PharmacoEconomics (2024) 42 (Suppl 1):S39–S55
https://doi.org/10.1007/s40273-023-01330-2

ORIGINAL RESEARCH ARTICLE

Comparative Psychometric Performance of Common Generic Paediatric Health-Related Quality of Life Instrument Descriptive Systems: Results from the Australian Paediatric Multi-Instrument Comparison Study

Renee Jones[1,2] · Rachel O'Loughlin[1,2,4] · Xiuqin Xiong[1] · Mina Bahrampour[3] · Nancy Devlin[1] · Harriet Hiscock[2,4,5] · Gang Chen[6] · Brendan Mulhern[3] · Kim Dalziel[1,2] on behalf of the Quality of Life in Kids: Key Evidence to Strengthen Decisions in Australia (QUOKKA) Project Team

Accepted: 11 October 2023 / Published online: 13 November 2023
© The Author(s) 2023

Abstract

Objective The aim of this study was to compare the psychometric performance of common generic paediatric health-related quality-of-life instrument descriptive systems (PedsQL generic core 4.0, EQ-5D-Y-3L, EQ-5D-Y-5L, Child Health Utility 9D [CHU9D], Assessment of Quality of Life 6D [AQoL-6D], and Health Utilities Index Mark 3 [HUI3]) by child age, report type, and health status.

Methods Data for children aged 5–18 years were from the Australian Paediatric Multi-Instrument Comparison study. Ceiling effects, test–retest reliability, known-group validity, convergent and divergent validity, and responsiveness were assessed in the total sample and by child age (5–12 years vs 13–18 years), report type (self- vs proxy report), and health status. Instruments were scored using an exploratory level sum score (LSS) approach.

Results Survey data were available for 5945 children, with follow-up data available for 2346 children. The EQ-5D-Y-3L demonstrated ceiling effects. The PedsQL, EQ-5D-Y-3L, EQ-5D-Y-5L, and CHU9D demonstrated acceptable test–retest reliability. All instruments demonstrated known-group, convergent, and divergent validity. The EQ-5D-Y-3L, EQ-5D-Y-5L, and CHU9D demonstrated responsiveness to improvements in health and the PedsQL, EQ-5D-Y-3L, EQ-5D-Y-5L, and CHU9D to worsening health. The AQoL-6D and HUI3 had inconclusive test–retest reliability and responsiveness evidence due to small sample size. Importantly, ceiling effects, test–retest reliability and responsiveness varied by subgroup.

Conclusion Results reflect instrument performance using LSSs, which may differ to utility scores. In the total sample, the EQ-5D-Y-5L and CHU9D descriptive systems demonstrated evidence of good performance (i.e., meeting prespecified criteria) across all psychometric attributes tested. Performance varied by child age and report type, indicating room for considerations by population and study.

1 Introduction

Paediatric health-related quality of life (HRQoL) instruments aim to aid in understanding how a child's health impacts different aspects of their life, such as physical, emotional and social wellbeing [1, 2]. Generic paediatric HRQoL instruments focus on aspects of health that are relevant to most children and can be used in any child population, enabling comparisons across conditions and settings [3]. Generic HRQoL instruments can be either accompanied by preference weights, a weighting algorithm used to score the HRQoL instrument into a single value anchored from 1 (full health) to 0 (equivalent to dead) derived from preferences of the general public, or not. Where accompanied by preference weights, generic HRQoL instruments can be used to generate quality-adjusted life-year (QALY) estimates, which are then used in economic evaluation [4]. This evidence is formally used in many countries to inform healthcare resource allocation [5, 6]. HRQoL instruments with or without accompanying preference weights can be used in

Brendan Mulhern, Kim Dalziel are Joint last authors.

The members of the Quality of Life in Kids: Key Evidence to Strengthen Decisions in Australia (QUOKKA) project team are mentioned in the Acknowledgments section.

Extended author information available on the last page of the article

Key Points for Decision Makers

There is currently a lack of evidence on which instruments should be used to measure health-related quality of life (HRQoL) in children and this study produces substantial evidence to support decision makers in using HRQoL evidence from the instruments.

This study compared the performance of the PedsQL, EQ-5D-Y-3L, EQ-5D-Y-5L, CHU9D, AQoL-6D, and HUI3 descriptive systems in a large diverse sample of Australian children using an exploratory level sum score approach, finding the EQ-5D-Y-5L and CHU9D performed well on *all* psychometric attributes evaluated in the total sample. Furthermore, the PedsQL, EQ-5D-Y-3L, AqoL-6D, and HUI3 performed well on *most* psychometric attributes evaluated in the total sample, with all instruments demonstrating evidence of known-group, convergent, and divergent validity.

Considering results for ceiling effects, test–retest reliability, and responsiveness varied by child age and report type, future instrument users should consider the decision-making context, target population, and study design when deciding which instrument will be most appropriate.

population studies, clinical registries, or routine clinical care as patient-reported outcome measures (PROMs) [7, 8]. Evidence from these uses of HRQoL instruments informs public funding decisions, access to treatments, and clinical care outcomes. Therefore, it is crucial to have good evidence on the measurement accuracy of generic HRQoL instruments.

An instrument's measurement accuracy can be informed by assessment of its psychometric properties. Although there is some evidence available regarding the individual psychometric performance of generic paediatric HRQoL instruments [9], there is a paucity of evidence on the comparative psychometric performance of these instruments (i.e., how these instruments perform compared with one another in the same sample) in terms of validity, reliability, and responsiveness [10, 11]. There are challenges measuring HRQoL in children when compared with adults, which may be why such evidence is still lacking [12]. Unlike adults, children have a broad developmental range, precluding the use of a single approach for measuring HRQoL in all children [13]. There have been five previous reviews of generic paediatric HRQoL instruments [9–11, 14, 15]. A 2015 systematic review examined HRQoL instruments accompanied by preference weights that were previously used in paediatric populations: the Adolescent Health Utility Measure, Assessment of Quality of Life (AQoL-6D), Child Health Utility (CHU9D), the EuroQol Group's EQ-5D-Y-3L, Health Utilities Index Mark 2 (HUI2) and Mark 3 (HUI3), 16D, 17D, and the Quality of Wellbeing Scale (QWB) [11]. The review concluded more evidence on the comparative validity of these instruments is required [11]. Another review was conducted in 2015 to assess the psychometric performance of a wider range of paediatric HRQoL instruments [15], including 35 generic paediatric HRQoL instruments with and without accompanying preference weights identified in a previous review [16]. The 2015 review highlighted the lack of evidence regarding the test–retest reliability and responsiveness of instruments, with no included studies having evaluated instrument responsiveness [15]. A 2021 systematic review examined the psychometric performance of key generic paediatric HRQoL instruments: the CHU9D, EQ-5D-Y-3L, HUI2, and HUI3 [10]. The review identified that there were key limitations in the current evidence precluding clear conclusions on the comparative performance of instruments [10]. Key limitations included the small sample size of some studies, and differences in study design, statistical methods and samples (age and condition) across studies, which limits accurate assessment of comparative performance [10]. A 2023 systematic review updated and expanded the 2021 review by including a wider set of generic instruments and studies [9]. The 2023 review explored the level of evidence available for each psychometric attribute, highlighting test–retest reliability, among others, as a psychometric attribute with the least evidence and where further research is required [9]. Except for the Janssens et al. 2015 review [15], a limitation of these reviews is their focus on HRQoL instruments that have accompanying preference weights, which misses the comparative performance of other generic paediatric HRQoL instruments that are commonly used in clinical care, population studies, and clinical registries, such as the Pediatric Quality of Life Inventory (PedsQL), which is commonly used in Australia.

The Australian Paediatric Multi-Instrument Comparison Study (P-MIC) was undertaken to address the limitations and gaps in the current evidence base for paediatric HRQoL instruments, with a focus on improving decision making in the Australian context [17]. The P-MIC study is the first of its kind to compare common paediatric HRQoL instruments head-to-head across a wide range of child ages and health statuses [17]. Results published in a separate manuscript outline the quality of the P-MIC study data and the acceptability and feasibility of the HRQoL instruments included, to children and their caregivers [18]. Using survey data from the P-MIC study, this paper aims to compare the psychometric performance of the PedsQL, EQ-5D-Y-3L, EQ-5D-Y-5L, CHU9D, AQoL-6D, and HUI3 descriptive systems by child age, report type (self- vs proxy report), and child health status.

2 Methods

P-MIC study data from children aged 5–18 years (inclusive) were used [17, 19]. P-MIC participants (children and their caregivers) were recruited between June 2021 and August 2022 into three samples: Sample (1) children with or without health conditions recruited via a large tertiary paediatric hospital based in Victoria, Australia; Sample (2) general population children recruited via an online panel available nationally (Pureprofile Australia); and Sample 3) children from nine condition-specific groups (attention-deficit/hyperactivity disorder (ADHD), anxiety and/or depression, autism spectrum disorder (ASD), asthma, eating disorder, epilepsy, recurrent abdominal pain, sleep problems, and tooth problems) recruited via the same online panel as above or—for rarer conditions—via patient organisations associated with the condition. P-MIC study data were from Data Cut 2, dated 10 August 2022. Data cut 2 includes approximately 94% of the total planned P-MIC participants.

2.1 Data Collection

All participants consented and completed an initial survey online via REDCap. Participants were then asked to complete a second online follow-up survey at 4 weeks. A small subset of participants from the online panel general population sample (Sample 2) were asked to complete the follow-up survey at 2 days to enable assessment of test–retest reliability.

All instruments were self-completed by the participant (i.e., no instruments were interviewer administered). Instruments were either proxy reported by the caregiver or self-reported by the child. Children aged 7 years or older who were deemed by their caregiver as currently able to complete questions about their health completed the HRQoL instruments themselves (child self-report), otherwise these were completed by the caregiver (proxy report). Where an instrument was proxy reported, the proxy was asked to rate the child's health from their perspective (i.e., from the caregiver's perspective).

