PharmacoEconomics (2024) 42 (Suppl 1):S57–S77
https://doi.org/10.1007/s40273-024-01354-2

ORIGINAL RESEARCH ARTICLE

Comparing the Psychometric Performance of Generic Paediatric Health-Related Quality of Life Instruments in Children and Adolescents with ADHD, Anxiety and/or Depression

Rachel O'Loughlin[1,2,3] · Renee Jones[1,2,3] · Gang Chen[4] · Brendan Mulhern[5] · Harriet Hiscock[2,3,6] · Nancy Devlin[1] · Kim Dalziel[1,3] · in collaboration with the Quality Of Life in Kids: Key evidence to strengthen decisions in Australia (QUOKKA) project team

Accepted: 11 October 2023 / Published online: 8 February 2024

Abstract

Objective The aim of this study was to examine the validity, reliability and responsiveness of common generic paediatric health-related quality of life (HRQoL) instruments in children and adolescents with mental health challenges.

Methods Participants were a subset of the Australian Paediatric Multi-Instrument Comparison (P-MIC) study and comprised 1013 children aged 4–18 years with attention-deficit/hyperactivity disorder (ADHD) ($n = 533$), or anxiety and/or depression ($n = 480$). Participants completed an online survey including a range of generic paediatric HRQoL instruments (PedsQL, EQ-5D-Y-3L, EQ-5D-Y-5L, CHU9D) and mental health symptom measures (SDQ, SWAN, RCADS-25). A subset of participants also completed the HUI3 and AQoL-6D. The psychometric performance of each HRQoL instrument was assessed regarding acceptability/feasibility; floor/ceiling effects; convergent validity; known-group validity; responsiveness and test–retest reliability.

Results The PedsQL, CHU9D, EQ-5D-Y-3L and EQ-5D-Y-5L showed similarly good performance for acceptability/feasibility, known-group validity and convergent validity. The CHU9D and PedsQL showed no floor or ceiling effects and fair–good test–retest reliability. Test–retest reliability was lower for the EQ-5D-Y-3L and EQ-5D-Y-5L. The EQ-5D-Y-3L showed the highest ceiling effects, but was the top performing instrument alongside the CHU9D on responsiveness to improvements in health status, followed by the PedsQL. The AQoL-6D and HUI3 showed good acceptability/feasibility, no floor or ceiling effects, and good convergent validity, yet poorer performance on known-group validity. Responsiveness and test–retest reliability were not able to be assessed for these two instruments. In subgroup analyses, performance was similar for all instruments for acceptability/feasibility, known-group and convergent validity, however, relative strengths and weaknesses for each instrument were noted for ceiling effects, responsiveness and test–retest reliability. In sensitivity analyses using utility scores, performance regarding known-group and convergent validity worsened slightly for the EQ-5D-Y-3L and CHU9D, though improved slightly for the HUI3 and AQoL-6D.

Conclusions While each instrument showed strong performance in some areas, careful consideration of the choice of instrument is advised, as this may differ dependent on the intended use of the instrument, and the age, gender and type of mental health condition of the population in which the instrument is being used.

Trial Registration ANZCTR—ACTRN12621000657820.

The Members of Quality Of Life in Kids: Key evidence to strengthen decisions in Australia (QUOKKA) are mentioned in "Acknowledgements".

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

1 Health Economics Unit, School of Population and Global Health, University of Melbourne, Melbourne, Victoria 3010, Australia

2 Health Services Research Unit, The Royal Children's Hospital, Parkville, VIC, Australia

3 Health Services and Economics, Murdoch Children's Research Institute, Parkville, VIC, Australia

4 Centre for Health Economics, Monash University, Caulfield East, VIC, Australia

5 Centre for Health Economics Research and Evaluation, University of Technology Sydney, Ultimo, NSW, Australia

6 Department of Paediatrics, University of Melbourne, Melbourne, VIC, Australia

Key Points for Decision Makers

Our results indicate that the CHU9D, PedsQL, EQ-5D-Y-3L and EQ-5D-Y-5L perform equally well in children and adolescents with ADHD or anxiety and/or depression, regarding acceptability/feasibility, known-group and convergent validity. However, relative strengths of the CHU9D and PedsQL were observed regarding their lack of ceiling effects, greater test–retest reliability, and consistently good performance across all psychometric properties. The CHU9D and EQ-5D-Y-3L were the most responsive to improvements in health status.

Careful consideration of the choice of instrument is advised, as the top performing instrument varied across the psychometric properties examined, and within some subgroups. The choice of which instrument is best to use may differ depending on the intended use of the data, and the age, gender, report type and type of mental health condition of the population in which the instrument is being used.

1 Introduction

Mental health and substance use disorders are the leading cause of disability in children and adolescents, globally, accounting for a quarter of all years lived with a disability [1]. In Australia, attention-deficit/hyperactivity disorder (ADHD), conduct disorders, anxiety and depression are the most common conditions affecting children aged 4–11 years [2]. It is well recognised that mental health problems early in life have a substantial impact on health-related quality of life (HRQoL) in childhood and adolescence [3, 4]. HRQoL is a multi-dimensional construct that captures the impact of health status on different aspects of physical, social and psychological functioning, either through self- or proxy report [5]. Information on children's HRQoL can be used in research and clinical settings to compare the relative impact of various health conditions and treatments on children's lives; identify groups of children with the greatest need; and aid in clinical decision making and treatment planning [5]. HRQoL has also been endorsed as important to include in an international overall paediatric health standard set (OPH-SS) of outcome measures in children and young people of all ages, and across all health conditions [6]. In addition to its clinical importance, HRQoL information holds significant value in health policy decision making. Most prominently, it is fundamental to the calculation of quality-adjusted life-years (QALYs), which combines measures of the quality and quantity of life into a single metric. Whilst a vast range of instruments exist to measure HRQoL [7], generic preference-weighted [8, 9] measures are preferred in economic analyses due to their ability to generate a single, weighted, summary metric, and facilitate comparisons across different types of conditions. The incorporation of valid and reliable HRQoL information within economic evaluations is a priority to ensure informed decisions are made regarding the comparative value of proposed or existing interventions. Ultimately, this would contribute to a health system that maximises outcomes for children, and maximises the value of children's mental health care. Problematically, very little evidence exists of the validity and reliability of the generic HRQoL instruments that would enable this kind of evaluation in child and adolescent mental health settings.

Perhaps contributing to this dearth of information are the myriad challenges that arise when measuring HRQoL in children and adolescents and specifically in those with mental health challenges. Comprehensive discussions of these issues have been published previously [10–12] and include the reliance on proxy reports of HRQoL in very young children or those unable to self-report; the inconsistency between child self-report and proxy-reported HRQoL; and the vast differences in developmental stage in the ages 0–18 years. Additionally, adult evidence has indicated that commonly used HRQoL instruments such as the EQ-5D and the SF-6D can be valid and reliable in general populations, yet show variable performance across different mental health populations [13, 14], where better performance has been observed in anxiety and depression than schizophrenia or bipolar disorder. In children and adolescents with mental health problems, these measurement complexities are compounded, requiring specific research efforts to determine the validity and reliability of HRQoL instruments in children of different ages, and across different mental health conditions.

A review by Mierau et al. [15] compared 22 generic HRQoL instruments using existing published literature, with the aim of determining suitable instruments for use in economic evaluation of child and adolescent mental health care. The authors concluded that none of the included instruments were 'perfect' for this use, based on each instrument's level of psychometric research/evidence; availability of a proxy version; suitability for young children (<8 years); availability of an age-specific value set for children under 18 years; and degree of focus on mental-health-related domains. However, of the existing instruments, the highest rated on the authors' scale was the CHU9D, followed by the EQ-5D-Y-3L and the PedsQL. However, as the authors noted, there is limited evidence of the validity and reliability of these

instruments, and other commonly used HRQoL instruments, in children and adolescents with mental health challenges.

The limitations of existing evidence are made clear by three recent systematic reviews [16–18] of the psychometric performance of generic preference-weighted HRQoL instruments for children. A review by Kwon et al. [18] provides a summary of existing psychometric evidence; however, performance of the instruments in mental health populations was not reported separately from general populations, leaving performance in children with mental health conditions unclear. Sequential reviews by Rowen et al. [16] and Tan et al. [17] found that only three, and subsequently four, studies had examined instrument performance in children with mental health difficulties. Since publication of these reviews, two further studies have been published in this area. Of the six existing studies, three examined only a single instrument. One study [19] examined convergent validity of the CHU9D in Australian children aged 5–17 years ($n = 200$) receiving mental health services, and another [20] examined construct validity and responsiveness of the CHU9D in Danish children aged 6–15 years ($n = 396$) with emotional or behavioural disturbances. Neither of these studies examined the test–retest reliability of the CHU9D. Another [21] examined acceptability and construct validity of the EQ-5D-Y-5L in a small sample ($n = 52$) of Swedish children aged 13–17 years with mixed mental health conditions. Responsiveness and test–retest reliability were not assessed. From these studies, it remains unclear how the CHU9D and EQ-5D-Y-5L perform comparatively with other generic HRQoL instruments.

Three of the six previous studies conducted a multi-instrument comparison, though each has limitations. Two of these studies were conducted in the United States [22, 23] using an overlapping sample of adolescents aged 13–17 years ($n = 392$) with and without depression. These studies examined the known-group validity [22] and responsiveness [23] of the HUI2/3, the PedsQL, the adult EQ-5D-3L, SF-6D and Quality of Wellbeing Scale (QWB). These studies only examined a limited number of psychometric properties of the measures, and within a single clinical population, leaving much of their psychometric performance unknown in teens with depression, and equally unknown whether instrument performance differs for younger children, or those with different mental health challenges.

Most recently, Mihalopoulos et al. [24] conducted a multi-instrument comparison of paediatric HRQoL instruments in children with mental health challenges in a clinical sample ($n = 426$) of children aged 7–18 years. The sample largely comprised children with internalising disorders (i.e. anxiety or depression, 75.8%), though with some externalising disorders (i.e. ADHD, conduct disorder, oppositional defiant disorder, 15.7%) or trauma/stress disorders (8.5%). This study examined the convergent validity and known-group differences based on severity of mental health of a range of generic paediatric HRQoL instruments (EQ-5D-Y-3L, HUI2/3, CHU9D, AQoL6D, and PedsQL). However, as the study was cross-sectional, the comparative performance of the instruments regarding responsiveness to change in health status, and test–retest reliability was not possible, and these areas remain a significant gap in this literature [18]. None of the existing studies examined whether performance of the HRQoL instruments varied by child sex.

There remains a significant gap in our understanding of the comparative psychometric performance of available generic HRQoL measures that may be suitable for use in children and adolescents with mental health challenges. Assessing the validity and reliability of these instruments in the most common mental health conditions is a priority to ensure the health resources currently being directed to these health conditions are appropriate and are producing the best possible outcomes for these children given the available healthcare funding and resources.

