



# Systematic Review of the Psychometric Performance of Generic Childhood Multi-attribute Utility Instruments

Joseph Kwon<sup>1</sup> · Sarah Smith<sup>2</sup> · Rakhee Raghunandan<sup>3</sup> · Martin Howell<sup>3</sup> · Elisabeth Huynh<sup>4</sup> ·  
Sungwook Kim<sup>1</sup> · Thomas Bentley<sup>5</sup> · Nia Roberts<sup>6</sup> · Emily Lancsar<sup>4</sup> · Kirsten Howard<sup>3</sup> · Germaine Wong<sup>3</sup> ·  
Jonathan Craig<sup>7</sup> · Stavros Petrou<sup>1</sup>

Accepted: 21 March 2023 / Published online: 3 May 2023  
© The Author(s) 2023

## Abstract

**Background** Childhood multi-attribute utility instruments (MAUIs) can be used to measure health utilities in children (aged  $\leq 18$  years) for economic evaluation. Systematic review methods can generate a psychometric evidence base that informs their selection for application. Previous reviews focused on limited sets of MAUIs and psychometric properties, and only on evidence from studies that directly aimed to conduct psychometric assessments.

**Objective** This study aimed to conduct a systematic review of psychometric evidence for generic childhood MAUIs and to meet three objectives: (1) create a comprehensive catalogue of evaluated psychometric evidence; (2) identify psychometric evidence gaps; and (3) summarise the psychometric assessment methods and performance by property.

**Methods** A review protocol was registered with the Prospective Register of Systematic Reviews (PROSPERO; CRD42021295959); reporting followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guideline. The searches covered seven academic databases, and included studies that provided psychometric evidence for one or more of the following generic childhood MAUIs designed to be accompanied by a preference-based value set (any language version): 16D, 17D, AHUM, AQoL-6D, CH-6D, CHSCS-PS, CHU9D, EQ-5D-Y-3L, EQ-5D-Y-5L, HUI2, HUI3, IQI, QWB, and TANDI; used data derived from general and/or clinical childhood populations and from children and/or proxy respondents; and were published in English. The review included ‘direct studies’ that aimed to assess psychometric properties and ‘indirect studies’ that generated psychometric evidence without this explicit aim. Eighteen properties were evaluated using a four-part criteria rating developed from established standards in the literature. Data syntheses identified psychometric evidence gaps and summarised the psychometric assessment methods/results by property.

**Results** Overall, 372 studies were included, generating a catalogue of 2153 criteria rating outputs across 14 instruments covering all properties except predictive validity. The number of outputs varied markedly by instrument and property, ranging from 1 for IQI to 623 for HUI3, and from zero for predictive validity to 500 for known-group validity. The more recently developed instruments targeting preschool children (CHSCS-PS, IQI, TANDI) have greater evidence gaps (lack of any evidence) than longer established instruments such as EQ-5D-Y, HUI2/3, and CHU9D. The gaps were prominent for reliability (test–retest, inter-proxy-rater, inter-modal, internal consistency) and proxy-child agreement. The inclusion of indirect studies ( $n = 209$  studies;  $n = 900$  outputs) increased the number of properties with at least one output of acceptable performance. Common methodological issues in psychometric assessment were identified, e.g., lack of reference measures to help interpret associations and changes. No instrument consistently outperformed others across all properties.

**Conclusion** This review provides comprehensive evidence on the psychometric performance of generic childhood MAUIs. It assists analysts involved in cost-effectiveness-based evaluation to select instruments based on the application-specific minimum standards of scientific rigour. The identified evidence gaps and methodological issues also motivate and inform future psychometric studies and their methods, particularly those assessing reliability, proxy-child agreement, and MAUIs targeting preschool children.

### Key Points for Decision Makers

A comprehensive systematic review and evaluation was conducted on the psychometric performance of 14 generic childhood multi-attribute utility instruments across 18 psychometric properties.

There is a high degree of variability surrounding psychometric performance across the instruments and the psychometric properties evaluated, with more recently developed instruments targeting preschool children having greater evidence gaps.

Identified gaps in psychometric evidence and common methodological issues in psychometric assessment can inform future research directions and methods in this topic area.

## 1 Background

Constraints on healthcare expenditure require the comparison of alternative interventions in terms of costs and consequences [1]. Many healthcare decision makers recommend cost-utility analysis (CUA) as a form of economic evaluation enabling such comparisons [2–5]. The preferred health outcome in CUA is the quality-adjusted life-year (QALY), which combines preference-based health state utility values or health utilities with the length of time in those states [1]. Health utilities derived from generic instruments enable comparison of interventions across health conditions [6]. One approach to measuring health utilities is the use of multi-attribute utility instruments (MAUIs), which describe self- or proxy-reported health status across multiple dimensions according to a prespecified classification system [7]. They are designed to be accompanied by value sets that reflect the stated preferences (typically of a representative sample of the general adult population) for the health states generated by the classification system [8]. The value set application produces health utilities anchored on a scale with 0 = dead and 1 = full health.

Unique challenges arise in measuring health utilities in childhood populations (aged  $\leq 18$  years). First, biopsychosocial development during childhood means that the relevant dimensions of health status undergo rapid change [9]. This implies that classification systems for adult populations may not be applicable to children, and that the systems should be tailored to different age groups within childhood [10]. Second, the instrument wording

and format (e.g., use of pictures, response scale levels) should be tailored to the comprehension level and attention span of the target childhood age group [11, 12]. Proxy respondents such as parents can be used, either when child self-report is not feasible (e.g., for very young preschool children) or when an alternative perspective on the child's health status and health needs is sought [11]. Accordingly, several reviews have investigated the level of proxy-child agreement for patient-reported outcome measures (PROMs), including MAUIs [13–15].

Given these challenges, recent methodological advances have included the development of a range of instruments with childhood-specific classification systems and child-friendly formats and childhood-derived value sets, as well as the use of childhood-compatible instruments [7, 16]. A recent systematic review identified 89 generic multidimensional PROMs designed specifically for or compatible with application in children, 14 of which were MAUIs designed to be accompanied by value sets that generate health utilities [12].

Psychometrics concerns the measurement properties of scales, originating from psychophysics research on subjective judgements about physical stimuli and later applied in healthcare to develop scientifically rigorous PROMs [17–19]. Key psychometric properties include content validity, reliability, construct validity, responsiveness, interpretability, robust translation (cross-cultural validity), and patient and investigator burden (acceptability) [18–21]. Each of these properties requires unique tests and criteria, and contributes to minimum scientific standards for the use of a given PROM in patient-centred outcomes research such that the PROM should demonstrate acceptable performance across all properties included in the minimum standard set [20]. Otherwise, the instrument cannot be considered as providing scientifically credible information [17].

This highlights the need for a comprehensive evaluation of the psychometric evidence of instruments prior to application. However, it is unlikely that a single study can address all psychometric properties of an instrument for all research contexts [20]. Investigators seeking to use an instrument must rely on the weight of evidence across multiple studies, evaluating the consistency of findings while considering the volume and methodological quality of the contributing studies [20]. Therefore, a comprehensive systematic review of the available psychometric evidence for generic childhood MAUIs is warranted to inform their selection based on their meeting the minimum standards of scientific rigour.

Previous reviews have covered only a limited number of instruments [22–28]. For example, Rowen and colleagues [26] focused on four instruments, namely the CHU9D, EQ-5D-Y (3L or 5L), HUI2 and HUI3; Tan and colleagues [28] similarly focused on five instruments, the above four plus

AQoL-6D. Janssens and colleagues [27] identified psychometric evidence (up to 2012) for eight instruments (16D, 17D, AQoL-6D, CHU9D, EQ-5D-Y, HUI2, HUI3 and CHSCS-PS) but not for those developed since 2012. The review by Janssens et al. also only included studies conducted on general childhood populations and excluded psychometric evidence from patient/clinical samples. Moreover, the above reviews excluded key psychometric properties from evaluation, including internal consistency and content validity [26], or cross-cultural validity [27]; the Tan review focused solely on known-group validity, convergent validity, test–retest reliability, and responsiveness [28]. The previous reviews also excluded indirect evidence from studies that did not explicitly aim to conduct psychometric assessment [26–28], even though such studies provide relevant psychometric evidence (e.g., evidence for responsiveness generated by clinical trials).

The aim of the current study was to conduct a systematic review that identifies and evaluates the psychometric evidence for all generic childhood MAUIs identified in a recent systematic review [12]. The review is intended to be comprehensive in terms of (1) the instruments covered; (2) the psychometric properties evaluated, drawing on diverse standards in the literature for both properties and evaluation criteria [17–21, 29–31]; (3) the coverage of evidence from general and clinical populations; and (4) inclusion of both ‘direct studies’ that explicitly aimed to assess psychometric properties and ‘indirect studies’ that provide evidence relevant to psychometric properties without this as a stated aim. Key review objectives were to:

1. create a catalogue of evaluated psychometric evidence that can aid in the selection of generic childhood MAUIs for application;
2. identify gaps in psychometric evidence to inform future psychometric research; and
3. summarise the commonly used psychometric assessment methods and the relative psychometric performance of instruments by property.

This review does not aim to arrive at conclusions on which MAUI is ‘optimal’. This would depend on the contextual factors for application, including, for example, country of application (determining relevant instrument language version and value set), the childhood population age and type (e.g., general population, clinical), and instrument respondent type. These in turn determine the relative importance of different psychometric properties and the appropriate level of consistency with which the instrument shows acceptable performance for each property in the accumulated evidence. These considerations comprise the minimum standard of

scientific rigour for MAUI application. The review provides the evidence base that analysts can use to select instruments according to their application-specific minimum standard and/or conduct further psychometric research to fill relevant evidence gaps.

## 2 Methods

A prespecified protocol outlining the systematic review methods was developed and registered with the Prospective Register of Systematic Reviews (PROSPERO; CRD42021295959). For reporting purposes, the review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guideline [32] (see the electronic supplementary material [ESM] for the PRISMA checklist). Figure 1 illustrates the systematic review process, terminology, and objectives, which are further discussed in the following sections.

### 2.1 Data Sources and Study Selection

The database searches aimed to identify studies that provide evidence for the psychometric performance of one or more of the following generic childhood MAUIs identified in a recent systematic review [12]: 16-Dimensional Health-Related Measure (16D) [33]; 17-Dimensional Health-Related Measure (17D) [34]; Adolescent Health Utility Measure (AHUM) [35]; Assessment of Quality of Life, 6-Dimensional, Adolescent (AQoL-6D Adolescent) [36]; Child Health—6 Dimensions (CH-6D) [37]; Comprehensive Health Status Classification System—Preschool (CHSCS-PS) [38]; Child Health Utility—9 Dimensions (CHU9D) [39, 40]; EuroQoL 5-Dimensional questionnaire for Youth 3 Levels (EQ-5D-Y-3L) [41]; EQ-5D-Y 5 Levels (EQ-5D-Y-5L) [42]; Health Utilities Index 2 (HUI2) [43]; Health Utilities Index 3 (HUI3) [44]; Infant health-related Quality of life Instrument (IQI) [45]; Quality of Well-Being scale (QWB) [46, 47]; and Toddler and Infant Health Related Quality of Life Instrument (TANDI; recently renamed EuroQoL Toddler and Infant Populations (EQ-TIPS) [48]) [10, 49]. It should be noted that the above MAUIs are designed to be accompanied by value sets with which health utilities can be derived, but not all (e.g., the recently developed TANDI) currently have a value set developed (see the list of value sets published before October 2020 in our previous review [12]). In this case, the psychometric performance of the classification system was evaluated.

The following databases were covered from database inception to 7 October 2021: MEDLINE (OvidSP) [1946–present]; EMBASE(OvidSP) [1974–present];

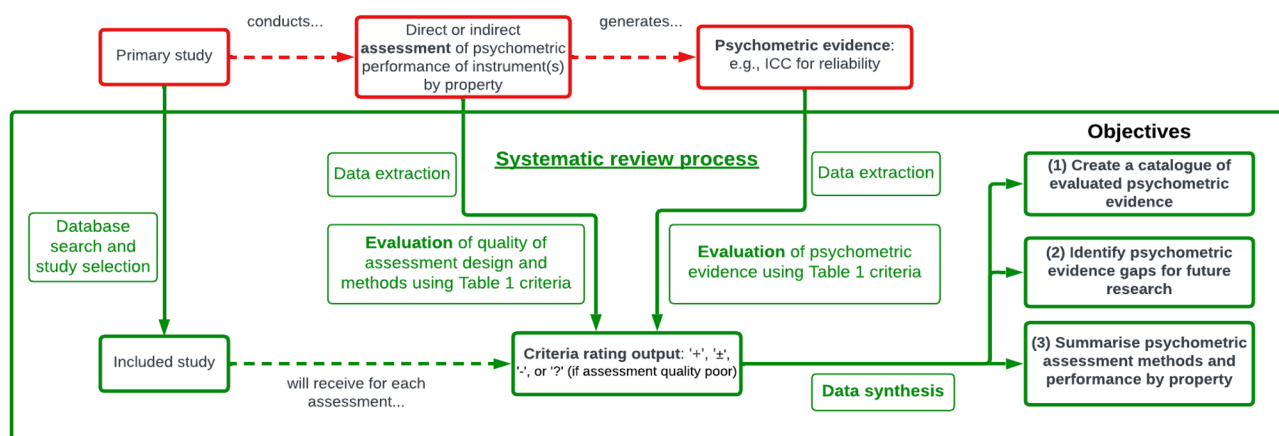


Fig. 1 Illustration of the systematic review process, terminology, and objectives. *ICC* intraclass correlation coefficient

PsycINFO (OvidSP) [1806–present]; EconLit (Proquest); CINAHL (EBSCOHost) [1982–present]; Scopus (Elsevier); and Science Citation Index (Web of Science Core Collection) [1945–present]. Tables A1–A7 in the ESM provide the search strategy for each database. References of the included studies were also searched. Endnote 20 was used to remove duplicates among the imported references. At least two researchers from a team of seven (JK, SK, TB, MH, EH, SP, and RR) independently reviewed the titles and abstracts and then the full texts on Covidence [50]. At each review stage, if an article received two approvals from any pair of the seven reviewers, it proceeded to the next stage (from title/abstract screening to full-text screening, then to data extraction), with disagreements referred to a third reviewer for the final assessment.