For further information on P-MIC study methodology, including details of participant recruitment (i.e., quotas), survey structure, instruments, survey questions, and statistical analysis plans, please see the technical methods paper [19].

2.2 Instruments

The PedsQL core generic version 4.0, EQ-5D-Y-3L and EQ-5D-Y-5L, CHU9D, AQol-6D adolescent, and HUI3 were included in both the initial and follow-up surveys. As per the prespecified protocol [17], the PedsQL EQ-5D-Y-3L and EQ-5D-Y-5L and CHU9D were included in the core set of instruments received by all participants, and the HUI3 and AQoL-6D were included as additional instruments that only some participants were randomised to receive. Although the study team wanted to include all instruments for all participants, feedback was received from the consumer group during the design phase of the study expressing concern about responder burden. Hence, efforts were made by the study team to reduce responder burden where possible. The HUI3 and AQoL-6D were not included in the sample recruited via hospital (Sample 1) to minimise responder burden (following patient feedback), and in the online panel samples (Samples 2 and 3), participants were randomised to receive either the HUI3 or the AQoL-6D or another generic instrument not included in this analysis. A summary of participants who received each instrument is available in Table 1, and characteristics of instruments included in analysis are available in Supplementary Table 1 (see electronic supplementary material [ESM]). The order of the core set of instruments (PedsQL, EQ-5D-Y-3L, EQ-5D-Y-5L, and CHU9D) was randomised to minimise order effects, and there was always another instrument between the EQ-5D-Y-3L and EQ-5D-Y-5L given their similarity. The AQoL-6D or HUI3 were completed after the other generic HRQoL instruments. Where participants were allocated to an instrument, they were required to answer all instrument questions; hence, there is no missing HRQoL instrument data.

The priority for which instruments to include in the study, and which instruments to include in the core set (all participants receive) or additional set (only some participants randomised to receive) was determined by the study team following a review of key literature available at the time of study design [10, 11, 15], and consultation with experts (including clinical, health technology assessment, health economist, government, and consumer experts). The decision was guided by the following factors: (1) instruments commonly used to measure HRQoL in children (instruments were prioritised if they had evidence of strong psychometric performance from single studies), (2) instruments used to measure HRQoL in children that had been recently developed and were likely to be commonly used in future, and (3) instruments that would be useful in informing policy and healthcare decision making in Australia. It was not a requirement that instruments had preference weights available, although the study team did consider which instruments had

Table 1 Participant characteristics by child age, report type, and child health status

Participant characteristic	*N* (% of those completed initial survey) or mean (SD)						
	Total sample	Child age		Report type		Child health status special healthcare need[a]	
		5–12 years	13–18 years	Proxy	Self	No	Yes
Sample characteristics							
Completed initial survey, *n* (%)	5945 (100)	3752 (100)	2193 (100)	2083 (100)	3862 (100)	3362 (100)	2583 (100)
Completed follow-up survey, *n* (%)	2346 (39.5)	1507 (40.2)	839 (38.3)	855 (41.1)	1491 (38.6)	1256 (37.4)	1090 (42.2)
Allocated to follow-up survey at 2 days and completed follow-up survey, n (% of all follow-up surveys completed)	169 (7.2)	99 (6.6)	70 (8.3)	81 (9.5)	88 (5.9)	151 (12.0)	18 (1.7)
Days between initial and follow-up survey for those allocated to 2-day follow-up, mean (SD)	4.6 (4.0)	4.7 (3.9)	4.4 (4.2)	5.1 (4.6)	4.1 (3.5)	4.6 (4.0)	4.1 (4.5)
Allocated to follow-up survey at 4 weeks and completed follow-up survey, n (% of all follow-up surveys completed)	2177 (92.8)	1408 (93.4)	769 (91.7)	774 (90.5)	1403 (94.1)	1105 (88.0)	1072 (98.4)
Days between initial and follow-up survey for those allocated to 4-week follow-up, mean (SD)	42.2 (22.6)	41.6 (21.8)	43.1 (24.0)	41.0 (18.8)	42.8 (24.4)	42.8 (23.9)	41.5 (21.1)
Recruited via hospital (sample 1), *n* (%)	759 (12.8)	506 (13.5)	253 (11.5)	306 (14.7)	453 (11.7)	187 (5.6)	572 (22.1)
Online panel general population (sample 2), *n* (%)	1531 (25.8)	913 (24.3)	618 (28.2)	536 (25.7)	995 (25.8)	1414 (42.1)	117 (4.5)
Online panel condition groups (sample 3), *n* (%)	3655 (61.5)	2333 (62.2)	1322 (60.3)	1241 (59.6)	2414 (62.5)	1761 (52.4)	1894 (73.3)
ADHD, *n* (%)	492 (8.3)	339 (9.0)	153 (7.0)	188 (9.0)	304 (7.9)	160 (4.8)	332 (12.9)
Anxiety or depression, *n* (%)	480 (8.1)	196 (5.2)	284 (13.0)	95 (4.6)	385 (10.0)	229 (6.8)	251 (9.7)
ASD, *n* (%)	510 (8.6)	337 (9.0)	173 (7.9)	183 (8.8)	327 (8.5)	115 (3.4)	395 (15.3)
Asthma, *n* (%)	487 (8.2)	363 (9.7)	124 (5.7)	185 (8.9)	302 (7.8)	295 (8.8)	192 (7.4)
Eating disorder, *n* (%)	186 (3.1)	0	186 (8.5)	41 (2.0)	145 (3.8)	51 (1.5)	135 (5.2)
Epilepsy, *n* (%)	272 (4.6)	180 (4.5)	92 (4.2)	96 (4.6)	176 (4.6)	78 (2.3)	194 (7.5)
Recurrent abdominal pain, *n* (%)	392 (6.7)	272 (7.3)	120 (5.5)	124 (6.0)	268 (6.9)	228 (6.8)	164 (6.4)
Sleep problems, *n* (%)	346 (5.8)	269 (7.2)	77 (3.5)	156 (7.5)	190 (4.9)	180 (5.4)	166 (6.4)
Tooth problems, *n* (%)	490 (8.2)	377 (10.1)	77 (3.5)	173 (8.3)	317 (8.2)	425 (12.6)	65 (2.5)
Completed core HRQoL instruments (CHU9D, PedsQL, EQ-5D-Y-3L, EQ-5D-Y-5L), *n* (%)	5945 (100)	3752 (100)	2193 (100)	2083 (100)	3862 (100)	3362 (100)	2583 (100)
Completed AQoL-6D, *n* (%)	1523 (25.6)	962 (25.6)	561 (25.6)	582 (27.9)	941 (24.4)	961 (28.6)	562 (21.8)
Completed HUI3, *n* (%)	1728 (29.1)	1075 (28.7)	653 (29.8)	604 (29.0)	1124 (29.1)	1083 (32.2)	645 (25.0)
Child characteristics							
Child age, mean (SD)	10.9 (3.9)	8.3 (2.2)	15.3 (1.6)	9.1 (4.2)	11.9 (3.4)	10.6 (3.9)	11.2 (3.9)
Child gender, *n* (%)							
Female	2737 (46.0)	1678 (44.7)	1059 (48.3)	935 (44.9)	1802 (46.7)	1628 (48.4)	1109 (42.9)
Male	3118 (52.5)	2049 (54.6)	1069 (48.8)	1121 (53.8)	1997 (51.7)	1707 (50.8)	1411 (54.6)
Transgender female	17 (0.3)	4 (0.1)	13 (0.6)	3 (0.1)	14 (0.4)	4 (0.1)	13 (0.5)
Transgender male	30 (0.5)	7 (0.2)	23 (1.1)	10 (0.5)	20 (0.5)	10 (0.3)	20 (0.8)
Not described or prefer not to say	43 (0.7)	14 (0.4)	29 (1.3)	14 (0.7)	29 (0.8)	13 (0.4)	30 (1.2)
Child of Aboriginal and/or Torres Strait Islander origin—Yes, *n* (%)	379 (6.4)	239 (6.4)	140 (6.4)	135 (6.5)	244 (6.3)	166 (4.9)	213 (8.3)
Child speaks language other than English spoken at home—Yes, *n* (%)	513 (8.6)	336 (9.0)	177 (8.1)	173 (8.3)	340 (8.8)	361 (10.7)	152 (5.9)
Child has chronic health condition or disability (lasting at least 6 months), *n* (%)	2537 (42.7)	1476 (39.3)	1061 (48.4)	879 (42.2)	1658 (42.9)	393 (11.7)	2144 (83.0)
Caregiver characteristics							
Caregiver age, mean (SD)	40.8 (8.5)	38.0 (7.3)	45.6 (8.2)	39.3 (8.8)	41.5 (8.2)	40.5 (8.6)	41.1 (8.4)
Caregiver highest education level—bachelor's degree or above, *n* (%)	2161 (36.4)	1418 (37.8)	743 (33.9)	804 (38.6)	1357 (35.1)	1298 (38.6)	863 (33.4)

ADHD attention-deficit/hyperactivity disorder, *ASD* autism spectrum disorder, *AQoL-6D* Assessment of Quality of Life 6D, *CHU9D* Child Health Utility 9D, *EQ-5D-Y-3L* and *EQ-5D-Y-5L* refer to the EuroQol Group's Youth instruments with 3 and 5 response options, respectively, *HRQoL* health-related quality of life, *HUI3* Health Utilities Index Mark 3, *PedsQL* Paediatric Quality of Life Inventory, *SD* standard deviation

[a] Special healthcare need is defined as per the Children with Special Health Care Needs (CSHCN) screener [22]

preference weights available at the time of study design and which may be likely to have preference weights available in future. For example, although the EQ-5D-Y-5L and PedsQL did not have preference weights available at the time of study design, they were considered instruments that would quite likely have preference weights available in future. Further details on the justification for the inclusion of each instrument are available in the published study protocol [17].

