PharmacoEconomics
https://doi.org/10.1007/s40273-024-01355-1

ORIGINAL RESEARCH ARTICLE

Psychometric Properties of Child Health Utility 9D (CHU9D) Proxy Version Administered to Parents and Caregivers of Children Aged 2–4 Years Compared with Pediatric Quality of Life Inventory™ (PedsQL)

Xiuqin Xiong[1] · Natalie Carvalho[1] · Li Huang[1] · Gang Chen[2] · Renee Jones[1] · Nancy Devlin[1] · Brendan Mulhern[3] · Kim Dalziel[1]

Accepted: 25 October 2023

Abstract

Objective This study examines the psychometric properties of the Child Health Utility 9D (CHU9D) proxy version administered to parents/caregivers of 2–4-year-old Australian children compared with Pediatric Quality of Life Inventory™ version 4.0 (PedsQL).

Methods Data collected in 2021/2022 from parents/caregivers of 2–4-year-olds from the Australian pediatric multi-instrument comparison study were used. Feasibility, ceiling/floor effects, test–retest reliability, convergent validity, known-group validity, and responsiveness were assessed.

Results A total of 842 caregivers completed the survey at baseline, with 513 completing the follow-up survey. The CHU9D did not demonstrate ceiling effects in the sample with special health care needs, with only 6% of respondents reporting best levels for all nine dimensions. CHU9D correlated with PedsQL moderately-to-strongly between comparable items (correlation coefficients 0.34–0.70). CHU9D was able to differentiate between groups with known health differences with moderate-to-large effect sizes (Cohen's *d* 0.58–2.03). Moderate test–retest reliability was found for CHU9D in those reporting no health change at a 2-day follow-up (ICC 0.52). A standard response mean (SRM) of 0.25–0.44 was found for children with changes in general health and a SRM of 0.72–0.82 for children who reported worsened health when developing new illnesses, indicating small-to-large responsiveness according to different definitions of health changes. Compared with PedsQL, CHU9D had similar known-group validity and responsiveness and slightly poorer test–retest reliability.

Conclusion The CHU9D was found to be valid and reliable to measure health-related quality-of-life in children aged 2–4 years, although with relatively low test–retest reliability in some dimensions. Further development and validation work is warranted.

1 Introduction

Children under 5 years of age are important users of health care services and have greater health service use than older children [1]. Many new healthcare technologies target early childhood diseases [2–4]. It is thus important to make wise health resource allocation decisions for this age group. The use of economic evaluation for childhood interventions to aid resource allocation decisions has increased in recent years, especially cost-utility analysis [5, 6]. However, there are few instruments appropriate and validated for utility measurement for young children [7–9]. A recent systematic review of 372 studies assessing the psychometric performance of pediatric utility instruments reported a prominent research gap in the validation of instruments for preschool-aged children [10].

Many economic evaluations for younger ages used utilities obtained from generic pediatric preference-accompanied measures developed for older age groups or adults [11, 12]. This is problematic [13] as there is evidence that children under 5 years old have different developmental stages and may have different quality of life dimensions or constructs compared with older populations [14]. It is questionable

✉ Kim Dalziel
kim.dalziel@unimelb.edu.au

[1] Centre for Health Policy, Melbourne School of Population and Global Health, The University of Melbourne, Parkville, VIC, Australia

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

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

Published online: 27 January 2024

Key Points for Decision Makers

There is a lack of established generic pediatric measures of health-related quality of life (HRQoL) appropriate for use in economic evaluation for young children despite young children being relatively high health system users.

The CHU9D proxy version for children under 5 years of age is a potential instrument for measuring HRQoL in economic evaluations. However, no psychometric evidence on it is available. This is the first study assessing the psychometric properties of the CHU9D proxy version completed by parents or caregivers of children aged 2–4 years old.

This study provides evidence that CHU9D is valid and reliable overall for use by parents of 2–4-year-olds compared with PedsQL, although with relatively low test–retest reliability in some dimensions. This evidence will be useful for those wishing to measure HRQoL for children aged 2–4 years including for incorporation in economic evaluation.

whether instruments having common health dimensions with versions for older children or adults are suitable for use in younger children directly; they usually have adapted wording (or added guidance notes) and different report types (e.g., proxy report or self-report), which often requires further validation evidence [15]. Health technology assessment authorities in Australia and the UK have also noted the lack of utilities used in pediatric economic evaluations and promote the use of concise, generic measures of pediatric HRQoL accompanied by relevant value sets [16, 17]. There is potential to unfairly penalize young children in the health technology assessment process due to poor quality, missing, or uncertain utility evidence [18]. It is therefore important to explore appropriate HRQoL measurement in young children.

The evaluation of the performance of HRQoL measures is important before their wide application. There are four important considerations in these assessments: feasibility, reliability, construct validity, and responsiveness [19]. Feasibility refers to the practicality and acceptability of the instrument to participants, such as the time required to complete the survey and whether the questions are difficult to understand. Reliability concerns the consistency of responses when health status remains unchanged. Psychologists usually examine three forms of consistency: over time (test–retest reliability or intra-rater reliability), across items (internal consistency), and between different assessors (inter-rater reliability). Validity refers to whether the instrument accurately measures the intended concepts. As no “gold standard” exists for HRQoL measures in young children, the typical approach employed is hypotheses testing for construct validity, including convergent validity (testing expected relationships with other measurement instruments, also referred to as concurrent validity) and known-group validity (testing expected differences between relevant groups). Responsiveness assesses the instrument’s ability to detect important changes in HRQoL over time. Reliability is a necessary but not sufficient criteria for validity [20]. Validation is context specific. In other words, one instrument may perform very well in discriminating some diseases but not others.

Recently, five HRQoL measures have become available with the potential for cost-utility analysis for children aged 2–4 years old, all with limited validation evidence. They include: the EuroQol Toddler and Infant Populations (for children aged 0–3 years) instrument [21], the Health Status Classification System for Pre-School Children (for children aged 2.5–5 years) [22], Health Utilities Preschool [23] [for children aged 2–4 years, which has been developed from the Health Utilities Index Mark 3 (HUI3)] [9], EQ-5D-Y adapted version, and Child Health Utility 9D (CHU9D) with guidance notes. The EuroQol Toddler and Infant Populations HRQoL instrument were assessed for convergent validity, known-group validity, and test–retest reliability [21, 24]. The Health Status Classification System for Pre-School Children was assessed for feasibility, known-group validity, convergent validity, test–retest reliability, and inter-rater reliability between parents and clinicians [22, 25]. The Health Utilities Preschool instrument was evaluated for inter-rater reliability, construct validity through hypothesis testing, interpretability, and acceptability [23]. The EQ-5D-Y [26] and CHU9D [27] are two established instruments originally for older children. They now have versions with either adapted wording or guidance notes, providing the potential for measurement of HRQoL in young children for cost-utility analysis. While measuring HRQoL for children aged 2–4 years old using instruments with the same constructs of HRQoL as older children would enable consistent HRQoL measurement throughout childhood, there is currently no validation evidence for the two adapted instruments. This current paper focuses on CHU9D.

CHU9D is a concise, generic measure of HRQoL, accompanied by utilities, which was developed specifically for children [28]. It has been well validated for use for children between 5 and 17 years of age, with good feasibility and validity, although relatively poor test–retest reliability [29–32]. CHU9D developers also offered a proxy version with guidance notes for measuring HRQoL for children aged 2–4 years old [33]. However, its psychometric performance remains unclear. The available research on the measurement of HRQoL for young children aged 2–4 years old is rather limited. There is no gold standard instrument for measuring HRQoL for children aged 2–4 years old. There are some nonpreference-based HRQoL measures for children under 5 years old including Infant Toddler Quality of Life Questionnaire [34] and the Pediatric quality of life inventory (PedsQL) 4.0 [35]. Reviews are available on their performance [36, 37]. Although being nonpreference based, they could be a useful comparison in validation studies for health utility measures. More specifically, the PedsQL is widely used and well established, with the toddler version for 2–4 years olds shown to be valid and acceptable for pediatric health research [38–40]. There is no validation evidence for PedsQL toddler version in Australia; however, no HRQoL tool for this age group has been validated in Australia.

The primary objective of this study is to assess the psychometric properties of CHU9D proxy version administered to parents or caregivers of Australian children aged 2–4 years compared with the PedsQL. Specifically, we aim to assess the CHU9D's feasibility, ceiling/floor effects, test–retest reliability, convergent and divergent validity, known-group validity, and responsiveness, compared with the PedsQL. We hypothesize that the CHU9D would show good convergent validity with PedsQL due to their similar constructs. Other tests are exploratory due to little previous evidence for measurement of HRQoL for children aged 2–4 years.