The aim of this study was to address this evidence gap using data from the Australian Paediatric Multi-Instrument Comparison (P-MIC) study—the largest of its kind, internationally. Specifically, we aimed to examine the relative psychometric performance (acceptability, validity, reliability and responsiveness) of a range of commonly used generic paediatric HRQoL instruments in a large sample of children with mental health challenges. We aimed to examine the comparative performance of the PedsQL, EQ-5D-Y-3L, EQ-5D-Y-5L and CHU9D overall, and across subgroups of (i) child age (4–6 years; 7–12 years; 13–18 years); (ii) gender; (iii) type of mental health condition (ADHD; anxiety and/or depression); and (iv) self- or proxy report. Where sample size allows, a secondary aim was to describe the performance of additional instruments (AQoL-6D and HUI3) across the same subgroups.

2 Methods

2.1 Study Design

Data were obtained from the *Quality of Life in Kids: Key evidence to strengthen decisions in Australia* (QUOKKA) Australian P-MIC study (data cut 2, dated 10 August 2022, see published technical methods [25]). The protocol and methods for the P-MIC study have been published in detail elsewhere [26, 27]. Briefly, the study involved the prospective and concurrent collection of a range of generic paediatric HRQoL instruments and condition-specific measures via an online survey. This was followed by a shorter follow-up survey at 4 weeks [25]. The P-MIC study received ethics approval via The Royal Children's Hospital (RCH) Human Research Ethics Committee (HREC/71872/RCHM-2021) and was prospectively registered with the

Australia New Zealand Clinical Trials Registry (ANZCTR) (ACTRN12621000657820). The study findings are reported in line with COSMIN guidelines [28].

2.2 Participants

Participants were a subset of the P-MIC study [26], recruited via an online panel survey company (PureProfile). The current study included the ADHD sample (aged 4–18 years) and the anxiety and/or depression sample (7–18 years). Different age ranges were used for each mental health sample due to the recommended age range for the condition-specific symptom measures relevant for each sample (see Sect. 2.3.2). Participants were included in these condition-specific samples if the caregiver answered 'yes' to the following questions: "Do you have a child aged 7–18 with anxiety or depression as diagnosed by a health professional?" or "Do you have a child aged 4–18 with [ADHD] as diagnosed by a health professional?"

Participants were either the child or their caregiver, as follows. Caregivers completed all sociodemographic measures. Children self-reported the HRQoL instruments if they were aged ≥ 7 years and were deemed by their caregiver to be capable of completing the survey. Alternatively, the caregiver completed all measures on behalf of the child (proxy report) if the child was (i) aged < 7 years, or (ii) aged ≥ 7 years but the caregiver deemed the child not able to complete the survey. Mental-health-specific measures were completed by the caregiver or child in the same way, with the exception of the ADHD-symptom measure, which only has a proxy report version available in this age group. Caregivers provided consent for their own participation and for their child, with additional consent provided by the child in instances of child self-report.

2.3 Instruments and Measures

2.3.1 Sociodemographic Measures

Child age (in years), was used to form three age bands for subgroup analyses: 4–6 years; 7–12 years; and 13–18 years. These age bands were chosen to distinguish between the different age groups available for each mental health sample (ADHD: 4–18 years; anxiety/depression: 7–18 years); to align with the first instance of child self-report in this study (≥ 7 years); and to broadly align with the major developmental stages of childhood and adolescence (i.e. pre-school; primary school; and high school).

Child gender was used to describe the sample and to determine individual symptom severity cut points on mental-health-specific instruments as per published norms (see Sect. 2.3.2). Available published norms do not specify appropriate cut points for children and adolescents who identify as transgender, non-binary, gender fluid or those of undisclosed gender, and we did not feel it was appropriate to apply gendered norms in these instances. While these children ($n = 16$) are included in the wider sample, they are not included in the known-group analyses that require the use of these symptom severity cut points.

Special health care needs (SHCN) was indicated based on the parents' response to a SHCN screening question: "*Child has a condition which has lasted or is expected to last for at least 12 months which causes them to use medicine prescribed by a doctor (other than vitamins) or more medical care, mental health or educational services. Yes/no*" [29]. Additional sociodemographic data were used solely to describe the sample, as shown in Table 1.

2.3.2 Mental Health and HRQoL Instruments and Cut Points

Details of mental health and HRQoL instruments are available in Table 2. Symptom severity cut points were calculated for the mental health symptom measures, as follows, to facilitate known-group validity testing.

Strengths and Difficulties Questionnaire (SDQ) The SDQ is a validated screening questionnaire used to assess a child's emotional and behavioural wellbeing [30, 31]. All participants completed the SDQ at baseline. Australian norms exist for 4–6 year olds [32] and 7–18 year olds [33] to classify children based on severity. The total score was used to determine symptom severity cut points individually for each child based on their age, gender, and self/proxy report, though were generally as follows: low level (< 12/40); borderline/query (12–16/40) or abnormal/of concern (17+/40).

Revised Children's Anxiety and Depression Scale, Short form (RCADS-25) All participants within (only) the anxiety/depression sample completed the RCADS-25 at baseline. From the raw sum scores for the instrument, total t-scores were calculated, adjusted for child gender, age and respondent using available syntax based on United States population norms [34]. These *T* scores were then used to classify children into three symptom levels: 'normal' (< 65); 'borderline' (65–69); and 'clinical' (≥ 70) [34].

Strengths and Weaknesses of Attention-Deficit/Hyperactivity Disorder Symptoms and Normal Behavior Scale (SWAN) All participants within (only) the ADHD sample completed the SWAN at baseline. *Z* scores were calculated from the instrument raw sum scores, adjusted for child gender and age. We used a known-group cut point, based on a previously published threshold, of $z > 0.74$ [35]. This cut point has previously shown good sensitivity and specificity (AUC = 0.85–0.88) in a large general population sample of children

Table 1 Sample characteristics

Characteristics	Combined sample *n* (%)	Subgroups of interest								
		Mental health condition		Age band			Gender		Report type	
		ADHD *n* (%)	Anx/Dep *n* (%)	4–6 y *n* (%)	7–12 y *n* (%)	13–18 y *n* (%)	Male *n* (%)	Female *n* (%)	Self *n* (%)	Proxy *n* (%)
Total completed initial survey	1013	533	480	132	444	437	566	431	689	324
Total completed follow-up survey at 4 weeks	284 (28.0)	128 (24.0)	156 (32.5)	30 (22.7)	127 (28.6)	127 (29.1)	154 (27.2)	128 (29.7)	202 (29.3)	82 (25.3)
Survey respondent										
Child self-report	689 (68.0)	304 (57.0)	385 (80.2)	*N/A*[a]	362 (81.5)	327 (74.8)	373 (65.9)	307 (71.2)		
Parent proxy report	324 (32.0)	229 (43.0)	95 (19.8)	132 (100)	82 (18.4)	110 (25.1)	193 (43.1)	124 (28.8)		
Child characteristics										
Age, mean (SD)	11.5 (4.1)	10.0 (4.0)	13.2 (3.4)	5 (0.8)	9.4 (1.6)	15.5 (1.6)	11.0 (4.1)	12.0 (4.0)	12.3 (3.3)	9.7 (4.8)
4–6 years	132 (13.0)	132 (24.8)	*N/A*[b]				91 (16.1)	41 (9.5)	*N/A*[a]	132 (40.7)
7–12 years	444 (43.8)	248 (46.5)	196 (40.8)				259 (45.8)	183 (42.5)	362 (52.5)	82 (25.3)
13–18 years	437 (43.1)	153 (28.7)	284 (59.2)				216 (38.2)	207 (48.0)	327 (47.4)	110 (33.9)
Gender										
Male	566 (55.9)	354 (66.4)	212 (44.2)	91 (68.9)	259 (58.3)	216 (49.4)			373 (54.1)	193 (59.5)
Female	431 (42.5)	172 (32.3)	259 (54.0)	41 (31.0)	183 (41.2)	207 (47.3)			307 (44.5)	124 (38.2)
Transgender female	1 (0.1)	1 (0.2)	0	0	0	1 (0.2)			0	1 (0.3)
Transgender male	6 (0.6)	4 (0.8)	2 (0.4)	0	1 (0.2)	5 (1.1)			3 (0.4)	3 (0.9)
Non-binary/gender fluid	6 (0.6)	1 (0.2)	5 (1.0)	0	1 (0.2)	5 (1.1)			4 (0.5)	2 (0.6)
Prefer not to answer	3 (0.3)	1 (0.2)	2 (0.4)	0	0	3 (0.6)			2 (0.2)	1 (0.3)
Aboriginal and/or Torres Strait Islander descent, *yes*	97 (9.6)	56 (10.5)	41 (8.5)	19 (14.3)	42 (9.4)	36 (8.2)	55 (9.7)	41 (9.5)	61 (8.8)	36 (11.1)
Language other than English spoken at home, *yes*	39 (3.8)	22 (4.1)	17 (3.5)	6 (4.5)	19 (4.2)	14 (3.2)	24 (4.2)	15 (3.5)	29 (4.2)	10 (3.0)
Mental and physical health comorbidities										
Mental health symptom severity										
SDQ, total score, *mean (SD)*	18.2 (6.3)	19.4 (6.3)	16.9 (6.0)	20.2 (5.9)	18.6 (6.3)	17.3 (6.3)	18.1 (6.5)	18.4 (6.1)	17.9 (6.3)	19.1 (6.2)
'low'	211 (21.1)	87 (16.5)	124 (26.3)	10 (7.5)	93 (21.0)	108 (25.5)	148 (26.1)	63 (14.6)	167 (24.5)	44 (13.8)
'borderline/query'	179 (17.9)	72 (13.6)	107 (22.7)	9 (6.8)	88 (19.9)	82 (19.3)	97 (17.1)	82 (19.0)	125 (18.3)	54 (17.0)
'abnormal/of concern'	607 (60.8)	367 (69.7)	240 (50.9)	113 (85.6)	261 (59.0)	233 (55.0)	321 (56.7)	286 (66.3)	388 (57.0)	219 (69.0)
RCADS-25, T-score, *mean (SD)*	55.8 (16.5)	*N/A*[c]	55.8 (16.5)	*N/A*[c]	50.6 (14.2)	59.4 (17.1)	56.5 (18.4)	55.2 (14.8)	53.3 (15.2)	66.2 (17.7)
'Normal'	347 (73.6)	*N/A*[c]	347 (73.6)	*N/A*[c]	171 (87.2)	176 (64.0)	152 (71.7)	195 (75.2)	304 (80.0)	43 (47.2)
'Borderline'	34 (7.2)	*N/A*[c]	34 (7.2)	*N/A*[c]	9 (4.5)	25 (9.0)	15 (7.0)	19 (7.3)	23 (6.0)	11 (12.0)
'Clinical'	90 (19.1)	*N/A*[c]	90 (19.1)	*N/A*[c]	16 (8.1)	74 (26.9)	45 (21.2)	45 (17.3)	53 (13.9)	37 (40.6)
SWAN, raw total score, *mean (SD)*	16.3 (17.1)	16.3 (17.1)	*N/A*[d]	17.0 (18.7)	18.7 (15.4)	11.6 (17.3)	17.2 (17.4)	14.2 (16.3)	15.6 (17.2)	17.1 (16.8)
'Low'	401 (76.2)	401 (76.2)	*N/A*[d]	96 (72.7)	187 (76.0)	118 (79.7)	266 (75.1)	135 (78.4)	231 (77.0)	170 (75.2)
High'	125 (23.7)	125 (23.7)	*N/A*[d]	36 (27.2)	59 (23.9)	30 (20.2)	88 (24.8)	37 (21.5)	69 (23.0)	56 (24.7)