2.3 Instrument Scoring Used for Analysis

The PedsQL total score was calculated by reverse scoring and linearly transforming raw item responses (0 = 100, 1 = 75, 2 = 50, 3 = 25, 4 = 0), then the sum of all item scores was divided by the number of items [20]. PedsQL domain scores were calculated using a similar approach, where raw items were linearly transformed, and the sum of all item scores in a domain was divided by the number of items in that domain [20]. An exploratory level sum score (LSS) approach was used to obtain an overall instrument total score for all other instruments. LSSs were calculated by summing together the numerical value attached to each item response (e.g., 1 for ‘no problems’ and 5 for ‘extreme problems’ in the EQ-5D-Y-5L) for all items in the instrument. The total score range possible for each instrument varies and is described in Supplementary Table 1 (see ESM). The LSS approach is considered exploratory. It has some advantages in providing an equally weighted score for comparison but other disadvantages, such as its non-normal distribution and inability to distinguish between health states that may be quite different from one another [21]. In addition, there were a lack of preference weights available for all instruments included in this study. Furthermore, the aim of this analysis is to understand the descriptive systems of each instrument.

2.4 Statistical Analysis

Analyses were completed in Stata Version 17 (StataCorp, Texas, US). Statistical tests, hypotheses, and thresholds were based on the statistical analysis protocol set *a priori* by the study team; the statistical analysis protocol is available in the technical methods paper [19]. Where appropriate, subgroup analyses were completed using the following prespecified subgroups: child age (5–12 years vs 13–18 years), report type (proxy vs self-report) and health status (children without a special healthcare need vs children with a special healthcare need) [19, 22]. The child age subgroups (5–12 years and 13–18 years) reflect key child development stages (pre-adolescence and adolescence), and this age cut point is consistent with the PedsQL instrument age versions, which is one of the most well-validated paediatric HRQoL instruments [23]. Adjusting for multiple comparisons was not required in the primary analyses as all statistical tests were hypothesis driven or included different samples. Adjusting for multiple comparisons may have been applicable for subgroup analyses; however, given this is not commonly performed in the research field of psychometric analysis, we opted for an approach where subgroup analyses were not adjusted for multiple comparisons.

2.4.1 Distribution of Responses

Distribution of responses was evaluated by descriptively assessing participant responses to each instrument item. The distributions of responses were visually inspected. Additionally, the total instrument ceiling and floor effects were assessed. As this study includes general population children and children with health conditions, ceiling effects were assessed only in children with a special healthcare need [22], as these children were expected to report health problems on HRQoL instruments. An instrument was considered to have a ceiling or floor effect if > 15% of participants with a special healthcare need reported the lowest severity (e.g., ‘no problems’) or highest severity category across all items. This 15% threshold is based on previous thresholds used in the literature [24, 25].

2.4.2 Test–Retest Reliability

Test–retest reliability was assessed by comparing instrument total scores between initial and follow-up measurements for participants who reported no change in health and were allocated to receive their first reminder for the follow-up survey at 2 days. Only participants in the online panel general population sample were allocated to receive the follow-up survey at 2 days. Test–retest reliability was assessed using intraclass correlation coefficient (ICC) estimates and corresponding 95% confident intervals. ICC estimates were calculated based on an absolute-agreement, two-way mixed-effects model [26]. As per Koo and Li (2016), an ICC of < 0.5 indicates poor reliability, 0.50–0.74 moderate reliability, 0.75–0.90 good reliability, and > 0.90 excellent reliability [26]. An ICC ≥ 0.5 (moderate reliability) was considered acceptable test–retest reliability. Primary analysis was completed using Koo and Li (2016) thresholds [26], however, it is acknowledged that other thresholds for interpreting ICC results exist. Cicchetti (1994) thresholds were applied in a sensitivity analysis. These thresholds state that an ICC of < 0.4 indicates poor agreement, 0.40–0.59 indicates fair agreement, 0.60–0.74 indicates good agreement, and ≥ 0.75 indicates excellent agreement [27].

2.4.3 Known-Group Validity

Known-group validity was assessed by comparing groups with expected differences in HRQoL, which were set *a priori* by the study team [19]. Group differences were assessed by comparing the mean instrument total score for each group, and effect sizes were estimated using Cohen's d [28]. Effect sizes of 0.2–0.49 were considered small, 0.5–0.79 moderate, and ≥ 0.8 large [28, 29]. A mean difference with a p value of < 0.05 and a large effect size (≥0.8) was considered acceptable. Children with a special healthcare need were considered a known group who were hypothesised to have differences in HRQoL compared with children without special healthcare needs [30]. Additionally, sensitivity analysis was conducted on other known groups: children with a chronic health condition, EQ VAS score ≤ 80 [31], PedsQL total score ≤ 69.7 (one standard deviation below the child self-reported population mean for children aged 5–18 years), and PedsQL total score ≤ 74.2 (child self-reported mean from a sample of children with chronic conditions) [20]. PedsQL known-group cut points were not used to assess the known-group validity of the PedsQL and were only used to assess known-group validity of other instruments.

2.4.4 Convergent and Divergent Validity

The assessment of an instrument's convergent or divergent validity usually requires a 'gold standard' to compare against, to see how much another instrument converges or diverges from this gold standard. Although there is currently no gold standard instrument for measuring quality of life in children, the PedsQL is a very commonly used instrument that has undergone extensive content validity testing [20, 23]. Hence, for the purposes of assessing the convergent and divergent validity of instruments, the PedsQL was chosen as the comparator instrument. Convergent and divergent validity were assessed by correlating each item in the EQ-5D-Y-3L, EQ-5D-Y-5L, CHU9D, AQoL-6D, and HUI3 with each item and domain in the PedsQL. Correlations were calculated using Spearman's correlation, as data were not normally distributed. Correlations of 0.1–0.29 were considered weak, 0.3–0.49 moderate, and ≥ 0.5 strong [28]. Through an *a priori* consensus approach involving members of the study team, different instrument item combinations were reviewed to assess if the study team hypothesised the item of one instrument would be at least moderately correlated with a PedsQL item (to assess for convergence) or not correlated at all with a PedsQL item (to assess for divergence) [19]. Hypotheses were based on similarity (convergence) or dissimilarity (divergence) of item wording [19]. The proportion of an instrument's items hypothesised to be at least moderately correlated with the PedsQL items and that resulted in at least a statistically moderate correlation were assessed to evaluate convergent validity. The proportion of an instrument's items hypothesised not to be correlated with the PedsQL items and that resulted in a statistically weak correlation were assessed to evaluate divergent validity.

2.4.5 Responsiveness

Responsiveness was assessed by comparing the mean difference in total instrument score between initial and follow-up surveys for children whose caregiver reported the child had a change in health between the initial and follow-up survey. Analysis focused on participants allocated to receive the follow-up survey at 4 weeks. Responsiveness was assessed by comparing the mean total score at initial and follow-up survey using a paired t-test. A mean difference in the expected direction with a p value of < 0.05 was considered acceptable and was used as the main indicator of responsiveness. Responsiveness was also assessed by calculating the standardised response mean (SRM) to provide a more detailed picture of instrument responsiveness [32]. SRM is a ratio of the mean change to the standard deviation of that change [32]. An SRM of 0.2–0.49 was considered small, 0.5–0.79 moderate, and ≥ 0.8 large [28, 29, 32]. Caregivers were asked to report their child's change in health in the follow-up survey. Change in health was calculated as follows:

1. *Change in general health* was reported as (1) much better, (2) somewhat better, (3) about the same, (4) somewhat worse, or 5) much worse. Responses were split into two categories for analysis: 'much better' and 'somewhat worse and much worse'.
2. For participants who reported a health condition in the initial survey, caregivers were asked to report their child's *change in 'main health condition'*. The same categorisation used for change in general health was applied.

Responsiveness was only assessed in those who reported a change in health as 'much better', rather than those who reported 'somewhat or much better'. It was felt that a more stringent classification would provide a clearer indication that a change in health had occurred. Due to small sample sizes in the number of children who had worsening health, this same stringent classification was not possible, and 'somewhat worse and much worse' were pooled together.

A sensitivity analysis was conducted whereby the responsiveness analysis described above was repeated in only participants recruited via hospital (Sample 1), as this sample had a higher follow-up survey response rate compared with other samples.

Table 2 Ceiling effects of each instrument, by child age and report type

Instrument	*N* (%) of children with special healthcare need[a] reporting 'no problems' or the equivalent on all instrument items within each group				
	Total sample	Child age		Report type	
		5–12 years	13–18 years	Proxy	Self
PedsQL (23 items)	11 (0.4)	10 (0.7)	1 (0.1)	1 (0.1)	10 (0.6)
EQ-5D-Y-3L (5 items)	459 (**17.8**)	287 (**18.8**)	172 (**16.2**)	161 (**17.7**)	298 (**17.8**)
EQ-5D-Y-5L (5 items)	383 (14.8)	241 (**15.8**)	142 (13.4)	141 (**15.5**)	242 (14.5)
CHU9D (9 items)	103 (4.0)	65 (4.3)	38 (3.6)	41 (4.5)	62 (3.7)
AQoL-6D (20 items)	6 (1.1)	4 (1.2)	2 (0.9)	1 (0.4)	5 (1.5)
HUI3 (8 items)	49 (7.6)	34 (9.4)	15 (5.3)	28 (12.7)	21 (5.0)