2 Method

2.1 Sample

Survey data were from a large Australian pediatric multi-instrument comparison study (P-MIC) conducted during June 2021 to September 2022; data cut 2 dated 10 August 2022 was used in this study, which includes approximately 94% of the total planned P-MIC participants [41, 42]. Any parent, caregiver, or guardian of a child aged 2–18 years (inclusive) at the time of study enrolment was eligible to take part. We included data from those parents/caregivers of children aged 2–4 years old in the current study. The sample was roughly divided as: (1) generally healthy sample and (2) sample with health condition(s). The generally healthy sample included the online general population sample and those recruited through the hospital who were not receiving care (e.g., small number of siblings of patients or children of staff). The sample with health condition(s) included online disease group samples and those recruited through the hospital who were receiving health care. We compared the characteristics of the generally healthy sample with a similar nationally representative sample, i.e., the Longitudinal Study of Australian Children, to check the general representativeness of our sample.

2.2 Survey

Detailed data collection methods were published elsewhere [41, 42]. Data were collected at two time points: the initial survey and a follow-up. There were two follow-up intervals; the first at 2 days (for a subset of the online general population sample) to assess test–retest reliability and the other at 4 weeks (for the remaining whole sample), mainly to assess responsiveness. Data were collected and stored on REDcap, an online survey system [43].

At the beginning of the initial survey, screening questions were presented to establish the eligibility of participants [42]. Respondents who consented would proceed with the survey. The survey then asked participants about their sociodemographic characteristics including age, gender, language, income, education, and general health status of their child. The survey asked if the child had any chronic conditions that have lasted or are likely to last for 6 months or more. If yes, then the caregivers would be prompted to select listed conditions. Only conditions with sample sizes equal to or larger than 30 were included in the analysis. The next survey section presented multiple HRQoL instruments including CHU9D and PedsQL, with the order of these instruments randomized to minimize order and survey fatigue effects [44]. The order of the instruments was the same for the initial and follow-up survey for each participant. Questions about changes in the child's health status since the first survey were included in the follow-up survey. Time to complete sections of the survey was also recorded on the online REDcap system.

2.3 HRQoL Instruments

The CHU9D has a proxy version with guidance notes for children under 5 years on which parents/caregivers report their child's HRQoL "today" [33]. The CHU9D consists of nine dimensions (worried, sad, pain, tired, annoyed, schoolwork/homework, sleep, daily routine, and able to join in activities), with five levels of responses for each dimension. The developer of CHU9D developed the guidance notes, with input from other health outcome researchers. The guidance notes provide additional instructions and adaptations on how to interpret schoolwork/homework, daily routine, and ability to join in activities questions for children aged under 5 years (Appendix Table S1). In this study, the CHU9D scoring algorithms, developed based on preferences obtained from Australian adolescents, were applied to calculate and report CHU9D utilities, with the UK adult weights used for sensitivity analysis [45, 46]; no specific value set was available for CHU9D proxy version with guidance notes for children under 5 years.

PedsQL^{TM} version 4.0 is an established, standardized, generic profile instrument for nonpreference-based HRQoL measurement for children aged 2–18 years old [39]. The toddler (ages 2–4 years) version contains 21 items and measures four health dimensions: physical, emotional, social, and school functioning (questions related to school or daycare if attended) [39]. The PedsQL toddler version asks, “please tell us how much of a problem each one has been for your child during the past one month.” This was completed by the study child’s parent/caregiver, who rated the frequency of each item in the past month on a 5-point Likert scale from 0 (never) to 4 (almost always). Items were reversed scored and linearly transformed to a 0–100 scale (0 = 100, 1 = 75, 2 = 50, 3 = 25, 4 = 0), with higher scores indicating better HRQoL [39].

2.4 Psychometric Analyses

Several subgroups were defined to facilitate analysis: subgroups defined by variables including general health status (excellent, very good, good, fair, poor), having special health care needs (yes, no), having a chronic health condition (yes, no), or general health status change (much better, somewhat better, about the same, somewhat worse, much worse). More details of classifications are available in the relevant sections below.

2.4.1 Acceptability and Feasibility

Acceptability and feasibility were measured by examining the time taken to complete the survey and respondents’ reported level of difficulty completing the instrument [47]. There were no established criteria for good feasibility. We assumed that it would be acceptable if completion time was less than 5 min, with more than 90% respondents reporting that the survey was “not difficult” to complete for the general population.

2.4.2 Ceiling/Floor Effects

The presence of ceiling and floor effects is often measured by the distribution of responses. The percentages of respondents choosing the highest/lowest levels in all items were calculated, with above 15% commonly considered high ceiling/floor effects [48]. The percentage of respondents choosing the highest level of each item was also calculated, with percentages > 70% considered potentially problematic [49]. The ceiling effect is often more of a concern when it appears in a patient or unwell sample, while less of a concern if present in a healthy sample where good health is expected.

2.4.3 Test–Retest Reliability

Participants who completed the 2-day follow-up survey and reported “about the same” (i.e., no change) on the general health status change indicator question were included when assessing the test–retest reliability. Intraclass correlation coefficients (ICCs), a widely used index for test–retest reliability, were calculated (using an absolute agreement, two-way mixed effects model) for overall scores of instruments [50]. It is suggested that ICC values < 0.5, 0.50–0.74, 0.75–0.90, > 0.90 are indicative of poor, moderate, good, and excellent reliability, respectively [50]. Weighted kappa coefficients were used to evaluate the test–retest reliability of ordinal responses for individual instrument items. These coefficients took into account differences in reported levels within items to provide a more accurate measure of agreement [51]. They were interpreted as follows: ≤ 0.2 for poor agreement, 0.21–0.40 for fair agreement, 0.41–0.60 for moderate agreement, 0.61–0.80 for substantial agreement, and ≥ 0.81 for almost perfect agreement [52]. Additionally, a larger sample (the 4-week follow-up with unchanged health) was used to calculate the weighted kappa and ICCs as a second measure of test–retest reliability.

2.4.4 Convergent and Divergent Validity

As the CHU9D and the PedsQL measure broadly the same concept (i.e., generic health-related quality of life), we hypothesized that their similar prespecified items (e.g., sad versus feeling sad; pain versus having hurts or aches) and overall scores should demonstrate moderate to high correlation (≥ 0.3) [53]. We hypothesized that their unrelated prespecified items (i.e., worried versus lift something; sad versus lift something) should demonstrate weak correlations (< 0.3). Using an a priori consensus method, the study team collaboratively examined various combinations of instrument items to determine whether they anticipated a moderate correlation between an item from CHU9D and a corresponding PedsQL item (to evaluate convergence) or no correlation at all (to evaluate divergence) [42]. These hypotheses were based on the likeness (convergence) or dissimilarity (divergence) of item wording [42]. Spearman’s rank correlation was applied to assess the correlation [54]. We adopted thresholds whereby 0.1–0.29 indicates low, 0.3–0.49 indicates moderate and 0.5 or above indicates high correlation [55].

2.4.5 Known-Group Validity

Known-group validity refers to the extent to which an instrument discriminates between groups with expected health differences. Groups were defined as: (1) children with any chronic health condition (yes/no); (2) children with special health care needs [56] (yes/no); (3) children with relatively poor health defined by general health status of being good, fair or poor (yes/no); and (4) children with a specific chronic condition (yes/no condition, for example, children with autism compared with children without any health condition). The difference between groups was tested using nonparametric Mann–Whitney *U* test as the overall indexes and responses for individual dimensions are not normally distributed [57]. Cohen's *d* between-subject design (mean difference divided by pooled standard deviation) [49] was estimated to assess effect sizes based on standard thresholds, with 0.2 to < 0.5, 0.5 to < 0.8, and 0.8 or more indicating small, medium, and large effect sizes, respectively [58].

2.4.6 Responsiveness

Responsiveness is used to demonstrate the extent to which an instrument's response reflects changes in underlying health status [19]. Caregivers were asked to report their child's general health status change, general health status change related specifically to the initially reported main condition, and health change related to new events occurring during follow up (e.g., new illness or treatment) at the follow-up survey. We identified two subgroups for the analysis of responsiveness: "improved" (answer of "much better") and "worsened" (answers of "somewhat worse" or "much worse" combined because of small sample size). Mean changes in scores between baseline and follow-up were tested by paired *t*-test in each group [59]. One sided *P* values were used as we had specific hypothesis for the direction of the changes [60]. Standard response mean (SRM) or Cohen's *d* within groups is another type of effect size and is widely used to assess responsiveness [61, 62]. The SRM was computed by dividing the mean score change by the standard deviation of the change. The magnitude of responsiveness was evaluated using conventional threshold according to Cohen, with < 0.2 deemed as trivial, 0.2 to < 0.5 as small, 0.5 to < 0.8 as medium, and ≥ 0.8 as large [55]. Both the SRM and Cohen's *d* are methods to calculate effect sizes, but they are typically used in different contexts. SRM is most used for within-group comparisons over time to assess instrument responsiveness, while Cohen's *d* is more versatile and used for between-group as well as within-group comparisons. Cohen's *d* within groups (or paired samples Cohen's *d*) shares the same formula as the SRM, and the two terms are sometimes used interchangeably.

