A REVIEW OF THE PSYCHOMETRIC PERFORMANCE OF CHILD AND ADOLESCENT PREFERENCE-BASED MEASURES USED TO GENERATE UTILITY VALUES FOR CHILDREN

REPORT BY THE DECISION SUPPORT UNIT

15th January 2020

Donna Rowen, Anju Keetharuth, Edith Poku, Ruth Wong, Becky Pennington, Allan Wailoo

School of Health and Related Research, University of Sheffield

Decision Support Unit, ScHARR, University of Sheffield, Regent Court, 30 Regent Street
Sheffield, S1 4DA

Tel (+44) (0)114 222 0734
E-mail dsuadmin@sheffield.ac.uk
Website www.nicedsu.org.uk
Twitter @NICE_DSU

ABOUT THE DECISION SUPPORT UNIT

The Decision Support Unit (DSU) External Assessment Centre is based at the University of Sheffield with members at York, Bristol, Leicester and the London School of Hygiene and Tropical Medicine. The DSU is commissioned by The National Institute for Health and Care Excellence (NICE) to provide a research and training resource to support the Institute's Centre for Health Technology Evaluation Programmes. Please see our website for further information www.nicedsu.org.uk.

The production of this document was funded by the National Institute for Health and Care Excellence (NICE) through its Decision Support Unit. The views, and any errors or omissions, expressed in this document are of the authors only. NICE may take account of part or all of this document if it considers it appropriate, but it is not bound to do so.

Acknowledgements

We would like to thank Sophie Cooper, Sarah Davis, Alan Lamb, Rosie Lovett and Tracey Young for commenting on previous drafts. We would also like to thank Donna Davis and Liz Metham for project management and formatting of the report.

This report should be referenced as follows:

Rowen D, Keetharuth A, Poku E, Wong R, Pennington B, Wailoo A. A review of the psychometric performance of child and adolescent preference-based measures used to generate utility values for children. NICE DSU Report. 2020.

EXECUTIVE SUMMARY

NICE needs to assess the suitability of different approaches for estimating health state utilities across the broad range of conditions that feature in its guidance producing programmes in order to recommend the preferred measure in most situations. When considering approaches for adults, this assessment has been informed by reviews of psychometric performance of preference-based measures in studies that span a wide range of health conditions. Psychometric performance includes assessments of validity, responsiveness, reliability, acceptability and feasibility. However, similar reviews of the psychometric performance of child and adolescent preference-based measures have not been performed. Generic child and adolescent preference-based measures that can be used to generate health state utilities for children and adolescents include AQoL-6D, CHU9D, EQ-5D-Y, HUI2 and HUI3.

This report aims to address this evidence gap. We review the psychometric performance of the main child and adolescent preference-based measures that could be used in submissions to NICE. This work is intended to help inform NICE's future considerations about recommendations for estimating child health utilities.

The study objectives are:

1. Identify published literature that reports on the psychometric properties of one or more measures of AQoL-6D, CHU9D, EQ-5D-Y, HUI2 and HUI3;
2. Review and critically examine the published evidence around the psychometric properties of one or more measures of AQoL-6D, CHU9D, EQ-5D-Y, HUI2 and HUI3;
3. Identify gaps in the available evidence with recommendations for further research.

Methods

A systematic search was conducted in Medline, PsycINFO and the Web of Science (Science Citation Index Expanded) from the date of database inception until March 2019 to identify studies reporting the psychometric performance of AQoL-6D, CHU9D, EQ-5D-Y, HUI2 and HUI3 in children and adolescents.

Summary data for each paper was extracted by one of two reviewers (EP or AK) and checked by one of two reviewers (DR, AK). Two reviewers independently double

extracted the psychometric analyses for 3 papers (DR, AK) and after comparing extractions, undertook single extraction of the psychometric data of the remaining papers (DR, AK). Data were extracted around: the preference-based measure(s) used; whether it was the English version of the measure; preference weights applied (where applicable); whether the paper assessed the index (i.e. the utility scores generated by the measure), dimensions or both index and dimensions; other health-related quality of life measures or clinical measures used; age of participants (mean age and age range); proportion of females; whether the sample consisted of members of the general population, patients or both; clinical area (where applicable); whether the measure was self-reported and/or proxy-reported by parents/caregivers or both; and sample size.