In the total sample, *N* = 2560 children with a special healthcare need or their proxies completed the PedsQL, EQ-5D-Y-3L, EQ-5D-Y-5L, and CHU9D; *N* = 562 completed the AQol-6D; and *N* = 645 completed the HUI3. *N* = 1508 children aged 5–12 years with a special healthcare need or their proxies completed the PedsQL, EQ-5D-Y-3L, EQ-5D-Y-5L, and CHU9D; *N* = 335 completed the AQol-6D; and *N* = 361 completed the HUI3. *N* = 1059 children aged 13–18 years with a special healthcare need or their proxies completed the PedsQL, EQ-5D-Y-3L, EQ-5D-Y-5L, and CHU9D; *N* = 227 completed the AQol-6D; and *N* = 284 completed the HUI3. The proxies of *N* = 911 children with a special healthcare need completed the PedsQL, EQ-5D-Y-3L, EQ-5D-Y-5L, and CHU9D; *N* = 236 completed the AQol-6D, and *N* = 221 completed the HUI3. *N* = 1672 children with a special healthcare need self-reported the PedsQL, EQ-5D-Y-3L, EQ-5D-Y-5L, and CHU9D; *N* = 326 self-reported the AQol-6D; and *N* = 424 self-reported the HUI3

[a] Special healthcare need is defined as per the Children with Special Health Care Needs (CSHCN) screener [22]

AQoL-6D Assessment of Quality of Life 6D, *CHU9D* Child Health Utility 9D, *EQ-5D-Y-3L* and *EQ-5D-Y-5L* refer to the EuroQol Group's Youth instruments with 3 and 5 response options, respectively, *HUI3* Health Utilities Index Mark 3, *PedsQL* Paediatric Quality of Life Inventory

Bold indicates responses above prespecified threshold of 15% indicating a ceiling effect

2.4.6 Summary of Psychometric Performance

The psychometric performance of all instruments was summarised by categorising each instrument as 1) having significant evidence of good performance (tick), 2) having significant evidence of poor performance (cross) or, 3) having inconclusive evidence of performance (question mark) for each psychometric attribute assessed. Significant evidence of good performance (tick) for each psychometric attribute was based on the following thresholds:

- *Response distribution (no ceiling effect)*, < 15% of participants with a special healthcare need report the lowest severity or frequency level (e.g., 'no problems') across all instrument items.
- *Test–retest reliability*, moderate, good, or excellent agreement (ICC ≥ 0.5).
- *Known-group validity,* mean difference with a value of < 0.05 and large effect size (Cohen's d effect size ≥ 0.8).
- *Convergent and divergent validity,* items at least moderately correlated (Spearman's correlation ≥ 0.3) with other instrument items where hypothesised to be correlated (convergent validity), and weakly correlated (Spearman's correlation < 0.3) where hypothesised not to be correlated (divergent validity).
- *Responsiveness,* significant mean difference (*p* value < 0.05).

An instrument was considered to have inconclusive evidence for a psychometric attribute if the sample size used to assess the psychometric attribute was too small (i.e., inadequate or doubtful according to the 2019 Consensus-based Standards for the selection of health Measurement Instruments [COSMIN] guidelines) [33], or the direction of evidence was unclear.

3 Results

3.1 Participant Characteristics

Supplementary Fig. 1 summarises the participant flow (see ESM). The sample characteristics of the 5945 children aged 5–18 years and their caregivers who completed the P-MIC initial survey are presented in Table 1.

3.2 Response Distribution

Supplementary Figs 1a-7d (see ESM) summarise the item response distribution for each instrument by child age, report type, and health status for each instrument. Table 2 summarises instrument ceiling effects. Only the EQ-5D-Y-3L had a ceiling effect in the overall sample, with more than 15% of children with a special healthcare need reporting the lowest severity level (i.e., 'no problems') across all items. No instruments had a floor effect.

Minimal variation in ceiling effects were observed by subgroup (child age and report type) for all instruments except the HUI3, which had a higher ceiling effect when proxy reported (12.7%) compared with self-reported by the child (5.0%).

3.3 Test–Retest Reliability

Mean days between initial and follow-up survey completion for participants included in the test–retest analysis was 4.6 days. Table 3 summarises results of the test–retest analyses, showing the ICC for each instrument total score (calculated using an exploratory LSS approach) by child age and report type for participants who were allocated to complete the follow-up survey at 2 days, completed the follow-up survey, and reported no change in health. In the total sample, the PedsQL, EQ-5D-Y-3L, and EQ-5D-Y-5L demonstrated 'good' agreement (ICC 0.75–0.9), with 95% confidence intervals within the 'moderate', 'good' and 'excellent' agreement range (ICC ≥ 0.5). The CHU9D demonstrated 'moderate' agreement (ICC 0.64), with a 95% confidence interval within the 'moderate' range (ICC 0.52–0.74). Samples sizes for the AQoL-6D and HUI3 may be too small to robustly assess test–retest reliability.

Minor variation in performance was observed by subgroup. Except for the EQ-5D-Y-5L when proxy reported (ICC 0.48) and the CHU9D in children aged 13–18 years (ICC 0.31), all instruments demonstrated at least 'moderate' agreement in all subgroups assessed (ICC ≥ 0.5).

Additionally, when alternate ICC thresholds from Cicchetti (1994) were applied in a sensitivity analysis, interpretation of results varied slightly, with instruments interpreted as having better agreement using the Cicchetti (1994) thresholds [27]. In the total sample, the CHU9D was considered to have 'good' agreement under the Cicchetti (1994) thresholds [27], compared with 'moderate' under the Koo and Li (2016) thresholds [26]. Again, in the total sample, the PedsQL, EQ-5D-Y-3L, and EQ-5D-Y-5L were considered to have 'excellent' agreement under the Cicchetti (1994) thresholds [27], compared with 'good' under the Koo and Li (2016) thresholds [26].

Table 3 Test–retest reliability of each instrument, by child age and report type

Instrument	Total sample			Child age						Report type					
				5–12 years			13–18 years			Proxy			Self		
	N	ICC	95% CI	*N*	ICC	95% CI	*N*	ICC	95% CI	*N*	ICC	95% CI	*N*	ICC	95% CI
PedsQL	114	0.80	0.73–0.86	63	0.82	0.72–0.89	51	0.77	0.64–0.86	55	0.74	0.59–0.84	59	0.88	0.81–0.93
EQ-5D-Y-3L	114	0.83	0.76–0.88	63	0.85	0.75–0.91	51	0.76	0.61–0.85	55	0.79	0.67–0.88	59	0.84	0.74–0.90
EQ-5D-Y-5L	113	0.78	0.69–0.84	63	0.77	0.65–0.86	50	0.80	0.67–0.88	55	0.48	0.24–0.66	58	0.94	0.89–0.96
CHU9D	114	0.64	0.52–0.74	63	0.81	0.71–0.88	51	0.31	0.05–0.54	55	0.67	0.48–0.79	59	0.63	0.45–0.76
AQoL-6D	34	0.86	0.74–0.93	19	0.86	0.62–0.95	15	0.86	0.63–0.95	19	0.85	0.65–0.94	15	0.86	0.65–0.95
HUI3	35	0.82	0.67–0.90	21	0.78	0.54–0.91	14	0.95	0.84–0.99	11	0.95	0.84–0.99	24	0.76	0.52–0.89

ICC is based on comparison of instrument total score at initial and follow-up survey for participants who completed follow-up 2 days after initial survey and reported no change in health

AQoL-6D Assessment of Quality of Life 6D, *CHU9D* Child Health Utility 9D, *CI* confidence interval, *EQ-5D-Y-3L* and *EQ-5D-Y-5L* refer to the EuroQol Group's Youth instruments with 3 and 5 response options, respectively, *HUI3* Health Utilities Index Mark 3, *ICC* intraclass correlation coefficient, *PedsQL* Paediatric Quality of Life Inventory

Green indicates 'good' or 'excellent' agreement (ICC ≥ 0.75). Yellow indicates 'moderate' agreement (ICC 0.5–0.74). White indicates 'poor' agreement (ICC < 0.5) [26]. Grey cells indicate a small sample size and hence results should be interpreted with caution

Table 4 Known-group validity of each instrument

Instrument	Child health status—special healthcare need [a]					
	N—No	Mean total score (SD)—No	*N*—Yes	Mean total score (SD)—Yes	*p*-value	Effect size (Cohen's *d*)
PedsQL	3362	75.6 (16.0)	2583	58.6 (18.9)	< 0.001	0.98
EQ-5D-Y-3L	3362	6.0 (1.3)	2583	7.4 (2.0)	< 0.001	− 0.84
EQ-5D-Y-5L	3362	6.4 (2.0)	2583	8.8 (3.5)	< 0.001	− 0.85
CHU9D	3362	13.8 (5.0)	2583	18.9 (7.1)	< 0.001	− 0.84
AQoL-6D	961	32.7 (8.7)	562	42.5 (11.4)	< 0.001	− 1.00
HUI3	1083	10.3 (3.0)	645	13.4 (4.0)	< 0.001	− 0.90

AQoL-6D Assessment of Quality of Life 6D, *CHU9D* Child Health Utility 9D, *EQ-5D-Y-3L* and *EQ-5D-Y-5L* refer to the EuroQol Group's Youth instruments with 3 and 5 response options, respectively, *HUI3* Health Utilities Index Mark 3, *PedsQL* Paediatric Quality of Life Inventory, *SD* standard deviation

[a] Special healthcare need is defined as per the Children with Special Health Care Needs (CSHCN) screener[22]

Green cells indicate large effect size (≥0.8), yellow indicates medium effect size (0.5–0.79), white indicates small effect size (0.2–0.49)

3.4 Known-Group Validity

Table 4 summarises the known-group validity for children with and without special healthcare needs using the total score of each instrument. All instrument total scores (calculated using an exploratory LSS approach) demonstrated statistically significant differences with large effects sizes, across all known groups. This finding was consistent across child age groups and report types (see Supplementary Table 2 in the ESM), as well as additional known groups tested (children with a chronic health condition, VAS score ≤ 80, PedsQL total score ≤ 69.7, and PedsQL total score ≤ 74.2) (See Supplementary Table 3 in the ESM).