The inclusion criteria were (1) the study provided evidence for at least one psychometric property (see below for list of properties) of at least one of the 14 instruments listed above using any language version; (2) the study obtained data from general or clinical childhood populations (sample mean age  $\leq 18$  years) or relevant proxy respondents; (3) where the study covered childhood and adult populations, results were reported separately for a childhood subgroup with a mean age  $\leq 18$  years; and (4) the study was published in English. Studies reporting the original instrument development (as identified previously [12]) were included for content validity evidence. Studies without the explicit aim of assessing psychometric properties (i.e., instrument application studies) but nonetheless contained relevant psychometric evidence (e.g., responsiveness within intervention trials) were included as ‘indirect’ assessment studies. Conference abstracts that met the inclusion criteria were included. Studies that used one or more of the 14 instruments as a criterion standard to test a new, yet-to-be validated instrument were excluded, as were

studies that developed and validated value sets for health utility derivation without assessing or providing evidence of the psychometric properties of the health utilities.

## 2.2 Data Extraction

Data from the included studies were extracted by JK, with 20% of studies double data extracted by the other reviewers (SK, TB, MH, EH, and RR). Disagreements from double extraction were resolved by discussion within the research team and by decision from senior authors (SS, SP) if consensus could not be reached. The solutions reached through the discussions were applied to the other 80% of single-extracted studies.

The following data were extracted according to proformas in Excel (Microsoft Corporation, Redmond, WA, USA): (1) first author name and year of publication; (2) study country(ies); (3) study design—e.g., cross-sectional survey; (4) whether the study explicitly aimed to assess psychometric properties (‘direct’) or not (‘indirect’); (5) psychometric property(ies) assessed or relevant evidence reported; (6) main methodological issues affecting psychometric assessment (described in the next section); (7) instrument(s) assessed; (8) instrument language version(s); (9) instrument component(s) assessed—e.g., index (after value set application), dimension score; (10) country and population (where relevant) where a value set was derived; (11) respondent type(s)—child self-report or proxy report; (12) administration mode(s)—e.g., by self, with interviewer; (13) study population type—healthy, clinical, or general childhood population(s)—and clinical characteristics; (14) target and sample age,—e.g., mean, range—and proportion of females; (15) target and actual sample size; and (16) intervention(s) assessed (if relevant).



### 2.3 Evaluation and Data Synthesis

Table 1 lists and defines the psychometric properties evaluated. These have been drawn from established standards developed and used by medical industry regulators, previous psychometric research projects, and professional communities who are collectively interested in the psychometric performance of PROMs. The standards included the Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) checklist [29–31]; the International Society for Quality of Life Research (ISOQOL) guideline [20]; the Food and Drug Administration (FDA) guideline [18]; the Medical Outcomes Trust (MOT) guidelines [21]; and the taxonomy of psychometric properties previously applied by Smith and colleagues [17]. Table 1 references the related guidelines for each property. There was a broad overlap in the properties covered by the guidelines and also a nontrivial variation, and this variation motivated the referral to multiple guidelines for comprehensiveness. Brazier and Deverill [19] have discussed the relevance of psychometric properties to health utility instruments, and their views were incorporated where relevant. We distinguished between proxy-child agreement and inter-rater agreement between different proxies given the emphasis placed on the former for paediatric PROM application [11].

To clarify the terminology used (see Fig. 1), a primary study directly or indirectly conducts an *assessment* of the psychometric performance of instrument(s) by property, which generates primary psychometric evidence. This review in turn conducts an *evaluation* of this assessment in terms of the quality of its design and methods and the resulting psychometric evidence. This evaluation produces a *criteria rating output* for each assessment. Assessments for different components of an instrument (e.g., index, dimension) are evaluated separately. As shown in Table 1, a four-part criteria rating for each property (except for interpretability, with a three-part set of criteria) was developed for the evaluation. References are given where the criteria are sourced from specific guidelines. In general, a criteria rating output of ‘+’ indicates psychometric evidence consistent with an a priori hypothesis (formulated by the primary study) in terms of clinical and psychometric expectation; ‘±’ partially consistent; and ‘–’ no evidence or evidence contrary to the a priori hypothesis. A ‘?’ output is given where the poor quality of assessment design and methods prevented a conclusion to be drawn by the evaluation concerning the psychometric performance for a property. General considerations for this quality evaluation across all properties included sample size, missing data level, and appropriateness of the statistical technique used (e.g., suited to continuous, binary,

or ordinal nature of instrument outcome) [30]. Thereafter, considerations for the appropriate assessment design and methods differed by psychometric property [30].

It should be noted that the referenced guidelines provide general recommendations for evaluation: the COSMIN checklist, for example, expects the adequate sample size to vary by application context (p. 545) [30]. Further context-specific judgements were thus required for evaluation and these are reported case-by-case in the Excel file in the ESM. A key source of between-context variation was the a priori hypothesis specified by the included study, which introduced between-study variation in the expected correlation, association, or change sought. For example, one study defined a moderate and therefore acceptable correlation as a coefficient above 0.3 [51], while another as above 0.4 [52]. Where criteria and thresholds were specified as part of a priori hypotheses, these were followed by the review. Finally, studies frequently performed multiple assessments for the same psychometric property; i.e., sub-assessments within the assessment for the property. These sub-assessments each received criteria rating outputs, which were then grouped together to produce a summary output for the property. Details on this process are given in footnote b of Table 1.

The criteria rating outputs were synthesised to address the three review objectives:

1. *Create a catalogue of evaluated psychometric evidence.* The Excel file in the ESM serves as the main catalogue where the criteria rating outputs are tabulated with other extracted variables (e.g., study design, sample size) and the main rationale for each rating. More condensed catalogues are presented in this manuscript.
2. *Identify gaps in psychometric evidence.* Two metrics were used to identify evidence gaps: (1) the number of cases where no criteria rating output was available to an instrument for a given property; and (2) the number of cases where no criteria rating output was available or where available outputs had no ‘+’ output (no ‘+’ or ‘±’ for interpretability, which had very few ‘+’). The metrics were calculated for the whole evidence base and for a subset of evidence from direct assessment studies. Comparing these two scenarios allows one to understand the impact of including evidence from indirect assessment studies.
3. *Summarise the commonly used psychometric assessment methods and the psychometric performance of instruments by property.* Common reasons for assessments receiving ‘?’ were noted by property. The relative performance of instruments was compared by property,

**Table 1** Definitions of psychometric properties assessed by the systematic review and their performance criteria rating

Psychometric property (label) <sup>a</sup>	Definition	Rating output	Criteria <sup>ab</sup>
<b>1. Reliability</b>			
1.1 Internal consistency (IC) [17, 18, 20, 21, 29–31]	The degree of the interrelatedness among items from the same scale	+	Cronbach alpha for summary scores $\geq 0.7$ AND item-total correlation $\geq 0.2$ if both reported [17]
		±	Mixed assessment results, e.g., Cronbach alpha for summary scores $\geq 0.7$ but item-total correlation $< 0.2$ if both reported
		–	Cronbach alpha for summary scores $< 0.7$ AND item-total correlation $< 0.2$ if both reported
		?	Inconclusive results due to assessment design and method issues, e.g., small sample size [30]
1.2 Test–retest reliability (TR) [17–21, 29–31]	The degree to which the instrument scores for patients are the same for repeated measurements over time, assuming no intervention or clinical change	+	High agreement, e.g., ICC $\geq 0.7$ [17, 20]
		±	Mixed agreement results where multiple assessments were conducted
		–	Low agreement, e.g., ICC $< 0.7$
		?	Inconclusive results due to issues in assessment design and methods, e.g., inappropriate time interval, unclear whether health construct of interest remained stable over time interval [30]
1.3 Inter-rater reliability (IR) [17–19, 21, 29–31]	The degree to which the instrument scores for patients are the same for ratings made by different (proxy) raters on the same occasion	+	High agreement, e.g., ICC $\geq 0.7$ [17]
		±	Mixed agreement results where multiple assessments were conducted
		–	Low agreement, e.g., ICC $< 0.7$
		?	Inconclusive results due to issues in assessment design and methods, e.g., unclear whether instrument applied to rater groups at a similar time
1.4 Inter-modal reliability (IM) [19, 21]	The degree to which the instrument scores for patients who have not changed are the same for measurements by different instrument administration modes (e.g., online, postal)	+	High agreement, e.g., ICC $\geq 0.7$
		±	Mixed agreement results where multiple assessments were conducted
		–	Low agreement, e.g., ICC $< 0.7$
		?	Inconclusive results due to issues in assessment design and methods, e.g., unclear whether instrument applied to modal groups at a similar time
<b>2. Proxy-child agreement (PC) [11]</b>			
	The extent of agreement in instrument scores between proxy respondent (e.g., parent) and child, where some discrepancies in measurement are expected due to different perspectives on childhood health	+	High agreement, e.g., ICC $\geq 0.7$ [17]
		±	Mixed agreement results where multiple assessments were conducted
		–	Low agreement, e.g., ICC $< 0.7$
		?	Inconclusive results due to issues in assessment design and methods, e.g., unclear whether the proxy is sufficiently aware of child's health status, excessively different administration mode between child and proxy

Table 1 (continued)

Psychometric property (label) <sup>a</sup>	Definition	Rating output	Criteria <sup>a,b</sup>
<b>3. Content validity (CV)</b> [11, 17–21, 29–31]	The degree to which the content of the instrument is an adequate reflection of the construct to be measured	+	Conducted the following steps in the original instrument development: (1) stated the conceptual framework for the purpose of measurement; (2) qualitative research with children or appropriate proxies; (3) cognitive interviews and pilot tests with children or appropriate proxies [11, 17, 20, 30] <sup>c</sup> Conducted two of the three steps above
<b>4. Structural validity (SV)</b> [17, 29–31]	The degree to which the within-scale item relationships are an adequate reflection of the dimensionality of the construct to be measured by the scale	± – ? +	Conducted one or none of the three steps above, OR other evidence that the instrument content does not reflect the construct to be measured Contained insufficient descriptions of the above assessments Exploratory factor analysis: items with factor-loading coefficient $\geq 0.4$ AND at least moderate correlations between scale scores (if both assessed) [17] Mixed assessment results from factor analysis. Exploratory factor analysis: items with factor-loading coefficient $< 0.4$ AND low correlations between scale scores (if both assessed) Inconclusive results due to issues in assessment design and methods, e.g., unclear rotation method
<b>5. Cross-cultural validity (CCV)</b> [20, 21, 29–31]	The degree to which the performance of the items on a translated or culturally adapted instrument are an adequate reflection of the performance of the items of the original version of the instrument	+	Rigorous translation process: expert involvement; independent forward and backward translations; committee review; comparison with original (conceptual and linguistic equivalence); pre-testing (e.g., cognitive interviews). Should be followed by psychometric assessment of the cross-cultural version and comparable result to the original version [21, 30] Mixed translation process, e.g., expert involvement but no pretesting Problematic translation process and/or poor psychometric performance compared with the original version
<b>6. Construct validity</b> [21]		± – ?	Provided insufficient detail on the translation process to reach a conclusion, e.g., insufficient description of the pretest sample [30]
<b>6.1 Known-group validity (KV)</b> [17, 18, 20, 29–31]	The degree to which the instrument scores can differentiate groups with expected differences in the constructs measured by the instrument	+	Description of subgroups delineated by clinical or sociodemographic variables and a priori hypotheses on instrument score differences. Statistically and clinically significant results consistent with hypotheses [17] <sup>d,e,f</sup>
		±	A priori hypotheses and mixed results where multiple between-group comparisons were conducted
		–	A priori hypotheses and results contrary to hypotheses

Table 1 (continued)

Psychometric property (label) <sup>a</sup>	Definition	Rating output	Criteria <sup>a,b</sup>
6.2 Hypothesis testing (HT) [20, 29–31]	The degree to which the instrument scores are associated with sociodemographic, clinical, and other variables according to hypotheses	? + ± – ?	No a priori hypothesis that can interpret the significant/non-significant results or inconclusive due to assessment design and method issues <sup>e</sup> Description of variables potentially associated with instrument score and a priori hypotheses on strength and direction of association. Statistically and clinically significant result consistent with hypotheses [20, 30] <sup>f</sup> A priori hypotheses and mixed results where multiple associations examined A priori hypotheses and results contrary to hypotheses No a priori hypothesis that can interpret the significant/non-significant results or inconclusive results due to assessment design and method issues <sup>g</sup>
6.3 Convergent validity (CNV) [17, 18, 29–31]	The degree to which the instrument scores are correlated with other measures of the same or similar constructs	+ ±	A priori hypotheses and results consistent with hypotheses in terms of correlation strength (e.g., Pearson correlation coefficient >0.4), <sup>h</sup> statistical significance, and direction [17] A priori hypotheses and mixed results where multiple correlations assessed
6.4 Discriminant validity (DV) [17, 18, 29–31]	The degree to which the instrument scores are not correlated with measures of different constructs	– ?	A priori hypotheses and results contrary to hypotheses No a priori hypothesis that can interpret the correlation or inconclusive results due to assessment design and method issues, e.g., statistical significance of correlation not reported
6.5 Empirical validity (EV) [19]	The degree to which the utility values generated by the preference-based instruments reflect people's preferences over health (e.g., self-reported health status)	+ ± – ?	A priori hypotheses and consistent results in terms of correlation strength (e.g., Pearson correlation coefficient <0.4), <sup>h</sup> statistical significance, and direction [17] A priori hypotheses and mixed results where multiple correlations assessed A priori hypotheses and results contrary to hypotheses No a priori hypothesis that can interpret the significant/non-significant results or inconclusive results due to assessment design and method issues
		+ ± – ?	A priori hypotheses and consistent results in terms of clinically and statistically significant associations A priori hypotheses and mixed assessment results where multiple associations assessed A priori hypotheses and results contrary to hypotheses No a priori hypothesis that can interpret the significant/non-significant results or inconclusive results due to assessment design and method issues <sup>i</sup>



Table 1 (continued)