Psychometric performance of the measures, including both the performance of the utility index and dimensions where this information was available, was assessed using an approach based on a previous review examining the psychometric performance of the adult generic preference-based measures assessing: known-group validity (ability to differentiate between groups of different severity or between people with and without the condition); convergent validity (strength of association between the measure of interest and other measures of health-related quality of life); responsiveness (ability to capture change over time when change is expected); reliability (ability to reproduce the same value on two administrations when there is no change in health); acceptability and feasibility (practicality of a measure for administration). Data were extracted separately for dimensions and the utility index where this was reported. Typically preference-based measures are scored using their value set to generate a utility index score. Whilst preference-based measures can be scored using summative scoring of dimensions and levels this is not typically recommended. Psychometric performance is reported both for the index score and the dimensions since examining the dimension performance is indicative of the performance of the index, and is independent of any country value set that is used to generate the index score.

Results

A total of 1,218 unique records were retrieved, with 8 additional records identified from reference lists. Of these, 102 records were examined in detail. Following the exclusion of 26 papers, 76 papers including 72 full-text articles and 4 conference abstracts were considered suitable for providing evidence for the psychometric assessment of EQ-5D-Y, CHU9D, HUI2, HUI3 and/or AQoL-6D.

Out of the 76 studies, 52 studies assess only one of the child and adolescent-specific preference-based measures analysed here. Nineteen studies assess both HUI2 and HUI3, two studies assess CHU9D and EQ-5D-Y, one assesses EQ-5D-Y and HUI2, one assesses CHU9D and AQoL-6D, and one assesses CHU9D and HUI2. Forty-two studies assess HUI3, 26 studies assess HUI2, 20 studies assess EQ-5D-Y, 12 studies assess CHU9D, and one study assesses AQoL-6D. In addition, one study compares the EQ-5D-Y 3 level and 5 level versions. The number of studies using the English language version of the measures are as follows: HUI3 (n = 34); HUI2 (n = 22); CHU9D (n = 11); EQ-5D-Y (n = 6); and AQoL (n =1).

There is variation in the value sets used across the studies. As there is no official value set available for the EQ-5D-Y most studies assess its performance by focussing upon the dimensions in the classification system. Nine studies apply UK value sets (one study also applies the UK EQ-5D value set to EQ-5D-Y). The majority of studies assess a clinical population (n=49), though some studies assess the measure using a general population sample (n=15) and other studies compare the general population and clinical population samples (n=12). A wide range of conditions are covered in the studies.

In total 30 studies administer the measures to children/adolescents using only self-report, and fourteen studies administer the measures using only proxy-report. Twenty-seven studies use both self-report and proxy-report for the same children, though for eleven of these studies restrictions are given around when self-complete was administered, for example a minimum age or only where the child was able to self-complete, and one of the studies administered the measures separately and then as a dyad. Three studies use either self or proxy report depending on the age of the child, and two studies do not report who completes the measure.

The age range of children and adolescents included in each study varies. Eleven studies include children aged below five which is below the recommended age for the measures included in these studies. Mean age varies from 6.4 to 16 years. Sample size varies considerably across the studies, from 7 to 9,949 subjects, with 28 studies having sample sizes below 100.

Across all of the studies, 48 studies assess known-group validity, 33 studies assess convergent validity, 14 studies assess responsiveness, 24 studies assess reliability, and 19 studies assess acceptability and feasibility.