3.5 Convergent and Divergent Validity

Supplementary Tables 4–8 provide the Spearman correlations for all instruments compared with the PedsQL at the item level. These tables demonstrate that most instruments, when correlated with the PedsQL at the item level, had > 50% of items moderately or strongly correlated, except for the HUI3 which had the least number of items correlated with the PedsQL. Almost all instrument items that were hypothesised to have at least a moderate correlation with a PedsQL item did, indicating convergent validity of all instruments. Additionally, almost all instrument items that were hypothesised not to be correlated with a PedsQL item were weakly correlated, indicating divergent validity of all instruments.

3.6 Responsiveness

Table 5 summarises the responsiveness of each instrument's total score (calculated using an exploratory LSS approach) for children whose caregiver reports a change in their general health or main health condition between initial and follow-up surveys. Where caregivers reported that their child's general health or main health condition was 'much better' at the follow-up survey compared with the initial survey, the EQ-5D-Y-3L, EQ-5D-Y-5L, and CHU9D total instrument scores all demonstrated a significant mean difference between initial and follow-up survey, in the direction expected, albeit with small effect sizes. Where caregivers reported that their child's general health or main health condition was 'somewhat worse or much worse' at the follow-up survey compared with the initial survey, the PedsQL, EQ-5D-Y-3L, EQ-5D-Y-5L, and CHU9D total instrument scores all demonstrated a significant mean difference between initial and follow-up survey, in the direction expected with small effect sizes. Sample sizes for the AQoL-6D and HUI3 may be too small to robustly assess responsiveness.

Table 5 Responsiveness of each instrument, by change in general health and change in main health condition

Instrument	Change in general health						Change in main health condition					
	N	Initial survey mean total score (SD)	Follow-up survey mean total score (SD)	Mean difference (SD)	*p*-value	Standardised response mean	*N*	Initial survey mean total score (SD)	Follow-up survey mean total score (SD)	Mean difference (SD)	*p*-value	Standardised response mean
Change in health—'much better'												
PedsQL	213	73.4 (19.1)	75.0 (19.6)	− 1.5 (15.5)	0.0740	− 0.10	178	73.2 (17.3)	74.3(17.8)	− 1.1 (16.9)	0.1925	− 0.07
EQ-5D-Y-3L	213	6.4 (1.7)	6.1 (1.7)	0.3 (1.6)	**0.0047**	0.18	178	6.5 (1.7)	6.1 (1.7)	0.4 (1.8)	**0.0028**	0.21
EQ-5D-Y-5L	209	7.3 (3.0)	6.7 (2.8)	0.6 (2.3)	**<0.001**	0.26	177	7.4 (2.9)	6.7 (2.5)	0.8 (2.4)	**<0.001**	0.33
CHU9D	213	15.0 (6.6)	13.1 (5.6)	1.9 (5.4)	**<0.001**	0.36	178	15.5 (6.4)	13.0 (5.0)	2.5 (6.1)	**<0.001**	0.41
AQoL-6D	39	35.3 (11.3)	35.8 (12.3)	− 0.5 (7.4)	0.6667	–	34	35.2 (10.3)	33.7 (10.9)	1.5 (8.2)	0.1428	–
HUI3	39	11.3 (3.3)	10.8 (4.7)	0.5 (3.6)	0.2029	–	37	10.7 (2.9)	10.7 (4.7)	0 (3.8)	0.5000	–
Change in health—'somewhat or much worse'												
PedsQL	107	54.9 (21.0)	52.0 (20.8)	2.9 (15.6)	**0.0304**	0.18	128	57.6 (19.8)	52.3 (19.6)	5.3 (15.4)	**< 0.001**	0.34
EQ-5D-Y-3L	107	8.3 (2.4)	8.8 (2.4)	− 0.5 (2.0)	**0.0071**	− 0.24	128	7.8 (2.2)	8.4 (2.4)	− 0.6 (1.8)	**< 0.001**	− 0.36
EQ-5D-Y-5L	106	9.9 (4.1)	11.5 (4.7)	− 1.6 (3.7)	**< 0.001**	− 0.42	125	9.4 (3.8)	10.6 (4.6)	− 1.2 (3.5)	**< 0.001**	− 0.34
CHU9D	107	21.3 (8.1)	23.3 (8.2)	− 2.0 (8.0)	**0.0053**	− 0.25	128	20.0 (7.5)	22.7 (8.2)	− 2.6 (6.8)	**< 0.001**	− 0.37
AQoL-6D	22	44.4 (12.8)	46.1 (11.2)	− 1.7 (10.0)	0.2190	− 0.17	23	43.3 (11.3)	45.3 (11.8)	− 2.0 (8.5)	0.1409	− 0.23
HUI3	15	13.4 (5.0)	13.7 (4.6)	− 0.3 (2.8)	0.3596	− 0.10	30	12.9 (4.2)	12.6 (4.0)	0.3 (2.5)	0.7179	0.11

AQoL-6D Assessment of Quality of Life 6D, *CHU9D* Child Health Utility 9D, *EQ-5D-Y-3L* and *EQ-5D-Y-5L* refer to the EuroQol Group's Youth instruments with 3 and 5 response options, respectively, *HUI3* Health Utilities Index Mark 3, *PedsQL* Paediatric Quality of Life Inventory, *SD* standard deviation

Bold indicates significant *p*-value < 0.05. Grey cells indicate a small sample size and hence results should be interpreted with caution

'-' indicates SRM was not reported as there was no significant difference between assessments

Sensitivity analyses were conducted using only the sample of children recruited via hospital (Sample 1) due to the lower follow-up response rates in the online panel samples (Samples 2 and 3) (See Supplementary Table 9 in the ESM for baseline characteristic differences between 4-week completers and non-completers), and these results are presented in Supplementary Table 10 (see ESM). Results from the sensitivity analysis are consistent with the main analysis, although effect sizes were larger with respect to responsiveness to improved health.

Responsiveness results varied by subgroup and these results are presented in Supplementary Table 11 (see ESM). The EQ-5D-Y-5L and CHU9D were the only instruments to demonstrate responsiveness across all subgroups to *improving health* (including main health condition and general health), in terms of a significant mean difference. The EQ-5D-Y-5L was the only instrument to demonstrate responsiveness across all subgroups for *worsening health* (including main health condition and general health), in terms of a significant mean difference. Among the subgroup of children aged 13–18 years with worsening general health and the subgroup of proxy report for both worsening main health condition and general health, sample sizes may be too small to robustly assess responsiveness.

3.7 Summary of Psychometric Performance

Table 6 summarises the performance of all instruments for all psychometric analyses completed. In the total sample, the EQ-5D-Y-5L and CHU9D demonstrated evidence of good performance, meeting the prespecified cut offs, for all psychometric analyses completed. The PedsQL also demonstrated evidence of good performance, with the only exception being lack of responsiveness for improving health and the EQ-5D-Y-3L demonstrated evidence of good performance with the only exception being ceiling effects. The AQoL-6D and HUI3 demonstrated evidence of good performance for all psychometric attributes able to be assessed; however, due to sample size limitations, conclusions regarding test–retest reliability and responsiveness were not able to be drawn.

Additionally, results for ceiling effects, test–retest reliability and responsiveness varied by child age and report type.

4 Discussion

We have generated new evidence on the comparative psychometric performance of the PedsQL, EQ-5D-Y-3L, EQ-5D-Y-5L, CHU9D, AQoL-6D, and HUI3 descriptive systems by child age (5–12 years vs 13–18 years), report type (self- vs proxy report), and child health status. The EQ-5D-Y-5L and CHU9D demonstrated evidence supporting their psychometric performance in the total sample, based on prespecified criteria, for all assessments conducted—ceiling and floor effects, test–retest reliability, known-group validity, convergent and divergent validity, and responsiveness. Only the EQ-5D-Y-3L had evidence of a ceiling effect among children with special healthcare needs; however, it is important to note these ceiling effects were mild, with the proportion only just over the 15% threshold (17.8%). No instruments had floor effects. The PedsQL, EQ-5D-Y-3L, EQ-5D-Y-5L, and CHU9D demonstrated moderate–excellent test–retest reliability in the total sample. All instruments demonstrated known-group validity with large effect sizes. All instrument items were correlated with PedsQL items in the hypothesised directions, demonstrating convergent and divergent validity. The EQ-5D-Y-3L, EQ-5D-Y-5L, and CHU9D demonstrated some responsiveness to improvements in health with small effect sizes. The PedsQL, EQ-5D-Y-3L, EQ-5D-Y-5L, and CHU9D demonstrated some responsiveness to worsening health with small effect sizes. Importantly, ceiling effect, test–retest reliability and responsiveness results varied by child age and report type.

The P-MIC study builds on previous research comparing adult HRQoL instruments in a Multi-Instrument Comparison (MIC) study [34], and is the first of its kind worldwide to collect common paediatric HRQoL instruments head-to-head concurrently across a large and diverse sample. It allows for a direct comparison of generic instruments across a range of child ages and health statuses within the same dataset, which is currently missing from the literature [10, 11]. The P-MIC study was designed with strong sample quality assessment procedures, resulting in high-quality data from both hospital and online samples to enable robust conclusions [18, 35].

Ceiling effects are statistically more likely with shorter instrument length and shorter recall period which is likely what is observed in this study, with higher ceiling effects for the EQ-5D-Y-3L [36]. These findings are consistent with previous studies [37–40]. In adults, the EQ-5D-5L is preferred over the EQ-5D-3L, given the reduction in overall ceiling effects [41]. Although a previous study noted no difference in ceiling effects between the EQ-5D-Y-3L and EQ-5D-Y-5L [38], this study demonstrated the EQ-5D-Y-5L had slightly lower overall ceiling effects compared with the 3L (14.8% vs 17.8%) in the total sample, which is consistent with the finding of a recent systematic review of EQ-5D-Y performance [42]. Information on ceiling effects can be useful when informing the choice of instrument for specific samples; for instance, future instrument users may wish to avoid the use of instruments with stronger ceiling effects, such as the EQ-5D-Y-3L, in relatively well populations of children where there is a higher risk of very large ceiling effects.