Statistical analyses were performed using Stata version 16 (Statacorp, Texas). Significance levels were set at $P = 0.05$.

3 Results

3.1 Basic Characteristics

The total sample had a generally even distribution of gender and age, with slightly more males (54%) and children aged 4 years (39%). The characteristics of the generally healthy sample were comparable with the estimates from population representative Australian data (Longitudinal Study of Australian Children), except that the study sample had higher parental education and income (Table 1).

3.2 Acceptability and Feasibility

Parents/carers took on average 1.1 and 1.4 min to complete CHU9D and PedsQL, respectively, for the total sample (Appendix Table S2). Most respondents found CHU9D and PedsQL easy to complete, with only 5.5% and 4.8% of the total sample reporting difficulty completing the two instruments, respectively (Appendix Fig. S1).

3.3 Ceiling/Floor Effects

Ceiling effects were not present for CHU9D in the total sample or the sample with special health care needs, with only 12.4% and 6.1% of respondents reporting best levels for all nine dimensions; 15.5% of respondents reported best levels for all nine dimensions in the sample with no special health care needs, just exceeding the ceiling effects threshold. PedsQL did not demonstrate ceiling effects in any sample, with only 3%, 4%, and 1% of respondents reporting best levels for all 21 items in the total sample, the sample with no health care needs, and the sample with special health care needs, respectively. No floor effects were found for any sample. In terms of CHU9D dimensions, pain dimension had over 70% of respondents reporting best level in the total sample (82.30%) and the sample with special health care needs (70.25%). In general, CHU9D had a distribution of different levels of response in the sample with special health care needs and the unwell sample (Fig. 1), which was similarly observed for the PedsQL (Appendix Fig. S2).

3.4 Test–Retest Reliability

The median days between initial and the follow-up survey completion for participants for the 2-day and 4-week follow-up were 3 days and 35 days, respectively. The CHU9D had moderate test–retest reliability overall, with estimated ICCs of 0.52 [95% confidence interval (CI) 0.21, 0.72] and 0.60 (95% CI 0.52, 0.67) for CHU9D Australian utilities in the

Table 1 Baseline characteristics

Baseline characteristics	Total sample (*N* = 842) [*N* (%)]	Generally healthy[b] (*N* = 465) [*N* (%)]	With health condition(s)[c] (*N* = 377) [*N* (%)]	LSAC[a] (%)
Child sex				
Male	453 (53.80)	243 (52.26)	210 (55.70)	51.66
Female	386 (45.84)	222 (47.74)	164 (43.50)	48.34
Other	3 (0.36)		3 (0.80)	
Child age (years)				
2	263 (31.24)	180 (38.71)	83 (22.02)	
3	247 (29.33)	144 (30.97)	103 (27.32)	
4	332 (39.43)	141 (30.32)	191 (50.66)	
Aboriginal or Torress Strait Islander				
No	791 (93.94)	442 (95.05)	349 (92.57)	97.35
Yes	49 (5.82)	23 (4.95)	26 (6.90)	2.65
Prefer not to say	2 (0.24)		2 (0.53)	
Child having a health condition or disability that lasted or are likely to last for 6 months or more				
No	529 (62.83)	368 (79.14)	161 (42.71)	
Yes	313 (37.17)	97 (20.86)	216 (57.29)	
Child having special health care needs				
No	563 (66.86)	391 (84.09)	172 (45.62)	87.26
Yes	279 (33.14)	74 (15.91)	205 (54.38)	12.74
Caregiver education—bachelor's degree or above				
Yes	407 (48.34)	227 (48.82)	180 (47.75)	34.95
No	435 (51.66)	238 (51.18)	197 (52.25)	65.05
Household weekly income before tax				
Less than \$500 per week (\$25,999 or less per year)	40 (4.75)	25 (5.38)	15 (3.98)	5.46
\$500–\$999 per week (\$26,000–\$51,999 per year)	151 (17.93)	81 (17.42)	70 (18.57)	16.81
\$1000–\$1999 per week (\$52,000–\$103,9799 per year)	314 (37.29)	172 (36.99)	142 (37.67)	48.17
\$2000 or more per week (\$104,000 or more per year)	320 (38.00)	183 (39.35)	137 (36.34)	29.57
Missing	17 (2.02)	4 (0.86)	13 (3.45)	
In general, how would you say the study child's current health is?				
Excellent	287 (34.09)	212 (45.59)	75 (19.89)	52.72
Very good	355 (42.16)	198 (42.58)	157 (41.64)	34.32
Good	150 (17.81)	49 (10.54)	101 (26.79)	10.99
Fair	47 (5.58)	5 (1.08)	42 (11.14)	1.86
Poor	3 (0.36)	1 (0.22)	2 (0.53)	0.12

[a] Longitudinal Study of Australian Children (LSAC) is a nationally representative survey of Australian children aged 0 to 18 years old. LSAC estimates here are based on LSAC 2–4 years old and used population weights

[b] The generally healthy sample is composed of the online general population sample and those not receiving health care from the hospital sample

[c] The sample with health condition(s) is composed of online disease groups and those receiving healthcare at The Royal Children's Hospital of the hospital sample

2-day and 4-week follow-ups, respectively. PedsQL also had moderate test–retest reliability, with ICCs of 0.63 (95% CI 0.34, 0.80) and 0.80 (95% CI 0.75, 0.84) for PedsQL total score in the 2-day and 4-week follow-ups, respectively. The 95% confidence intervals for ICCs at 2-day follow-up were wide due to a small sample size of 53 (Appendix Table S3).

The test–retest reliability for individual dimensions were diverse for CHU9D, with four dimensions (worried, pain, annoyed, and schoolwork) having moderate agreement (weighted kappa ranging 0.44–0.48) and the remaining five dimensions (sad, tired, sleep, daily routine, and joining activities) having fair agreement (weighted kappa ranging 0.19–0.29) (Table 2). Results using the 4-week follow-up without health change sample had generally similar or larger agreement except the "worried," "sad," and "pain" dimensions. PedsQL