For AQoL-6D the single study identified in the review only found evidence of known-group validity and no other psychometric properties were assessed. For CHU9D the review found evidence of known-group validity and convergent validity, mixed evidence of responsiveness and acceptability and feasibility, but the only study assessing test-retest reliability did not find evidence of reliability. For EQ-5D-Y the review found evidence for its dimensions of known group validity, convergent validity, responsiveness, test-retest reliability, acceptability and feasibility, but the only study assessing inter-rater reliability did not find evidence of reliability. There is no evidence available around the psychometric performance of potential UK utility values since there is no UK value set, nor any official value set for any country, for the EQ-5D-Y. For HUI2 the review found evidence of test-retest reliability and mixed evidence of known-group validity, convergent validity, responsiveness, inter-rater reliability, acceptability and feasibility, as good performance was not found unanimously across these aspects of psychometric performance. For HUI3 the review found mixed evidence of known-group validity, convergent validity, responsiveness, inter-rater reliability, test-retest reliability and acceptability and feasibility, with a proportion of studies not demonstrating evidence of known group validity, responsiveness or reliability.

Discussion

This is a review of available published evidence on the psychometric performance of a selection of child and adolescent-specific preference-based measures. Due to the limited number and heterogeneity of published studies, the evidence is based on a relatively small number of studies across a range of countries, a range of different

populations and conditions, using different study designs, different languages, different value sets and many different statistical techniques. The wide variation in studies makes it difficult to synthesise the evidence to generate a consistent picture of the overall performance of each measure. From the current evidence, EQ-5D-Y has the largest amount of evidence of good psychometric performance in proportion to the number of studies that have examined its psychometric performance. The majority of the evidence related to EQ-5D-Y is based on dimensions. The CHU9D is assessed in fewer studies, but the majority of studies find evidence of good psychometric performance. The evidence for HUI2 and HUI3 are more mixed, and for AQoL-6D the evidence is based on only one study. HUI3 has the largest proportion of studies that do not report good psychometric performance. However, for HUI2 and HUI3 the studies are more limited in their sample sizes and statistical power and this is likely to have impacted on their performance.

Overall the evidence is limited in the number of studies conducted in each condition, the number of studies that include patients (rather than general population), at times in the sample size of the study (in particular for HUI2 where 15 of 26 studies assessing performance had sample sizes below 100 and for HUI3 where 18 of 42 studies had sample sizes below 100), and the lack of studies administering more than one preference-based measure to provide comparative assessments of measures. The review is also limited in that the comparisons across measures do not take into account the differences in studies, since good psychometric performance may not have been observed due to sample size issues or design issues of the study. Relatively few studies use UK value sets to generate utility values. Comparisons of EQ-5D and EQ-5D-Y were beyond the remit of this review, though there are published papers available where both measures are administered to the same people at the same time (though EQ-5D is not designed for use in children).

For EQ-5D-Y there is no official value set, and the good psychometric performance that is observed is based mainly on the performance on the dimensions. Whilst it could be anticipated that a UK utility index would have the same psychometric performance, this can only be confirmed through data analyses. The value set may not have sufficiently large differences in utility decrements for different severity levels of each dimension.

There is a concern raised across all measures around their reliability. Only HUI2 performs strongly for test-retest reliability. None of the measures perform strongly for inter-rater reliability between child self-report and parent proxy-report (though AQoL-6D and CHU9D are not assessed). The findings suggest that there is reason for concern around the comparability of self-report and proxy responses to measure HRQOL of children and adolescents.

Suggested points for consideration by NICE:

The review has highlighted that there is limited published evidence around the psychometric performance of EQ-5D-Y, CHU9D, HUI2, HUI3 and AQoL-6D. The evidence is further limited in particular for NICE in that:

1) the AQoL-6D and EQ-5D-Y studies do not involve use of a UK value set, since there are no UK value sets currently available;

2) Only eight CHU9D studies use the UK value set;

3) Only two HUI2 studies use the UK value set.

Different value sets can have different psychometric properties, and drawing conclusions about the performance of an instrument based on the classification system alone may be misleading.