Table 6 Summary of instrument psychometric performance, by child age and report type

Instrument	Response distribution (No ceiling effect)[a]	Test–retest reliability[b]	Known-group validity[c]	Convergent & divergent validity[d]	Responsiveness[e]
Total sample					
PedsQL	✓	✓	✓	✓	? (✗ for improving health, ✓for worsening health)
EQ-5D-Y-3L	✗	✓	✓	✓	✓
EQ-5D-Y-5L	✓	✓	✓	✓	✓
CHU9D	✓	✓	✓	✓	✓
AQoL-6D	✓	? (small sample size)	✓	✓	? (small sample size)
HUI3	✓	? (small sample size)	✓	✓	? (small sample size)
By child age					
5–12 years					
PedsQL	✓	✓	✓	n/a	? (✗ for improving health, ✓for worsening main health condition, ✗ for worsening general health)
EQ-5D-Y-3L	✗	✓	✓	n/a	✓
EQ-5D-Y-5L	✗	✓	✓	n/a	✓
CHU9D	✓	✓	✓	n/a	? (✓for improving health, ✓ for worsening main health condition, ✗ for worsening general health)
AQoL-6D	✓	? (small sample size)	✓	n/a	? (small sample size)
HUI3	✓	? (small sample size)	✓	n/a	? (small sample size)
13–18 years					
PedsQL	✓	✓	✓	n/a	? (✗for improving health, ✓ for worsening main health condition, ? for worsening general health [small sample size])
EQ-5D-Y-3L	✗	✓	✓	n/a	? (✓ for improving main health condition, ✗ for improving general health, ✓ for worsening main health condition, ? for worsening general health [small sample size])
EQ-5D-Y-5L	✓	✓	✓	n/a	✓ improving main health, ✓ for worsening main health condition, ? for worsening general health (small sample size)
CHU9D	✓	✗	✓	n/a	✓ improving main health, ✓ for worsening main health condition, ? for worsening general health (small sample size)
AQoL-6D	✓	? (small sample size)	✓	n/a	? (small sample size)
HUI3	✓	? (small sample size)	✓	n/a	? (small sample size)
By report type					
Proxy report					
PedsQL	✓	✓	✓	n/a	✗for improving health, ? for worsening health (small sample size)
EQ-5D-Y-3L	✗	✓	✓	n/a	✗for improving health, ? for worsening health (small sample size)
EQ-5D-Y-5L	✗	✗	✓	n/a	✓ for improving health, ? for worsening health (small sample size)
CHU9D	✓	✓	✓	n/a	✓ for improving health, ? for worsening health (small sample size)
AQoL-6D	✓	? (small sample size)	✓	n/a	? (small sample size)
HUI3	✓	? (small sample size)	✓	n/a	? (small sample size)
Self-report					
PedsQL	✓	✓	✓	n/a	✓
EQ-5D-Y-3L	✗	✓	✓	n/a	✓
EQ-5D-Y-5L	✓	✓	✓	n/a	✓

Table 6 (continued)

Instrument	Response distribution (No ceiling effect)[a]	Test–retest reliability[b]	Known-group validity[c]	Convergent & divergent validity[d]	Responsiveness[e]
CHU9D	✓	✓	✓	n/a	✓
AQoL-6D	✓	? (small sample size)	✓	n/a	? (small sample size)
HUI3	✓	? (small sample size)	✓	n/a	? (small sample size)

AQoL-6D Assessment of Quality of Life 6D, *CHU9D* Child Health Utility 9D, *EQ-5D-Y-3L* and *EQ-5D-Y-5L* refer to the EuroQol Group's Youth instruments with 3 and 5 response options, respectively, *HUI3* Health Utilities Index Mark 3, *n/a* not applicable, *PedsQL* Paediatric Quality of Life Inventory

✓ Evidence of significant performance; ✗ No evidence of significant performance; ? Inconclusive evidence. Convergent validity was not assessed by subgroups

[a]✓ Less than 15% of participants with a special healthcare need reported the lowest severity or frequency level (i.e., 'no problems') across all items

[b]✓ ICC (moderate, good, or excellent) ≥ 0.5

[c]✓ Effect sizes ≥ 0.8 and mean difference with a p-value of < 0.05

[d]✓ Items at least moderately correlated (Spearman's correlation ≥ 0.3) with other instrument items where hypothesised to be correlated (convergent validity) and weakly correlated (Spearman's correlation < 0.3) where hypothesised not to be correlated (divergent validity)

[e]✓ Mean difference with a p-value of < 0.05

This study found evidence of good test–retest reliability for the PedsQL, EQ-5D-Y-3L, and EQ-5D-Y-5L, and evidence of moderate test–retest reliability for the CHU9D in the total sample. A previous review found no significant evidence of test–retest reliability for CHU9D and mixed evidence of test–retest reliability for the EQ-5D-Y-3L and HUI3 [10]. The previous review includes only one study assessing test–retest reliability of the CHU9D, which assesses the test–retest reliability of the dimensions rather than the whole instrument; hence, results may not be comparable to this study [43]. However, other studies have reported significant evidence of test–retest reliability for the CHU9D in China and Sweden [44, 45]. This study also found some instruments demonstrated better test–retest reliability than others in different subgroups, with only the PedsQL and EQ-5D-Y-3L demonstrating evidence of good or excellent test–retest reliability across all subgroups. These results highlight the importance of assessing the psychometric properties of instruments across a wide range of participants, as instrument performance can vary depending on the population it is assessed in. Additionally, given the lack of evidence previously available regarding test–retest reliability [9, 14, 15], the newly generated evidence from this study helps to fill an important gap in the literature.

Although almost all instrument items hypothesised to be at least moderately correlated with PedsQL items were at least moderately correlated, indicating convergent validity of instruments, not all were strongly correlated. Our *a priori* hypotheses were driven by the wording of each item, thus moderate correlations might be observed rather than strong correlations due to other differences between the items, for example, the response options (severity vs frequency); the item format (statements vs Likert scale items) and the recall period (today vs 1 week vs 1 month). Additionally, instrument items hypothesised not to be at least moderately correlated with PedsQL items were found largely not to be correlated, indicating divergent validity of instruments. This may be due to instruments measuring different aspects of HRQoL to the PedsQL. For example, the HUI3 had the least number of items correlated with PedsQL items and this finding is consistent with expectations, as the HUI3 had the fewest items hypothesised to be correlated with PedsQL items. It is likely that the HUI3 measures different aspects of HRQoL to the PedsQL, such as vision, hearing, speech, and dexterity. This has been explored in more detail in a companion paper assessing the dimensionality of instruments.

In the total sample, the EQ-5D-Y-3L, EQ-5D-Y-5L, and CHU9D were responsive to improvements in health, and the PedsQL, EQ-5D-Y-3L, EQ-5D-Y-5L, and CHU9D were responsive to worsening health; however, despite demonstrating significant mean differences, effect sizes were small. When the responsiveness analysis was restricted only to participants recruited via hospital (Sample 1), results were consistent with those from the main analysis; however, effect sizes were slightly larger with respect to responsiveness to improved health, although were still in the small–moderate range. This may be because participants from the sample recruited via hospital (Sample 1) report poorer HRQoL in the initial survey compared with participants recruited from the online panel (Sample 2 and 3); hence may have more capacity for improved HRQoL at the follow-up survey. A previous review of common paediatric HRQoL instruments found evidence of responsiveness for EQ-5D-Y-3L and mixed evidence of responsiveness for CHU9D and HUI3

[10]. This finding is consistent with that of another more recent review that also found mixed evidence of responsiveness for the CHU9D and HUI3 [9]. This more recent review also found evidence of responsiveness for the EQ-5D-Y-5L [9]; however, this was only from one study. In the P-MIC study, researchers were reliant on a change in health happening naturally between the two surveys, as well as caregivers being able to accurately report this change. This may be less clear than defining a change in health based on a clinical intervention or external clinical information, which may be why responsiveness effect sizes are smaller than those previously published. Previous literature assessing responsiveness is limited [14, 15], meaning that the new knowledge generated through this study is a significant addition.

This study has several limitations. Firstly, sample sizes for the AQoL-6D and HUI3 are too small to robustly assess test–retest reliability and responsiveness results; hence, these results have not been presented. This limitation arises due to these instruments being presented in randomised blocks to reduce responder burden. Secondly, the reliance on online report means that we do not have any non-participant reported information, such as clinical tests or clinical reports; hence, we have no external indicator against which to compare the responses given by participants. We dealt with this using strong data quality control procedures [18], but acknowledge uncertainties with online data collection. Thirdly, the use of LSS to assess known-group validity, test–retest reliability, and responsiveness is a limitation. Although the total score on the PedsQL is an accepted way to score responses [23], using LSS for the EQ-5D-Y-3L, EQ-5D-Y-5L, HUI3, and CHU9D is more challenging. The LSS is a very crude measure and may miss the complexity and relative importance of the different dimensions and levels that are captured by utility scores [46]. Future work on the P-MIC data will explore the psychometric performance of instruments using preference weights. LSS was used due to the lack of established preference weights for all instruments included in this study. Further to this point, a limitation for the implementation of these results is that not all instruments have available value sets to enable scoring for use in economic evaluations, or these are not available across many countries, precluding their use in country-specific contexts, which is an area for future research. It is noted that there is a mismatch between strongest psychometric performance and availability of value sets, which will practically limit those wishing to use instruments to produce utilities. For example, the EQ-5D-Y-5L was considered a top performing instrument in this study, however, there are currently no value sets available for this instrument, which limits its applicability in an economic evaluation. Fourthly, the P-MIC study includes a select range of commonly used generic paediatric HRQoL instruments; although as many instruments were included in the study as was feasible, it is important to acknowledge that other instruments exist that are not captured in this study. Finally, although this study provides evidence of performance in an Australian population, it is unknown if these results are generalisable across countries.