Psychometric property (label) <sup>a</sup>	Definition	Rating output	Criteria <sup>a,b</sup>
<b>7. Criterion-related validity</b> [20, 21]			
7.1 Concurrent validity (CRV) [17, 18, 29–31]	The degree to which the instrument scores adequately reflect a 'gold standard' criterion measured at the same time, i.e., correlation of scores with those of a criterion measure	+  ±  –  ?	Described the 'gold standard' criterion and results consistent with a priori hypotheses in terms of correlation strength and direction between instrument and criterion scores [17, 30]  A priori hypotheses and mixed assessment results where multiple correlations assessed  A priori hypotheses and results contrary to hypotheses  No a priori hypothesis that can interpret the significant/non-significant results or inconclusive results due to assessment design and method issues <sup>d</sup>  Described the 'gold standard' criterion and results consistent with a priori hypotheses in terms of associations between predicted instrument and criterion scores [17, 30]
7.2 Predictive validity (PV) [17, 18, 29–31]	The degree to which the instrument scores adequately reflect a 'gold standard' criterion measured in the future	+  ±	A priori hypotheses and mixed assessment results where multiple predicted associations assessed
<b>8. Responsiveness (RE)</b> [17, 18, 20, 21, 29–31]	The extent to which the instrument can identify differences in scores over time in individuals or groups who have changed with respect to the measurement concept	–  ?	A priori hypotheses and results contrary to hypotheses  No a priori hypothesis that can interpret the significant/non-significant results or inconclusive results due to assessment design and method issues, e.g., flaws in the statistical method of prediction
<b>9. Acceptability (AC)</b> [17, 19–21]	The level of data quality, assessed by its completeness and score distributions.	+  –  ?	Change in the instrument score consistent with: (1) a priori hypotheses on direction and strength of score change in response to a change of interest (e.g., intervention); and (2) direction and strength of change in reference (e.g., disease-specific) measure [17, 18, 21, 30]. If the MIC is specified and justified, and score change is greater than the MIC, then consistency with (2) is not required  Mixed assessment results, e.g., consistent with changes in one reference measure but not with another  Score changes that are contrary to a priori hypotheses and/or change in the reference measure(s)  No a priori hypothesis or reference measure that can interpret the significant/non-significant results or inconclusive due to assessment design and method issues, e.g., randomisation bias, inadequate power, unclear time interval between measurements, unclear description of intervention and expected impact [30]  Conducted at least two assessments and met two or more of the following criteria: (1) missing data for instrument scores <5%; <sup>b</sup> (2) floor and ceiling effects <10%; <sup>1</sup> (3) high score distribution (e.g., number of unique health states >0.4 per respondent); and (4) other criteria concerning the ease of use (e.g., response time of <15 min) [17, 21]

Table 1 (continued)

Psychometric property (label) <sup>a</sup>	Definition	Rating output	Criteria <sup>a,b</sup>
		±	Conducted at least two assessments and met one of the four criteria above
		–	Conducted at least two assessments and met none of the four criteria above
		?	Conducted less than two assessments or unclear how many of the criteria are met due to poor reporting <sup>k</sup>
<b>10. Interpretability (ITR)</b> [20, 21, 29–31]	The degree to which one can assign qualitative meaning (i.e., clinical or commonly understood connotations) to the instrument scores and change in scores and/or quantitative descriptions of MID	+	Conducted one of the following: (1) established instrument score mean and standard deviation for the normative childhood reference population; (2) quantified MIC or MID [20, 30]
		±	Compared instrument scores with reference population scores (not necessarily established population norms) and/or with MIC/MID derived from external studies of childhood populations rather than primary derivation
		?	Inconclusive results due to poor reporting or comparison features, e.g., comparison with MIC/MID derived in external studies of adult populations

ICC intraclass correlation coefficient; MIC minimal clinically important change; MID minimal clinically important difference

<sup>a</sup> For all psychometric properties, sample size, low missing data, and appropriate statistical techniques were considered in evaluating the quality of assessment design and methods [30]. References to guidelines are provided alongside each label in the first column; these guidelines discussed the given property in terms of its definition and its relevance to determining an instrument's scientific credibility

<sup>b</sup> Studies frequently contained multiple sub-assessments within the property assessment (e.g., KV assessment using multiple subgroup delineators), each of which received a criteria rating output. For a mix of '+' and '±' outputs, the higher output '+' was presented as the summary, and likewise '±' for a mix of '+' and '-'. A mix of '+' and '-' was presented as '±'. A mix of '+/±/-' and '?' was presented as '+/±/-' ; i.e., an assessment received '?' only if *all* sub-assessments had issues in design and methods. References to the guidelines that offer specific criteria for the evaluation of psychometric performance is also provided

<sup>c</sup> The review also considered CV evidence from non-original development studies. If so, reasonable criteria formulated by the primary study authors were used for evaluation

<sup>d</sup> Few studies prespecified *which* dimensions were expected to differ across groups beyond the general hypothesis that some dimension-level differences were expected. In this case, if  $\geq 75\%$  of dimensions (e.g., four or five dimensions out of five in EQ-5D-Y) showed significant differences, then '+' was given;  $< 75\%$  and  $> 25\%$  '+'; and  $\leq 25\%$  '-'

<sup>e</sup> Where the study used Bonferroni correction for multiple comparisons (e.g., comparisons for multiple dimensions), the adjusted statistical significance threshold was used for evaluation.

<sup>f</sup> *Adjusted* between-group comparisons through multivariate regression models were evaluated under HT; only unadjusted comparisons under KV

<sup>g</sup> Unadjusted comparison between sociodemographic subgroups without *a priori* statement of expected difference was evaluated under HT and given '?' for the *ad hoc* hypothesis testing

<sup>h</sup> A different threshold was used if specified as part of *a priori* hypothesis [20]

<sup>i</sup> Where it was uncertain whether the subgroup delineator reflected people's preference over health, this assessment was evaluated under KV or HT

<sup>j</sup> If there was insufficient justification for a measure (other than the instrument) being a 'gold standard' criterion, the assessment was evaluated under CNV

<sup>k</sup> Effort was made to distinguish between missing data due to poor survey design (i.e., low response rate) from that due to the instrument itself (i.e., missing response from survey participants) Where this distinction was not possible due to poor reporting, the assessment was given '?'

<sup>l</sup> Ceiling and floor effects at dimension level were not evaluated; hence, ceiling/floor effect meant top/bottom level on all dimensions

using the proportion of ‘+’ as the performance metric, while also considering the number of outputs over which the proportion was calculated.

### 3 Results

#### 3.1 Search Results

Figure 2 presents the PRISMA flow diagram. After screening of titles/abstracts and full texts, 372 studies were included. Studies excluded at full-text screening stage ( $n = 167$ ) are listed in Table B in the ESM, alongside the main reason for exclusion. Table C in the ESM details co-authors’ contributions to the screening stages.

#### 3.2 Characteristics of the Included Studies

Table D in the ESM reports the characteristics of the included studies and these are summarised in Table 2. The full extracted information on the included studies is available in the Excel file in the ESM. There were 163 (43.8%) studies that were judged to have conducted direct assessments of psychometric properties. Ninety-six (25.8%) studies assessed two or more instruments, 48 of them as direct assessments. Most studies ( $n = 210$ ;

56.5%) concerned clinical paediatric populations. The largest target age category was that combining pre-adolescents and adolescents spanning the age range of 5–18 years ( $n = 145, 39.0%$ ).

#### 3.3 Psychometric Evaluation Results

Table 3 summarises the characteristics of the psychometric assessments from which the criteria rating outputs were derived. The full extracted information is contained in the Excel file in the ESM, which serves as the main catalogue that meets the review objective (1).

Table E in the ESM provides a condensed catalogue of the evaluated psychometric evidence by study, instrument, component, and respondent type. There were 2153 criteria rating outputs and 1598 (74.2%) excluding ‘?’ outputs. All outputs and rationale for ratings, alongside the study and assessment characteristics, are reported in the Excel file in the ESM as the main evidence catalogue. Figure A in the ESM displays the output numbers and proportions by instrument, component, and psychometric property, except for AHUM, CH-6D, and IQI, which had four or fewer outputs.

#### 3.4 Psychometric Evidence Gaps

This section addresses the second review objective and identifies psychometric evidence gaps. Table 4 shows the number

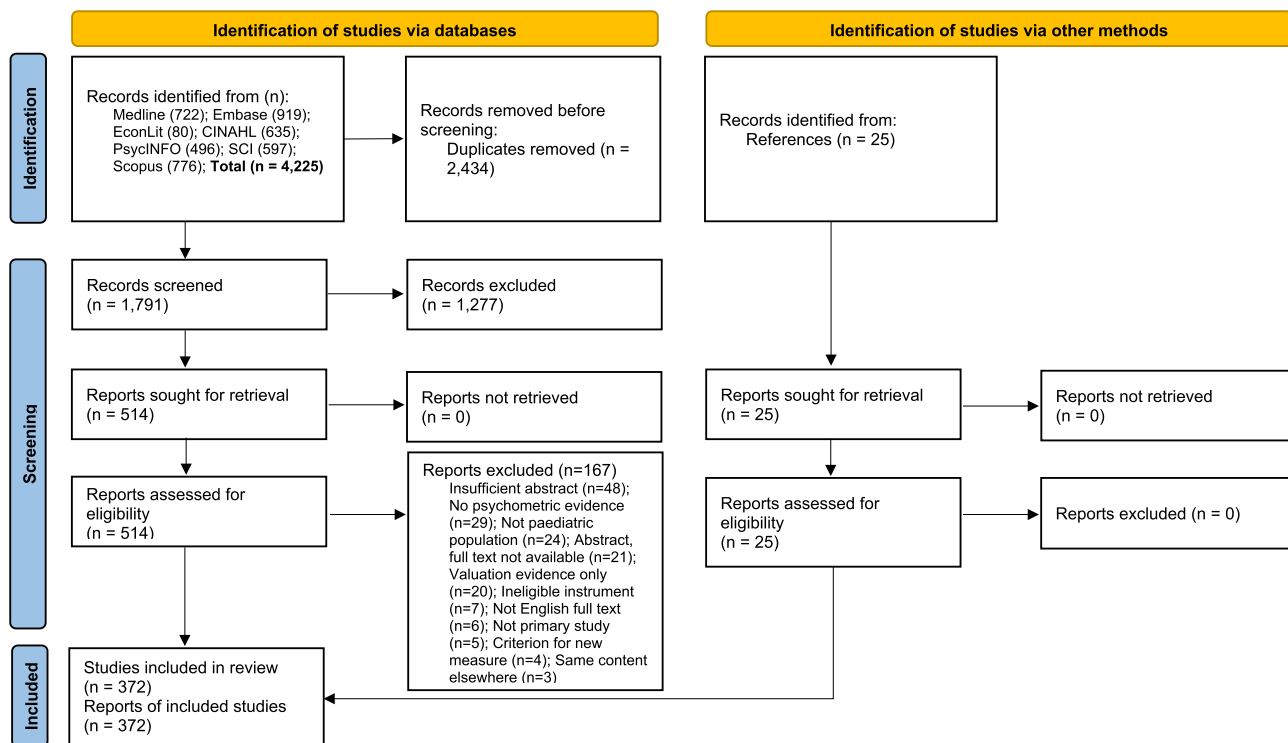


Fig. 2 Preferred Reporting Items for Systematic Reviews and Meta-Analyses flow diagram

**Table 2** Characteristics of the included studies

	<i>N</i>	%
Explicit aim to assess psychometric performance		
Yes: Direct assessment study	163	43.8
No: Indirect assessment study	209	56.2
Total	372	100.0
Instrument(s) assessed		
16D	10	2.7
17D	6	1.6
AHUM	1	0.3
AQoL-6D	4	1.1
CH-6D	1	0.3
CHSCS-PS	3	0.8
CHU9D	43	11.6
EQ-5D-Y-3L	68	18.3
EQ-5D-Y-5L	6	1.6
HUI2	38	10.2
HUI3	78	21.0
IQI	1	0.3
QWB	14	3.8
TANDI	3	0.8
16D; 17D	10	2.7
AQoL-6D; CHU9D	2	0.5
CHU9D; EQ-5D-Y-3L	9	2.4
CHU9D; HUI2	1	0.3
CHU9D; HUI2; HUI3	3	0.8
CHU9D; HUI3	1	0.3
EQ-5D-Y-3L; EQ-5D-Y-5L	5	1.3
EQ-5D-Y-3L; HUI2; HUI3	2	0.5
HUI2; HUI3	59	15.9
HUI2; HUI3; QWB	2	0.5
HUI3; QWB	2	0.5
Total	372	100.0
Childhood population type		
General/healthy	74	19.9
Clinical	210	56.5
General/healthy and clinical	88	23.7
Total	372	100.0
Target childhood age group		
(1) Infants and preschool children aged < 5 years	14	3.8
(2) Pre-adolescents aged 5–11 years	66	17.7
(3) Adolescents aged 12–18 years	85	22.8
(1) and (2)	11	3.0
(2) and (3)	145	39.0
(1), (2), and (3)	45	12.1
Unclear	6	1.6
Total	372	100.0

*16D* 16-Dimensional Health-Related Measure, *17D* 17-Dimensional Health-Related Measure, *AHUM* Adolescent Health Utility Measure, *AQoL-6D* Assessment of Quality of Life, 6-Dimensional, Adolescent, *CH-6D* Child Health—6 Dimensions, *CHSCS-PS* Comprehensive Health Status Classification System—Preschool, *CHU9D* Child Health Utility—9 Dimensions, *EQ-5D-Y-3L/5L* EuroQoL 5-Dimensional questionnaire for Youth 3/5 Levels, *HUI2/3* Health Utilities Index 2/3, *IQI* Infant health-related Quality of life Instrument, *QWB* Quality of Well-Being scale, *TANDI* Toddler and Infant Health-Related Quality of Life Instrument

**Table 3** Characteristics of psychometric assessments

	<i>N</i>	%
Whether the study had an explicit aim to assess psychometric performance		
Yes: Direct assessment study evidence	1253	58.2
No: Indirect assessment study evidence	900	41.8
Total	2153	100.0
Childhood population type		
General/healthy	391	18.2
Clinical	1194	55.5
General/healthy and clinical	568	26.4
Total	2153	100.0
Target childhood age group		
(1) Infants and preschool children aged <5 years	66	3.1
(2) Pre-adolescents aged 5–11 years	340	15.8
(3) Adolescents aged 12–18 years	441	20.5
(1) and (2)	29	1.3
(2) and (3)	1006	46.7
(1), (2), and (3)	251	11.7
Unclear	20	0.9
Total	2153	100.0
Respondent type		
Involved child self-report	1258	58.4
Proxy report only	858	39.9
Unclear	37	1.7
Total	2153	100.0
Administration mode		
Self-administered by child or proxy	1367	63.5
Interviewer-administered	509	23.6
Mix of self- and interviewer-administered	82	3.8
Unclear	195	9.1
Total	2153	100.0

of criteria rating outputs by instrument and property. Seven properties—internal consistency, inter-modal reliability, and structural, discriminant, empirical, concurrent and predictive validities—had fewer than 30 outputs, while three had more than 300: known-group validity ( $n = 500$ ), acceptability ( $n = 352$ ), and hypothesis testing ( $n = 309$ ). Similarly, there was high variation in output numbers across instruments, with the HUI2 ( $n = 472$ ) and HUI3 ( $n = 623$ ) comprising nearly half of all outputs.

