheswaran H, Verstraete J, Petrou S, Niessen L, Smith SC. Psychometric performance of the Chichewa versions of the EQ-5D-Y-3L and EQ-5D-Y-5L among healthy and sick children and adolescents in Malawi. *J Patient Rep Outcomes*. 2023;7(1):22.
15. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol*. 2010;63(7):737–745.
16. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Qual Life Res*. 2010;19(4):539–549.
17. Mokkink LB, de Vet HCW, Prinsen CAC, et al. COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. *Qual Life Res*. 2018;27(5):1171–1179.
18. Elsman EBM, Butcher NJ, Mokkink LB, et al. Study protocol for developing, piloting and disseminating the PRISMA-COSMIN guideline: a new reporting guideline for systematic reviews of outcome measurement instruments. *Syst Rev*. 2022;11(1):121.
19. Mokkink LB, Prinsen C, Patrick DL, et al. COSMIN methodology for systematic reviews of patient-reported outcome measures (PROMs). *User manual*. 2018;27(5):1147–1157.
20. Higgins JP, Green S. *Cochrane Handbook for Systematic Reviews of Interventions*. Chichester, UK: John Wiley & Sons; 2011.
21. Paez A. Gray literature: an important resource in systematic reviews. *J Evid Based Med*. 2017;10(3):233–240.
22. The EndNote Team. *EndNote*. 20th ed. Philadelphia, PA: Clarivate; 2013.
23. Kreimeier S, Åström M, Burström K, et al. EQ-5D-Y-5L: developing a revised EQ-5D-Y with increased response categories. *Qual Life Res*. 2019;28(7):1951–1961.
24. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71.
25. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*. 2012;22(3):276–282.
26. Tacconelli E. Systematic reviews: CRD's guidance for undertaking reviews in health care. *Lancet Infect Dis*. 2010;10(4):226.
27. Aspden T, Bradshaw SA, Playford ED, Riazi A. Quality-of-life measures for use within care homes: a systematic review of their measurement properties. *Age Ageing*. 2014;43(5):596–603.
28. Buchholz I, Janssen MF, Kohlmann T, Feng YS. A systematic review of studies comparing the measurement properties of the three-level and five-level versions of the EQ-5D. *Pharmacoeconomics*. 2018;36(6):645–661.
29. Shannon CE. Communication theory of secrecy systems. *Bell Syst Tech J*. 1949;28(4):656–715.
30. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. Abingdon, UK: Routledge; 2013.
31. Fayers PM, Machin D. *Quality of Life: the Assessment, Analysis and Interpretation of Patient-Reported Outcomes*. Chichester, UK: John Wiley & Sons; 2013.
32. Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Med Res Methodol*. 2013;13:61.
33. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15(2):155–163.
34. Guyatt GH, Oxman AD, Schunemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. *J Clin Epidemiol*. 2011;64(4):380–382.

35. Fitriana TS, Purba FD, Rahmatika R, et al. Comparing measurement properties of EQ-5D-Y-3L and EQ-5D-Y-5L in paediatric patients. *Health Qual Life Outcomes*. 2021;19(1):256.
36. Lin J, Wong CKH, Cheung JPY, Cheung PWH, Luo N. Psychometric performance of proxy-reported EQ-5D youth version 5-level (EQ-5D-Y-5L) in comparison with three-level (EQ-5D-Y-3L) in children and adolescents with scoliosis. *Eur J Health Econ*. 2022;23(8):1383–1395.
37. Perez-Sousa MA, Olivares PR, Gusi N. Psychometric properties of the Spanish versions of EQ-5D-Y-3L and EQ-5D-Y-5L in children with cancer: a comparative study. *Int J Environ Res Public Health*. 2022;19(18):11420.
38. Verstraete J, Amien R, Scott D. Comparing measurement properties of the English EQ-5D-Y 3-level version with the 5-level version in South Africa. *Value Health Reg Issues*. 2022;30:140–147.
39. Verstraete J, Marthinus Z, Dix-Peek S, Scott D. Measurement properties and responsiveness of the EQ-5D-Y-5L compared to the EQ-5D-Y-3L in children and adolescents receiving acute orthopaedic care. *Health Qual Life Outcomes*. 2022;20(1):28.
40. Verstraete J, Scott D. The performance of the EQ-5D-Y-5L compared to the EQ-5D-Y-3L in children and adolescents with cerebral palsy (CP). *Dialogues Health*. 2022;1:100032.
41. Verstraete J, Scott D. Comparison of the EQ-5D-Y-5L, EQ-5D-Y-3L and PedsQL in children and adolescents. *J Patient Rep Outcomes*. 2022;6(1):67.
42. Xu RH, Zhu L, Sun R, et al. Investigating the psychometric properties of the EQ-5D-Y-3L, EQ-5D-Y-5L, CHU-9D, and PedsQL in children and adolescents with osteogenesis imperfecta. *Eur J Pediatr*. 2022;181(12):4049–4058.
43. Fitriana TS, Purba FD, Stolk E, Busschbach JJV. EQ-5D-Y-3L and EQ-5D-Y-5L proxy report: psychometric performance and agreement with self-report. *Health Qual Life Outcomes*. 2022;20(1):88.
44. Wang P, Yue S, Zhi-Hao Y, Ruo-Yu Z, Bin W, Nan L. Testing measurement properties of two EQ-5D youth versions and KIDSCREEN-10 in China. *Eur J Health Econ*. 2021;22(7):1083–1093.
45. 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;22(7):1103–1113.
46. Wong CKH, Cheung PWH, Luo N, Cheung JPY. 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;20(5):647–656.
47. Wong CKH, Cheung PWH, Luo N, Lin J, Cheung JPY. Responsiveness of 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(21):1507–1514.
48. EuroQoL Research Foundation. EQ-5D-5L user guide: basic information on how to use the EQ-5D-5L instrument. version 3.0. https://euroqol.org/wp-content/uploads/2023/11/EQ-5D-5LUserguide-23-07.pdf; Published September 2019. Accessed October 1, 2024.
49. Cheng LJ, Pan T, Chen LA, et al. The ceiling effects of EQ-5D-3L and 5L in general population health surveys: a systematic review and meta-analysis. *Value Health*. 2024;27(7):986–997.
50. 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(6):675–697.
51. Herdman M, Gudex C, Lloyd A, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res*. 2011;20(10):1727–1736.
52. Rencz F, Janssen MF. Analyzing the pain/discomfort and anxiety/depression composite domains and the meaning of discomfort in the EQ-5D: A mixed-methods study. *Value Health*. 2022;25(12):2003–2016.
53. Bahrampour M, Devlin N, Jones R, Dalziel K, Mulhern B. On behalf of the QUOKKA (Quality of Life in Kids: Key Evidence for Decision Makers in Australia) Team. A comparison of the psychometric properties of the EQ-5D-Y-3L and EQ-5D-Y-5L using paediatric multi-instrument comparison (P-MIC) study data. *Pharmacoeconomics*. 2024;42(suppl 1):95–111.
54. Wyse AE. How days between tests impacts alternate forms reliability in computerized adaptive tests. *Educ Psychol Meas*. 2021;81(4):644–667.
55. Roydhouse JK, Cohen ML, Eshoj HR, et al. The use of proxies and proxy-reported measures: a report of the international society for quality of life research (ISOQOL) proxy task force. *Qual Life Res*. 2022;31(2):317–327.