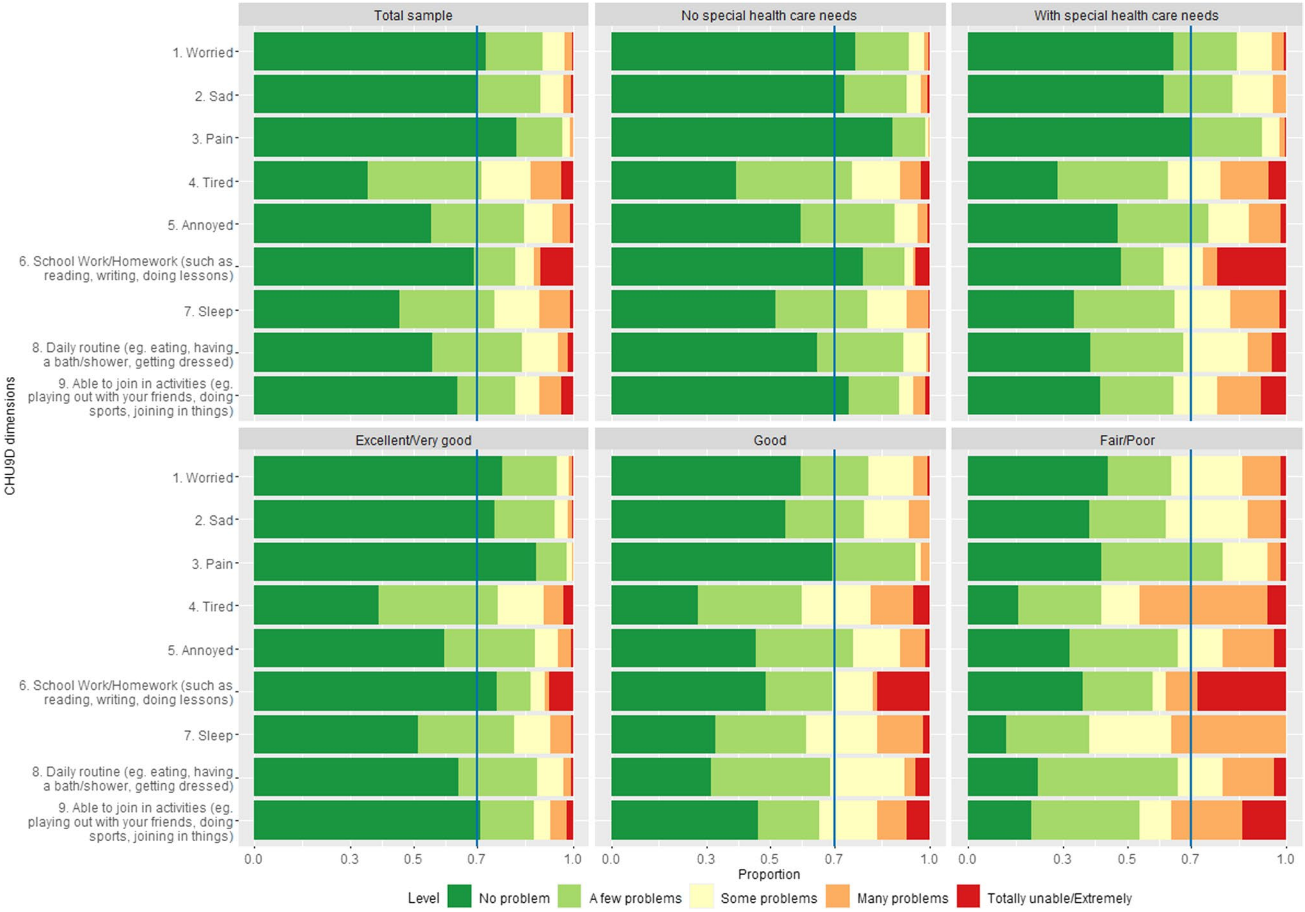

Fig. 1 Distribution of CHU9D response in different samples. (1) In all dimensions, level 1 always indicates the best state of health, while level 5 always indicates the worst state. (2) CHU9D proxy version for under 5 years has same wording as older version, only with added guidance notes for dimensions "school," "daily routine," and "able to join activities" as appropriate for their age. For example, Dimension "school" asks parents to think about activities such as coloring, looking at books/reading, and concentrating, as appropriate for their child's age if their children did not go to any preschool/nursery/kindergarten. (3) The groups are defined by a variable asking about special health care needs (yes, no) and a variable asking about the general health status of the child, with responses of excellent, very good, good, fair, and poor

generally had better test–retest reliability for individual items than CHU9D, with 13 (out of total 21) items demonstrating moderate agreement (kappa above 0.4). PedsQL generally showed similar results using the two follow-ups.

3.5 Convergent and Divergent Validity

As hypothesized, CHU9D utilities strongly correlated with PedsQL total scores ($r = 0.63$). In addition, CHU9D and PedsQL displayed moderate correlations (r = 0.3–0.5) across all hypothesized correlated items, except for "sleep" and "trouble sleeping," which had a high correlation ($r = 0.7$). Weak correlations were found in items hypothesized not to be correlated ($r < 0.3$) (Table 3).

3.6 Known Group Validity

The CHU9D and PedsQL were both able to discriminate between groups with health difference defined as presence versus not of any chronic health conditions, or with special health care needs versus without, or having versus not having good/fair/poor general health status (Table 4). The group mean differences of CHU9D utilities and PedsQL total scores were all significant, with medium-to-large Cohen's d effect sizes. Known-group validity was also tested in 15 specific health conditions identified in this study compared with those with no health conditions. CHU9D performed well in discriminating individual chronic conditions compared with those with no health conditions, with significant utility differences (0.16–0.36) and large effect sizes (0.86–2.03). The top five conditions with largest effect size were behavioral/cognitive/

Table 2 Weighted-kappa of CHU9D dimensions compared with PedsQL for children reporting no health changes at different follow-ups

Dimensions	Dimensions/Items	Weighted kappa (95% CI)	
		2-day follow-up (*N* = 53)	4-week follow up (*N* = 265)
	CHU9D		
	1. Worried	0.45 (0.27, 0.64)	0.27 (0.18, 0.36)
	2. Sad	0.26 (0.07, 0.46)	0.22 (0.13, 0.31)
	3. Pain	0.47 (0.24, 0.70)	0.35 (0.25, 0.45)
	4. Tired	0.21 (0.04, 0.39)	0.32 (0.24, 0.40)
	5. Annoyed	0.44 (0.24, 0.63)	0.38 (0.29, 0.47)
	6. School work	0.48 (0.29, 0.67)	0.44 (0.35, 0.54)
	7. Sleep	0.28 (0.09, 0.48)	0.36 (0.27, 0.45)
	8. Daily routine	0.19 (−0.03, 0.41)	0.50 (0.41, 0.58)
	9. Able to join in activities	0.29 (0.13, 0.45)	0.47 (0.38, 0.56)
	PedsQL		
Physical function	1. Walking	0.61 (0.42, 0.80)	0.61 (0.52, 0.71)
	2. Running	0.59 (0.38, 0.80)	0.60 (0.50, 0.69)
	3. Participating in active play or exercise	0.41 (0.20, 0.61)	0.53 (0.44, 0.62)
	4. Lifting something heavy	0.44 (0.25, 0.64)	0.50 (0.41, 0.59)
	5. Bathing	0.39 (0.19, 0.58)	0.51 (0.42, 0.60)
	6. Helping to pick up his or her toys	0.41 (0.23, 0.59)	0.43 (0.34, 0.51)
	7. Getting aches and pains	0.31 (0.11, 0.50)	0.43 (0.34, 0.52)
	8. Having a low energy level	0.28 (0.07, 0.48)	0.45 (0.36, 0.54)
Emotional function	1. Feeling afraid or scared	0.43 (0.23, 0.63)	0.38 (0.29, 0.46)
	2. Feeling sad	0.36 (0.15, 0.56)	0.37 (0.28, 0.45)
	3. Feeling angry	0.40 (0.23, 0.58)	0.46 (0.38, 0.54)
	4. Having trouble sleeping	0.52 (0.31, 0.72)	0.52 (0.44, 0.59)
	5. Worrying	0.54 (0.35, 0.73)	0.53 (0.45, 0.62)
Social function	1. Playing with other children	0.48 (0.31, 0.65)	0.52 (0.44, 0.61)
	2. Other children not wanting to play with him or her	0.28 (0.10, 0.46)	0.43 (0.34, 0.51)
	3. Getting teased by other children	0.26 (0.05, 0.46)	0.50 (0.41, 0.60)
	4. Not being able to do things that other children his or her age can do	0.43 (0.23, 0.63)	0.66 (0.58, 0.75)
	5. Keeping up when playing with other children	0.35 (0.18, 0.52)	0.57 (0.48, 0.66)
School function*	1. Doing the same school activities as other children his or her age	0.36 (0.19, 0.53)	0.50 (0.40, 0.59)
	2. Missing school because of not feeling well	0.59 (0.37, 0.81)	0.39 (0.30, 0.48)
	3. Missing school to go to the doctor or hospital	0.49 (0.26, 0.71)	0.52 (0.42, 0.62)

Unchanged health is defined using self-reported general health change variable with answer of "about the same." Landis and Koch's guidelines, with coefficients ≤ 0.2: poor agreement, 0.21–0.40: fair agreement, 0.41–0.60: moderate agreement, 0.61–0.80: substantial agreement, and ≥ 0.81: almost perfect agreement. *PedsQL school function is only available for children going to school/kindergarten/preschool (2-day unchanged health: *n* = 46, 46, 44 for school dimensions 1, 2, 3; 4-week unchanged health: *n* = 228, 228, 226 for school dimensions 1, 2, 3)

emotional problems, autism, genetic condition, soiling, and developmental delay. CHU9D had similar or better known-group validity compared with the PedsQL using all different definitions of health differences.

The effect sizes varied across CHU9D and PedsQL dimensions (Appendix Table S4.2). For example, children with anxiety compared with healthy children had a large effect size for "worried" but smaller effect size for "pain." In addition, for both CHU9D and PedsQL, the effect sizes for parents of children aged 2 years old were generally larger than parents of children aged 3 and 4 years (Appendix Table S4.3).

3.7 Responsiveness

In the sample with health condition(s), CHU9D had small effect sizes of responsiveness to general health change and health change to initially reported condition, with a SRM of 0.25–0.30 in the "improved" group and a SRM of 0.21–0.44 in