Of the 270 cells in Table 4 defined by instrument and property, 116 (43.0%) were empty, indicating absence of evidence, while 146 (54.1%) were either empty or had 0% ‘+’ output. The psychometric evidence gap varied greatly by instrument, with the longer established instruments having smaller gaps: EQ-5D-Y-3L had the least number of cells ( $n = 2$ ) that were empty or had 0% ‘+’ output, followed

**Table 4** Evaluation of psychometric evidence gaps by psychometric property and instrument from all included studies (*n* = 372)

<i>N</i> (% of '+')	IC	TR	IR	IM	PC	CV	SV	CCV	KV	HT	CNV	DV	EV	CRV	PV	RE	AC	ITR <sup>a</sup>	Total
16D	4 (75.0)	2 (100)			1 (0.0)	1 (0.0)	1 (0.0)	2 (100)	29 (24.1)	10 (30.0)	4 (75.0)					8 (25.0)	5 (20.0)	16 (75.0)	83
17D	3 (66.7)	2 (100)			1 (0.0)	1 (0.0)	1 (100)	1 (100)	27 (25.9)	11 (18.2)	3 (100)					3 (33.3)	6 (16.7)	11 (100)	69
AHUM					1 (0.0)	1 (0.0)			1 (100)	1 (100)						1 (0.0)			3
AQoL-6D					1 (0.0)	1 (0.0)			4 (75.0)	5 (40.0)	3 (33.3)		2 (100)				2 (0.0)	1 (100)	18
CH-6D					1 (0.0)	1 (0.0)			1 (100)	1 (0.0)	1 (100)								4
CHSCS-PS		1 (0.0)	3 (66.7)		1 (100)	1 (100)			1 (100)	3 (33.3)				1 (100)		2 (50.0)			12
CHU9D	7 (85.7)	8 (0.0)			3 (33.3)	5 (40.0)	3 (100)	4 (25.0)	48 (52.1)	49 (32.7)	51 (29.4)	1 (0.0)	18 (94.4)	3 (100)		11 (45.5)	37 (29.7)	14 (78.6)	262
EQ-5D-Y-3L	3 (66.7)	20 (45.0)	4 (50.0)	2 (100)	13 (23.1)	6 (0.0)	2 (50.0)	5 (60.0)	53 (35.8)	42 (35.7)	36 (25.0)	2 (50.0)	1 (100)	5 (40.0)		22 (40.9)	67 (10.4)	15 (80.0)	298
EQ-5D-Y-5L		5 (40.0)			1 (0.0)	1 (0.0)		3 (33.3)	8 (62.5)	6 (33.3)	6 (33.3)	1 (100)				1 (100)	11 (9.1)	3 (33.3)	46
EQ-5D-Y VAS <sup>b</sup>		12 (75.0)	4 (0.0)	2 (100)	11 (9.1)	1 (0.0)		3 (66.7)	37 (40.5)	29 (34.5)	23 (39.1)	1 (100)		2 (0.0)		13 (30.8)	39 (12.8)	9 (100)	186
HUI2	3 (33.3)	4 (25.0)	34 (11.8)	2 (0.0)	48 (31.3)	7 (0.0)	9 (33.3)	9 (33.3)	127 (34.6)	53 (17.0)	27 (22.2)		5 (80.0)	3 (66.7)		30 (46.7)	77 (14.3)	43 (32.6)	472
HUI3	2 (50.0)	7 (14.3)	30 (6.7)	5 (0.0)	68 (25.0)	5 (0.0)	3 (0.0)	7 (42.9)	155 (36.1)	81 (28.4)	39 (15.4)	2 (50.0)	3 (100)	2 (100)		47 (38.3)	98 (12.2)	69 (23.2)	623
IQI					1 (100)	1 (100)													1
QWB	2 (50.0)	1 (100)	2 (100)		3 (0.0)	1 (0.0)			8 (62.5)	15 (33.3)	5 (20.0)					11 (18.2)	6 (33.3)	4 (75.0)	58
TANDI	1 (100)	2 (100)			2 (100)	2 (100)	1 (100)		2 (100)	3 (33.3)	4 (0.0)			1 (0.0)			2 (0.0)		18
Total	25	64	77	11	149	35	10	34	500	309	202	7	29	17	0	147	352	185	2153

The number in the cells indicates the number of criteria rating outputs; the number in parentheses is the percentage of outputs that is '+' for the given property and instrument

IC internal consistency, TR test-retest reliability, IR inter-rater reliability, IM inter-modal reliability, PC proxy-child agreement, CV content validity, SV structural validity, CCV cross-cultural validity, KV known-group validity, HT hypothesis testing, DV discriminant validity, EV empirical validity, CRV concurrent validity, PV predictive validity, RE responsiveness, AC acceptability, ITR interpretability

<sup>a</sup> Proportion of '+' and '±' ratings combined

<sup>b</sup> From the EQ-5D-Y-3L and -5L versions

by HUI3 ( $n = 4$ ), HUI2 ( $n = 5$ ), and CHU9D ( $n = 5$ ). By contrast, the more recently developed instruments targeting preschool children had high numbers of empty or 0% '+' output cells: CHSCS-PS ( $n = 12$ ); IQI ( $n = 17$ ); and TANDI ( $n = 12$ ).

Another noticeable feature is the frequent evidence gaps for reliability (IC, TR, IR, and IM), which had 32 of 56 (57.1%) cells that were empty or had 0% '+' output. This compares with 24 of 70 (34.3%) for construct validity (KV, HT, CNV, DV, and EV) and 5 of 42 (11.9%) for KV, HT, and CNV combined. Proxy-child agreement is a property specifically relevant to childhood MAUIs, yet its evidence gap was similarly high: 10 of 14 (71.4%) cells were empty or had 0%, and likewise 7 of 11 (63.6%) after excluding CHSCS-PS, IQI, and TANDI, which target preschool children and hence likely rely solely on proxy report.

Table F in the ESM shows the number of criteria rating outputs and proportion of '+' from the subset of 163 direct assessment studies. There were 118 (43.7%) empty cells in Table F, only two more than in Table 4. This suggests that the further inclusion of indirect assessment studies in Table 4 expanded the volume but not the type of psychometric evidence. However, there was a noticeable difference in the number of empty or 0% '+' cells, with 160 (59.3%) in Table F compared with 146 (54.1%) in Table 4. Hence, the inclusion of indirect assessment studies reduced the number of cases where there was no '+' output. This was particularly so for 16D, 17D, and AQoL-6D, for which the number of empty or 0% '+' cells declined from 12, 13, and 16, respectively, in Table F, to 9, 9, and 13, respectively, in Table 4.

### 3.5 Psychometric Assessment Methods and Performance by Property

This section addresses the third review objective and describes the common psychometric assessment methods and the relative performance of instruments by property, except for predictive validity, which had no criteria rating output.

#### 3.5.1 Internal Consistency

Internal consistency criteria rating outputs ( $n = 25$ ) were available for eight instruments. Most evidence consisted of Cronbach's alpha ( $n = 19$ , 76.0%), which fell below the acceptable threshold of 0.7 only twice, once each for CHU9D (from seven outputs) [53] and EQ-5D-Y-3L (from three) [54]. The proportion of '+' was highest for TANDI at 100% (from one output), followed by CHU9D at 85.7% (from seven outputs).

#### 3.5.2 Test-Retest Reliability

Test-retest reliability criteria rating outputs ( $n = 64$ ) were available for 10 instruments. The interval between initial test and retest varied significantly from 24 hours to 1 year. Nineteen outputs from eight studies involved test-retest intervals longer than 4 weeks [55–62]; of these, five studies limited the test-retest sample to those who maintained stable health conditions in the interval [55, 57, 60–62]. Statistical methods included intraclass correlation coefficient (ICC) for continuous outcomes (e.g., index, visual analogue scale [VAS]) and the Kappa coefficient and percentage agreement for categorical outcomes (e.g., ordinal item response). Figure B in the ESM shows the criteria rating outputs by instrument. QWB and TANDI had 100% '+' outputs, although the total numbers per instrument were small (one and two, respectively). 16D and 17D had 100% '+' from four outputs. Among others, EQ-5D-Y VAS (from the 3L and 5L versions) had the highest proportion of '+' (75.0%) from 12 outputs.

#### 3.5.3 Inter-Rater Reliability

Inter-rater (proxies) reliability criteria rating outputs ( $n = 77$ ) were available for five instruments. Proxies involved in comparison included parents, physicians, nurses, and therapists. Statistical methods included ICC, Kappa coefficient, and percentage agreement. Figure C in the ESM shows the criteria rating outputs by instrument. QWB had 100% '+', followed by CHSCS-PS at 66.7%, but both instruments had few outputs (two and three, respectively). HUI2/3 had more outputs than other instruments, but the proportions of '+' were low at 11.8% and 6.7%, respectively.

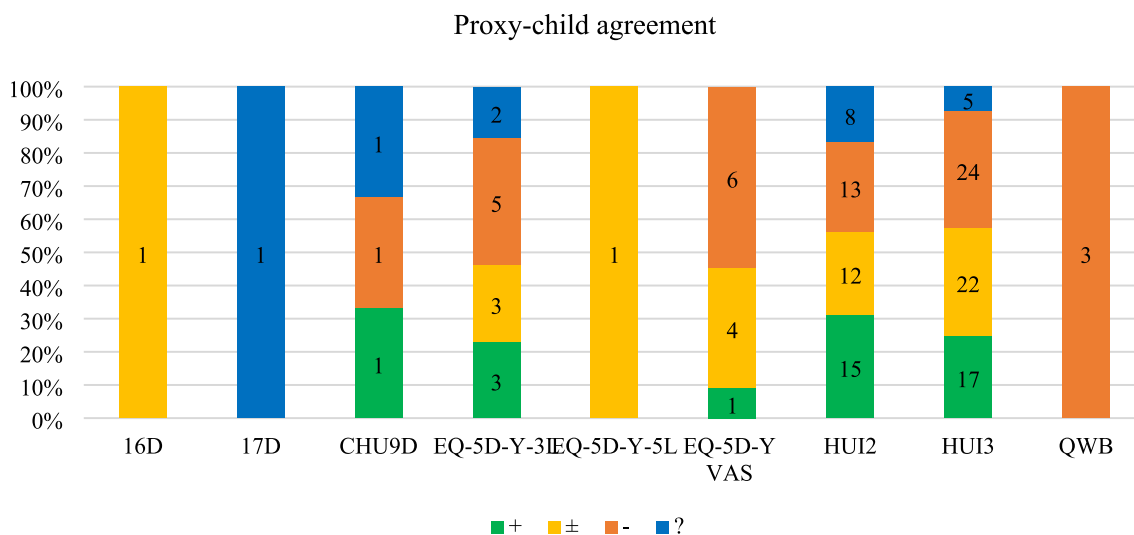
#### 3.5.4 Inter-Modal Reliability

Inter-modal reliability criteria rating outputs ( $n = 11$ ) were only available for three instruments and from three studies [63–65]. Modal comparisons of interest were between home- and clinic-based administration [63], between paper and online versions [64], and between telephone interview, face-to-face interview, and postal self-administration [65]. EQ-5D-Y-3L dimension and VAS scores had '+' outputs twice each: HUI2 '±' twice, and HUI3 '±' twice and '-' three times.

#### 3.5.5 Proxy-Child Agreement

Proxy-child agreement criteria rating outputs ( $n = 149$ ) were available for eight instruments. Proxies included caregivers,





**Fig. 3** Proxy-child agreement criteria rating outputs by instrument. Note: Absolute numbers of criteria rating outputs are displayed within each bar

parents, physicians, nurses, teachers, and therapists. Statistical methods included ICC, Kappa coefficient, and percentage agreement. Figure 3 shows the outputs by instrument. Agreements were generally low, with 24.8% (37 of 149) of outputs being ‘+’. CHU9D had the highest proportion of ‘+’ at 33.3%, followed by HUI2 at 31.3% from a considerably larger number of outputs.

### 3.5.6 Content Validity

At least one content validity criteria rating output was available for all instruments ( $n = 35$ ). Around half of the outputs ( $n = 18$ ) were reported in the original development of instruments, i.e., whether a conceptual framework for the purpose of measurement was stated, whether qualitative research with children or appropriate proxies were conducted for dimension/item elicitation, and whether cognitive interviews and piloting were conducted. An output of ‘+’ for original development was given to CHSCS-PS [38], CHU9D [39, 40], IQI [45], and TANDI [10, 49]; ‘±’ to 17D [34], AHUM [35], AQoL-6D [36], EQ-5D-Y-3L [41, 66], EQ-5D-Y-5L [42], HUI2 [43], HUI3 [44], and QWB [46]; and ‘-’ to 16D [33] and CH-6D [37].

Among other outputs ( $n = 17$ ), some concerned content validation of modified versions ( $n = 7$ ): cognitive bolt-on to EQ-5D-Y-3L [67] and the ‘child-friendly’ version of HUI2/3 [68]; the latter suggests conceptual issues in applying the standard HUI2/3 with children (aged 5–18 years in the study). Although not counted in the rating scale, 69 of 105 studies (65.7%) that directly or indirectly assessed HUI2 removed the fertility dimension. The remainder ( $n = 10$ ) concerned post-development content

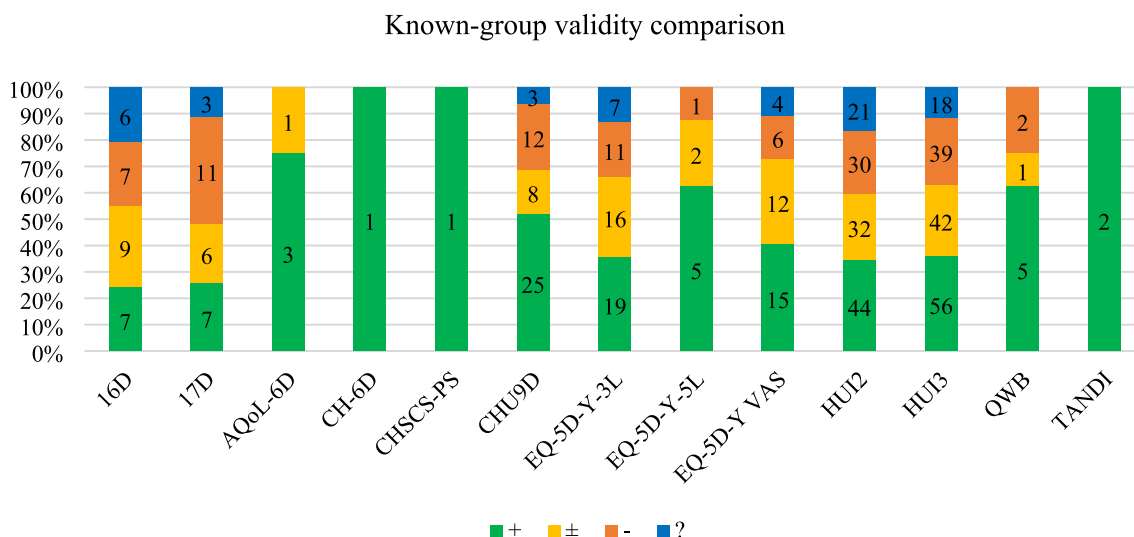
validations of CHU9D [69, 70], EQ-5D-Y-3L [70, 71], HUI2 [72–74], and HUI3 [75]. These studies conducted surveys and qualitative research with children and proxies to evaluate the instruments’ conceptual relevance to the childhood health construct of interest. Only the more recently developed instruments targeting preschool children—CHSCS-PS, IQI, and TANDI—received 100% ‘+’ outputs. Among instruments for older children, only CHU9D had an above-zero percentage of ‘+’ (40%).