The following points are suggested for consideration:

- Given the paucity of evidence comparing measures, and the limitations relating much of the evidence that does exist, NICE must consider whether it is appropriate to recommend a specific instrument at this time.
- This review does not cover all available child and adolescent-specific generic preference-based measures, as the following also are potential candidates for use: AHUM; QWB; 16D; 17D. However, the review included the currently available measures the authors consider as most appropriate for use to inform UK policy using criteria around: intended and worded appropriately for use in children and adolescents; applicability across conditions using a generic classification system; development (or validation) with an English-speaking population; potential availability and feasibility of inclusion in datasets used to inform UK policy.

- Overall given the evidence available examining the psychometric performance of EQ-5D-Y, CHU9D, HUI2, HUI3 and AQoL-6D, the EQ-5D-Y has the largest amount of evidence of good psychometric performance in proportion to the number of studies that have examined its psychometric performance, followed by CHU9D. Any choice of measure for recommendation for use to inform policy would require additional considerations including but not limited to: content validity of the dimensions and severity levels in the measure; the appropriateness of the methods used to generate the value set; projected usage in trials and other relevant studies used to inform health technology assessment; relationship to adult EQ-5D since models often require utility values into adulthood.
- Though a large number of conditions are assessed in studies included in the review, not all conditions are assessed and many are only assessed in one study. New evidence may be needed to demonstrate the performance of a measure when it is applied in a patient population where it has not previously been validated.

Recommendations for future research:

The following are potential research questions that would be informative around the psychometric performance of the main generic child and adolescent-specific preference-based measures:

- What is the comparative psychometric performance of the main generic child and adolescent-specific preference-based measures, when administered to the same patients? Answering this research question could involve:
    - Primary data collection of the main child and adolescent-specific preference-based measures of interest administered to patients, preferably with a range of conditions across different ICD classifications. This would enable psychometric analyses to be undertaken across different measures using the same sample and applying the same statistical methods. In particular data collection could focus upon reliability where the evidence is mixed for EQ-5D-Y and limited for CHU9D. In addition, data collection could be linked to an intervention, and/or clinical measures, to determine responsiveness.

    - Accessing existing datasets of one or more of the main child and adolescent-specific preference-based measures of interest administered to patients to conduct independent analyses on these datasets, particularly where some of these datasets may not have had psychometric analyses published.
- Do the main generic child and adolescent-specific preference-based measures have content validity of dimensions and severity levels across the age range of respondents that they are recommended for?
- What is the impact of using self-report EQ-5D-Y versus proxy-report EQ-5D? Since many economic evaluations in children and adolescents use adult EQ-5D values in their economic model, this would be informative around the impact of using child and adolescent EQ-5D-Y over adult EQ-5D. This could include a review of studies comparing both the results and psychometric performance of EQ-5D and EQ-5D-Y. This could be extended to other adult preference-based measures and/or other child and adolescent preference-based measures (for example CHU9D).
- When, and at what ages, should self-report and proxy-report administrations of a measure be used to generate utility values to inform the economic model?
- Do any new UK value sets have good psychometric performance (note that CHU9D and EQ-5D-Y are expected to have new value sets in the next few years)? This could be assessed using either new or existing datasets.
- Does new evidence around the psychometric performance of the main child and adolescent-specific preference-based measures confirm the findings of this review? This could involve regular annual updates to the excel spreadsheet associated with the review that summarises all studies assessing the psychometric performance of selected child and adolescent preference-based measures (for example EQ-5D-Y and CHU9D).
- Do the findings of the review differ if a quality assessment is undertaken of the studies included in the review that assess psychometric performance of the main child and adolescent-specific preference-based measures?