Table 1 (continued)

Characteristics	Combined sample *n* (%)	Subgroups of interest								
		Mental health condition		Age band			Gender		Report type	
		ADHD *n* (%)	Anx/Dep *n* (%)	4–6 y *n* (%)	7–12 y *n* (%)	13–18 y *n* (%)	Male *n* (%)	Female *n* (%)	Self *n* (%)	Proxy *n* (%)
Mental health condition(s)										
Anxiety	619 (61.1)	182 (34.2)	437 (91.0)	28 (21.2)	271 (61.0)	320 (73.2)	295 (52.1)	311 (72.2)	469 (68.0)	150 (46.3)
ADHD	569 (56.2)	533 (100)	36 (7.5)	132 (100)	263 (59.2)	174 (39.8)	374 (66.1)	186 (43.2)	332 (48.1)	237 (73.1)
Autism spectrum disorder	78 (7.7)	54 (10.1)	24 (5.0)	16 (12.1)	40 (9.0)	22 (5.0)	46 (8.1)	28 (6.5)	44 (6.3)	34 (10.4)
Behavioural, cognitive, emotional	279 (27.5)	178 (33.4)	101 (21.0)	49 (37.1)	139 (31.3)	91 (20.8)	161 (28.4)	112 (26.0)	178 (25.8)	101 (31.1)
Depression	241 (23.8)	58 (10.9)	183 (38.1)	2 (1.5)	60 (13.5)	179 (40.9)	105 (18.6)	126 (29.2)	190 (27.5)	51 (15.7)
Developmental delay	75 (7.4)	62 (11.6)	13 (2.7)	22 (16.6)	39 (8.7)	14 (3.2)	52 (9.2)	22 (5.1)	41 (5.9)	34 (10.4)
Eating disorder	24 (2.4)	9 (1.7)	15 (3.1)	4 (3.0)	6 (1.3)	14 (3.2)	10 (1.8)	12 (2.8)	15 (2.1)	9 (2.7)
Problems with psych development	34 (3.4)	26 (4.9)	8 (1.7)	8 (6.0)	17 (3.8)	9 (2.0)	22 (3.9)	11 (2.6)	21 (3.0)	13 (4.0)
Chronic physical health condition(s)[e], *yes*	433 (42.7)	209 (39.2)	224 (46.6)	47 (35.6)	197 (44.3)	189 (43.2)	224 (39.5)	200 (46.4)	313 (45.4)	120 (37.0)
Special healthcare needs[f], *yes*	606 (59.8)	355 (66.6)	251 (52.2)	77 (58.3)	257 (57.8)	272 (62.2)	350 (61.8)	243 (56.3)	411 (59.6)	195 (60.1)
Family characteristics										
Caregiver age, mean (SD)	40.2 (8.9)	38.2 (8.5)	42.5 (8.8)	33.5 (6.9)	37.4 (7.6)	45.0 (8.0)	39.5 (8.6)	41.0 (9.1)	40.9 (8.6)	38.5 (9.1)
Caregiver gender, female	829 (81.8)	430 (80.7)	399 (83.1)	103 (78.0)	371 (83.5)	355 (81.2)	448 (79.2)	370 (85.9)	563 (81.7)	266 (82.1)
Parent education, 'Bachelor or above', *yes*	302 (29.8)	154 (28.9)	148 (30.8)	39 (29.5)	126 (28.3)	137 (31.3)	167 (29.5)	130 (30.2)	199 (28.8)	103 (31.7)
SEIFA IRSAD[g], mean (SD)	992 (70.7)	989 (66.9)	946 (74.7)	985 (67.6)	990 (67.6)	995 (74.5)	993 (73.8)	989 (66.7)	993 (72.4)	989 (66.7)
Resides in major city of Australia[h], *yes*	709 (70.0)	372 (69.8)	337 (70.2)	99 (75.0)	306 (68.9)	304 (69.5)	396 (70.0)	300 (69.6)	486 (70.5)	223 (68.8)

ADHD attention-deficit/hyperactivity disorder, *Anx/Dep* anxiety and/or depression, *IRSAD* Index of Relative Socioeconomic Advantage and Disadvantage, *RCADS-25* Revised Children's Anxiety and Depression Scale Short Form, *SDQ* Strengths and Difficulties Questionnaire, *SEIFA* Socioeconomic Indexes for Area, *SWAN* Strengths and Weaknesses of Attention-Deficit/Hyperactivity Disorder Symptoms and Normal Behavior Scale

[a]All instruments for the 4- to 6-year-olds were completed by parent proxy report; no child self-report data are available for this age band

[b]The anxiety/depression sample was limited to children aged 7–18 years; data for 4- to 6-year-olds are not available for this sample

[c]The RCADS-25 instrument was only completed by the anxiety/depression sample; no RCADS-25 data are available for the ADHD sample or the 4- to 6-year-old age band (as no children aged 4–6 years were recruited into the anxiety/depression sample)

[d]The SWAN instrument was only completed by the ADHD sample; no SWAN data are available for the anxiety/depression sample

[e]Includes parent report of one or more of the following conditions: asthma; bone, joint or muscle problems; chronic fatigue; diabetes mellitus; epilepsy/seizures; vision problems; recurrent pain (abdominal, headaches, chest, back or other parts of the body); problems with hearing; overweight/obese; or physical disability

[f]Classified yes/no based on the special healthcare needs (SHCN) screener [29]

[g]Socioeconomic measure based on home postcode, as per Socioeconomic Indexes for Area (SEIFA) Index of Relative Socioeconomic Advantage and Disadvantage (IRSAD) [43]. This index has a national mean of 1000 and SD of 100

[h]Classification based on the Australian Statistical Geography Standard levels of remoteness [44]

Table 2 Mental health and health-related quality of life instruments

Instrument (no. of items)	Recommended age	Domains	Response format	Total score possible range	Recall period	General description
Mental Health Symptom Measures						
Strengths and Difficulties Questionnaire **(SDQ)** (25 items)	2–17 years	Internalising subscale: 1. emotional symptoms 2. peer problems Externalising subscale: 3. conduct problems 4. hyperactivity/inattention [*Not included in total score*]: 5. prosocial behaviour	0 'Not true' 1 'Somewhat true' 2 'Certainly true'	0 (low symptoms) to 40 (high symptoms)	Past 6 months	The SDQ is a validated screening questionnaire used to assess a child's emotional and behavioural wellbeing [30, 31]. Developmentally appropriate self- and proxy report versions of the SDQ exist for children aged 2–17 years. Higher scores represent more problems
Revised Children's Anxiety and Depression Scale, Short form **(RCADS-25)** (25 items)	7–18 years	Total anxiety subscale: 1. generalised anxiety disorder 2. obsessive compulsive disorder 3. panic disorder 4. separation anxiety disorder 5. social phobia Total depression subscale: 6. major depressive disorder	0 'Never' 1 'Sometimes' 2 'Often' 3 'Always'	0 (low symptoms) to 75 (high symptoms)	Not explicitly stated	The RCADS-25 was completed by participants in the anxiety/depression sample, by self- or proxy report. It is a symptom-based scale measuring anxiety and depression in children [45]. The RCADS-25 has been validated in children aged 7–18 years [46]. Higher scores represent greater symptoms of anxiety and depression
Strengths and Weaknesses of Attention-Deficit/Hyperactivity Disorder Symptoms and Normal Behavior Scale **(SWAN)** (18 items)	6–18 years	Inattentive subscale (9 items) Hyperactive/impulsive subscale (9 items)	−3 'Far above average' −2 'Above average' −1 'Somewhat above average' 0 'Average' +1 'Somewhat below average' +2 'Below average' +3 'Far below average'	−54 (low symptoms) to +54 (high symptoms)	Past 1 month	The SWAN [36] was completed by participants in the ADHD sample, by proxy report. It is an ADHD symptom scale based on the ADHD criteria from the Diagnostic and Statistical Manual of Mental Disorders, 4th edition (DSM-IV [47]). Strengths are scored as negative and weaknesses are scored as positive (i.e. higher scores represent greater ADHD traits)

Table 2 (continued)

Instrument (no. of items)	Recommended age	Domains	Response format	Total score possible range	Recall period	General description
Health-Related Quality of Life Instruments						
Pediatric Quality of Life Inventory **(PedsQL)** (23 items)	2–18 years	1. Physical functioning 2. Emotional functioning 3. Social functioning 4. School functioning	0 'Never' 1 'Almost never' 2 'Sometimes' 3 'Often' 4 'Almost always'	0 (*worst* health) to 100 (*best* health)	Past 1 month	The **PedsQL** is a generic HRQoL instrument [48] with validated versions available for children aged 2–18 years, and self-report available for children ≥8 years [49]. Items were reverse scored and linearly transformed to a 0–100 scale, and a total score was calculated as the sum of each item divided by the number of items answered at each time point, provided less than half of the items were missing [49]
EQ-5D-Y-3L (5 items)	4–18 years	1. Mobility 2. Self-care 3. Usual activities 4. Pain/discomfort 5. Anxiety/depression	1 'No problems' 2 'Some problems' 3 'A lot of problems'	5 (best health) to 15 (worst health)	Today	The **EQ-5D-Y-3L** [50] and recently developed **EQ-5D-Y-5L** [51] are youth versions adapted from the EQ-5D adult version. Both the EQ-5D-Y-3L and EQ-5D-Y-5L are generic, preference-weighted HRQoL instruments differentiated only by the response options on each instrument (either a 3-point or 5-point response scale). Validated versions of the EQ-5D-Y exist for children aged 4–18 years. Where parent proxy report was provided, this was using the Proxy 1 version, i.e. the parent provided a rating of the child's HRQoL in their (the parent's) opinion
EQ-5D-Y-5L (5 items)	4–18 years	1. Mobility 2. Self-care 3. Usual activities 4. Pain/discomfort 5. Anxiety/depression	1 'No problems' 2 'A little bit of a problem' 3 'Some problems' 4 'A lot of problems' 5 'Unable to/ extreme'	5 (best health) to 25 (worst health)	Today	*[As per EQ-5D-Y-3L description above]*

Table 2 (continued)