Table 3 Convergence between CHU9D and PedsQL in total sample

PedsQL	PedsQL	CHU9D								
Dimensions	**Items**	Worried	Sad	Pain	Tired	Annoyed	School	Sleep	Daily routine	Activities
Physical function	Walking	0.13	0.16	0.28	0.19	0.19	0.30	0.18	0.26	0.29
	Running	0.16	0.18	0.25	0.20	0.19	0.29	0.20	0.28	0.32
	Participating in sports activities or exercise	0.18	0.23	0.26	0.25	0.24	0.38	0.24	0.36	**0.45**
	Lifting something	***0.14***	***0.08***	0.20	0.18	0.17	0.26	0.20	0.28	0.28
	Bathing	0.18	0.17	0.22	0.24	0.23	0.34	0.24	**0.43**	0.33
	Helping pick up toys	0.13	0.14	0.21	0.27	0.28	0.32	0.26	**0.41**	0.35
	Having hurts or aches	0.20	0.20	**0.35**	0.28	0.20	0.17	0.24	0.30	0.22
	Low energy levels	0.28	0.23	0.26	**0.34**	0.22	0.23	0.31	0.30	0.29
Emotional function	Feeling afraid or scared	**0.30**	0.27	0.22	0.22	0.22	0.17	0.27	0.30	0.26
	Feeling sad	0.32	**0.43**	0.26	0.28	0.32	0.25	0.23	0.29	0.27
	Feeling angry	0.25	0.26	0.15	0.28	**0.47**	0.27	0.23	0.36	0.32
	Trouble sleeping	0.20	0.25	0.26	0.39	0.24	0.27	**0.70**	0.40	0.26
	Worrying	**0.43**	0.32	0.24	0.28	0.30	0.30	0.27	0.30	0.28
Social function	Playing with other children	0.24	0.25	0.21	0.23	0.27	0.33	0.20	0.35	**0.41**
	Other children not wanting to play with him or her	0.22	0.22	0.12	0.18	0.28	0.37	0.19	0.33	**0.40**
	Getting teased	0.23	0.20	0.13	0.16	0.21	0.30	0.14	0.23	0.24
	Not able to do things that other children their age can do	0.19	0.22	0.24	0.18	0.26	0.44	0.19	**0.44**	**0.46**
	Keeping up when playing with other children	0.16	0.18	0.23	0.18	0.19	0.33	0.17	0.32	**0.35**
School function	Keeping up with school activities	0.16	0.16	0.22	0.16	0.20	**0.37**	0.16	0.35	0.41
	Missing school because not well	0.18	0.17	0.28	0.19	0.16	**0.27**	0.21	0.28	0.27
	Missing school to go to doctor or hospital	0.18	0.19	0.33	0.20	0.20	**0.35**	0.23	0.32	0.32

High correlations, ≥ 0.5 (green); moderate correlations, 0.3–0.49 (yellow); low correlation, < 0.3 (white). All correlation significant at 0.05 level.

Bold indicates expected moderate or high correlations ($r \geq 0.3$) based on highly similar items in line with published technical guide. Bolditalic indicate items hypothesized not to be correlated or weak correlations ($r < 0.3$). Correlation coefficients were calculated by Spearman rank correlation.

the "worsened" group (Table 5). The results in the "worsened" group need to be treated with caution considering the small sample sizes (n = 14 and 16). PedsQL had small effect sizes (SRM 0.26–0.41) in the "improved" group and trivial effect size (SRM 0.15–0.18) in the "worsened" group. The supplementary results demonstrated that the CHU9D was able to reflect health changes in those who reported worsened health when developing new illness, with medium-to-large effect sizes (SRM 0.72-0.82). PedsQL was able to reflect this health change with medium effect size (SRM = 0.50) (Appendix Table S5.2).

The test–retest reliability, known-group validity, and responsiveness results were similar using CHU9D UK weights (Appendix Table S3, Table S4.1, Table S5.1). There were relatively large differences in the mean utilities when using the Australian- and UK-derived CHU9D utility weights for the same groups, which was expected (Appendix Table S4.1).

4 Discussion

4.1 Overview

Our study showed that the CHU9D with guidance notes proxy reported for 2–4 year old Australian children was easy to complete, had no ceiling effects in a sample with special health care needs, had moderate to high correlation with PedsQL prespecified similar items, medium-to-large effect sizes of known-group validity, overall moderate test–retest reliability (with diverse results for individual dimensions), and showed some responsiveness to meaningful health changes over time (with small to large effect sizes using different definitions of health change). Compared with the PedsQL, CHU9D had similar feasibility, known-group validity, responsiveness, and slightly poorer test–retest reliability.

4.2 Distribution of Responses

CHU9D did not exhibit ceiling effects except in the sample with no special health care needs. However, the ceiling effects issue was minor as the percentage of those reporting best levels in all dimensions (15.5%) just exceeded the criteria (15%). In addition, it may be less of a concern as good health was expected in the generally healthy sample with no special health care needs. Most CHU9D dimensions had a good distribution across different levels in the sample of children with impaired health. However, 70.3% of respondents reported the best level for dimension

Table 4 Known group validity (Cohe's *d* effect size) of CHU9D and PedsQL for different health difference groups

Groups	Sample size	CHU9D utilities Australia adolescents (range 0–1, lower utility reflects more health problems)				PedsQL total score (range 0–100, lower score reflects more health problems)			
		Mean	Diff.	*P*-value	Cohen's *d* ES	Mean	Diff.	*P*-value	Cohen's *d* ES
Any medical condition or disabilities lasting for 6 months or more	Yes = 313	0.65	− 0.13	< 0.001	0.58	69.03	− 12.12	< 0.001	0.74
	No = 529	0.78				81.15			
Special health care needs	Yes = 279	0.62	− 0.16	< 0.001	0.75	67.19	− 14.14	< 0.001	0.88
	No = 563	0.78				81.33			
General health status (good/fair/poor)	Yes = 200	0.57	− 0.21	< 0.001	1.02	65.24	− 14.96	< 0.001	0.92
	No = 642	0.78				80.20			
Healthy (children with no chronic conditions, as comparison)	267	0.84				84.25			
Behavioral, cognitive, emotional problems	98	0.49	− 0.36	< 0.001	2.03	60.15	− 24.10	< 0.001	1.57
Autism	54	0.48	− 0.36	< 0.001	2.00	55.56	− 28.69	< 0.001	1.95
Genetic condition	30	0.50	− 0.34	< 0.001	1.92	58.39	− 25.86	< 0.001	1.66
Soiling	33	0.51	− 0.33	< 0.001	1.91	59.87	− 24.39	< 0.001	1.58
Developmental delay	96	0.53	− 0.31	< 0.001	1.69	58.29	− 25.96	< 0.001	1.63
Bone, joint or muscle problem	40	0.56	− 0.28	< 0.001	1.59	61.60	− 22.65	< 0.001	1.43
ADHD	64	0.56	− 0.28	< 0.001	1.56	65.11	− 19.15	< 0.001	1.27
Ear infection	44	0.59	− 0.25	< 0.001	1.45	71.58	− 12.67	< 0.001	0.83
Sleep problems	190	0.58	− 0.26	< 0.001	1.37	67.37	− 16.88	< 0.001	1.06
Anxiety	42	0.60	− 0.24	< 0.001	1.34	61.87	− 22.38	< 0.001	1.47
Constipation	62	0.60	− 0.24	< 0.001	1.27	65.96	− 18.29	< 0.001	1.12
Hay fever	56	0.64	− 0.20	< 0.001	1.18	69.69	− 14.56	< 0.001	0.98
Asthma	108	0.68	− 0.16	< 0.001	0.87	73.89	− 10.36	< 0.001	0.69
Eczema	149	0.68	− 0.16	< 0.001	0.87	74.58	− 9.67	< 0.001	0.61
Food or digestive allergies	84	0.68	− 0.16	< 0.001	0.86	74.60	− 9.65	< 0.001	0.62

The diseases included in this table have sample sizes ≥ 30. Standard thresholds 0.2 to < 0.5, 0.5 to < 0.8, and 0.8 or more denote small, medium, and large effect sizes, respectively. Cohen's *d* ES (effect sizes): the numerator is the difference between means of the two groups, and the denominator is the pooled standard deviation. Healthy sample = with no chronic condition

P-values are obtained from Man–Whitney *U* test as scores were not normally distributed. The health conditions are ordered from large to small by their effect sizes calculated using Australia adolescent weights

"pain" even in the sample with special health care needs, which indicates that this item may not distinguish children well. This is consistent with the results from the EuroQoL Toddler and Infant Populations, with 73% and 88% respondents reporting the best level for "pain" in acute and chronic condition samples [21]. This may be because pain is infrequent for young children or is difficult for parent/caregivers to observe. Recommended observable pain related behavior in young children includes grimacing, restless movement, and inconsolable crying [21]. These, or other behaviors, could be added as guidance notes for the "pain" dimension in CHU9D to help improve the sensitivity of this item.

4.3 Test–Retest Reliability

In our study, CHU9D shows overall moderate test–retest reliability, with ICCs of 0.52 and 0.60 for 2-day and 4-week follow-ups, although the reliability for individual dimensions were more diverse (kappa 0.19–0.47). The test–retest reliability is similar or slightly poorer compared with previous studies of CHU9D or other similar pediatric HRQoL measures in older children [21, 26, 30, 53]. For example, Yang et al. found an ICC of 0.653 for the CHU9D utility score, and kappa estimates ranging 0.20–0.53 for different CHU9D dimensions in 232 school children aged 8–17 years old who completed a retest survey 2 weeks post the initial survey [53]. Ravens et al. found satisfactory ICC (0.82–0.83) and fair-to-moderate kappa estimates up to 0.67 in children aged 8-19 years old who completed the retest for EQ-5D-Y 7–10 days after the first examination [26].