This study addresses major gaps in current evidence for child HRQoL instruments and provides an overview of the performance of commonly used child HRQoL instruments, however, several gaps remain and should be the focus of future research—assessment of performance by more granular child ages, assessment of performance using utility scores (once available), assessment of performance by different child health conditions, test–retest reliability and responsiveness of the AQoL-6D and HUI3, and assessment of performance in different countries.

5 Conclusion

The results presented here have implications for the use and choice of instruments in the design of studies, and use of HRQoL data in decision making. In the total sample, the EQ-5D-Y-5L and CHU9D descriptive systems demonstrated evidence of significant performance, based on prespecified criteria, across all psychometric attributes tested—ceiling and floor effects, test–retest reliability, known-group validity, convergent and divergent validity, and responsiveness (improving and worsening health). Additionally, the PedsQL, EQ-5D-Y-3L, AQoL-6D, and HUI3 descriptive systems performed well on most psychometric attributes evaluated in the total sample. Importantly, instrument performance varied by child age and report type for ceiling effect, test–retest reliability and responsiveness results, hence future instrument users should consider their intended population and study design when selecting an instrument. For example, when looking at the CHU9D and EQ-5D-Y-5L, the EQ-5D-Y-5L did not demonstrate acceptable test–retest reliability when proxy reported, the CHU9D did not demonstrate acceptable test–retest reliability in children aged 13–18 years, and the CHU9D did not demonstrate significant responsiveness to worsening general health in children aged 5–12 years. Furthermore, different instruments may capture different aspects of HRQoL and including only one instrument may result in important aspects of HRQoL potentially not being measured. The choice of instrument in studies may be guided by the population being assessed and aspects of HRQoL that are expected to be impacted by a condition or treatment. Those designing studies may consider including multiple instruments to ensure a more holistic assessment of HRQoL and to enable sensitivity analysis of HRQoL inputs in economic evaluation.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40273-023-01330-2.

Acknowledgements We would like to thank the QUOKKA study investigators, research fellows and research assistants: Richard Norman, Rosalie Viney, Julie Ratcliffe, Deborah Street, Cate Bailey, Christine Mpundu-Kaambwa, Tessa Peasgood, Kristy McGregor and Shilana Yip. We would also like to thank the children and families for taking the time to complete the study as well as members of the consumer advisory group who helped pilot, test and refine different elements of this study. Finally, we would like to sincerely thank Dr Oliver Rivero-Arias for providing guidance and feedback during the development of this manuscript.

Declarations

Funding/Support Open Access funding enabled and organized by CAUL and its Member Institutions. This research was funded by an Australian Government Medical Research Futures Fund (MRFF) grant (1200816) and a EuroQol Research Foundation Grant (361-RA). RJ and RO are supported by a Research Training Program Scholarship provided by the Australian Commonwealth Government and the University of Melbourne. RJ is also supported by a EuroQol Research Foundation PhD Grant (330-PhD). HH is funded by a National Health and Medical Research Council (NHMRC) Practitioner Fellowship (1136222). The Murdoch Children's Research Institute is supported by the Victorian Government's Operational Infrastructure Support Program.

Competing Interests KD, ND, BM, HH, and RJ have all received previous or current funding from the EuroQol Research Foundation, who is the developer and copyright holder of some instruments included in this study. ND and BM are members of the EuroQol Group. The EuroQol Research Foundation contributed to funding for this research. Views expressed in this paper are those of the authors and are not necessarily those of the EuroQol Research Foundation.

Ethics Approval This study was approved by The Royal Children's Hospital (RCH) Human Research Ethics Committee (HREC/71872/RCHM2021) on 20 May 2021.

Registration The study was prospectively registered with the Australia New Zealand Clinical Trials Registry (ANZCTR) (ACTRN12621000657820).

Data Availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Consent for Publication Not applicable.

Code Availability Not applicable.

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

Author Contributions Conceptualisation: KD, ND, BM, HH, GC. Methodology: all authors. Project administration: RJ. Formal analysis: RJ. Validation: RO, XX, MB. First manuscript draft writing: RJ. Draft review and editing: all authors. Supervision: KD, ND, BM, HH. All authors read and approved the final manuscript.

Disclosure This article is published in a journal supplement wholly funded by the Australian Government MRFF funded QUOKKA research grant 1200816, the University of Melbourne, the EuroQol Research Foundation, the University of Technology Sydney and Flinders University.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License, which permits any non-commercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc/4.0/.

References

1. De Civita M, Regier D, Alamgir AH, Anis AH, FitzGerald MJ, Marra CA. Evaluating health-related quality-of-life studies in paediatric populations. Pharmacoeconomics. 2005;23(7):659–85. https://doi.org/10.2165/00019053-200523070-00003.
2. Wallander JL, Koot HM. Quality of life in children: A critical examination of concepts, approaches, issues, and future directions. Clin Psychol Rev. 2016;45:131–43. https://doi.org/10.1016/j.cpr.2015.11.007.
3. Churruca K, Pomare C, Ellis LA, Long JC, Henderson SB, Murphy LED, et al. Patient-reported outcome measures (PROMs): a review of generic and condition-specific measures and a discussion of trends and issues. Health Expect. 2021;24(4):1015–24. https://doi.org/10.1111/hex.13254.
4. Kromm SK, Bethell J, Kraglund F, Edwards SA, Laporte A, Coyte PC, et al. Characteristics and quality of pediatric cost-utility analyses. Qual Life Res. 2012;21(8):1315–25. https://doi.org/10.1007/s11136-011-0049-7.
5. Kennedy-Martin M, Slaap B, Herdman M, van Reenen M, Kennedy-Martin T, Greiner W, et al. Which multi-attribute utility instruments are recommended for use in cost-utility analysis? A review of national health technology assessment (HTA) guidelines. Eur J Health Econ. 2020;21(8):1245–57. https://doi.org/10.1007/s10198-020-01195-8.
6. Rencz F, Gulácsi L, Drummond M, Golicki D, Prevolnik Rupel V, Simon J, et al. EQ-5D in Central and Eastern Europe: 2000–2015. Qual Life Res. 2016;25(11):2693–710. https://doi.org/10.1007/s11136-016-1375-6.
7. Black N. Patient reported outcome measures could help transform healthcare. BMJ Br Med J. 2013;346: f167. https://doi.org/10.1136/bmj.f167.
8. Al Sayah F, Jin X, Johnson JA. Selection of patient-reported outcome measures (PROMs) for use in health systems. J Patient-Report Outcomes. 2021;5(Suppl 2):99. https://doi.org/10.1186/s41687-021-00374-2.
9. Kwon J, Smith S, Raghunandan R, Howell M, Huynh E, Kim S, et al. Systematic review of the psychometric performance of generic childhood multi-attribute utility instruments. Appl Health Econ Health Policy. 2023. https://doi.org/10.1007/s40258-023-00806-8.
10. 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 in Health. 2021;24(3):443–60. https://doi.org/10.1016/j.jval.2020.09.012.
11. Chen G, Ratcliffe J. A review of the development and application of generic multi-attribute utility instruments for paediatric