Instrument (no. of items)	Recommended age	Domains	Response format	Total score possible range	Recall period	General description
Child Health Utility Instrument (**CHU9D**) (9 items)	6–17 years	1. Worried 2. Sad 3. Pain 4. Tired 5. Annoyed 6. Schoolwork 7. Sleep 8. Daily routine 9. Activities	1 'No problems' 2 'A few problems' 3 'Some problems' 4 'Many problems' 5 'Can't do/extreme'	9 (best health) to 45 (worst health)	Today/last night	The CHU9D is a generic, preference-based instrument developed specifically for the purpose of measuring HRQoL in children and young people [52]. It was initially developed for children aged 7–11 years, however, it has been validated in children up to 17 years [53, 54]
Assessment of Quality of Life (**AQoL-6D**) Adolescent (20 items)	11–18 years	1. Independent living 2. Relationships 3. Mental health 4. Coping 5. Pain 6. Senses	[*Varies for each of the 20 questions, ranging from 4 to 6 response levels]*	20 (best health)to 99 (worst health)	Past week	The AQoL-6D Adolescent is a generic, preference-based HRQoL instrument [55]. It was developed with adolescents aged 12–18 years, however, has been used in children aged 11 years [56]
Health Utilities Index Mark 2/3 (**HUI 2/3**) (15 items)	5–18 years	HUI 3 classification system 1. Vision 2. Hearing 3. Speech 4. Ambulation 5. Dexterity 6. Emotion 7. Cognition 8. Pain	Varies for each domain, as follows: 1. 6 levels 2. 6 levels 3. 5 levels 4. 6 levels 5. 6 levels 6. 5 levels 7. 6 levels 8. 5 levels	8 (best health) to 45 (worst health)	'Usual' ability/level	The HUI 2/3 is a single instrument that can be scored according to the HUI2 or HUI3 classification system [57]. We used the HUI3 classification system for all analyses except for the feasibility and acceptability analysis which is based on the completion of the 15-item HUI 2/3 instrument. The HUI 2/3 is recommended for use in children ≥5 years, with self-report recommended for children ≥8 years

aged 6–17 years ($n = 15{,}560$) [35]. This cut point was subsequently validated in a clinical sample [35], and found to be a more accurate classification than the cut point of 1.65 SD recommended by Swanson et al. [36]. As the sample of children used in this analysis only includes children with ADHD, the mean score will be much higher than that of a general population or mixed clinical sample [35]. Hence this standardised scoring method is used as a symptom severity cut point to identify the top 25% most severe cases from our sample.

2.4 Instrument Completion

In response to patient feedback aiming to reduce response burden, not all respondents were offered all instruments. For detail on how instruments were chosen as the 'main' or 'additional' instruments for assessment within the QUOKKA study, please see Jones et al. (within this Special Supplement – *reference to be updated in editing stages*) and the QUOKKA Study Technical Methods Guide [25]. All participants aged 5–18 years completed all four main HRQoL instruments (PedsQL, EQ-5D-Y-3L, EQ-5D-Y-5L and CHU9D). Participants aged 4 years completed the PedsQL and the CHU9D, but were randomised to complete *either* the EQ-5D-Y-3L *or* the EQ-5D-Y-5L. All participants were then randomised to complete *either* the AQoL-6D *or* the HUI3 in addition to the four main instruments. All participants completed the SDQ and the mental health symptom measure relevant to their sample (i.e. *either* the RCADS-25 *or* the SWAN). See published technical methods [25] for further details on instrument randomisation.

2.5 Psychometric Analyses

See Table 3 for a description of each of the psychometric analyses performed; relevant thresholds for interpreting results; and a priori hypotheses. Analyses encompassed feasibility and acceptability; floor and ceiling effects; known-group validity; convergent validity; responsiveness; and test–retest reliability. All statistical analyses were conducted using StataSE 16 (Statacorp, Texas, US). Statistical methods, subgroups and thresholds for interpretation were prespecified and are reported in a statistical analysis plan which is available in the technical methods paper [25].

2.6 Subgroup and Sensitivity Analyses

All validity, reliability and responsiveness analyses described in Table 3 were assessed using the combined sample, and within the following subgroups: (i) mental health condition (ADHD; anxiety and/or depression); (ii) age band (4–6 years; 7–12 years; 13–18 years); (iii) gender (male; female); and (iv) report type (self/proxy report). Note that subgroup analyses for gender are male/female only due to low sample size ($n = 16$) for transgender/non-binary/gender fluid/undisclosed children.

The preference-weighted HRQoL instruments were designed to be scored using preference weights to give a 'utility score' (ranging from 0 to 1) and are predominantly used in this way. However, in the absence of utility weights, for the purpose of preliminary assessment of psychometric properties, the instruments can be scored by summing the response score for each item to give a 'level sum score' (LSS) [9]. The LSS is the total score with equal weight for each item (e.g. for a child reporting no problems on the EQ-5D-Y-3L, this would be $1 + 1 + 1 + 1 + 1 = 5$; see Table 2 for possible total sum score range for each instrument). Given the development of preference weights is still underway for the PedsQL and EQ-5D-Y-5L, to allow a comparison across all HRQoL instruments, our analyses use individual items or the instrument LSS. While this LSS may be useful in clinical settings, for descriptive systems it has limitations linked to the interpretation of equal sum scores that can be derived from quite different combinations of responses. Furthermore, preference elicitation studies have shown that in practice respondents place different importance on each item within HRQoL instruments. Despite these limitations, the LSS approach has recently been shown to form a strong Mokken scale [37] and was deemed by Feng et al. to be a meaningful measurement, particularly in samples with health conditions. Given the different interpretations arising from these two scoring approaches, a sensitivity analysis was performed that repeated the psychometric analyses—where appropriate—using utility scores as the outcome variable. This analysis was performed for instruments that have a currently available value set (i.e. the EQ-5D-Y-3L, CHU9D, AQoL-6D and HUI3). Further sensitivity analysis methods and results, including the value sets used for these analyses, are available in Supplementary Table S10 (see electronic supplementary material [ESM]).

2.7 Instrument Comparison

A summary of results was undertaken to compare the relative performance of the PedsQL, EQ-5D-Y-3L, EQ-5D-Y-5L and CHU9D across the psychometric properties, within the combined sample and all subgroups. Data for the AQoL-6D and the HUI3 are presented in the ESM only due to a lower total sample size for these instruments which precludes a direct comparison of results. For details on how to interpret each cell, see footnotes for Table 4. A total performance score was calculated for each instrument, which weights each psychometric property equally; however, individual section scores can be referred to when a combined equally weighted summary is not considered helpful.

Table 3 Description of psychometric analyses

Psychometric property	Description of analysis
Feasibility and acceptability	Self-reported difficulty of completing each instrument was measured following each instrument, as rated on a 5-point scale from 1 'very difficult' to 5 'very easy'. The proportion of respondents was calculated for each response category for the total sample, and for each subgroup of interest. The proportion of participants who found each measure 'somewhat' or 'very' easy to complete is reported as an indicator of good performance
Floor and ceiling effects	Floor and ceiling effects refer to the proportion of respondents who report being in the poorest possible health state (e.g. 5,5,5,5,5 on the EQ-5D-Y-5L) ('floor effects') or the best possible health state (e.g. 1,1,1,1,1 on the EQ-5D-Y-5L) ('ceiling effects') as measured by each instrument. An instrument was considered to have a significant floor/ceiling effect if more than 15% of respondents reported being in the poorest/best possible health state, respectively [58, 59]. Additionally, we visually inspected the distribution of responses for each item on each measure for the total sample, and by subgroups of interest
Construct validity	Construct validity refers to the degree to which scores on each HRQoL instrument are consistent with hypotheses based on the assumption that the instrument validly measures the construct to be measured [60]. There is no gold standard for measuring HRQoL in mental health; the diversity of the types of mental health conditions that exist means it is difficult for a generic instrument to capture the range of ways that the mental health condition might affect a person's life [14]. However, if the instruments are in fact valid, we would expect that instruments aiming to measure HRQoL should be correlated with one another and—to some degree—with measures of mental health, as all include at least one domain related to emotional/psychosocial functioning. We formally tested the construct validity of each instrument through the known-group and convergent validity tests described below
Known-group validity	Known-group validity refers to the degree to which the instruments can distinguish between a priori determined known groups. The ability of each instrument to detect group differences was assessed by comparing the mean total sum score (i.e. mean HRQoL) for each known group. Effect sizes were estimated using Cohen's D, where effect sizes of 0.2–0.49 were considered small, 0.5–0.79 moderate, and ≥ 0.8 large [61]. Group differences were examined using a conservative threshold of $p < 0.01$ due to the large number of comparisons. The sample size for each analysis was considered in line with the COSMIN guidelines [40], where: $n \geq 100$ per group is very good; $n = 50$–99 per group is adequate; $n = 30$–49 per group is doubtful; $n < 30$ per group is inadequate. Results for all analyses are presented in ESM Table S6, however, only results based on 'very good' or 'adequate' sample sizes are interpreted in text Group differences were assessed for the following a priori determined groups: (i) Published symptom severity cut points on the SDQ, as detailed in Methods section. Assessed in both mental health samples (ii) Published symptom severity cut points on the SWAN, as detailed in Methods section. Assessed in ADHD sample only (iii) Published symptom severity cut points on the RCADS-25, as detailed in Methods section. Assessed in anxiety/depression sample only; and (iv) Whether the child has special health care needs (SHCN), (yes/no), based on a published special healthcare needs screener [29]
Convergent validity	In the absence of a 'gold standard' HRQoL instrument to compare with in this population, convergent validity is assessed by the degree to which the instruments are correlated with one another following an expected pattern, and the degree to which each instrument is correlated with gold standard mental health symptom measures. Correlations were assessed using Spearman's rank correlation coefficient (rho, ρ) as the response data were not normally distributed. Correlations of 0.1–0.29 were considered weak, 0.3–0.49 moderate, and ≥ 0.5 strong [61]. The sample size for each analysis was considered in line with the COSMIN guidelines [40], where: $n \geq 100$ is very good; $n = 50$–99 is adequate; $n = 30$–49 is doubtful; $n < 30$ is inadequate. Results for all analyses are presented in ESM Table S7, however, only results based on 'very good' or 'adequate' sample sizes are interpreted in text. Hypothesised correlations were set a priori, as follows: (i) Total sum scores on each HRQoL instrument, in all combinations, were hypothesised to be strongly correlated, as all are measuring the construct of 'HRQoL'; and (ii) Total sum scores on each HRQoL instrument were hypothesised to be at least moderately correlated with total sum scores on each mental-health-specific instrument, as they are measuring related but distinct constructs (i.e. mental health and HRQoL) and all HRQoL instruments have at least one mental health item to capture this area of health

Table 3 (continued)