### 3.5.7 Structural Validity

Only 10 criteria rating outputs were available for structural validity. Exploratory or confirmatory factor analysis was conducted for 16D [76] ( $n = 1$ ), CHU9D [77, 78] ( $n = 3$ ), EQ-5D-Y-3L with cognitive bolt-on [67] ( $n = 1$ ), and TANDI [49] ( $n = 1$ ). One study conducted multi-trait analyses of the correlations between dimension scores and dimension-relevant item responses for the HUI3 and found hypothesised correlations for all dimensions except cognition [68]. Another study estimated the between-item correlations for the EQ-5D-Y-3L, finding low–moderate correlations [79]. CHU9D and TANDI received 100% ‘+’ from three and one outputs, respectively.

### 3.5.8 Cross-Cultural Validity

There were 34 cross-cultural validity criteria rating outputs for 16D translations into French [80] and Norwegian [81]; 17D into French [80]; CHU9D into Chinese [51], Danish [82], Dutch [83], and Canadian French [84]; EQ-5D-Y-3L into English, German, Italian, Spanish, and Swedish [41, 79,



**Fig. 4** Known-group validity criteria rating outputs by instrument. Note: Absolute numbers of criteria rating outputs are displayed within each bar

85, 86] and Japanese [87]; EQ-5D-Y-5L into Chinese for the mainland [88] and Hong Kong [89]; HUI2 and HUI3 into Dutch [90], Canadian French [72, 91], German [92], Russian [93, 94], Spanish [95, 96], Swedish [97], and Turkish [98]. EQ-5D-Y (3L and 5L) thus reported on the highest number of translated versions, likely aided by the established Euro-QoL translation guideline. However, the overall proportion of ‘+’ outputs was 54.5% (6 of 11). Deficits in rating were generally due to a lack of backward translation or pilot testing with children. One study assessed the feasibility of a Spanish version of HUI2 and found that translation errors in emotion items resulted in a high volume of missing parental responses [96].

### 3.5.9 Known-Group Validity

Known-group validity criteria rating outputs ( $n = 500$ ) were available for all instruments except AHUM and IQI. Group comparisons included those between patients and healthy/general controls, between disease presence/absence, severities, and diagnosis types, and between sociodemographic groups where expected differences were specified *a priori* (e.g., by parental education, employment status or socioeconomic status [52, 99]). Statistical methods included the Mann–Whitney U and Kruskal–Wallis tests for continuous outcomes and the Chi-square test for categorical outcomes. Figure 4 shows the outputs by instrument. CH-6D, CHSCS-PS, and TANDI had 100% ‘+’ from one or two outputs, and AQoL-6D, EQ-5D-Y-5L, and QWB likewise had high proportions of ‘+’ outputs (75.0%, 62.5%, and 62.5%, respectively) from less than 10. Among others, CHU9D had the highest proportion of ‘+’ outputs at 52.1%.

### 3.5.10 Hypothesis Testing

Hypothesis test criteria rating outputs ( $n = 309$ ) were available for all instruments except IQI. Statistical methods included multivariate linear regression to estimate the adjusted association between the instrument score and a variable of interest (e.g., disease severity) and comparison against the *a priori* hypothesis. A high proportion (101 of 309, 32.7%) of outputs were ‘?’ since *ad hoc* tests without *a priori* hypotheses (e.g., by age group and sex) were placed under this property. Figure 5 shows the outputs by instrument. Other than AHUM with a single ‘+’ and AQoL-6D with 40% ‘+’ from five outputs, EQ-5D-Y-3L had the highest proportion of ‘+’ outputs at 35.7%, followed by EQ-5D-Y VAS at 34.5%.

### 3.5.11 Convergent Validity

Convergent validity criteria rating outputs ( $n = 202$ ) were available for all instruments except AHUM, CHSCS-PS, and IQI. The main statistical method involved correlations between comparable dimensions or between index utilities/total scores that measured similar constructs (e.g., between EQ-5D-Y-3L pain and PedsQL physical functioning dimensions [100], or between CHU9D index and PedsQL total score [101]), and assessing whether they were significantly different from zero and higher than prespecified thresholds of strength (e.g.,  $r > 0.40$ ). Figure 6 shows the outputs by instrument. CH-6D had 100% ‘+’ from only one output, while 16D and 17D had 85.7% ‘+’ from seven combined and EQ-5D-Y VAS had 39.1% ‘+’ from 23.

### 3.5.12 Discriminant Validity

Only seven discriminant validity criteria rating outputs were available for four instruments. Studies prespecified hypotheses of negligible correlations between dimensions and index/total scores measuring different constructs, e.g., between HUI3 index and quality-of-life dimensions of being, belonging, and becoming [102]. EQ-5D-Y-5L dimensions and VAS had one ‘+’ each; EQ-5D-Y-3L dimensions had one ‘+’ and one ‘±’, as did the HUI3 index; and CHU9D dimensions had one ‘?’ due to poor reporting.

### 3.5.13 Empirical Validity

Empirical validity criteria rating outputs ( $n = 29$ ) were available for five instruments. They all concerned differences in index utility scores across groups defined by self-reported general health status. Most outputs ( $n = 18$ ) concerned the CHU9D, all but one of which was ‘+’ (94.4%). Other instruments also had high proportions of ‘+’: AQoL-6D, 100% from two; EQ-5D-Y-3L, 100% from one; HUI2, 80% from five; and HUI3, 100% from three. This suggests acceptable empirical validity for the five instruments and particularly for the CHU9D.

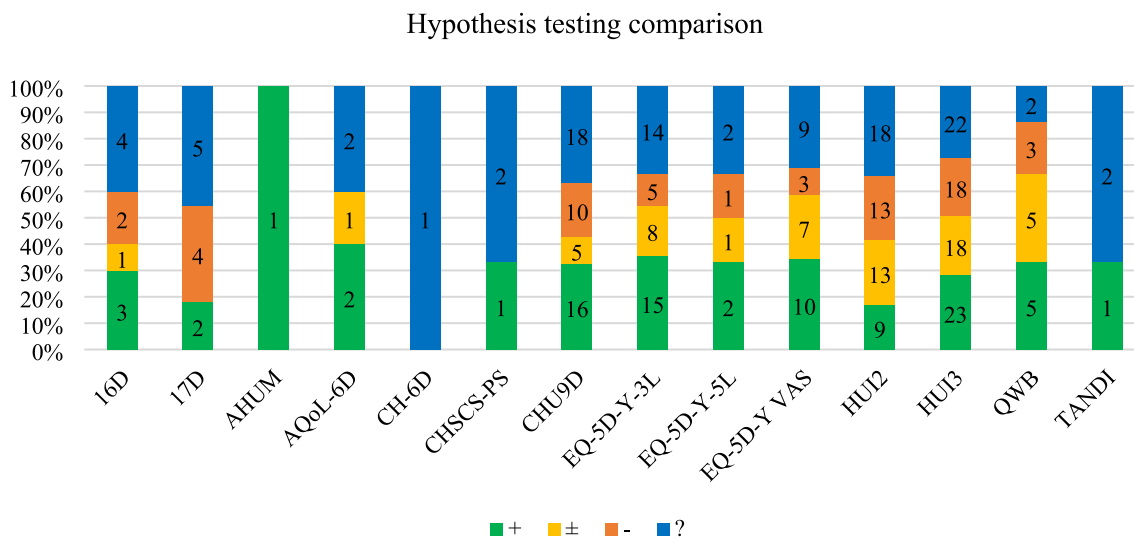


Fig. 5 Hypothesis testing criteria rating outputs by instrument. Note: Absolute numbers of criteria rating outputs are displayed within each bar

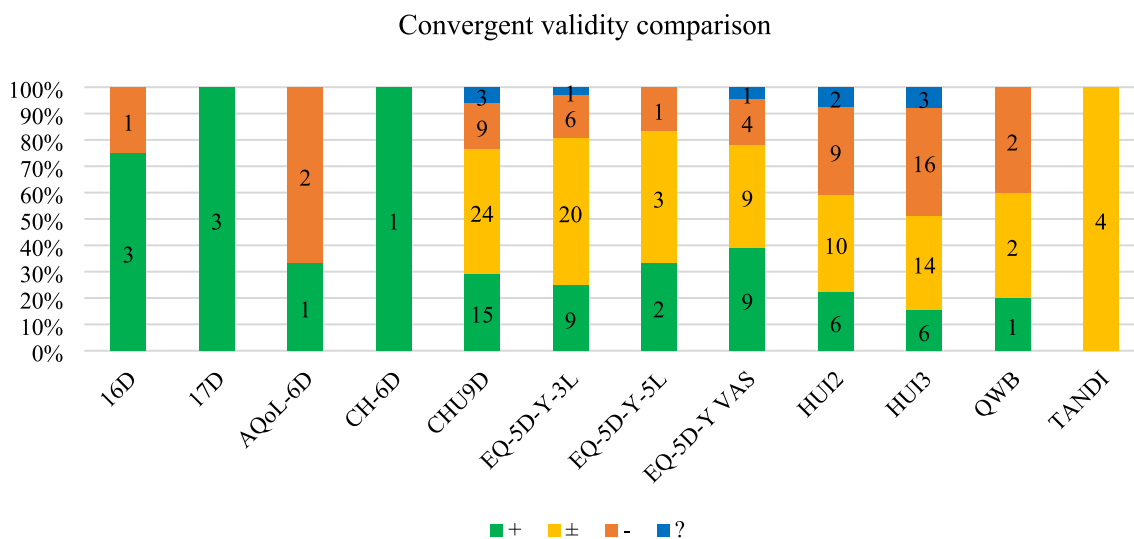


Fig. 6 Convergent validity criteria rating outputs by instrument. Note: Absolute numbers of criteria rating outputs are displayed within each bar

### 3.5.14 Concurrent Validity

Concurrent validity criteria rating outputs ( $n = 17$ ) were available for six instruments. Studies prespecified ‘gold standard’ measures and expected correlation strength and direction between the gold standard and the given instrument. EQ-5D-Y-3L had the highest number of outputs ( $n = 5$ ) with 40% ‘+’, while CHSCS-PS ( $n = 1$ ), CHU9D ( $n = 3$ ), and HUI3 ( $n = 2$ ) had 100% ‘+’.

### 3.5.15 Responsiveness

Responsiveness criteria rating outputs ( $n = 147$ ) were available for nine instruments. Studies assessed the statistical and clinical significance of changes in instrument score over time and whether these were consistent with *a priori* hypotheses and changes in the reference measure score (e.g., a disease-specific measure or clinical outcome). A significant proportion of outputs (64 of 147, 43.5%) was ‘?’ due to the non-inclusion of a reference measure. Figure 7 shows the outputs by instrument. EQ-5D-Y-5L had 100% ‘+’ from one output. Among others, HUI2 had the highest proportion of ‘+’ at 46.7%, followed by CHU9D at 45.5%.

### 3.5.16 Acceptability

Acceptability criteria rating outputs ( $n = 352$ ) were available for all instruments except AHUM, CH-6D, and IQI. Assessed features included missing data rates (due to instrument, not survey, design), ceiling effects (no study reported significant floor effects), time for completion, comprehensibility, and number of unique health states per sample respondent. The prespecified criterion that studies evaluating

only one acceptability feature receive ‘?’ resulted in a high proportion of outputs being ‘?’ (175 of 352, 49.7%). Figure 8 shows the outputs by instrument. CHSCS-PS had the highest proportion of ‘+’ at 50% but from only two outputs, followed by QWB at 33.3%, similarly from few outputs ( $n = 6$ ). Among instruments with 10 or more outputs, CHU9D had the highest proportion of ‘+’ at 29.7%.

Synthesising individual acceptability features separately, there were 183 assessments of ceiling effect, 254 of high missing data, 52 of response time, and 78 of other features (e.g., number of unique health states). Concerning the two most assessed features, namely ceiling effect and high missing data (see Table 1 for thresholds used), Tables G and H in the ESM compare their respective occurrences by instrument. In Table G, ceiling effect occurrences were tabulated separately for all childhood populations and samples including patients only. The ceiling effect occurrences did not diminish for patient populations even though a lower ceiling effect would be expected. As for the high missing data occurrence in Table H, AQoL-6D, CHSCS-PS, and TANDI had no occurrence, although only from one or two assessments each. Among others, EQ-5D-Y-3L and -5L had the lowest percentages of samples with high missing data: 11.5% and 0%, respectively.