CONTENTS

TABLE OF TABLES

TABLE OF FIGURES

ABBREVIATIONS AND DEFINITIONS

ADOS	Autism Diagnostic Observation Schedule
ADQOL	Atopic dermatitis-specific preference-based measure
AHUM	Adolescent Health Utility Measure
AQoL-6D	Assessment of Quality of Life- 6 Dimensions
CBCL	Child Behaviour Checklist
CDRSR	Child Depression Rating Scale-Revised
CHU9D	Child Health Utility 9 Dimensions
CHQ	Child Health Questionnaire
CFQ	Cystic Fibrosis Questionnaire
CUA	Cost-utility analysis
DCE	Discrete Choice Experiment
eGFR,	estimated Glomerular Filtration Rate
EQ-5D	EuroQol- 5 Dimensions
EQ-5D-Y	EuroQol- 5 Dimensions Youth version
FAE	Foetal alcohol effects
FAS	Foetal alcohol syndrome
FASD	Foetal alcohol syndrome disorder
GMFCS	Gross Function Motor Classification System
HOQ	Hydrocephalus Outcome Questionnaire
HS-FOCUS	Hunter Syndrome-Functional Outcomes for Clinical Understanding Scale
HRQoL	Health-related quality of life
HTA	Health Technology Assessment
HUI	Health Utilities Index
HUI2	Health Utilities Index Mark II
HUI3	Health Utilities Index Mark III
NDD	Neurodevelopmental difficulties
PAQLQ	Paediatric Asthma Quality of Life Questionnaire
PedsQL	Paediatric Quality of Life Inventory
POEM	Patient Oriented Eczema Measure
POQOLS	Paediatric Oncology Quality of Life Scale

QALY	Quality-adjusted life year
QWB	Quality of Well-being Scale
ScHARR	School of Health and Related Research
SDQ	Strengths and Difficulties Questionnaire
SRH	Self-rated health questionnaire
SSEN	Statement of Special Educational Needs
TRF	Teacher's report form, teachers' version of the CBCL
VAS	Visual Analogue Scale
VLBW	Very low birth weight
VP	Very pre-term
WAItE	Weight-specific Adolescent Instrument for Economic evaluation
Wee-FIM	Researcher-reported measure capturing functional independence

1. INTRODUCTION

1.1. BACKGROUND

Resource allocation decisions are increasingly important in the existence of constrained resources and large demands on a healthcare system. Health technology assessment (HTA) can be used as a tool for informing resource allocation decisions by assessing the cost-effectiveness of interventions and enabling comparisons of relative cost-effectiveness across a range of interventions for different conditions and populations. The Quality Adjusted Life Year (QALY) is commonly used to capture the benefit of interventions for use in HTA. The QALY is calculated by quality weighting survival using a quality adjustment which is often generated using an off-the-shelf generic preference-based measure. A preference-based measure consists of a classification system and a value set that is used to score responses to the classification system. The classification system contains dimensions with severity levels. Responses to the classification system are used to assign people to a health state. A value set is then used to score the relative value of the health state to generate a utility value, also known as an index score, on the 1-0 full health-dead scale, with values below zero indicating that the health state is worse than being dead. There are many different preference-based measures available, and these can be condition-specific or generic, and population-specific (for example for adults or children) or suitable across many populations.

Measures for estimating adult health utilities are often assessed by reference to the psychometric performance of measures, for example assessing their known-group validity, content validity, face validity and responsiveness in particular populations. EQ-5D, for example, is a generic preference-based measure for adults that has been found to have good psychometric performance in many disease areas[1] including urinary incontinence[2] and conditions in skin and subcutaneous tissues[3], but has more questionable psychometric performance in some other conditions such as schizophrenia[4] and hearing impairments[5] which challenge the appropriateness of the use of EQ-5D in those conditions. The psychometric performance of the main generic preference-based measures including EQ-5D and SF-6D have been assessed widely in the published literature, and there is a published review of reviews around their performance[1]. This means that both researchers and decision makers have

knowledge around the appropriateness of the utility values generated by these measures across a range of conditions and also around whether the measure would be expected to identify a statistically significant change in that population. This can provide valuable information around the confidence in the utility estimates and interpretation of changes in utility values. However, to our knowledge there is no review of the published literature examining the psychometric performance of the child and adolescent preference-based measures.