Psychometric property	Description of analysis
Responsiveness	Responsiveness refers to the ability of an instrument to detect a change in health status. We assessed the responsiveness of the HRQoL instruments by examining the magnitude of change in total sum scores on each HRQoL instrument between the initial survey and the follow-up survey administered at approximately 4 weeks. Change in health was assessed in two ways: (i) for children who reported a change in their general health; and (ii) those who reported a change in their mental health condition (i.e. ADHD or anxiety/depression). Respondents reported these changes in response to two separate questions, "How would you rate the Study Child's health in general now?" and "In relation to the child's {condition group}, how would you say this is going now compared to when you completed the first survey?" Both questions were rated on a 5-point scale: 1 'Much better', 2 'Somewhat better', 3 'About the same', 4 'Somewhat worse', 5 'Much worse' Each HRQoL instrument was determined to be responsive to change if a significant mean difference was detected for those who reported either an improvement (i.e. somewhat or much better) or a deterioration (i.e. somewhat or much worse), related to their general health or their main mental health condition. Responsiveness was thus assessed in a 2 × 2 pattern, resulting in four analyses for each instrument: (1) a change in general health for (a) the better; (b) the worse; and (2) a change in health related to the main mental health condition for (a) the better; (b) the worse Statistically significant ($p < 0.01$) mean differences were examined using t-tests. The magnitude of the change was assessed using standardised response means (SRM), calculated by dividing the mean difference on the instrument by the standard deviation of the change. An SRM of 0.2–0.49 is considered small, 0.5–0.79 is moderate, and ≥ 0.8 is large [62]. The sample size for each analysis was considered in line with the COSMIN guidelines [40], where $n \geq 100$ is very good; $n = 50$–99 is adequate; $n = 30$–49 is doubtful; $n < 30$ is inadequate. Results for all analyses are presented in ESM Table S8, however, only results based on 'very good' or 'adequate' sample sizes are interpreted in text
Test–retest reliability	Test–retest reliability refers to the extent to which scores on each HRQoL instrument were consistent across repeated measurement, i.e. when no change was reported in the child's general health and mental health condition at the 4-week follow-up survey. The sample size for each analysis was considered in line with the COSMIN guidelines [40], where $n \geq 100$ is very good; $n = 50$–99 is adequate; $n = 30$–49 is doubtful; $n < 30$ is inadequate. Results for all analyses are presented in ESM Table S9, however, only results based on 'very good' or 'adequate' sample sizes are interpreted in text The intraclass correlation coefficient (ICC) is the most suitable and most commonly used reliability parameter for continuous measures [58]. The ICC and 95% CIs were calculated using a two-way mixed-effects model for each HRQoL instrument, based on absolute agreement [39, 58]. We acknowledge there are no accepted thresholds for interpreting ICC results, therefore ICC values were interpreted using two separate recommended thresholds as follows: (i) *Using thresholds recommended by Cicchetti et al.* [38]: ICC values of < 0.4 indicate poor agreement; 0.40–0.59 fair agreement; 0.60–0.74 good agreement; and ≥0.75 excellent agreement (ii) *Using thresholds recommended by Koo et al.* [39]: ICC values of < 0.5 indicate poor agreement; 0.50–0.74 moderate agreement; 0.75–0.90 good agreement; and > 0.90 excellent agreement

ESM electronic supplementary material, *HRQoL* health-related quality of life

3 Results

3.1 Sample Characteristics

Survey data were collected for a total of $n = 1013$ children and adolescents aged 4–18 years (mean age 11.5 years, SD 4.1). Surveys were completed largely by child self-report ($n = 689$, 68.0%), and approximately one-third by parent-proxy report ($n = 324$, 32.0%). A total of $n = 533$ children were included in the ADHD sample and $n = 480$ in the anxiety/depression sample. The sample included slightly more boys ($n = 566$, 55.9%) than girls ($n = 431$, 42.5%). The response rate for the follow-up survey at approximately 4 weeks was 28.0% ($n = 284$ completed surveys). A subset of participants completed the AQoL-6D ($n = 330$) or the HUI3 ($n = 370$). Sample characteristics for these additional instruments were similar to the total combined sample who completed all other instruments: age ($m = 11.8$, SD = 3.9, $m = 10.8$, SD = 4.4, respectively); self-report (69.7%, 62.1%, respectively); sample (ADHD 51.9%, anx/dep 48.1%; ADHD 56.0%, anx/dep 44.0%, respectively); gender (male 59.3%, male 54.0%, respectively). Sample characteristics for the total sample are described in Table 1.

The number of respondents who completed each instrument at baseline and follow-up, for the combined sample, and within each subgroup, is described in ESM Table S1. Partial completion was not permitted by the survey platform, such that no submitted surveys had missing data. Baseline clinical and demographic characteristics between those who did or did not complete the follow-up showed no significant differences, with the exception that a greater proportion of children were from the anxiety/depression sample at follow-up ($n = 156$, 54.9%) compared with baseline ($n = 324$, 44.4%; $p = 0.003$). See ESM Table S2.

Table 4 Summary of psychometric performance of each health-related quality of life (HRQoL) instrument for the combined sample and by subgroup

Psychometric properties	PedsQL (23 items)	EQ-5D-Y-3L (5 items)	EQ-5D-Y-5L (5 items)	CHU9D (9 items)
ACCEPTABILITY / FEASIBILITY				
Combined Sample	✔ 68.0%	✔ 70.6%	✔ 72.5%	✔ 71.2%
ADHD Sample	✔ 68.7%	✔ 69.1%	✔ 73.6%	✔ 70.4%
Anx / Dep Sample	✔ 67.3%	✔ 70.8%	✔ 71.3%	✔ 72.1%
4-6 Year Olds	✔ 69.7%	✔ 68.5%	✔ 73.2%	✔ 71.2%
7-12 Year Olds	✔ 69.1%	✔ 72.7%	✔ 76.6%	✔ 75.2%
13-18 Year Olds	✔ 66.4%	✔ 67.5%	✔ 68.2%	✔ 67.0%
Males	✔ 69.1%	✔ 71.2%	✔ 75.0%	✔ 72.1%
Females	✔ 66.6%	✔ 68.4%	✔ 69.2%	✔ 70.1%
Self-report	✔ 66.9%	✔ 69.5%	✔ 72.1%	✔ 71.1%
Proxy-report	✔ 70.4%	✔ 71.0%	✔ 73.4%	✔ 71.3%
Subtotal score	**68%**	**70%**	**72%**	**71%**
CEILING EFFECTS				
Combined Sample	✔ 0.3%	✘ 17.9%	✔ 14.7%	✔ 3.1%
ADHD Sample	✔ 0.3%	✘ 22.0%	✘ 18.5%	✔ 3.5%
Anx / Dep Sample	✔ 0.2%	✔ 13.5%	✔ 10.6%	✔ 2.7%
4-6 Year Olds	✔ 0.0%	✘ 18.0%	✘ 16.0%	✔ 3.7%
7-12 Year Olds	✔ 0.2%	✘ 17.1%	✔ 14.1%	✔ 2.4%
13-18 Year Olds	✔ 0.4%	✘ 18.7%	✔ 14.8%	✔ 3.6%
Males	✔ 0.3%	✘ 20.6%	✘ 15.9%	✔ 3.7%
Females	✔ 0.2%	✔ 14.1%	✔ 13.4%	✔ 2.5%
Self-report	✔ 0.4%	✘ 16.9%	✔ 13.5%	✔ 3.0%
Proxy-report	✔ 0.0%	✘ 20.1%	✘ 17.3%	✔ 3.4%
Subtotal score	**100%**	**20%**	**60%**	**100%**
KNOWN GROUPS				
Combined Sample	✔ 4/4	✔ 4/4	✔ 4/4	✔ 4/4
ADHD Sample	✔ 3/3	✔ 3/3	✔ 3/3	✔ 3/3
Anx / Dep Sample	✔ 3/3	✔ 3/3	✔ 3/3	✔ 3/3
7-12 Year Olds	✔ 3/3	✔ 3/3	✔ 3/3	✔ 3/3
13-18 Year Olds	✔ 3/3	✔ 3/3	✔ 3/3	✔ 3/3
Males	✔ 3/3	✔ 3/3	✔ 3/3	✔ 2/3
Females	✔ 2/2	✔ 2/2	✔ 2/2	✔ 2/2
Self-report	✔ 4/4	✔ 4/4	✔ 3/4	✔ 4/4
Proxy-report	✔ 2/3	✘ 1/2	✔ 2/3	✔ 2/3
Subtotal score	**96%**	**96%**	**93%**	**93%**
CONVERGENT VALIDITY				
Combined Sample	✔ 7/8	✔ 7/8	✔ 7/8	✔ 7/8
ADHD Sample	✔ 6/7	✔ 6/7	✔ 6/7	✔ 6/7
Anx / Dep Sample	✔ 7/7	✔ 7/7	✔ 7/7	✔ 7/7
4-6 Year Olds	✔ 5/6	✔ 5/6	✔ 5/6	✔ 5/6
7-12 Year Olds	✔ 7/8	✔ 7/8	✔ 7/8	✔ 7/8
13-18 Year Olds	✔ 8/8	✔ 8/8	✔ 7/8	✔ 8/8
Males	✔ 7/8	✔ 7/8	✔ 7/8	✔ 8/8
Females	✔ 7/8	✔ 6/8	✔ 7/8	✔ 7/8
Self-report	✔ 7/8	✔ 7/8	✔ 7/8	✔ 7/8
Proxy-report	✔ 7/8	✔ 6/8	✔ 7/8	✔ 7/8

Table 4 (continued)

Subtotal score	89%	87%	88%	91%
RESPONSIVENESS				
Combined Sample	✔ 2/2	✔ 2/2	✘ 1/2	✔ 2/2
Anx / Dep Sample	✔ 2/2	✔ 2/2	✔ 2/2	✔ 2/2
Males	✘ 1/2	✔ 2/2	✘ 0/2	✘ 1/2
Females	✘ 0/1	✘ 0/1	✘ 0/1	✔ 1/1
Self-report	✔ 2/2	✔ 2/2	✔ 2/2	✔ 2/2
Subtotal score	**78%**	**89%**	**56%**	**89%**
TEST-RETEST RELIABILITY				
Combined Sample	✘ 0/1	✘ 0/1	✘ 0/1	✔ 1/1
ADHD Sample	✔ 1/1	✘ 0/1	✘ 0/1	✘ 0/1
Anx / Dep Sample	✘ 0/1	✔ 1/1	✘ 0/1	✔ 1/1
7-12 Year Olds	✔ 1/1	✘ 0/1	✔ 1/1	✘ 0/1
13-18 Year Olds	✔ 1/1	✔ 1/1	✘ 0/1	✔ 1/1
Males	✔ 1/1	✘ 0/1	✔ 1/1	✔ 1/1
Females	✘ 0/1	✔ 1/1	✘ 0/1	✘ 0/1
Self-report	✔ 1/1	✔ 1/1	✘ 0/1	✔ 1/1
Subtotal score	**63%**	**50%**	**25%**	**63%**
	PedsQL	**EQ-5D-Y-3L**	**EQ-5D-Y-5L**	**CHU9D**
OVERALL TOTAL*	**82.4%**	**68.9%**	**65.7%**	**84.4%**