populations. Pharmacoeconomics. 2015;33(10):1013–28. https://doi.org/10.1007/s40273-015-0286-7.
12. Germain N, Aballéa S, Toumi M. Measuring the health-related quality of life in young children: how far have we come? J Mark Access Health Policy. 2019;7(1):1618661. https://doi.org/10.1080/20016689.2019.1618661.
13. Ungar W. Economic evaluation in child health. Oxford University Press; 2009.
14. Tan RL, 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):155–74. https://doi.org/10.1007/s40273-022-01205-y.
15. Janssens A, Rogers M, Thompson Coon J, Allen K, Green C, Jenkinson C, et al. A systematic review of generic multidimensional patient-reported outcome measures for children, part II: evaluation of psychometric performance of English-language versions in a general population. Value Health J Int Soc Pharmacoecon Outcomes Res. 2015;18(2):334–45. https://doi.org/10.1016/j.jval.2015.01.004.
16. Janssens A, Thompson Coon J, Rogers M, Allen K, Green C, Jenkinson C, et al. A systematic review of generic multidimensional patient-reported outcome measures for children, part I: descriptive characteristics. Value Health J Int Soc Pharmacoecon Outcomes Res. 2015;18(2):315–33. https://doi.org/10.1016/j.jval.2014.12.006.
17. Jones R, Mulhern B, McGregor K, Yip S, O'Loughlin R, Devlin N, et al. Psychometric performance of HRQoL Measures: an Australian Paediatric Multi-Instrument Comparison Study Protocol (P-MIC). Children (Basel, Switzerland). 2021. https://doi.org/10.3390/children8080714.
18. Jones R, O'Loughlin R, Xiong X, Bahrampour M, McGregor K, Yip S, et al. Collecting paediatric health-related quality of life data: assessing the feasibility and acceptability of the australian paediatric multi-instrument comparison (P-MIC) study. Children (Basel, Switzerland). 2023;10(10):1604. https://doi.org/10.3390/children10101604.
19. Jones R, Mulhern B, Devlin N, Hiscock H, Chen G, O'Loughlin R, Xiong X, Bahrampour, McGregor K, Yip S, Dalziel K. Australian Paediatric Multi-Instrument Comparison (P-MIC) Study: Technical Methods Paper [Online]. Melbourne, Australia; 2023. https://www.quokkaresearchprogram.org/project-1-1. Accessed 24 May 2023
20. Varni JW, Burwinkle TM, Seid M, Skarr D. The PedsQL 4.0 as a pediatric population health measure: feasibility, reliability, and validity. Ambul Pediatr. 2003;3(6):329–41. https://doi.org/10.1367/1539-4409(2003)003%3c0329:TPAAPP%3e2.0.CO;2.
21. Feng YS, Jiang R, Pickard AS, Kohlmann T. Combining EQ-5D-5L items into a level summary score: demonstrating feasibility using non-parametric item response theory using an international dataset. Qual Life Res. 2022;31(1):11–23. https://doi.org/10.1007/s11136-021-02922-1.
22. Bethell CD, Read D, Stein RE, Blumberg SJ, Wells N, Newacheck PW. Identifying children with special health care needs: development and evaluation of a short screening instrument. Ambul Pediatr. 2002;2(1):38–48. https://doi.org/10.1367/1539-4409(2002)002%3c0038:icwshc%3e2.0.co;2.
23. Varni JW, Seid M, Kurtin PS. PedsQL™ 4.0: reliability and validity of the pediatric Quality of Life Inventory™ Version 4.0 generic core scales in healthy and patient populations. Med Care. 2001. https://doi.org/10.1097/00005650-200108000-00006.
24. Terwee CB, Bot SDM, de Boer MR, van der Windt DAWM, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60(1):34–42. https://doi.org/10.1016/j.jclinepi.2006.03.012.
25. McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? Qual Life Res. 1995;4(4):293–307. https://doi.org/10.1007/BF01593882.
26. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155–63. https://doi.org/10.1016/j.jcm.2016.02.012.
27. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. American Psychological Association; 1994. p. 284–90. https://doi.org/10.1037/1040-3590.6.4.284.
28. Cohen J. A power primer. Psychol Bull. 1992;112(1):155–9. https://doi.org/10.1037//0033-2909.112.1.155.
29. Sullivan GM, Feinn R. using effect size—or why the P value is not enough. J Grad Med Educ. 2012;4(3):279–82. https://doi.org/10.4300/JGME-D-12-00156.1.
30. Chen H-Y, Cisler RA. Assessing health-related quality of life among children with special health care needs in the United States. Child Health Care. 2011;40(4):311–25. https://doi.org/10.1080/02739615.2011.617239.
31. Peasgood T, Mukuria C, Brazier J, Marten O, Kreimeier S, Luo N, et al. Developing a new generic health and wellbeing measure: psychometric survey results for the EQ-HWB. Value in Health. 2022;25(4):525–33. https://doi.org/10.1016/j.jval.2021.11.1361.
32. Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. J Clin Epidemiol. 2000;53(5):459–68. https://doi.org/10.1016/s0895-4356(99)00206-1.
33. Mokkink LB, Prinsen C, Patrick DL, Alonso J, Bouter LM, De Vet H, et al. COSMIN Study Design checklist for Patient-reported outcome measurement instruments. Amsterdam; 2019, p. 1–32. https://www.cosmin.nl/wp-content/uploads/COSMIN-study-designing-checklist_final.pdf.
34. Richardson J, Iezzi A, Khan MA, Maxwell A. Cross-national comparison of twelve quality of life instruments: MIC paper 1: background, questions, instruments. Research paper 76 [Online]. Melbourne: Centre for Health Economics, Monash University; 2012. https://www.aqol.com.au/papers/researchpaper76.pdf. Accessed 8 May 2023.
35. Gagnier JJ, Lai J, Mokkink LB, Terwee CB. COSMIN reporting guideline for studies on measurement properties of patient-reported outcome measures. Qual Res. 2021;30(8):2197–218. https://doi.org/10.1007/s11136-021-02822-4.
36. Cunillera O, Tresserras R, Rajmil L, Vilagut G, Brugulat P, Herdman M, et al. Discriminative capacity of the EQ-5D, SF-6D, and SF-12 as measures of health status in population health survey. Qual Life Res. 2010;19(6):853–64. https://doi.org/10.1007/s11136-010-9639-z.
37. Ravens-Sieberer U, Wille N, Badia X, Bonsel G, Burström K, Cavrini G, et al. Feasibility, reliability, and validity of the EQ-5D-Y: results from a multinational study. Qual Life Res. 2010;19(6):887–97. https://doi.org/10.1007/s11136-010-9649-x.
38. Fitriana TS, Purba FD, Rahmatika R, Muhaimin R, Sari NM, Bonsel G, 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. https://doi.org/10.1186/s12955-021-01889-4.
39. Sung L, Greenberg ML, Doyle JJ, Young NL, Ingber S, Rubenstein J, et al. Construct validation of the Health Utilities Index and the Child Health Questionnaire in children undergoing cancer chemotherapy. Br J Cancer. 2003;88(8):1185–90. https://doi.org/10.1038/sj.bjc.6600895.
40. Le HND, Petersen S, Mensah F, Gold L, Wake M, Reilly S. Health-related quality of life in children with low language or

congenital hearing loss, as measured by the PedsQL and Health Utility Index mark 3. Value in Health. 2020;23(2):164–70. https://doi.org/10.1016/j.jval.2019.07.019.

41. 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):675–97. https://doi.org/10.1007/s40273-018-0623-8.
42. Golicki D, Młyńczak K. Measurement properties of the EQ-5D-Y: a systematic review. Value Health. 2022. https://doi.org/10.1016/j.jval.2022.05.013.
43. Canaway AG, Frew EJ. Measuring preference-based quality of life in children aged 6–7 years: a comparison of the performance of the CHU-9D and EQ-5D-Y—the WAVES Pilot Study. Qual Life Res. 2013;22(1):173–83. https://doi.org/10.1007/s11136-012-0119-5.
44. Yang P, Chen G, Wang P, Zhang K, Deng F, Yang H, et al. Psychometric evaluation of the Chinese version of the Child Health Utility 9D (CHU9D-CHN): a school-based study in China. Qual Life Res. 2018;27(7):1921–31. https://doi.org/10.1007/s11136-018-1864-x.
45. Lindvall K, Vaezghasemi M, Feldman I, Ivarsson A, Stevens KJ, Petersen S. Feasibility, reliability and validity of the health-related quality of life instrument Child Health Utility 9D (CHU9D) among school-aged children and adolescents in Sweden. Health Qual Life Outcomes. 2021;19(1):193. https://doi.org/10.1186/s12955-021-01830-9.
46. Devlin N, Parkin D, Janssen B. Analysis of EQ-5D Profiles. In: Devlin N, Parkin D, Janssen B, editors. Methods for analysing and reporting EQ-5D data. Cham: Springer International Publishing; 2020. p. 23–49.
47. Wille N, Badia X, Bonsel G, Burström K, Cavrini G, Devlin N, et al. Development of the EQ-5D-Y: a child-friendly version of the EQ-5D. Qual Life Res. 2010. https://doi.org/10.1007/s11136-010-9648-y.
48. Stevens K. Developing a descriptive system for a new preference-based measure of health-related quality of life for children. Qual Life Res. 2009. https://doi.org/10.1007/s11136-009-9524-9.
49. Stevens K. Assessing the performance of a new generic measure of health-related quality of life for children and refining it for use in health state valuation. Appl Health Econ Health Policy. 2011. https://doi.org/10.2165/11587350-000000000-00000.
50. Ratcliffe J, Stevens K, Flynn T, Brazier J, Sawyer MG. Whose values in health? An empirical comparison of the application of adolescent and adult values for the CHU-9D and AQOL-6D in the Australian adolescent general population. Value Health. 2012. https://doi.org/10.1016/j.jval.2012.04.005.
51. Moodie M, Richardson J, Rankin B, Iezzi A, Sinha K. Predicting time trade-off health state valuations of adolescents in four Pacific countries using the Assessment of Quality-of-Life (AQoL-6D) instrument. Value Health. 2010. https://doi.org/10.1111/j.1524-4733.2010.00780.x.
52. Feeny D, Furlong W, Torrance GW, Goldsmith CH, Zhu Z, DePauw S, et al. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. Med Care. 2002. https://doi.org/10.1097/00005650-200202000-00006.
53. Torrance GW, Feeny DH, Furlong WJ, Barr RD, Zhang Y, Wang Q. Multiattribute utility function for a comprehensive health status classification system: health utilities index mark 2. Med Care. 1996. https://doi.org/10.1097/00005650-199607000-00004.
54. Horsman J, Furlong W, Feeny D, Torrance G. The Health Utilities Index (HUI): concepts, measurement properties and applications. Health Qual Life Outcomes. 2003. https://doi.org/10.1186/1477-7525-1-54.

Authors and Affiliations

Renee Jones[1,2] · Rachel O'Loughlin[1,2,4] · Xiuqin Xiong[1] · Mina Bahrampour[3] · Nancy Devlin[1] · Harriet Hiscock[2,4,5] · Gang Chen[6] · Brendan Mulhern[3] · Kim Dalziel[1,2] on behalf of on behalf of the Quality of Life in Kids: Key Evidence to Strengthen Decisions in Australia (QUOKKA) Project Team

✉ Renee Jones
reneej1@student.unimelb.edu.au

Rachel O'Loughlin
oloughlin.r@unimelb.edu.au

Xiuqin Xiong
xiuqinx@student.unimelb.edu.au

Mina Bahrampour
Mina.Bahrampour@uts.edu.au

Nancy Devlin
nancy.devlin@unimelb.edu.au

Harriet Hiscock
Harriet.Hiscock@rch.org.au

Gang Chen
gang.chen@monash.edu

Brendan Mulhern
Brendan.Mulhern@uts.edu.au

Kim Dalziel
kim.dalziel@unimelb.edu.au

[1] Health Economics Unit, Centre for Health Policy, The University of Melbourne, Melbourne, VIC, Australia

[2] Health Services and Economics, Murdoch Children's Research Institute, Melbourne, VIC, Australia

[3] Centre for Health Economics Research and Evaluation, University of Technology Sydney, Ultimo, NSW, Australia

[4] Health Services Research Unit, Royal Children's Hospital, Melbourne, VIC, Australia

[5] Department of Paediatrics, The University of Melbourne, Melbourne, VIC, Australia

[6] Centre for Health Economics, Monash Business School, Monash University, Melbourne, VIC, Australia