One existing review examined the development and application of generic preference-based measures available for use in paediatric populations[6], finding nine measures and concluding that further empirical analyses are required to examine the relative performance of these measures. Another recent review found that six of these preference-based measures had been commonly used internationally in paediatric populations: EQ-5D-Y, CHU9D, HUI2, 15D/16D/17D, QWB and AQoL-6D[7]. Another review reviewed the valuation methods used to generate the values sets of the preference-based measures[8]. Of the more commonly used measures, the CHU9D and HUI2 have UK value sets, and the EQ-5D-Y can be scored using EQ-5D-3L (adult measure) utility values (though this is not recommended by the EuroQol group). The Kwon et al. review[7] provides a fully comprehensive source of published utility values from the existing literature across a range of conditions. However, the review did not assess the psychometric performance of the measures used to generate the utility values, nor can this information be inferred from the extraction spreadsheet or the appendices provided with the paper.

The National Institute for Health and Care Excellence (NICE) have clear recommendations around the generation, source and usage of utility values for adults. NICE have a clear recommendation in the NICE Guide to the methods of technology appraisal 2013[9] around the use of EQ-5D to generate utility values for adults. However, there is no specific guidance around the measure that should be used to generate adolescent and child utility values:

“When necessary, consideration should be given to alternative standardised and validated preference-based measures of health-related quality of life that have been designed specifically for use in children. The standard version of the EQ-5D has not

been designed for use in children. An alternative version for children aged 7 to 12 years is available, but a validated UK valuation set is not yet available"[9] (page 42).

Therefore, the 2013 NICE methods guide acknowledges that the adult EQ-5D may not always be suitable for use in children and adolescents, but does not recommend alternative measure(s) that should be used instead. It does not explicitly state that adult EQ-5D should be used to generate utility values for children and adolescents. However, a previous NICE DSU project reviewed all 31 NICE appraisals that were published as part of the Technology Appraisal (TA) and NICE Highly Specialised Technology (HST) evaluation programmes since inception, where the licensed indication for the technology included people aged under 18[10]. The review found that most appraisals included utility values generated using EQ-5D scored using the UK adult tariff (n=27), though it was unclear from the TAs and HST evaluations whether this was adults completing the measure for their own health or whether EQ-5D had been used to measure the health directly of children/adolescents. The review found limited use of child and adolescent population-specific measures to generate health state utility values for children and adolescents in technology appraisals submitted to NICE, where only seven appraisals used a child and adolescent population-specific measure to generate utility values, and all of these also used an adult measure. Four appraisals used HUI2, one appraisal used a child and adolescent-specific preference-based measure for atopic dermatitis, and three appraisals used EQ-5D-Y predicted by statistical mapping from another measure and subsequently valued using the UK EQ-5D adult tariff. This raises the questions of 1) why child and adolescent-specific preference-based measures were not used more frequently to generate utility values in the TAs and HSTs submitted to NICE, 2) which preference-based measure(s) could and should be used to generate utility values for children and adolescents; and 3) how child and adolescent preference-based measures perform both in comparison to each other and in comparison to adult measures including the EQ-5D.

NICE needs to assess the suitability of different approaches for estimating health state utilities across the broad range of conditions that feature in its guidance producing programmes in order to recommend the preferred measure in most situations. When considering approaches for adults, this assessment has been informed by reviews of

psychometric performance in studies that span a wide range of health conditions. However, similar reviews of the psychometric performance of child and adolescent preference-based measures have not been performed. This report aims to address this evidence gap.

1.1. AIMS AND OBJECTIVES

The purpose of this project is to review the psychometric performance of the main child and adolescent preference-based measures that could be used in submissions to NICE. This work is intended to help inform NICE's future considerations about recommendations for estimating child health utilities.