✔ = Good performance (proportion of tests passed is >50%)

✘ = Poor performance (proportion of tests passed is ≤50%)

Some subgroups are not included in this summary table due to inadequate sample sizes for these analyses. All results for all subgroups are provided in the electronic supplementary material (ESM)

Green shading = top performing instrument(s) in each section

Subtotal scores = average for all results within that section (i.e. results for each analysis, including the combined sample). Data are presented in this table for comparison of specific metrics of performance where available

*Overall total = average score from each section (i.e. each psychometric property), weighted equally

Key for interpreting each cell in the table:

Note for all cells: Counts for all cells exclude any analyses that had a doubtful (n = 30–49) or inadequate (n < 30) sample size (i.e. these analyses are not included in numerator or denominator of the cell)

Acceptability/feasibility: Proportion of respondents who reported the instrument was either 'somewhat' or 'very' easy to complete

Ceiling effects: Proportion of respondents who reported being in the best possible health state for each instrument. A ceiling effect is considered to be detected if >15% of participants scored in the best possible health state. *Subtotal* score here shows the proportion of tests where the instrument performed well, i.e. where no ceiling effect was detected

Known groups: Number of known groups correctly differentiated (numerator), of the total number of possible known groups (denominator)

Convergent validity: Number of intercorrelations between HRQoL instruments that were 'strong', as hypothesised, and number of correlations between HRQoL instruments and mental health measures that were 'strong/moderate', as hypothesised (numerator), of the total number of intercorrelations possible (denominator)

Responsiveness: Number of correctly detected changes in health status (either improvements or deterioration in general health or mental health condition) (numerator), of the total number of possible differences (denominator)

Test–retest reliability: Counted as 1/1 if the ICC value was 'good' or 'excellent', according to the thresholds recommended by Cicchetti et al. [38]; counted as 0/1 if the ICC value was 'poor' or 'fair'

3.2 Psychometric Performance

3.2.1 Acceptability/Feasibility

There were no major differences observed in the acceptability/feasibility of the instruments. The majority of respondents (~ 70%) found all instruments 'somewhat' or 'very' easy to complete. In addition, ~ 20% of participants across all groups rated each instrument as 'neither easy nor difficult' to complete. The EQ-5D-Y-5L was most consistently rated as 'somewhat' or 'very' easy to complete (68.1–76.5%), followed by the CHU9D (67.0–75.2%). Ease of completion was similar between self- and proxy report for the EQ-5D-Y-3L, EQ-5D-Y-5L and CHU9D

(all within 1.5%), however the PedsQL was considered slightly easier to complete by proxy compared with self-report (70.4% vs 66.9%). For full results, see ESM Tables S3.1–S3.10.

3.2.2 Floor and Ceiling Effects

No floor effects were detected for any instrument in the combined sample or any subgroups; moreover, no respondents reported being in the worst possible health state on any instrument.

No ceiling effects were detected in the combined sample or within any subgroups for the PedsQL (proportion of respondents in best possible health state, all < 1%) or CHU9D (all < 4%). In contrast, for the EQ-5D-Y-5L, ceiling effects were detected in the ADHD sample (18.5%), 4-to 6-year-olds (16.0%), boys (15.9%), and by proxy report (17.3%). Ceiling effects were also detected in these subgroups using the EQ-5D-Y-3L, with further ceiling effects identified for the combined sample (17.9%), 7- to 12-year-olds (17.1%), 13- to 18-year olds (18.7%) and by self-report (16.9%). Where ceiling effects were apparent, they were more likely to occur in the following subgroups: ADHD, aged 4–6 years, boys, and by proxy report. For full results, see ESM Table S4.

Of note, and as expected, ceiling effects were less likely to occur in longer instruments (statistically less likely); as well as less likely in instruments that included more items expected to be of concern for children in our sample, for example, school problems, paying attention, or cognitive domains not included in shorter instruments. This is particularly highlighted through no ceiling effects being detected for any instrument in the anxiety/depression sample, as all instruments included at least one item related to sadness, worry, or emotional problems. See ESM Table S5 for figures displaying the distribution of responses for each item on each HRQoL instrument for the combined sample, and for all subgroups.

3.2.3 Construct Validity—Known-group Validity

The four instruments in the main comparison—the PedsQL, CHU9D, EQ-5D-Y-3L and EQ-5D-Y-5L—performed almost equally well across all known-group analyses, with total scores only differing by 1 of 28 known-group comparisons. In the combined sample, the largest effect sizes were observed for differences in severity on the RCADS-25 (range: large ES = 1.08–1.49), followed by severity on the SDQ (moderate–large ES = 0.69–1.16); severity on the SWAN (small–moderate ES = 0.42–0.63) and presence of SHCN (small ES = 0.27–0.39). This same pattern of effect sizes was observed in all subgroup analyses. The instruments were equally able to identify known groups across each subgroup, with the exception that known groups were better identified via self-report than proxy report for all instruments. For full results, see ESM Table S6.

3.2.4 Construct Validity—Convergent Validity

In the combined sample, the intercorrelations between the four main HRQoL instruments were all 'strong' (Spearman ρ, range: 0.62–0.73, all $p < 0.001$), and in the expected direction. Next, examining correlations between generic HRQoL instruments and the mental health symptom measures revealed the expected pattern of 'moderate/strong' correlations between generic HRQoL instruments and the SDQ ($\rho = 0.42$–0.60; $p < 0.001$). However, relationships were stronger than hypothesised between generic HRQoL instruments and the RCADS-25, all 'strong' correlations ($\rho = 0.53$–0.67; $p < 0.001$); and weaker than hypothesised between generic HRQoL instruments and the SWAN, all 'weak' correlations ($\rho = 0.20$–0.25; $p < 0.001$). Subgroup analyses were largely in line with the combined sample, with a notable exception regarding the SWAN, where correlations were strengthened to 'moderate' in subgroups of 13- to 18-year-olds and boys. For full results, see ESM Table S7.

3.2.5 Responsiveness

Tests of responsiveness were hindered due to sample size more than any other psychometric property; notably, we were unable to examine responsiveness of the instruments to *deterioration* in health, therefore reporting is limited to *improvements* in general health and/or the mental health condition.

In the combined sample, the four main instruments all performed well; able to detect improvements in general and mental health status, though with small effect sizes (SRM range: 0.26–0.39), and with the exception of the EQ-5D-Y-5L which did not detect improvements in general health. The CHU9D was the only instrument of the four to detect improvements related to the child's mental health condition in girls (SRM = 0.40, $p = 0.006$); though the EQ-5D-Y-3L was better able to detect improvements in boys' general and mental health (SRM = 0.46, SRM = 0.38; $p < 0.01$, respectively). Samples sizes were inadequate ($n < 30$) or doubtful ($n < 50$) to examine responsiveness in subgroups of ADHD, all age bands, and proxy report. For full results, see ESM Table S8.

3.2.6 Test–Retest Reliability

In the combined sample, the 95% confidence interval (CI) ranged from 'fair' to 'good' for all instruments (intraclass correlation coefficient [ICC] 95% CI range 0.41–0.70).

Using the ICC thresholds recommended by Cicchetti [38], test–retest reliability in the combined sample was good for the CHU9D (ICC 0.60, 95% CI 0.47–0.70), and fair for the PedsQL, EQ-5D-Y-3L and EQ-5D-Y-5L (ICC 0.59, 0.57, 0.55, respectively), though all estimates were within 0.05 of one another. Overall, the CHU9D and PedsQL most consistently showed good test–retest reliability across each subgroup, though subgroup analyses revealed relative strengths for each instrument. For example, the CHU9D was more reliable in the anxiety/depression than the ADHD sample, though conversely the PedsQL was more reliable in the ADHD compared with the anxiety/depression sample. In the anxiety/depression (ICC 0.73, 95% CI 0.56–0.83) and 13- to 18-year olds samples (ICC 0.76, 95% CI 0.63–0.85), the EQ-5D-Y-3L outperformed the CHU9D and the PedsQL. Both instruments were more reliable in boys than girls. Sample sizes were inadequate in 4- to 6-year-olds and by proxy report for each instrument.

Using the more restrictive thresholds recommended by Koo and Li [39], only the EQ-5D-Y-3L reached 'good' reliability, and this was only within the 13- to 18-year-old subgroup (ICC 0.76, 95% CI 0.63–0.85). No instrument reached the 'excellent' reliability threshold of ICC >0.90. For full results, see ESM Table S9.

3.3 Sensitivity Analyses

In sensitivity analyses using utility scores in place of instrument total sum scores for the EQ-5D-Y-3L and CHU9D, results were unchanged from the main analysis with the following exceptions. In convergent validity, correlations between mental health measures and the EQ-5D-Y-3L and CHU9D were weakened, becoming only moderate for the CHU9D and SDQ ($\rho = -0.48$; $p < 0.001$). In known-group testing, effect sizes weakened for the EQ-5D-Y-3L and CHU9D for known-group differences on the SWAN (both weak effects, $d = 0.43$; $p < 0.001$). For full results, see ESM Table S10.

3.4 Psychometric Performance of the HUI3 and AQoL-6D

Full results for the HUI3 and AQoL-6D are available in the ESM, Tables S1–S10, and are not included in the comparison with other instruments due to the lower sample size for these instruments. Briefly, approximately 70% of participants rated both instruments as 'somewhat' or 'very easy' to complete, though the AQoL-6D was rated easier to complete via proxy report compared with self-report (72.0% vs 63.5%). No floor or ceiling effects were detected for either instrument. Where sample size allowed, both instruments were able to detect differences in severity on the SDQ with moderate to large effect sizes, but were less consistently able to detect differences in severity on the SWAN, or children with SHCN. Correlations with other HRQoL instruments and mental health symptom measures were largely as hypothesised. Sample sizes were inadequate to assess responsiveness or test–retest reliability of these instruments.

In sensitivity analyses using utility scores for the AQoL-6D and HUI3, results were unchanged from the main analysis with the following exceptions. In convergent validity, correlations between mental health instruments and the AQoL-6D and HUI3 were strengthened, including a change to moderate correlations with the SWAN ($\rho = -0.31$ [for both]; $p < 0.001$). In known-group testing, improvements were seen for the HUI3, which was able to detect known-group differences with moderate effects sizes for severity of ADHD using the SWAN (Cohen's $d = 0.67$; $p < 0.001$); and children with SHCN ($d = 0.51$; $p < 0.001$).

3.5 Instrument Comparison

Table 4 provides a high-level summary of the psychometric performance of the PedsQL, EQ-5D-Y-3L, EQ-5D-Y-5L and CHU9D in the combined sample and for each subgroup. All results and results summaries in text and in Table 4 are based on results with 'adequate' or 'very good' sample sizes based on the COSMIN guidelines [40]. Results based on 'doubtful' or 'inadequate' sample sizes are shown in the ESM tables for completeness of information to the reader only, and are not interpreted in text or in Table 4.