### 3.5.17 Interpretability

Interpretability criteria rating outputs ( $n = 185$ ) were available for nine instruments. Only two studies sought to establish childhood population norm data, one for EQ-5D-Y-3L index in Japan [103] and the other for EQ-5D-Y-5L dimensions and VAS in Hong Kong, China [104]. Two studies calculated the minimal clinically important difference (MID) based on effect size for the EQ-5D-Y-3L index [105] and unweighted sum of

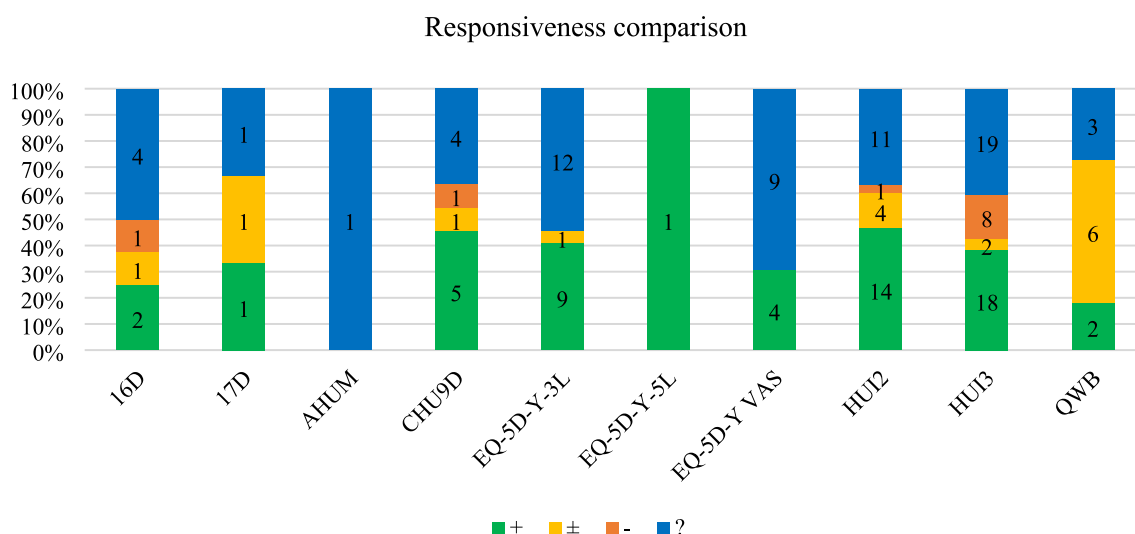


Fig. 7 Responsiveness criteria rating outputs by instrument. Note: Absolute numbers of criteria rating outputs are displayed within each bar

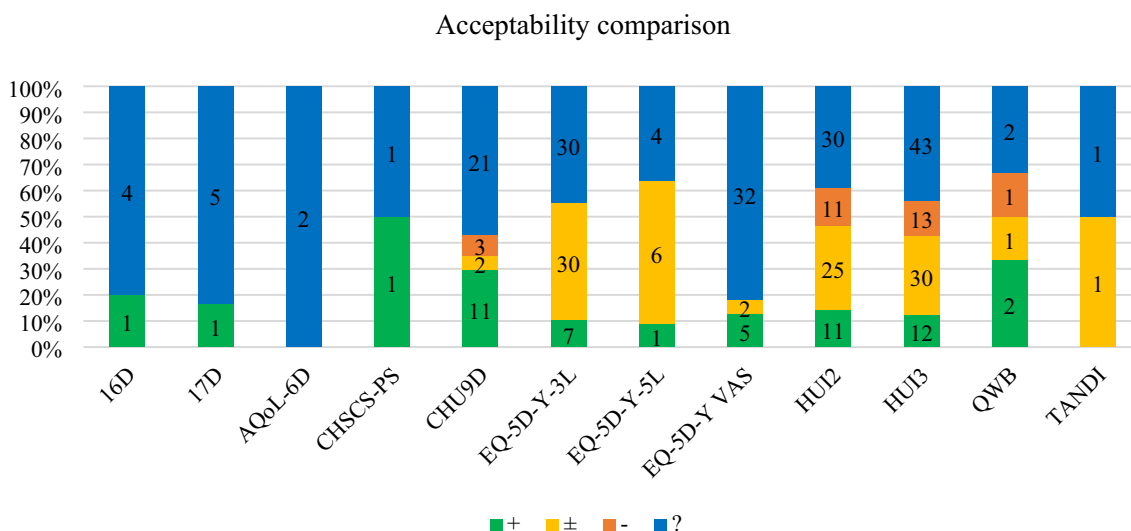


Fig. 8 Acceptability criteria rating outputs by instrument. Note: Absolute numbers of criteria rating outputs are displayed within each bar

item scores [106], while another estimated it as half of baseline standard deviation for EQ-5D-Y-3L unweighted sum and VAS [107].

Other studies used external data from the literature to interpret results, e.g., comparing the instrument scores with those of general/healthy childhood samples (not necessarily population norms), for which the output was evaluated as ‘±’. Studies referencing external MIDs repeatedly used those derived from non-childhood populations, for which the output was ‘?’. For example, studies for HUI2/3 referenced MIDs of 0.03 for index and 0.05 for dimension score from Horsman and colleagues [108], which were not based on primary derivation from childhood population but on expert

opinion. Figure 9 shows the outputs by instrument. Comparing the proportion of ‘±’ and ‘+’ combined (given the low incidence of ‘+’), 16D/17D had the highest proportion (85.2%) among instruments with 10 or more outputs.

### 4 Discussion

This review evaluated the psychometric performance of 14 generic childhood MAUIs [12], drawing on directly or indirectly provided psychometric evidence from 372 primary studies. Psychometric performance criteria were drawn from established standards in the literature to generate criteria

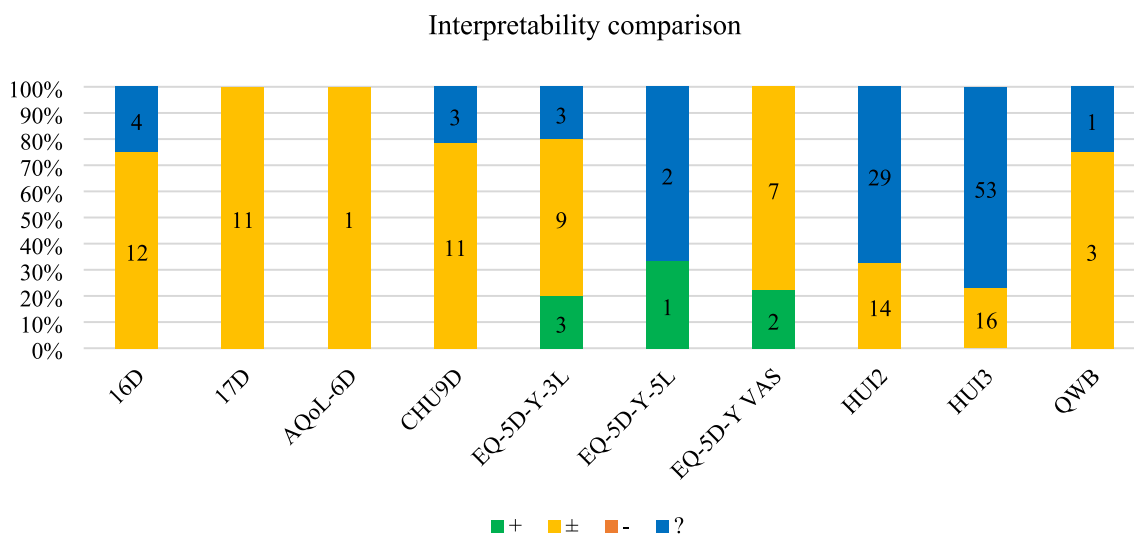


Fig. 9 Interpretability criteria rating outputs by instrument. Note: Absolute numbers of criteria rating outputs are displayed within each bar

rating outputs for 18 psychometric properties. The resulting 2153 outputs constitute a comprehensive psychometric evidence catalogue for generic childhood MAUIs, which analysts can use to inform MAUI selection for application. Further synthesis identified psychometric evidence gaps (e.g., instruments targeting young children below the age of 3 years) and summarised the range of psychometric assessment methods and results by property (e.g., whether a reference measure was included to test the instrument responsiveness against the measured health status change). No one instrument outperformed others across all properties, and the comparisons had to consider both the proportion of ‘+’ and the number of outputs from which the proportion is calculated.

As noted above, the evidence catalogue is necessary but insufficient to finalise the optimal MAUI selection for application. Considering established standards in the literature [17–21, 30], stakeholders involved in instrument selection should establish a minimum standard of scientific rigour for the given research setting. This would involve decisions on (1) which properties are relevant—proxy-child agreement, for example, would not be relevant for applications relying solely on proxy report; (2) the relative importance of different properties—responsiveness, for example, would be most important for application in clinical trials; and (3) the appropriate level of consistent acceptable performance for each relevant property within the accumulated evidence. As noted, the decision on (3) would involve setting thresholds for both the proportion of ‘+’ and the number of outputs from which the proportion is calculated: several instruments had very high proportions of ‘+’ but from a small pool of outputs. We note that the small pool of outputs may be due to recent development of an instrument and/or limited evidence generation relating to a specific psychometric property for that instrument. Moreover, the application will most likely target a specific childhood population, meaning that only a subset of the accumulated evidence (defined, for example, in terms of country, instrument language version, age group, and clinical characteristics) will be relevant. By extracting a wide range of study- and assessment-related variables, the current catalogue should aid analysts in retrieving the necessary psychometric evidence for the given minimum standard. Indeed, without application-specific target populations and minimum standards, reaching a conclusion on the overall performance of each measure from the highly heterogeneous evidence base is difficult [26].

For the decision on (1), a further question is whether the properties regularly assessed for non-preference-based PROMs and health status measures are relevant for preference-based MAUIs [19, 26]. Brazier and Deverill [19] argue that construct validity and responsiveness are not relevant to health utilities (although relevant to responses on the classification system prior to value set

application) since they concern *preferences* over health level/change, not the level/change itself. They instead recommend empirical validity, one test of which is whether utility scores reflect hypothetical preferences over health/disease states expressed as self-rated health status [19, 109]. Moreover, internal consistency for MAUIs should concern the relevance of items to societal preferences, not between-item correlations within the same dimension [19]. Structural validity should concern orthogonality of the items rather than their conformity to hypothesised dimensionality [110]. Likewise, content validity for MAUIs should concern the classification system’s relevance to individuals’ utility function, not its comprehensiveness in reflecting the health construct of interest [19].

In practice however all the above properties are included in psychometric assessments of MAUIs by primary studies and in their evaluations by psychometric reviews. For example, the recent review by Rowen and colleagues [26] did not distinguish between empirical validity and construct validity, while the review by Janssens and colleagues [27] covered internal consistency, structural validity, and content validity. There was likewise a nontrivial volume of evidence for these properties from primary studies, e.g.,  $n = 25$  for internal consistency (particularly for CHU9D) and  $n = 35$  for content validity (for all instruments, which contradicts the finding of Tan and colleagues [28] that content validity evidence was insufficient to be synthesised). These suggest that the psychometric properties specified in Table 1 remain relevant to generic childhood MAUIs. Analysts could therefore draw on the Table 1 criteria for the psychometric assessment or evaluation of MAUIs (generic or disease-specific, childhood or adult), particularly regarding the psychometric performance of both the utility index and the underlying health classification system, while noting the methodological ambiguities present in this space intersected by the disciplines of psychometric research and health economics.

Another key strength of the Table 1 criteria is the incorporation of the ‘?’ criteria rating output, given to 25.8% of 2153 psychometric assessments, which identified psychometric assessment designs and methods of poor quality. This quality evaluation was not incorporated in the Rowen review [26]. The Janssens review [27] incorporated the ‘?’ output but at the study level rather than at the more granulated assessment level. It nevertheless noted “significant methodological limitations” in several included studies (p. 342) [27]. The Tan review [28] highlighted the methodological limitations in the assessment of test-retest reliability and responsiveness in particular. In this review, the commonly observed reasons for the ‘?’ output were narrated in Sect. 3.5 by property, e.g., (1) the reliance on external MIDs derived from expert opinion or non-childhood populations [108, 111]; (2) assessment of only one acceptability feature out of many available; (3) lack of a reference measure or MID



to help interpret responsiveness evidence; and (4) lack of a clear *a priori* hypothesis in testing associations, particularly between instrument scores and sociodemographic variables. These methodological gaps, further detailed case-by-case in the Excel catalogue, should help inform the design of future psychometric assessments of existing and new measures.

A significant finding of this review is the high volume of psychometric evidence available from indirect studies (900 of 2153, or 41.8% of all criteria rating outputs) that did not explicitly aim to conduct psychometric assessments. Thirty percent of outputs from indirect studies were '?', which was higher, but not substantially so, than the 22.1% from direct studies. More than half of outputs for known-group validity, hypothesis testing, and responsiveness were drawn from indirect studies. These findings, i.e., the general non-inferiority and the nontrivial quantity of the psychometric evidence from indirect studies, justify the inclusion of both study types. This is a key strength of our review relative to previous reviews that covered direct studies only and which reported the scarcity of responsiveness evidence in particular [26–28]. It was found that the inclusion of indirect studies, while expanding the volume of evidence, did not increase the range of psychometric properties for which there was evidence. It did however expand the number of properties with at least one '+' output. Therefore, indirect study evidence supplements that of direct studies, which should enable future primary studies to focus on those properties for which little direct or indirect evidence currently exists, particularly reliability (internal consistency, test-retest, inter-rater, inter-modal) and proxy-child agreement as found in Sect. 3.4. Given that the scarcity of reliability evidence and the uncertainty over proxy-child agreement have already been emphasised in previous reviews [26–28], greater research attention on these properties is warranted.

This review has further key strengths. First, it is an up-to-date and comprehensive review on the psychometric performance of generic childhood MAUIs. Previous reviews predate the development of some notable instruments, including the EQ-5D-Y-5L targeting school-aged children, and TANDI and IQI targeting preschool children [22–27, 112, 113]. The Rowen and Tan reviews deliberately focused on EQ-5D-Y, HUI2/3, AQoL-6D, and CHU9D [26, 28], thereby excluding instruments targeting young children below the age of 3 years. By contrast, the significant psychometric evidence gaps for these instruments were a key finding in Sect. 3.4, with no evidence available for inter-modal reliability, cross-cultural validity, discriminant validity, empirical validity, responsiveness, and interpretability. Their inclusion in our review will help with planning future studies addressing these evidence gaps. Otherwise, they would face greater difficulty than the longer established ones in meeting a given minimum standard of scientific rigour. The current review

is also comprehensive in terms of the range of properties covered: the Rowen review, for example, excluded internal consistency and content validity [26]; the Janssens review excluded cross-cultural validity [27]; while the Tan review focused on known-group and convergent validities, test-retest, and responsiveness, and chose not to extract data on content validity despite this being a prespecified research objective [28].

The review nevertheless has several limitations. First, the review did not quantify (e.g., by using the COSMIN checklist [30]) a methodological/reporting quality score for each study and property. Such quantification for all 372 studies was deemed impractical in terms of time and research resources. This nevertheless neglects variation in methodological quality among assessments that avoided the '?' output. It is likely, for example, that assessments with larger sample sizes (all other factors being equal) provide more robust psychometric evidence than those with smaller sizes, even if both sets meet the minimum level of methodological quality. Second, the criteria rating applied uniform performance thresholds (e.g.,  $p$  value < 0.05 for statistical significance and ICC > 0.7 for proxy-child agreement being '+') unless specified otherwise and justified by primary studies. This created situations where an instrument found to be favourable relative to another within a study nevertheless received a negative rating and vice versa. For example, in the study by Ungar and colleagues [61], the child-parent dyad approach to questionnaire administration produced higher ICCs for proxy-child agreement than independent administrations for the HUI2 index and emotion dimension score and the HUI3 ambulation score. However, the ICC remained below 0.7, yielding a '-' rating for the dyad approach. Third, the presentation of the evidence synthesis results in Sects. 3.4 and 3.5 were not disaggregated to each component of the instrument (e.g., index, dimension, VAS) as done in the Rowen review [26], but the case-by-case evaluation results are available in the Excel catalogue in the ESM. Distinctions could also be drawn between different variable types derived from dimension and index scores, e.g., index utility as a continuous variable versus categorised into disability levels (mild, moderate, and severe).