The kappa should be interpreted with caution as it is also impacted by other factors such as the distribution of different levels for each dimension [63]. ICC results also relate to the variation in participant characteristics and study sample

Table 5 Responsiveness of CHU9D and PedsQL in sample with health condition(s)

Index	Health status change[c]	Sample size	Baseline (mean, SD)	At 4-week follow-up (mean, SD)	Paired difference (mean, SD)	P[d]	SRM[a]
General health change[b]							
CHU9D utility Australia adolescents (higher score reflects better health)	Improved	33	0.76 (0.23)	0.83 (0.23)	0.07 (0.26)	0.080	0.25
	Worsened	14	0.42 (0.22)	0.32 (0.21)	− 0.09 (0.22)	0.063	−0.44
PedsQL total score (higher score reflects better health)	Improved	33	74.83 (21.51)	80.88 (17.59)	6.05 (14.58)	0.012	0.41
	Worsened	14	53.97 (18.89)	50.48 (27.28)	− 3.49 (22.67)	0.287	−0.15
Health change related to initially reported condition[2]							
CHU9D utility Australia adolescents (higher score reflects better health)	Improved	32	0.76 (0.23)	0.84 (0.20)	0.08 (0.26)	0.052	0.30
	Worsened	16	0.44 (0.23)	0.38 (0.24)	− 0.06 (0.27)	0.207	−0.21
PedsQL total score (higher score reflects better health)	Improved	32	78.22 (20.75)	82.11 (17.08)	3.88 (14.66)	0.072	0.26
	Worsened	16	54.53 (22.57)	50.98 (26.64)	− 3.56 (20.13)	0.245	−0.18

[a] SRM, standard response mean; dividing the mean score change (i.e., follow-up minus baseline) by the standard deviation of the change. The interpretation for SRM were defined as trivial for < 0.2, small for ≥ 0.2 and < 0.5, medium for > 0.5 and < 0.8, and large for ≥ 0.8

[b] General health change: how would you rate the study child's health in general now? Health change related to initially reported condition: thinking about the study child's main health condition, how would you say this is going now compared with when you completed the first survey for this study? (with answers: much better, somewhat better, about the same, somewhat worse, much worse)

[c] "improved" includes *much better*, "worsened" includes *somewhat worse* and *much worse*

[d] *P*-values were one sided *P* from paired *t*-test

sizes [50, 64]. Similarly, Ravens et al. reported concerns that high ceiling effects in EQ-5D-Y impacted the test–retest reliability results and that the kappa coefficient was of limited value (kappa = −0.003) as nearly all retest responses were in the "no problems" category [26]. As the 2-day follow-up retest sample in our study was only from the online general population [41], the lack of variance of responses and high ceiling effects might contribute to the low kappa and ICC estimates.

4.4 Convergent and Divergent Validity

The CHU9D displayed convergent validity with PedsQL, confirming that the same latent construct of HRQoL was being measured by these two instruments. Our correlation coefficients (0.34–0.70 for similar items and 0.62–0.65 for overall scores) were generally similar with previous studies, with some slight differences. Petersen et al. found that correlations between CHU9D and PedsQL for related dimensions and overall scores were 0.40–0.50 and 0.69, respectively, for a Danish high school student sample and 0.28–0.46 and 0.63, respectively, for an Australian adolescent sample [65, 66]. Our stronger correlation coefficients compared with previous studies may be because previous studies calculated correlations between CHU9D items with PedsQL summary functions instead of with PedsQL individual items. Only a small number of potentially divergent items were prespecified and divergence was identified for each (PedsQL lifting something and CHU9D sad, worried). More item pairs could have been selected for divergence (such as bathing/picking up toys and sad/worried) however it was felt that in children 2–4 years of age both bathing and chores could be accompanied by an emotional response, especially with a "today" recall period for the CHU9D. It is also worth noting that CHU9D and PedsQL have different recall periods; CHU9D asks about today while PedsQL asks about the past month, which may reduce the correlation between similar constructs.

4.5 Known-Group Validity

The CHU9D was able to discriminate between groups with known health differences, with medium-to-large effect sizes, regardless of which scoring algorithm was applied. The utility difference between those "with and without chronic conditions or disabilities" was 0.13 using an Australian adolescent algorithm and 0.07 using the UK adult scoring algorithm, with differences similar to previous studies [65, 66]. Peterson et al. found that the utility differences between "with and without chronic conditions or disabilities" in a Danish high school student sample were 0.11 and 0.06 for Australian adolescent and UK adult scoring algorithms, respectively [65]. In a similar study conducted with an Australian adolescent sample using Australian adolescent weights, the utility difference between "with and without chronic conditions or disabilities" is 0.15 [66]. Neither of the prior studies reported Cohen's *d* effect sizes; however, the utility differences are similar to our findings. This suggests that CHU9D may have comparable known-group validity in children 2–4 years old compared with older age groups.

CHU9D utilities showed large effect sizes (range 0.86–2.13) for 15 health conditions (identified in this study with sample sizes larger than 30) compared with those reported no conditions, indicating that CHU9D can be applied in a variety of disease groups with good known-group validity in children aged 2–4 years old. There was a large difference in mean utilities when different value sets are applied. Nevertheless, the effect sizes for known-group validity remained very similar between the two value sets, emphasizing that the conclusion was not influenced by the choice of the value set.

4.6 Responsiveness

In our study, CHU9D demonstrated responsiveness to health changes over time, with mainly small effect sizes (SRM 0.25–0.44) according to different definitions of health change in 2–4-year-olds, except in those who developed new illness where large effect size was found (SRM 0.82). To our best knowledge, only one study has investigated the responsiveness of CHU9D, with no studies investigating young children. Wolf et al. (2021) examined the responsiveness of the proxy-reported CHU9D in 396 Danish children aged 6–15 years with mental health problems and found a SRM of 0.634–0.654 for children who experienced clinically significant improvements [67]. Our study had smaller magnitude of responsiveness in terms of SRM (0.25–0.55) for those self-reporting changes in general health status, although it was difficult clearly understanding why there was a change in health. The magnitude of responsiveness for those developing new illnesses was instead much larger (SRM 0.72–0.82); the results needed to be treated with caution considering the small sample size. This suggests that the context of health change may matter in assessing responsiveness and caution should be paid to the comparability of responsiveness between different studies or instruments.

4.7 Implications and Limitations

Our study provides consistent measurement of child health using the CHU9D across child age which could be important for measurement within pediatric clinical trials or in routine clinical care. Further development and validation work is warranted given the limitations and discussion as below.

Several limitations have been identified. First, missing data were not permitted for CHU9D based on a structural decision to not allow skipping items. This had its advantage such as reducing the percentage of missing data but might have forced people to randomly select an answer even when they thought the answer was not rational or suitable. We thus lacked the ability to assess the content validity of CHU9D for this age group through observation of missing data. Canaway et al found no missing values in CHU9D responses with interviewer-administered data collection (questions being read to the child) in slightly older children aged 6–7 years, which reduces our concerns [30]. Despite having good psychometric information on the CHU9D with guidance notes we are unable to determine the impact that the guidance notes themselves had on respondent's cognitive processing. This could usefully be explored in a follow-up study. Another limitation is that the sample size for the 2-day follow-up test–retest reliability was only 53 respondents from the general population. Despite being small this is still deemed adequate according to consensus-based standards for the selection of health measurement instruments (COSMIN) study design checklist [19]. In the responsiveness analysis, the sample sizes of the groups reporting health changes were also small, especially for the "worsened" group. However, this evidence is difficult to obtain given the low probability of serious health states and worsening health in children and in those populations the low tolerance for survey burden. Further studies in clinical studies with populations having severe health states or studies targeting children aged 2–4 years old with larger sample sizes might be beneficial. The *P* value of the paired difference may be of limited value considering the small sample sizes in some subgroups and therefore the SRM results were mainly reported. The SRM provided useful indication of potential effect sizes of responsiveness of CHU9D for future users. However, it is acknowledged that it might not be appropriate to report effect sizes if the differences were not significant and the effect sizes for nonsignificant differences were shown for illustrative purpose only. There are potential methodological limitations in applying scoring algorithms developed for older children to calculate CHU9D utilities in children aged 2–4 years old. For example, the preferences for health states of different age groups may differ. However, there is no alternative until a value set for this young age group is developed or the validity of the existing value set is confirmed for this purpose. There is a need to understand and test appropriate preference-weighted scoring for this instrument in this age group, which will further allow utility values to be accurately and consistently produced by the CHU9D in children as young as 2 years old for economic evaluation. Obtaining preference-weighted scores for CHU9D proxy version with guidance notes or developing mapping algorithms to other existing scoring systems could be important next steps to facilitate use of the CHU9D in economic evaluation and resultant policy decisions for this age group.