The project will involve a review of the psychometric properties of commonly used preference-based measures in paediatric populations: AQoL-6D, CHU9D, EQ-5D-Y, HUI2 and HUI3. The authors selected these measures, after consultation with NICE staff, because they are considered to be the measures most appropriate to inform UK policy using criteria around: intended and worded appropriately for use in children and adolescents; applicability across conditions using a generic classification system of dimensions and levels; development (or validation) with an English-speaking population; potential availability in datasets used to inform UK policy.

The objectives are:

1. Identify published literature that reports on the psychometric properties of one or more measures of AQoL-6D, CHU9D, EQ-5D-Y, HUI2 and HUI3;
2. Review and critically examine the published evidence around the psychometric properties of AQoL-6D, CHU9D, EQ-5D-Y, HUI2 and HUI3;
3. Identify gaps in the available evidence with recommendations for further research.

2. SUMMARY OF CHILD AND ADOLESCENT PREFERENCE-BASED MEASURES

This section provides a summary of generic preference-based measures for children and adolescents: AQoL-6D; CHU9D; EQ-5D-Y; HUI2; and HUI3. The summary is not exhaustive of all measures, and does not include the Adolescent Health Utility Measure (AHUM), the Quality of Well-Being scale (QWB), 16D or 17D (for a recent overview see [6]).

2.1. AQoL-6D

The AQoL-6D adolescent measure has six dimensions: independent living; relationships; mental health; coping; pain; senses[11]. Each dimension has between four and six severity levels. The adolescent measure was generated through adapting the adult AQoL-6D measure using focus groups with adolescents, though the adaptation seemed to mostly cover cultural and linguistic translation to be appropriate for valuation by adolescents in Australia, Fiji, New Zealand and Tonga[12]. Value sets have been generated for Australia, Fiji, New Zealand and Tonga generated using time trade-off with adolescents from the general population [12].

2.2. CHU9D

The CHU9D has nine dimensions each with five severity levels: worry; sadness; pain; tiredness; annoyance; school; sleep; daily routine; activities. The dimensions and severity levels were developed using qualitative research with children aged 7 to 11 years, and hence were designed for this age group, but can be completed via parent/guardian proxy for children aged 4 to 7 years and have been used in adolescents aged 12 to 18 years. Value sets exist for the UK[13], Australia[14-17], the Netherlands[18] and China[19]. The UK value set was generated using standard gamble with members of the adult general population who were asked to imagine themselves in the health state [13]. For the Netherlands value set a discrete choice experiment with duration was used with members of the adult general population[18], and for Australia[14-17] and China[19] a general population sample of adolescents provided values using best-worst scaling and these were anchored onto the 1-0 full

health-dead using time trade-off values elicited from young adults members of the general population.

2.3. EQ-5D-Y

The EQ-5D-Y is the youth version of the EQ-5D. The EQ-5D-Y was generated through adapting the adult EQ-5D to ensure relevance and clarity for children and adolescents[20-22]. The EQ-5D-Y has five dimensions each with three levels of severity: mobility; looking after myself; doing usual activities; having pain or discomfort; feeling worried, sad or unhappy. There is no officially accepted value set for the EQ-5D-Y, though there is a published value set for the US which was generated using a discrete choice experiment with members of the adult general population. The non-standard discrete choice experiment involves problems with one dimension for *x* years followed by full health for *y* years, and generates modelled latent scale values that are argued to be directly anchored on a 1-0 scale[23]. Recent research has found that current EQ-5D value sets cannot be appropriately used to value EQ-5D-Y health states[24, 25]. The EuroQol Group is currently developing an international valuation protocol for the development of country-specific EQ-5D-Y value sets, and a 5-level youth version of the EQ-5D, the EQ-5D-5L-Y.