Strong overall performance was observed across psychometric properties for the CHU9D (85.5%) and the PedsQL (81.9%), with some differences observed at the subgroup level favouring different instruments. As expected, the shorter instruments (the EQ-5D-Y-3L and EQ-5D-Y-5L) showed ceiling effects, however, these instruments showed strong performance for feasibility/acceptability, and convergent and known-group validity. In addition, the EQ-5D-Y-3L showed good responsiveness to improvements in health.

4 Discussion

In this study, we examined the psychometric performance of a range of generic paediatric HRQoL instruments in a large sample of children with anxiety and/or depression or ADHD. Overall, we found strong performance by the CHU9D, followed closely by the PedsQL, and more variable performance by the EQ-5D-Y-3L and EQ-5D-Y-5L. The PedsQL, CHU9D, EQ-5D-Y-3L and EQ-5D-Y-5L showed similarly good performance for acceptability/feasibility, known-group validity and convergent validity. The CHU9D and PedsQL showed no floor or ceiling effects and fair–good test–retest reliability. Test–retest reliability was lower for the EQ-5D-Y-3L and EQ-5D-Y-5L. The EQ-5D-Y-3L showed the highest

ceiling effects, but was the top performing instrument alongside the CHU9D on responsiveness to improvements in health status, followed by the PedsQL. In the smaller subsample, the AQoL-6D and HUI3 showed good acceptability/feasibility, no floor or ceiling effects, and good convergent validity, yet poorer performance on known-group validity. Responsiveness and test–retest reliability were not able to be assessed for these two instruments. In subgroup analyses, performance was similar for all instruments for acceptability/feasibility, known-group and convergent validity; however, relative strengths and weaknesses for each instrument were noted for ceiling effects, responsiveness and test–retest reliability. In sensitivity analyses using utility scores, performance regarding known-group and convergent validity worsened slightly for the EQ-5D-Y-3L and CHU9D, though improved slightly for the HUI3 and AQoL-6D.

Our finding of validity and reliability of the CHU9D and PedsQL is in line with the literature review by Mierau et al. [15]. Together, our findings suggest these instruments may be the most suitable—of existing HRQoL instruments—for use in economic evaluation of child and adolescent mental healthcare. However, an advantage of the CHU9D over the PedsQL is the existence of an adolescent- and Australian-specific value set for the instrument, where the PedsQL (and equally the EQ-5D-Y-3L or EQ-5D-Y-5L) have no validated value set for Australia, which currently limits their usefulness in child and adolescent mental healthcare evaluations [15].

One of the most useful metrics of HRQoL instruments is their ability to detect differences between known subgroups within a patient sample, which can be useful in developing economic models. In line with the previous study by Mihalopoulos et al. [24], we found the EQ-5D-Y instruments were able to detect known-group differences based on severity of mental health conditions. However, in our study, performance of the CHU9D and PedsQL was still very high. In line with Mihalopoulos et al., we observed poorer known-group validity using the HUI3. The similarity of our findings with Mihalopoulos et al. is notable given the use of different mental health symptom measures in our study, and the use of utility scores instead of sum scores, which leads to results with a different interpretation.

Arguably, also amongst the most important psychometric properties of HRQoL instruments for health economic analyses and description of health profiles is the ability of the instrument to detect a change in health status and reliability of scores across repeated measurement. Our findings are novel in this regard for children with anxiety and/or depression and ADHD, and suggest the CHU9D is the most responsive of the instruments to an improvement in the child's general health and their mental health, and also had the highest test–retest reliability estimate in the combined sample. The PedsQL also performed well, though showed lower responsiveness to changes in health status. Notably, however, the follow-up survey was completed at 4 weeks, which may impact responsiveness and reliability estimates. We did not have adequate sample size to examine responsiveness to deteriorations in health, and this will be a crucial area of future research, particularly given the variable performance observed for the instruments in detecting improvements in health.

Our findings are in line with others that have found the CHU9D performs well in children and adolescents with mental health challenges [15, 19, 20]. However, with regard to the implications this has for clinical and health policy decision making, the choice of HRQoL instrument should also consider the intended use and subgroup in which the instrument would be used; the best performing instrument in each instance may differ. Others have equally noted varying strengths of different HRQoL instruments across different psychometric properties [18]. Furthermore, in addition to the psychometric properties explored here, other practical considerations should be considered in the choice of instrument, such as the length of the instrument (i.e. the PedsQL has 23 items, whereas the CHU9D and EQ-5D-Y instruments are much shorter at 9 items and 5 items, respectively); licensing fees; technology and resources available; and the availability of a country-specific value set, etc. [41]. The balance of these considerations may vary for commercially sponsored drug trials, routine clinical use, research purposes or health economic analyses. As we and others have noted, these instruments have different properties and are accompanied by utility algorithms that are fundamentally different. This, combined with the potential need to choose different instruments for different populations or research questions, has profound implications for the ability to compare QALY estimates generated in each scenario.

Tests of known-group and convergent validity revealed a consistent pattern observed for all HRQoL instruments, where each was more closely aligned with measures of internalising disorders (i.e. anxiety or depression) than that of externalising disorders (i.e. ADHD). This pattern been noted previously [15] and was apparent in our results through larger effect sizes and correlations being observed between HRQoL instruments and the RCADS-25 (measuring anxiety/depression symptoms); followed by the SDQ (measuring a combination of internalising and externalising symptoms); and lastly the SWAN (measuring ADHD symptoms). This pattern of results appears to be consistent for all HRQoL instruments, regardless of the number of mental-health-related items included in the instrument. This pattern could arise for a number of reasons: (i) the instruments are simply more valid and reliable in internalising conditions [13, 14]; (ii) children and adolescents with internalising conditions have poorer HRQoL than those with externalising conditions [4, 24], making the differences between groups larger and

easier to detect; or (iii) the HRQoL instruments themselves are measuring internalising symptoms more so than externalising symptoms, making a relationship between poorer HRQoL and greater internalising symptoms a tautology [12, 42].

Further to this point, interestingly, in our sensitivity analyses using utility scores, the performance of some HRQoL instruments changed in relation to the ADHD-symptom measure. Specifically, using utility scores *improved* the functioning of the HUI3, which saw larger differences between known groups of ADHD symptoms, yet performance *worsened* for the EQ-5D-Y-3L and CHU9D, which saw smaller differences between these groups. This suggests that preference weightings (derived in different countries) that are used to generate the utility scores for each instrument, differentially weight problems that are impacted by ADHD. This highlights that the differences in performance between the instruments are ultimately a product of the measurement properties of both the descriptive systems and value sets, and can vary between instruments depending on the constructs being measured, and the characteristics of the value sets, including the valuation approaches used. This overlap of mental health and HRQoL instruments, and the types of mental health symptoms captured by HRQoL instruments when scored either by total sum scores or utility scores, warrants further attention in future research.

Strengths of the study include that it is the largest of its kind, internationally, and close control of data quality was maintained (see technical methods [25]). This provides the best evidence to date on the comparative acceptability, validity, reliability and responsiveness of paediatric generic HRQoL instruments for use in child and adolescent mental healthcare, and the first multi-instrument comparison examining responsiveness and test–retest reliability in both internalising and externalising mental health populations. An additional key strength of the study is the use of validated mental health symptom measures to assess symptom severity for use in known-group validity testing. There are limitations of the study. Children's mental health diagnosis and changes related to this condition at follow-up were reported by their caregivers. While we were able to measure children's anxiety/depression and ADHD symptoms with valid instruments, mental health diagnoses were not confirmed through an independent clinical diagnosis at either time point. As is common in longitudinal research, we found it difficult to identify large numbers of children with declining health at follow-up. Many children are in treatment, or have a natural history of disease that leads to improvements over time. This is an ongoing issue for validation of responsiveness for these instruments. Our findings may not be generalisable to other mental health conditions, given variable performance of HRQoL instruments has been observed across different mental health conditions in adult literature [13, 14], and now also through our study. While we made careful efforts to maximise the quality of the online sample [25], limitations arise due to the use of online panel recruitment, including the potential for sampling bias, self-selection into the survey by participants, and the inability to verify if self-report occurred or if there was parental influence in children's self-report. It is also important to note that the 'total performance score' for each instrument is constructed by the authors, which may mean that overall instrument performance could be calculated and interpreted in other ways.

In summary, our results indicate that the CHU9D, PedsQL, EQ-5D-Y-3L and EQ-5D-Y-5L perform equally well on acceptability/feasibility, known-group and convergent validity. However, relative strengths of the CHU9D and PedsQL were observed regarding their lack of ceiling effects, and greater test–retest reliability. Relative strengths were also observed for the CHU9D and EQ-5D-Y-3L regarding responsiveness to improvements in health. While each instrument showed strong performance in some areas, the CHU9D and PedsQL showed the most consistent performance across all psychometric properties. Instrument performance varied across subgroups, particularly for ceiling effects, responsiveness and test–retest reliability, thus careful consideration of the choice of instrument is advised, as this may differ depending on the intended use of the data, and the age, sex, report type and type of mental health condition of the population in which the instrument is being used. In addition, the closer relationship of these HRQoL instruments with internalising symptoms compared with externalising symptoms warrants targeted attention in future research.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40273-024-01354-2.

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. Additionally, we would like to sincerely thank members of the QUOKKA Consumer Advisory Group and others who helped pilot, test and refine different elements of this study. 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.

Declarations

Funding 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). RO and RJ 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). The Murdoch Children's Research Institute is

supported by the Victorian Government's Operational Infrastructure Support Program.

Conflicts of interest RJ, KD, ND, BM, HH have all received previous or current funding from the EuroQol Research Foundation, which is the developer of some instruments included in this study. ND and BM are members of the EuroQol Group. Views expressed in this paper are those of the authors and are not necessarily those of the EuroQol Research Foundation.

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

Ethics approval This study was approved by The Royal Children's Hospital (RCH) Human Research Ethics Committee (HREC/71872/RCHM-2021) on the 20th May 2021.

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

Consent for publication Not applicable.

Code availability Not applicable.

Author contributions Substantial contribution to the conception or design: KD, ND, BM, HH, GC. Main contribution to data analysis: RO. Substantial contributions to the acquisition, analysis or interpretation of data: All authors. Drafted the manuscript: RO. Reviewed and critically revised the manuscript: all authors. All authors approved the final version to be published and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

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.