There were also several caveats with the review methods applied. First, some studies applied instruments to childhood groups outside the instruments' intended target age ranges (e.g., HUI2/3 on children aged 3–6 years [114]), but assessments from such studies were not assigned a '?' output. The rationale was to compare the rating outputs (not conducted in this manuscript) between instrument applications on intended and unintended age groups, which would not be possible if the latter were uniformly assigned '?'. Second, evidence on cross-cultural validity was difficult to obtain via peer-reviewed publications. For example, according to the HUI website, HUI2/3 have at least 30 translated

versions [115], but the development of only seven versions were identified by this review. It is plausible that evidence on other psychometric properties might have been similarly missed by limiting the search to published articles and conference abstracts (e.g., instrument development steps in grey literature containing content validity evidence). Third, no study was identified that applied modern psychometric test theories such as Rasch analysis and item response theory [116]. This corroborates the finding of the Janssens review [27]. However, given the increasing use of modern theories for adult instrument development (e.g., EQ-5D-5L [117]), future psychometric reviews of childhood instruments will need to include criteria relevant to these theories.

## 5 Conclusion

This systematic review provides comprehensive and up-to-date evidence on the psychometric performance of generic childhood MAUIs that are designed to be accompanied by preference-based value sets. The inclusion of indirect evidence from studies without the explicit aim of psychometric assessment increased the comprehensiveness of the review. The catalogue of evaluated psychometric evidence provides a valuable resource for researchers and policymakers, particularly those involved in cost-effectiveness analysis, modelling, and decision-making, in selecting instruments for specific applications. The candidate instruments should meet a minimum standard of scientific rigour defined by the established criteria and consideration of the application setting. The final instrument(s) selected from the candidates should be that (those) with the most consistent performance according to the accumulated evidence in the review. The identified psychometric evidence gaps also motivate future psychometric studies, particularly the gaps on reliability and proxy-child agreement and on MAUIs targeting preschool children. The commonly observed issues in assessment design and methods, such as the statement of the *a priori* hypothesis for testing associations and changes, should likewise inform future psychometric studies.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s40258-023-00806-8>.

## Declarations

**Funding** This research has been funded by the Australian Government's Medical Research Future Fund under Grants MRF1200816 and MRF1199902. SP receives support as a National Institute for Health Research (NIHR) Senior Investigator (NF-SI-0616-10103) and from the NIHR Applied Research Collaboration Oxford and Thames Valley. The views expressed are those of the authors and not necessarily those of the Australian Government, the NIHR or the Department of Health and Social Care in the UK.

**Conflicts of interest** Joseph Kwon, Sarah Smith, Rakhee Raghunandan, Martin Howell, Elisabeth Huynh, Sungwook Kim, Thomas Bentley, Nia Roberts, Emily Lancsar, Kirsten Howard, Germaine Wong, Jonathan Craig, and Stavros Petrou have no conflict of interest to declare.

**Availability of data and material** The Excel file containing the individual criteria rating outputs and the rationale is available online.

**Author contributions** Conceptualisation and methodology: All authors. Database search, study selection and data extraction: JK, EH, MH, SK, TB, SP and RR. Data synthesis: JK, EH, MH, SK, TB and RR; SS and SP for verification. First manuscript draft writing: JK. Draft review and editing: All authors. All authors read and approved the final manuscript.

**Ethics approval** Not applicable.

**Consent to participate** Not applicable.

**Consent for publication (from patients/participants)** Not applicable.

**Code availability** Not applicable.

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

## References

1. Drummond MF, Sculpher MJ, Claxton K, Stoddart GL, Torrance GW. Methods for the economic evaluation of health care programmes. Oxford: Oxford University Press; 2015.
2. National Institute for Health and Care Excellence. Guide to the methods of technology appraisal 2013. PMG92013.
3. Canadian Agency for Drugs and Technologies in Health. Guidelines for the economic evaluation of health technologies: Canada 4th Edition. Canadian Agency for Drugs and Technologies in Health; 2017.
4. Pharmaceutical Benefits Advisory Committee. Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee (Version 5). Pharmaceutical Benefits Advisory Committee; 2016.
5. Scottish Medicines Consortium. Working with SMC—a guide for manufacturers. Scottish Medicines Consortium; 2017.
6. Brazier J, Ratcliffe J, Saloman J, Tsuchiya A. Measuring and valuing health benefits for economic evaluation. Oxford: Oxford University Press; 2017.
7. Chen G, Ratcliffe J. A review of the development and application of generic multi-attribute utility instruments for paediatric populations. *Pharmacoeconomics*. 2015;33(10):1013–28. <https://doi.org/10.1007/s40273-015-0286-7>.

8. Torrance GW. Measurement of health state utilities for economic appraisal: a review. *J Health Econ.* 1986;5(1):1–30.
9. Petrou S. Methodological issues raised by preference-based approaches to measuring the health status of children. *Health Econ.* 2003;12(8):697–702. <https://doi.org/10.1002/hec.775>.
10. 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):1–15.
11. Matza LS, Patrick DL, Riley AW, Alexander JJ, Rajmil L, Pleil AM, et al. Pediatric patient-reported outcome instruments for research to support medical product labeling: report of the ISPOR PRO good research practices for the assessment of children and adolescents task force. *Value Health.* 2013;16(4):461–79.
12. 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.* 2022:1–53.
13. Ungar WJ. Challenges in health state valuation in paediatric economic evaluation: are QALYs contraindicated? *Pharmacoeconomics.* 2011;29(8):641–52. <https://doi.org/10.2165/11591570>.
14. Khadka J, Kwon J, Petrou S, Lancsar E, Ratcliffe J. Mind the (inter-rater) gap. An investigation of self-reported versus proxy-reported assessments in the derivation of childhood utility values for economic evaluation: a systematic review. *Soc Sci Med.* 2019;240: 112543.
15. Eiser C, Morse R. Can parents rate their child's health-related quality of life? Results of a systematic review. *Qual Life Res.* 2001;10(4):347–57.
16. 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 in Health.* 2012;15(5):730–6.
17. Smith S, Lamping D, Banerjee S, Harwood R, Foley B, Smith P, et al. Measurement of health-related quality of life for people with dementia: development of a new instrument (DEMQOL) and an evaluation of current methodology. *Health Technol Assess (Winchester, England).* 2005;9(10):1–iv.
18. Food and Drug Administration. Patient reported outcome measures: use in medical product development to support labelling claims. Washington DC: Food and Drug Administration; 2009.
19. Brazier J, Deverill M. A checklist for judging preference-based measures of health related quality of life: learning from psychometrics. *Health Econ.* 1999;8(1):41–51.
20. Reeve BB, Wyrwich KW, Wu AW, Velikova G, Terwee CB, Snyder CF, et al. ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. *Qual Life Res.* 2013;22(8):1889–905.
21. Lohr KN. Assessing health status and quality-of-life instruments: attributes and review criteria. *Qual Life Res.* 2002;11(3):193–205.
22. Eiser C, Morse R. A review of measures of quality of life for children with chronic illness. *Arch Dis Child.* 2001;84(3):205–11.
23. Davis E, Waters E, Mackinnon A, Reddihough D, Graham HK, Mehmet-Radji O, et al. Paediatric quality of life instruments: a review of the impact of the conceptual framework on outcomes. *Dev Med Child Neurol.* 2006;48(4):311–8.
24. Grange A, Bekker H, Noyes J, Langley P. Adequacy of health-related quality of life measures in children under 5 years old: systematic review. *J Adv Nurs.* 2007;59(3):197–220.
25. Solans M, Pane S, Estrada MD, Serra-Sutton V, Berra S, Herdman M, et al. 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.
26. Rowen D, Keetharuth AD, Poku E, Wong R, Pennington B, Wailoo A. A review of the psychometric performance of selected child and adolescent preference-based measures used to produce utilities for child and adolescent health. *Value Health.* 2020.
27. Janssens A, Rogers M, Coon JT, Allen K, Green C, Jenkinson C, et al. A systematic review of generic multidimensional patient-reported outcome measures for children, part II: evaluation of psychometric performance of English-language versions in a general population. *Value Health.* 2015;18(2):334–45.
28. Tan RL-Y, Soh SZY, Chen LA, Herdman M, Luo N. Psychometric properties of generic preference-weighted measures for children and adolescents: a systematic review. *Pharmacoeconomics.* 2022:1–20.
29. 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.
30. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Qual Life Res.* 2010;19(4):539–49.
31. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. *BMC Med Res Methodol.* 2010;10(1):1–8.
32. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Bmj.* 2021;372.
33. Apajasalo M, Sintonen H, Holmberg C, Sinkkonen J, Aalberg V, Pihko H, et al. Quality of life in early adolescence: a sixteen-dimensional health-related measure (16D). *Qual Life Res.* 1996;5(2):205–11.
34. Apajasalo M, Rautonen J, Holmberg C, Sinkkonen J, Aalberg V, Pihko H, et al. Quality of life in pre-adolescence: a 17-dimensional health-related measure (17D). *Qual Life Res.* 1996;5(6):532–8.
35. Beusterien KM, Yeung J-E, Pang F, Brazier J. Development of the multi-attribute adolescent health utility measure (AHUM). *Health Qual Life Outcomes.* 2012;10(1):1–9.
36. Moodie M, Richardson J, Rankin B, Iezzi A, Sinha K. Predicting time trade-off health state valuations of adolescents in four Pacific countries using the Assessment of Quality-of-Life (AQoL-6D) instrument. *Value Health.* 2010;13(8):1014–27. <https://doi.org/10.1111/j.1524-4733.2010.00780.x>.
37. Kang E. Validity of Child Health-6 Dimension (Ch-6d) for Adolescents. *Value Health.* 2016;19(7):A854-A. <https://doi.org/10.1016/j.jval.2016.08.458>.
38. Saigal S, Rosenbaum P, Stoskopf B, Hault L, Furlong W, Feeny D, et al. Development, reliability and validity of a new measure of overall health for pre-school children. *Qual Life Res.* 2005;14(1):243–52.
39. 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.
40. 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.
41. 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.

42. Kreimeier S, Åström M, Burström K, Egmar A-C, 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.
43. Torrance GW, Feeny DH, Furlong WJ, Barr RD, Zhang Y, Wang Q. Multiattribute utility function for a comprehensive health status classification system: Health Utilities Index Mark 2. *Medical care.* 1996;702-22.
44. Furlong WJ, Feeny DH, Torrance GW, Barr RD. The Health Utilities Index (HUI®) system for assessing health-related quality of life in clinical studies. *Ann Med.* 2001;33(5):375–84.
45. Jabrayilov R, van Asselt AD, Vermeulen KM, Volger S, Detzel P, Dainelli L, et al. A descriptive system for the Infant health-related Quality of life Instrument (IQI): Measuring health with a mobile app. *PLoS ONE.* 2018;13(8): e0203276.
46. Kaplan RM, Bush JW, Berry CC. Health status: types of validity and the index of well-being. *Health Serv Res.* 1976;11(4):478–507.
47. Kaplan RM, Sieber WJ, Ganiats TG. The quality of well-being scale: comparison of the interviewer-administered version with a self-administered questionnaire. *Psychol Health.* 1997;12(6):783–91.
48. Verstraete J, Amien R. Cross-cultural adaptation and validation of the EuroQoL toddler and infant populations instrument Into Afrikaans for South Africa. *Value Health Regional Issues.* 2023;35:78–86.
49. 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-Reported Outcomes.* 2020;4(1):1–14.
50. Veritas Health Innovation. Covidence systematic review software. Veritas Health Innovation, Melbourne
51. Yang P, Chen G, Wang P, Zhang K, Deng F, Yang H, et al. Psychometric evaluation of the Chinese version of the Child Health Utility 9D (CHU9D-CHN): a school-based study in China. *Qual Life Res.* 2018;27(7):1921–31.
52. Zanganeh M, Adab P, Li B, Frew E. An assessment of the construct validity of the Child Health Utility 9D-CHN instrument in school-aged children: evidence from a Chinese trial. *Health Qual Life Outcomes.* 2021;19(1):1–14.
53. Foster Page LA, Beckett DM, Cameron CM, Thomson WM. Can the Child Health Utility 9D measure be useful in oral health research? *Int J Paediatr Dent.* 2015;25(5):349–57. <https://doi.org/10.1111/ipd.12177>.
54. Gaitan-Lopez DF, Correa-Bautista JE, Vinaccia S, Ramirez-Velez R. Self-report health-related quality of life among children and adolescents from Bogota, Colombia. The FUPRECOL study. *Colombia medica (Cali, Colombia).* 2017;48(1):12–8.
55. Bashir NS, Walters TD, Griffiths AM, Ungar WJ. An assessment of the validity and reliability of the pediatric child health utility 9D in children with inflammatory bowel disease. *Children.* 2021;8(5):343.
56. Hsu C-N, Lin H-W, Pickard AS, Tain Y-L. EQ-5D-Y for the assessment of health-related quality of life among Taiwanese youth with mild-to-moderate chronic kidney disease. *Int J Qual Health Care.* 2018;30(4):298–305.
57. Juniper EF, Guyatt GH, Feeny DH, Griffith LE, Ferrie PJ. Minimum skills required by children to complete health-related quality of life instruments for asthma: comparison of measurement properties. *Eur Respir J.* 1997;10(10):2285–94. <https://doi.org/10.1183/09031936.97.10102285>.
58. Lee JM, Rhee K, O'grady MJ, Basu A, Winn A, John P, et al. Health utilities for children and adults with type 1 diabetes. *Med Care.* 2011;49(10):924.
59. Livingston MH, Rosenbaum PL. Adolescents with cerebral palsy: stability in measurement of quality of life and health-related quality of life over 1 year. *Dev Med Child Neurol.* 2008;50(9):696–701.
60. Mayoral K, Rajmil L, Murillo M, Garin O, Pont A, Alonso J, et al. Measurement properties of the online EuroQoL-5D-youth instrument in children and adolescents with type 1 diabetes mellitus: questionnaire study. *J Med Internet Res.* 2019;21(11): e14947.
61. Ungar WJ, Boydell K, Dell S, Feldman BM, Marshall D, Willan A, et al. A parent-child dyad approach to the assessment of health status and health-related quality of life in children with asthma. *Pharmacoeconomics.* 2012;30(8):697–712.
62. Wong CKH, Cheung PWH, Luo N, Lin J, Cheung JPY. Responsiveness of EQ-5D youth version 5-level (EQ-5D-5L-Y) and 3-level (EQ-5D-3L-Y) in patients with idiopathic scoliosis. *Spine.* 2019;44(21):1507–14.
63. Glaser A, Davies K, Walker D, Brazier D. Influence of proxy respondents and mode of administration on health status assessment following central nervous system tumours in childhood. *Quality Life Res.* 1997;6(1):0.
64. Robles N, Rajmil L, Rodriguez-Arjona D, Azuara M, Codina F, Raat H, et al. Development of the web-based Spanish and Catalan versions of the Euroqol 5D-Y (EQ-5D-Y) and comparison of results with the paper version. *Health Qual Life Outcomes.* 2015;13(1):1–9.
65. Verrips G, Stuifbergen M, Den Ouden A, Bonsel G, Gemke R, Paneth N, et al. Measuring health status using the Health Utilities Index: agreement between raters and between modalities of administration. *J Clin Epidemiol.* 2001;54(5):475–81.
66. Burström K, Egmar A-C, Lugné A, Eriksson M, Svartengren M. A Swedish child-friendly pilot version of the EQ-5D instrument—the development process. *Eur J Pub Health.* 2011;21(2):171–7.
67. Ludwig K, Surmann B, Racker E, Greiner W. Developing and testing a cognitive bolt-on for the EQ-5D-Y (Youth). *Qual Life Res.* 2022;31(1):215–29.
68. Le Galès C, Costet N, Gentet JC, Kalifa C, Frappaz D, Edan C, et al. Cross-cultural adaptation of a health status classification system in children with cancer. First results of the French adaptation of the Health Utilities Index Marks 2 and 3. *Int J Cancer.* 1999;83(S12):112–8.
69. 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;13:22. <https://doi.org/10.1186/s12955-015-0218-4>.
70. 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.
71. Verstraete J, Lloyd A, Scott D, Jelsma J. How does the EQ-5D-Y Proxy version 1 perform in 3, 4 and 5-year-old children? *Health Qual Life Outcomes.* 2020;18:1–10.
72. Trudel J, Rivard M, Dobkin P, Leclerc J-M, Robaey P. Psychometric properties of the Health Utilities Index Mark 2 system in paediatric oncology patients. *Qual Life Res.* 1998;7(5):421–32.
73. Midgley DE, Bradlee TA, Donohoe C, Kent KP, Alonso EM. Health-related quality of life in long-term survivors of pediatric liver transplantation. *Liver Transpl.* 2000;6(3):333–9.
74. Miller TR, Steinbeigle R, Wicks A, Lawrence BA, Barr M, Barr RG. Disability-adjusted life-year burden of abusive head trauma at ages 0–4. *Pediatrics.* 2014;134(6):e1545–50.
75. Hinds PS, Burghen EA, Zhou Y, Zhang L, West N, Bashore L, et al. The Health Utilities Index 3 invalidated when completed by nurses for pediatric oncology patients. *Cancer Nurs.*