There is an ongoing debate on the validity of proxy-reported HRQoL, particularly due to poor agreement between self-report by older children and proxy-report by adults [68]. While proxy reports are discouraged when children can self-report, they remain the only option for very young or cognitively challenged individuals. Parents of young children under 5 years old, who usually spend more time caring for their children, may serve as better proxies due to their close observations and connections. There is evidence that agreement is stronger in the youngest age group (5.5–6.5 years) than older age groups (6.5–8.5 years) [68]. Concerns regarding proxy-report, especially for more subjective dimensions such as "worried", "sad," and "pain" may be addressed by including externally observable indicators. Evaluating the validity of proxy HRQoL measures is controversial. Nevertheless, the pressure to include young children and their QALYs for cost-utility analysis and the existence of valid preference-based measures for older children continues to underscore the practical value of these investigations for younger children.

It is acknowledged that children under 5 years of age may have different health dimensions of HRQoL and it may not be suitable to directly apply HRQoL measures designed for older children to this younger age group. This study was unable to evaluate the fundamental construct validity of CHU9D to measure HRQOL for this 2–4-year-old age group, i.e., to explore whether the included dimensions were appropriate and/or whether dimensions were missing. Developing a new instrument that incorporates literature reviews and qualitative research would be the ideal way to guarantee the appropriate construct of HRQoL for a new age group [28]. However, the time and expenses associated with this development task mean that it is worthwhile to better understand the performance of existing options and smaller modifications.

Adding guidance notes is assumed to enhance the applicability of CHU9D for children under 5 years old. However, uncertainty remains regarding its suitability. While it is ideal to conduct qualitative research to assess the content validity of these guidance notes first, in this case, we proceeded with testing as the CHU9D with guidance notes are already widely in use. Our study serves a crucial role in evaluating these guidance notes relative to the validated but nonpreference-based PedsQL, with findings offering valuable insights to further refine CHU9D to better suit this age group. Future qualitative research aimed at testing and improving the CHU9D would be highly beneficial.

5 Conclusion

CHU9D proxy version with guidance notes demonstrated good psychometric performance overall for measuring HRQoL for 2–4-year-old Australian children and shows potential as a valid and reliable instrument for assessing the HRQoL for this population.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40273-024-01355-1.

Acknowledgements We thank all participants who completed the questionnaire and Harriet Hiscock, Kristy McGregor, Shilana Yip, and other members of the P-MIC project team in the QUOKKA project who have designed and advised on recruitment, data collection, and study design.

Declarations

Funding/Support Open Access funding enabled and organized by CAUL and its Member Institutions. This study was supported by the Australian Medical Research Future Fund (grant number MRF1200816). X.X. is supported by the China Scholarship Council (201906010310). R.J. is supported by a EuroQol Research Foundation PhD grant (361-RA, 330-PhD).

Role of the Funder/Sponsor The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or the decision to submit the manuscript for publication.

Competing Interests K.D. reports grants from the Medical Research Futures Fund in Australia and the EuroQol Research Foundation during the conduct of the study and has received other related research grants from the EuroQol Foundation who owns some of the instruments not included in this paper but also collected in the P-MIC study. N.D. and B.M. are members of the EuroQol Group, and B.M. is chair of the descriptive systems working group.

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

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

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

Consent for Publication Not applicable.

Code Availability Not applicable.

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

Author Contributions Concept and design: KD; acquisition of data: RJ, ND, BM, KD; analysis and interpretation of data: all authors; drafting of the manuscript: XX; critical revision of the paper for important intellectual content: all authors; statistical analysis: XX; provision of study materials or patients: KD; obtaining funding: ND, BM, KD; administrative, technical, or logistic support: RJ, KD. Supervision: KD, NC, LH, ND, GC, and BM.

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. Freed GL, Gafforini S, Carson N. Age distribution of emergency department presentations in Victoria. Emerg Med Australas. 2015;27(2):102–7.
2. Shaker M, Chan ES, Protudjer JLP, Soller L, Abrams EM, Greenhawt M. The cost-effectiveness of preschool peanut oral immunotherapy in the real-world setting. J Allergy Clin Immunol Pract. 2021;9(7):2876-2884.e2874.
3. Wang L, Congdon N, Hogg RE, Zhang S, Li M, Shi Y, Jin L, He F, Wang H, Boswell M, Iyer M. The cost-effectiveness of alternative vision screening models among preschool children in rural China. Acta Ophthalmol. 2019;97(3):e419–25.
4. Tanaka M, Okubo R, Hoshi S-L, Ishikawa N, Kondo M. Cost-effectiveness of pertussis booster vaccination for preschool children in Japan. Vaccine. 2022;40(7):1010–8.
5. Sullivan SM, Tsiplova K, Ungar WJ. A scoping review of pediatric economic evaluation 1980–2014: do trends over time reflect changing priorities in evaluation methods and childhood disease? Expert Rev Pharmacoecon Outcomes Res. 2016;16(5):599–607.
6. Pediatric Economic Database Evaluation (PEDE). Trends in economic evaluation. http://pede.ccb.sickkids.ca/pede/trends.jsp. Accessed 1 May 2023.
7. Kromm SK, Bethell J, Kraglund F, Edwards SA, Laporte A, Coyte PC, Ungar WJ. Characteristics and quality of pediatric cost-utility analyses. Quality Life Res. 2012;21(8):1315–25.
8. Rowen D, Rivero-Arias O, Devlin N, Ratcliffe J. Review of valuation methods of preference-based measures of health for economic evaluation in child and adolescent populations: where are we now and where are we going? Pharmacoecon. 2020;38(4):325–40.
9. Kwon J, Freijser L, Huynh E, Howell M, Chen G, Khan K, Daher S, Roberts N, Harrison C, Smith S, Devlin N, Howard K, Lancsar E, Bailey C, Craig J, Dalziel K, Hayes A, Mulhern B, Wong G, Ratcliffe J, Petrou S. Systematic review of conceptual, age, measurement and valuation considerations for generic multidimensional childhood patient-reported outcome measures. Pharmacoecon. 2022;40(4):379–431.
10. Kwon J, Smith S, Raghunandan R, Howell M, Huynh E, Kim S, Bentley T, Roberts N, Lancsar E, Howard K, Wong G, Craig J, Petrou S. Systematic review of the psychometric performance of generic childhood multi-attribute utility instruments. Appl Health Econ Health Policy. 2023;21:559–84.
11. Kwon J, Kim SW, Ungar WJ, Tsiplova K, Madan J, Petrou S. Patterns, trends and methodological associations in the measurement and valuation of childhood health utilities. Qual Life Res. 2019;28(7):1705–24.
12. Wolstenholme JL, Bargo D, Wang K, Harnden A, Räisänen U, Abel L. Preference-based measures to obtain health state utility values for use in economic evaluations with child-based populations: a review and UK-based focus group assessment of patient and parent choices. Qual Life Res. 2018;27(7):1769–80.
13. Kind P, Klose K, Gusi N, Olivares PR, Greiner W. Can adult weights be used to value child health states? Testing the influence of perspective in valuing EQ-5D-Y. Quality Life Res. 2015;24(10):2519–39.
14. Verstraete J, Ramma L, Jelsma J. Item generation for a proxy health related quality of life measure in very young children. Health Qual Life Outcomes. 2020;18(1):11.
15. Kreimeier S, Oppe M, Ramos-Goñi JM, Cole A, Devlin N, Herdman M, Mulhern B, Shah KK, Stolk E, Rivero-Arias O, Greiner W. Valuation of EuroQol Five-Dimensional Questionnaire, Youth Version (EQ-5D-Y) and EuroQol Five-Dimensional Questionnaire, Three-Level Version (EQ-5D-3L) health states: the impact of wording and perspective. Value Health. 2018;21(11):1291–8.
16. Nancy Devlin RN, Ratcliffe J, Mulhern B, Dalziel K, Chen G, Viney R. Do child QALYs = adult QALYs? Five reasons why they might not. 2020. https://www.ohe.org/news/do-child-qalys-adult-qalys-five-reasons-why-they-might-not. Accessed 1 May 2023.
17. Department of Health and Aged Care, A. G. Preventive and Public Health Research initiative. From https://www.health.gov.au/initiatives-and-programs/preventive-and-public-health-research-initiative. Accessed 1 May 2023.
18. Ungar WJ, Prosser LA, Burnett HF. Values and evidence colliding: health technology assessment in child health. Expert Rev Pharmacoecon Outcomes Res. 2013;13(4):417–9.
19. Mokkink LB, Prinsen CA, Patrick DL, Alonso J, Bouter LM, De Vet H, Terwee CB. COSMIN study design checklist for patient-reported outcome measurement instruments. Amsterdam; 2019. p. 1–32.
20. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med. 2006;119(2):166.e167-166.e116.
21. Verstraete J, Ramma L, Jelsma J. Validity and reliability testing of the Toddler and Infant (TANDI) Health Related Quality of Life instrument for very young children. J Patient Rep Outcomes. 2020;4(1):94.
22. Saigal S, Rosenbaum P, Stoskopf B, Hoult L, Furlong W, Feeny D, Hagan R. Development, reliability and validity of a new measure of overall health for pre-school children. Qual Life Res. 2005;14(1):243–57.
23. Furlong W, Rae C, Feeny D, Ghotra S, Breakey VR, Carter T, Pai N, Pullenayegum E, Xie F, Barr R. Generic health-related quality of life utility measure for preschool children (health utilities preschool): design, development, and properties. Value Health. 2022;26:251–60.
24. Verstraete J, Amien R. Cross-cultural adaptation and validation of the EuroQoL Toddler and Infant Populations Instrument into Afrikaans for South Africa. Value Health Reg Issues. 2023;35:78–86.
25. Fang X, Bai G, Windhorst DA, Feeny D, Saigal S, Duijts L, Jaddoe VWV, Hu S, Jin C, Raat H. Feasibility and validity of the Health Status Classification System-Preschool (HSCS-PS) in a large community sample: the Generation R study. BMJ Open. 2018;8(12): e022449.
26. Ravens-Sieberer U, Wille N, Badia X, Bonsel G, Burström K, Cavrini G, Devlin N, Egmar A-C, Gusi N, Herdman M, Jelsma J, Kind P, Olivares PR, Scalone L, Greiner W. Feasibility, reliability, and validity of the EQ-5D-Y: results from a multinational study. Qual Life Res. 2010;19(6):887–97.
27. 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. Quality Life Res. 2012;21(4):717–25.
28. Stevens KJ. Working with children to develop dimensions for a preference-based, generic, pediatric, health-related quality-of-life measure. Qual Health Res. 2010;20(3):340–51.
29. Frew EJ, Pallan M, Lancashire E, Hemming K, Adab P. Is utility-based quality of life associated with overweight in children? Evidence from the UK WAVES randomised controlled study. BMC Pediatr. 2015;15(1).
30. Canaway AG, Frew EJ. Measuring preference-based quality of life in children aged 6–7 years: a comparison of the performance