References

1. Erskine HE, Moffitt TE, Copeland WE, Costello EJ, Ferrari AJ, Patton G, et al. A heavy burden on young minds: the global burden of mental and substance use disorders in children and youth. Psychol Med. 2015;45(7):1561–3.
2. Lawrence D, Hafekost J, Johnson SE, Saw S, Buckingham WJ, Sawyer MG, et al. Key findings from the second Australian Child and Adolescent Survey of Mental Health and Wellbeing. Aust N Z J Psychiatry. 2016;50(9):876–86.
3. Bastiaansen D, Ferdinand RF, Koot HM. Predictors of quality of life in children and adolescents with psychiatric disorders. Child Psychiatry Hum Dev. 2020;51(1):104–13.
4. O'Loughlin R, Hiscock H, Pan T, Devlin N, Dalziel K. The relationship between physical and mental health multimorbidity and children's health-related quality of life. Qual Life Res. 2022;31(7):2119–31.
5. 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.
6. Algurén B, Ramirez JP, Salt M, Sillett N, Myers SN, Alvarez-Cote A, et al. Development of an international standard set of patient-centred outcome measures for overall paediatric health: a consensus process. Arch Dis Child. 2021;106(9):868–76.
7. Kwon J, Freijser L, Huynh E, Howell M, Chen G, Khan K, et al. Systematic review of conceptual, age, measurement and valuation considerations for generic multidimensional childhood patient-reported outcome measures. Pharmacoeconomics. 2022;40(4):379–431.
8. Devlin N. 'Preference-based measure' is misleading—can we agree on something better? [Internet]. The Academic Health Economists' Blog. 2020. p. August 12. https://aheblog.com/2020/08/12/preference-based-measure-is-misleading-can-we-agree-on-something-better/. Accessed 5 May 2023.
9. Devlin N, Parkin D, Janssen B. Methods for analysing and reporting EQ-5D data. Cham: Springer Nature; 2020.
10. Jonsson U, Alaie I, Löfgren Wilteus A, Zander E, Marschik PB, Coghill D, et al. Annual Research Review: Quality of life and childhood mental and behavioural disorders—a critical review of the research. J Child Psychol Psychiatry. 2017;58(4):439–69.
11. Coghill D, Danckaerts M, Sonuga-Barke E, Sergeant J. Practitioner Review: Quality of life in child mental health—conceptual challenges and practical choices. J Child Psychol Psychiatry. 2009;50(5):544–61.
12. Katschnig H. Quality of life in mental disorders: challenges for research and clinical practice. World Psychiatry. 2006;5(3):139–45.
13. Brazier J, Connell J, Papaioannou D, Mukuria C, Mulhern B, Peasgood T, et al. A systematic review, psychometric analysis and qualitative assessment of generic preference-based measures of health in mental health populations and the estimation of mapping functions from widely used specific measures. Health Technol Assess (Rockv). 2014;18(34).
14. Mulhern B, Mukuria C, Barkham M, Knapp M, Byford S, Soeteman D, et al. Using generic preference-based measures in mental health: psychometric validity of the EQ-5D and SF-6D. Br J Psychiatry. 2014;205(3):236–43.
15. Mierau JO, Kann-Weedage D, Hoekstra PJ, Spiegelaar L, Jansen DEMC, Vermeulen KM, et al. Assessing quality of life in psychosocial and mental health disorders in children: a comprehensive overview and appraisal of generic health related quality of life measures. BMC Pediatr. 2020;20:329.
16. Rowen D, Keetharuth A, Poku E, Wong R, Pennington B, Wailoo A. A review of the psychometric performance of child and adolescent preference-based measures used to generate utility values for children. Sheffield; 2020.
17. Tan RLY, 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.
18. 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;03 May 2023 [EPub Ahead of Print].
19. Furber G, Segal L. The validity of the Child Health Utility instrument (CHU9D) as a routine outcome measure for use in child and adolescent mental health services. Health Qual Life Outcomes. 2015;13(1):1–14.
20. Wolf RT, Ratcliffe J, Chen G, Jeppesen P. The longitudinal validity of proxy-reported CHU9D. Qual Life Res. 2021;30(6):1747–56.
21. Åström M, Krig S, Ryding S, Cleland N, Rolfson O, Burström K. EQ-5D-Y-5L as a patient-reported outcome measure in psychiatric inpatient care for children and adolescents—a cross-sectional study. Health Qual Life Outcomes. 2020;18(1):1–14.
22. Lynch FL, Dickerson JF, Feeny DH, Clarke GN, MacMillan AL. Measuring health-related quality of life in teens with and without depression. Med Care. 2016;54(12):1089–97.
23. Dickerson JF, Feeny DH, Clarke GN, MacMillan AL, Lynch FL. Evidence on the longitudinal construct validity of major generic and utility measures of health-related quality of life in teens with depression. Qual Life Res. 2018;27(2):447–54.
24. Mihalopoulos C, Chen G, Scott JG, Bucholc J, Allen C, Coghill D, et al. Assessing outcomes for cost-utility analysis in children and adolescents with mental health problems: Are multi-attribute utility instruments (MAUIs) fit for purpose? Value Heal. 2022;[EPub ahead of print].
25. Jones R, Mulhern B, Devlin N, Hiscock H, Chen G, O'Loughlin R, et al. Australian Paediatric Multi-Instrument Comparison (P-MIC) Study: Technical Methods Paper [Online]. [Internet]. Melbourne, Australia; 2023. https://www.quokkaresearchprogram.org/project-1-1. Accessed 5 May 2023.
26. Jones R, Mulhern B, Mcgregor K, Yip S, Loughlin RO, Devlin N, et al. Psychometric performance of HRQoL measures: an Australian Paediatric Multi-Instrument Comparison Study Protocol (P-MIC). Children. 2021;8:714.
27. 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. 2023;10:1604.
28. Gagnier JJ, Lai J, Mokkink LB, Terwee CB. COSMIN reporting guideline for studies on measurement properties of patient—reported outcome measures. Qual Life Res. 2021;30(8):2197–218.
29. Bethell CD, Read D, Stein REK, 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.
30. Goodman R, Ford T, Simmons H, Gatward R, Meltzer H. Using the Strengths and Difficulties Questionnaire (SDQ) to screen for child psychiatric disorders in a community sample. Br J Psychiatry. 2000;177(6):534–9.
31. Stone LL, Otten R, Engels RCME, Vermulst AA, Janssens JMAM. Psychometric properties of the parent and teacher versions of the strengths and difficulties questionnaire for 4- to 12-Year-olds: A review. Clin Child Fam Psychol Rev. 2010;13(3):254–74.
32. Kremer P, De Silva A, Cleary J, Santoro G, Weston K, Steele E, et al. Normative data for the Strengths and Difficulties Questionnaire for young children in Australia. J Paediatr Child Health. 2015;51(10):970–5.
33. Mellor D. Normative data for the Strengths and Difficulties Questionnaire in Australia. Aust Psychol. 2005;40(3):215–22.
34. Child FIRST UCLA. SPSS Syntax for Batch Scoring. [Internet]. RCADS Scoring Programs. 2022. p. Updated June 2022. https://www.childfirst.ucla.edu/wp-content/uploads/sites/163/2022/06/RCADS-and-RCADSP-Tscore-syntax-2022-06-29.txt. Accessed 5 May 2023.
35. Burton CL, Wright L, Shan J, Xiao B, Dupuis A, Goodale T, et al. SWAN scale for ADHD trait-based genetic research: a validity and polygenic risk study. J Child Psychol Psychiatry. 2019;60(9):988–97.
36. Swanson JM, Schuck S, Porter MM, Carlson C, Hartman CA, Sergeant JA, et al. Categorical and dimensional definitions and evaluations of symptoms of ADHD: history of the SNAP and the SWAN Rating Scales. Int J Educ Psychol Assess. 2012;10(1):51–70.
37. 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.
38. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess. 1994;6(4):284–90.
39. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155–63.
40. Mokkink LB, Prinsen CA, Patrick D, Alonso J, Bouter LM, de Vet HC, et al. COSMIN Study Design checklist for Patient-reported outcome measurement instruments. Department of Epidemiology and Biostatistics Amsterdam Public Health research institute Amsterdam University Medical Centers. 2019.
41. 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(2):1–6.
42. Danckaerts M, Sonuga-Barke EJS, Banaschewski T, Buitelaar J, Döpfner M, Hollis C, et al. The quality of life of children with attention deficit/hyperactivity disorder: a systematic review. Eur Child Adolesc Psychiatry. 2010;19(2):83–105.
43. Australian Bureau of Statistics. 2033.0.55.001 - Census of Population and Housing: Socio-Economic Indexes for Areas (SEIFA), Australia, 2016. [Internet]. Canberra: Australian Bureau of Statistics. 2018 [cited 2021 Mar 23]. https://www.abs.gov.au/ausstats/abs@.nsf/Lookup/bySubject/2033.0.55.0012016MainFeaturesIRSAD20. Accessed 5 May 2023.
44. Australian Bureau of Statistics. 1270.0.55.005 - Australian Statistical Geography Standard (ASGS): Volume 5 - Remoteness Structure, July 2016. Table 3: Correspondence 2017 Postcode to 2016 Remoteness Area. [Internet]. Canberra: Australian Bureau of Statistics. 2018 [cited 2021 Mar 23]. https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1270.0.55.005July2016?OpenDocument. Accessed 5 May 2023.
45. Chorpita BF, Yim L, Moffitt C, Umemoto LA, Francis SE. Assessment of symptoms of DSM-IV anxiety and depression in children: a revised child anxiety and depression scale. Behav Res Ther. 2000;38(8):835–55.
46. Ebesutani C, Korathu-Larson P, Nakamura BJ, Higa-McMillan C, Chorpita B. The Revised child anxiety and depression scale 25–parent version: scale development and validation in a school-based and clinical sample. Assessment. 2017;24(6):712–28.
47. American Psychiatric Association. Diagnostic and statistical manual of mental disorders (4th ed., text rev.). Washington, DC: American Psychiatric Association; 2000.
48. 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 Author (s): James W. Varni, Michael Seid and Paul S. Kurtin Published by : Lippincott Williams. Med Care. 2001;39(8):800–12.
49. 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.
50. 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;19(6):875–86.

51. Kreimeier S, Åström M, Burström K, Egmar AC, Gusi N, Herdman M, et al. EQ-5D-Y-5L: developing a revised EQ-5D-Y with increased response categories. Qual Life Res. 2019;28(7):1951–61.
52. Stevens K. Developing a descriptive system for a new preference-based measure of health-related quality of life for children. Qual Life Res. 2009;18(8):1105–13.
53. Ratcliffe J, Stevens K, Flynn T, Brazier J, Sawyer M. An assessment of the construct validity of the CHU9D in the Australian adolescent general population. Qual Life Res. 2012;21(4):717–25.
54. Stevens K, Ratcliffe J. Measuring and valuing health benefits for economic evaluation in adolescence: an assessment of the practicality and validity of the Child Health Utility 9D in the Australian Adolescent Population. Value Heal. 2012;15(1):1092–9.
55. 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 Heal. 2010;13(8):1014–27.
56. 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 Heal. 2012;15(1):730–6.
57. 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;40(2):113–28.
58. 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;60:34–42.
59. McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? Qual Life Res. 1995;4:293–307.
60. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol 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):737–45.
61. Cohen J. A power primer. Psychol Bull. 1992;112(1):155–9.
62. Cohen J. Statistical power analysis for the behavioral sciences (2nd edn). Erlbaum; 1988.