- 2007;30(3):169–77. <https://doi.org/10.1097/01.NCC.0000270700.11425.4d>.
76. Granö N, Kiesseppä T, Karjalainen M, Roine M. Exploratory factor analysis of a 16D Health-Related Quality of Life instrument with adolescents seeking help for early psychiatric symptoms. *Nord J Psychiatry*. 2016;70(2):81–7.
  77. Lindvall K, Vaezghasemi M, Feldman I, Ivarsson A, Stevens KJ, Petersen S. Feasibility, reliability and validity of the health-related quality of life instrument Child Health Utility 9D (CHU9D) among school-aged children and adolescents in Sweden. *Health Qual Life Outcomes*. 2021;19(1):1–12.
  78. Mpundu-Kaambwa C, Chen G, Huynh E, Russo R, Ratcliffe J. Does the study population and the use of proxy respondent have an effect on the latent quality of life constructs measured by the CHU9D And the Pedsqtlm 4.0? An exploratory factor analysis. *Value Health*. 2017;20(9):A503–4.
  79. Otto C, Barthel D, Klasen F, Nolte S, Rose M, Meyrose A-K, et al. Predictors of self-reported health-related quality of life according to the EQ-5D-Y in chronically ill children and adolescents with asthma, diabetes, and juvenile arthritis: longitudinal results. *Qual Life Res*. 2018;27(4):879–90.
  80. Rondeau É, Desjardins L, Laverdière C, Sinnett D, Haddad É, Sultan S. French-language adaptation of the 16D and 17D Quality of Life measures and score description in two Canadian pediatric samples. *Health Psychol Behav Med*. 2021;9(1):619–35.
  81. Aas E, Iversen T, Holt T, Ormhaug SM, Jensen TK. Cost-effectiveness analysis of trauma-focused cognitive behavioral therapy: a randomized control trial among Norwegian youth. *J Clin Child Adolesc Psychol*. 2019;48(sup1):S298–311.
  82. Petersen KD, Ratcliffe J, Chen G, Serles D, Frøsig CS, Olesen AV. The construct validity of the child health utility 9D-DK instrument. *Health Qual Life Outcomes*. 2019;17(1):1–12.
  83. Rowen D, Mulhern B, Stevens K, Vermaire JH. Estimating a Dutch value set for the pediatric preference-based CHU9D using a discrete choice experiment with duration. *Value Health*. 2018;21(10):1234–42.
  84. Poder TG, Carrier N, Mead H, Stevens KJ. Canadian French translation and linguistic validation of the child health utility 9D (CHU9D). *Health Qual Life Outcomes*. 2018;16(1):1–7.
  85. Scalone L, Tommasetto C, Matteucci MC, Selleri P, Broccoli S, Pacelli B, et al. Assessing Quality of life in children and adolescents: development and validation of the Italian version of EQ-5D-Y. *Italian J Public Health*. 2011;8(4):331–41.
  86. Perez Sousa MÁ, Sánchez-Toledo PO, Fuertea NG. Parent-child discrepancy in the assessment of health-related quality of life using the EQ-5D-Y questionnaire. *Arch Argent Pediatr*. 2017;115(6):541–6.
  87. Shiroiwa T, Fukuda T, Shimozuma K. Psychometric properties of the Japanese version of the EQ-5D-Y by self-report and proxy-report: reliability and construct validity. *Qual Life Res*. 2019;28(11):3093–105.
  88. Pei W, Yue S, Zhi-Hao Y, Ruo-Yu Z, Bin W, Nan L. Testing measurement properties of two EQ-5D youth versions and KID-SCREEN-10 in China. *Eur J Health Econ*. 2021;22(7):1083–93.
  89. Wong CKH, Cheung PWH, Luo N, Cheung JPY. A head-to-head comparison of five-level (EQ-5D-5L-Y) and three-level EQ-5D-Y questionnaires in paediatric patients. *Eur J Health Econ*. 2019;20(5):647–56.
  90. Gemke RJ, Bonsel GJ. Reliability and validity of a comprehensive health status measure in a heterogeneous population of children admitted to intensive care. *J Clin Epidemiol*. 1996;49(3):327–33.
  91. Nixon Speechley K, Maunsell E, Desmeules M, Schanzer D, Landgraf JM, Feeny DH, et al. Mutual concurrent validity of the child health questionnaire and the health utilities index: an exploratory analysis using survivors of childhood cancer. *Int J Cancer*. 1999;83(S12):95–105.
  92. Felder-Ptuig R, Frey E, Sonnleithner G, Feeny D, Gadner H, Barr RD, et al. German cross-cultural adaptation of the Health Utilities Index and its application to a sample of childhood cancer survivors. *Eur J Pediatr*. 2000;159(4):283–8.
  93. Gorinova Y, Samsonova M, Simonova O, Vinyarskaya I, Chernikov V. 309 First results of health status assessment in children with cystic fibrosis using Russian version of HUI Questionnaire. *J Cyst Fibros*. 2012;11:S136.
  94. Simonova O, Gorinova Y, Vinyarskaya I, Chernikov V. Validation of Russian Version of Health Utility Index Questionnaire in Children with Cystic Fibrosis. *Value in Health*. 2014;17(7):A731.
  95. Szecket N, Medin G, Furlong WJ, Feeny DH, Barr RD, Depauw S. Preliminary translation and cultural adaptation of Health Utilities Index questionnaires for application in Argentina. *Int J Cancer*. 1999;83(S12):119–24.
  96. Fu L, Talsma D, Baez F, Bonilla M, Moreno B, Ah-Chu M, et al. Measurement of health-related quality of life in survivors of cancer in childhood in Central America: feasibility, reliability, and validity. *J Pediatr Hematol Oncol*. 2006;28(6):331–41.
  97. Philipsson A, Duberg A, Möller M, Hagberg L. Cost-utility analysis of a dance intervention for adolescent girls with internalizing problems. *Cost Effectiveness Resour Alloc*. 2013;11(1):1–9.
  98. Boran P, Horsman J, Tokuc G, Furlong W, Muradoglu PU, Vagas E. Translation and cultural adaptation of health utilities index with application to pediatric oncology patients during neutropenia and recovery in Turkey. *Pediatr Blood Cancer*. 2011;56(5):812–7. <https://doi.org/10.1002/psc.22835>.
  99. 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.
  100. 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.
  101. Fantaguzzi C, Allen E, Miners A, Christie D, Opondo C, Sadique Z, et al. Health-related quality of life associated with bullying and aggression: a cross-sectional study in English secondary schools. *Eur J Health Econ*. 2017. <https://doi.org/10.1007/s10198-017-0908-4>.
  102. Rosenbaum PL, Livingston MH, Palisano RJ, Galuppi BE, Russell DJ. Quality of life and health-related quality of life of adolescents with cerebral palsy. *Dev Med Child Neurol*. 2007;49(7):516–21.
  103. Shiroiwa T, Fukuda T. EQ-5D-Y population norms for Japanese children and adolescents. *Pharmacoeconomics*. 2021;39(11):1299–308.
  104. Wong CK, Wong RS, Cheung JP, Tung KT, Yam J, Rich M, et al. Impact of sleep duration, physical activity, and screen time on health-related quality of life in children and adolescents. *Health Qual Life Outcomes*. 2021;19(1):1–13.
  105. Wu X, Ohinmaa A, Johnson J, Veugelers P. Assessment of children's own health status using visual analogue scale and descriptive system of the EQ-5D-Y: linkage between two systems. *Qual Life Res*. 2014;23(2):393–402.
  106. Moula Z, Powell J, Karkou V. An investigation of the effectiveness of arts therapies interventions on measures of quality of life and wellbeing: a pilot randomized controlled study in primary schools. *Front Psychol*. 2020:3591.
  107. Trigg A, Brohan E, Cocks K, Jones A, Monfared AAT, Chabot I, et al. Health-related quality of life in pediatric patients with partial onset seizures or primary generalized tonic-clonic seizures

- receiving adjunctive perampanel. *Epilepsy Behav.* 2021;118:107938.
108. Horsman J, Furlong W, Feeny D, Torrance G. The Health Utilities Index (HUI®): concepts, measurement properties and applications. *Health Qual Life Outcomes.* 2003;1(1):1–13.
109. Petrou S, Hockley C. An investigation into the empirical validity of the EQ-5D and SF-6D based on hypothetical preferences in a general population. *Health Econ.* 2005;14(11):1169–89.
110. Young T, Yang Y, Brazier JE, Tsuchiya A, Coyne K. The first stage of developing preference-based measures: constructing a health-state classification using Rasch analysis. *Qual Life Res.* 2009;18(2):253–65.
111. Drummond M. Introducing economic and quality of life measurements into clinical studies. *Ann Med.* 2001;33(5):344–9.
112. Harding L. Children's quality of life assessments: a review of generic and health related quality of life measures completed by children and adolescents. *Clin Psychol Psychotherapy Int J Theory Pract.* 2001;8(2):79–96.
113. Cremeens J, Eiser C, Blades M. Characteristics of health-related self-report measures for children aged three to eight years: a review of the literature. *Qual Life Res.* 2006;15(4):739–54.
114. McNamara HC, Wood R, Chalmers J, Marlow N, Norrie J, MacLennan G, et al. STOPPIT Baby Follow-up Study: the effect of prophylactic progesterone in twin pregnancy on childhood outcome. *PLoS ONE.* 2015;10(4): e0122341.
115. Health Utilities Inc. Questionnaire Development, Translations and Support 2018 [12th September 2022]. <http://www.healthutilities.com/>.
116. Baylor C, Hula W, Donovan NJ, Doyle PJ, Kendall D, Yorkston K. An introduction to item response theory and Rasch models for speech-language pathologists. 2011.
117. Bilbao A, Martín-Fernández J, García-Pérez L, Mendezona JI, Arrasate M, Candela R, et al. Psychometric properties of the EQ-5D-5L in patients with major depression: factor analysis and Rasch analysis. *J Ment Health.* 2022;31(4):506–16.

## Authors and Affiliations

Joseph Kwon<sup>1</sup>  · Sarah Smith<sup>2</sup>  · Rakhee Raghunandan<sup>3</sup>  · Martin Howell<sup>3</sup>  · Elisabeth Huynh<sup>4</sup>  ·  
 Sungwook Kim<sup>1</sup>  · Thomas Bentley<sup>5</sup>  · Nia Roberts<sup>6</sup>  · Emily Lancsar<sup>4</sup>  · Kirsten Howard<sup>3</sup>  · Germaine Wong<sup>3</sup>  ·  
 Jonathan Craig<sup>7</sup>  · Stavros Petrou<sup>1</sup> 

✉ Stavros Petrou  
stavros.petrou@phc.ox.ac.uk

Joseph Kwon  
joseph.kwon@phc.ox.ac.uk

Sarah Smith  
sarah.smith@lshtm.ac.uk

Rakhee Raghunandan  
rakhee.raghunandan@sydney.edu.au

Martin Howell  
martin.howell@sydney.edu.au

Elisabeth Huynh  
elisabeth.huynh@anu.edu.au

Sungwook Kim  
sungwook.kim@phc.ox.ac.uk

Thomas Bentley  
thomas.bentley@gtc.ox.ac.uk

Nia Roberts  
nia.roberts@bodleian.ox.ac.uk

Emily Lancsar  
emily.lancsar@anu.edu.au

Kirsten Howard  
kirsten.howard@sydney.edu.au

Germaine Wong  
germaine.wong@health.nws.gov.au

Jonathan Craig  
jonathan.craig@flinders.edu.au

<sup>1</sup> Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK

<sup>2</sup> Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, London, UK

<sup>3</sup> School of Public Health, University of Sydney, Sydney, NSW, Australia

<sup>4</sup> Department of Health Services Research and Policy, Australian National University, Canberra, ACT, Australia

<sup>5</sup> Medical Sciences Division, University of Oxford, Oxford, UK

<sup>6</sup> Bodleian Health Care Libraries, University of Oxford, Oxford, UK

<sup>7</sup> College of Medicine and Public Health, Flinders University, Adelaide, SA, Australia