of the CHU-9D and EQ-5D-Y–the WAVES pilot study. Qual Life Res. 2013;22(1):173–83.

31. 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 Health. 2012;15(8):1092–9.
32. Stevens K. Assessing the performance of a new generic measure of health-related quality of life for children and refining it for use in health state valuation. Appl Health Econ Health Policy. 2011;9(3):157–69.
33. Measuring & Valuing Health. A brief overview of the Child Health Utility 9D (CHU9D). https://licensing.sheffield.ac.uk/product/CHU-9D. Accessed 1 May 2023.
34. Klassen AF, Landgraf JM, Lee SK, Barer M, Raina P, Chan HW, Matthew D, Brabyn D. Health related quality of life in 3 and 4 year old children and their parents: preliminary findings about a new questionnaire. Health Qual Life Outcomes. 2003;1:81.
35. Varni JW, Seid M, Kurtin PS. PedsQL 4.0: reliability and validity of the Pediatric Quality of Life Inventory version 4.0 generic core scales in healthy and patient populations. Med Care. 2001;39(8):800–12.
36. Solans M, Pane S, Estrada M-D, Serra-Sutton V, Berra S, Herdman M, Alonso J, Rajmil L. Health-related quality of life measurement in children and adolescents: a systematic review of generic and disease-specific instruments. Value Health. 2008;11(4):742–64.
37. Paltzer J, Barker E, Witt WP. Measuring the health-related quality of life (HRQoL) of young children in resource-limited settings: a review of existing measures. Qual Life Res. 2013;22(6):1177–87.
38. Gheissari A, Farajzadegan Z, Heidary M, Salehi F, Masaeli A, Mazrooei A, Varni JW, Fallah Z, Zandieh F. Validation of Persian version of PedsQL™ 4.0™ generic core scales in toddlers and children. Int J Prev Med. 2012;3(5):341–50.
39. 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.
40. Buck D. The PedsQL™ as a measure of parent-rated quality of life in healthy UK toddlers: Psychometric properties and cross-cultural comparisons. J Child Health Care. 2012;16(4):331–8.
41. Jones R, Mulhern B, McGregor K, Yip S, O'Loughlin R, Devlin N, Hiscock H, Dalziel K, On Behalf Of The Quality Of Life In Kids Key Evidence To Strengthen Decisions In Australia Quokka Project, T. Psychometric performance of hrqol measures: an Australian paediatric multi-instrument comparison study protocol (P-MIC). Children (Basel, Switzerland). 2021;8(8):714.
42. Renee Jones BM, Devlin N, Hiscock H, Chen G, O'Louglin R, Xiong X, Bahrampour M, McGregor K, Yip S, Dalziel K, on behalf of the Quality Of Life in Kids Key evidence to strengthen decisions in Australia (QUOKKA) project team. Australian Paediatric Multi-Instrument Comparison (P-MIC) Study: Technical Methods Paper [Online]. 2023. https://www.quokkaresearchprogram.org/project-1-1. Accessed 1 May 2023.
43. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, McLeod L, Delacqua G, Delacqua F, Kirby J, Duda SN. The REDCap consortium: Building an international community of software platform partners. J Biomed Inform. 2019;95: 103208.
44. Breakwell GM, Hammond SE, Fife-Schaw CE, Smith JA. Research methods in psychology. Sage Publications Inc.; 2006.
45. Ratcliffe J, Flynn T, Terlich F, Stevens K, Brazier J, Sawyer M. Developing adolescent-specific health state values for economic evaluation. Pharmacoecon. 2012;30(8):713–27.
46. Stevens K. Valuation of the child health utility 9D index. Pharmacoecon. 2012;30(8):729–47.
47. Dalziel K, Catchpool M, Garcia-Lorenzo B, Gorostiza I, Norman R, Rivero-Arias O. Feasibility, validity and differences in adolescent and adult eq-5d-y health state valuation in Australia and Spain: an application of best-worst scaling. Pharmacoeconomics. 2020;38:499–513.
48. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, Bouter LM, de Vet HC. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60(1):34–42.
49. Peasgood T, Mukuria C, Brazier J, Marten O, Kreimeier S, Luo N, Mulhern B, Greiner W, Pickard AS, Augustovski F, Engel L, Gibbons L, Yang Z, Monteiro AL, Kuharic M, Belizan M, Bjørner J. Developing a new generic health and wellbeing measure: psychometric survey results for the EQ health and wellbeing. Value Health. 2022;25:525–33.
50. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155–63.
51. Brenner H, Kliebsch U. Dependence of weighted kappa coefficients on the number of categories. Epidemiol. 1996;7(2):199–202.
52. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–74.
53. Yang P, Chen G, Wang P, Zhang K, Deng F, Yang H, Zhuang G. Psychometric evaluation of the Chinese version of the Child Health Utility 9D (CHU9D-CHN): a school-based study in China. Qual Life Res. 2018;27(7):1921–31.
54. Daniel WW, Cross CL. Biostatistics: a foundation for analysis in the health sciences. Wiley; 2018.
55. Cohen J. Statistical power analysis for the behavioral sciences. Routledge; 2013.
56. Huang L, Freed GL, Dalziel K. Children with special health care needs: how special are their health care needs? Acad Pediatr. 2020;20(8):1109–15.
57. Hollander M, Wolfe DA, Chicken E. Nonparametric statistical methods. Wiley; 2013.
58. Cohen J. A power primer. Psychol Bull. 1992;112(1):155–9.
59. Xu M, Fralick D, Zheng JZ, Wang B, Tu XM, Feng C. The differences and similarities between two-sample t-test and paired t-test. Shanghai Arch Psychiatry. 2017;29(3):184–8.
60. Ludbrook J. Should we use one-sided or two-sided P values in tests of significance? Clin Exp Pharmacol Physiol. 2013;40(6):357–61.
61. Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. J Clin Epidemiol. 2000;53(5):459–68.
62. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front Psychol. 2013;4.
63. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther. 2005;85(3):257–68.
64. Froberg DG, Kane RL. Methodology for measuring health-state preferences—II: Scaling methods. J Clin Epidemiol. 1989;42(5):459–71.
65. Petersen KD, Ratcliffe J, Chen G, Serles D, Frosig CS, Olesen AV. The construct validity of the Child Health Utility 9D-DK instrument. Health Qual Life Outcomes. 2019;17(1):187.
66. Petersen KD, Chen G, Mpundu-Kaambwa C, Stevens K, Brazier J, Ratcliffe J. Measuring health-related quality of life in adolescent populations: an empirical comparison of the CHU9D and the PedsQL(TM) 4.0 Short Form 15. Patient. 2018;11(1):29–37.
67. Wolf RT, Ratcliffe J, Chen G, Jeppesen P. The longitudinal validity of proxy-reported CHU9D. Qual Life Res. 2021;30(6):1747–56.
68. Cremeens J, Eiser C, Blades M. Factors influencing agreement between child self-report and parent proxy-reports on the Pediatric Quality of Life Inventory™ 4.0 (PedsQL™) generic core scales. Health Qual Life Outcomes. 2006;4(1):58.