2.4. HUI2

The HUI2 has seven dimensions: sensation; mobility; emotion; cognition; self-care; pain; and fertility[26]. Each dimension has between three and five levels. The measure was originally developed for use in childhood cancer, but is widely used as a generic measure, although the fertility dimension is rarely used. The HUI2 has a UK value set[27] and a Canadian value set[26]. The HUI2 value sets were generated using standard gamble and visual analogue scale with adults, who were asked to imagine a child aged 10 years was in the health state. The UK value set was generated using members of the adult general population[27], whereas the Canadian sample involved parents of children[26]. The HUI2 can be used to measure health of children and adults aged 5 and over. HUI2 and HUI3 are typically administered using a single set of 15 self-administered questions, which are then used to generate both HUI2 and HUI3 utilities. Interviewer administration of the set of items used to generate both HUI2 and HUI3 utilities involves between 13 and 39 questions.

2.5. HUI3

The HUI3 has eight dimensions: vision; hearing; speech; ambulation; dexterity; emotion; cognition; and pain[28]. Each dimension has between five and six levels. The HUI3 has only a Canadian value set, generated using standard gamble and visual analogue scale with adults, who were asked to imagine themselves in the health state[28]. The HUI3 can be used to measure health of children and adults aged 5 and over.

3. METHODS

3.1. SEARCH STRATEGY

A systematic search was conducted in Medline (Ovid) PsycINFO (Ovid) and the Web of Science Core Collection Science Citation Index Expanded (Clarivate Analytics) from the date of database inception until March 2019 to identify studies reporting the psychometric performance of EQ-5D-Y, CHU9D, HUI2 and AQoL-6D in children and adolescents. Terms for the measure (e.g. 'EQ-5D-Y' 'AQoL', 'HUI', CHU9D') were combined with 'child' population terms derived from a recently published systematic review of child utilities (that does not assess psychometric performance of measures)[7]. The search strategy was translated across each database and limits for human studies and English language were applied. No study type or publication date limits were applied. Following NICE's request, additional searches for HUI3 using a similar approach as above was undertaken in September 2019.

Supplementary grey literature searches include the conference abstract websites in the last three years (The International Society for Pharmacoeconomics and Outcomes Research and International Society for Quality of Life Research), Web of Science Cited Reference Search, keyword searching using Google Scholar search engine and examination of reference lists of included studies.

3.2. SELECTION OF PAPERS

Eligible papers (full-text articles and abstracts without available free full versions online) were selected. A summary of inclusion and exclusion criteria and final selection of relevant studies are presented in Table 1 and Figure 1. Citations were screened by one of three reviewers (DK, EP or DR). A ten percent randomly-selected sample of titles and abstracts was double-checked by a second researcher (DR) to minimise error and bias in interpreting the eligibility criteria. All potentially relevant evidence (included abstracts and full text articles) were independently checked by both researchers to ensure that eligible papers were included in the final set.

Table 1: Study eligibility criteria

	Inclusion criteria	Exclusion criteria	Additional notes relating to study eligibility
Population	Paediatric, i.e. participants age< 18 years Includes paediatrics and adults, but analyses reported separately for paediatrics and adults	Only adults, i.e. all participants age≥ 18 years	Include if data can be extracted for participants age< 18 years
Outcome	Primary outcome: Assess the validity (face, known-group, construct or convergent) OR responsiveness OR reliability OR acceptability OR feasibility of EQ-5D-Y, CHU9D, HUI2, HUI3 and/or AQoL-6D obtained from paediatric populations or relevant parents/caregivers acting as proxies for children	Incomplete, unclear or no data assessing the validity (face, known-group, construct or convergent), OR responsiveness OR reliability OR acceptability OR feasibility of EQ-5D-Y, CHU9D, HUI2 and/or AQoL-6D. Only nurse or clinician report data	Relevant data may include other preference-based measures and clinical outcomes for assessing psychometric properties
Study design	Randomised controlled trials Cohort or observational (cross-sectional or longitudinal) retrospective or prospective	Case studies	Include human studies only
Language	English	Non-English	Studies using non-English versions of the measure are included

3.3 DATA EXTRACTION

Summary data for each paper was extracted by one of two reviewers (EP or AK) and checked by one of two reviewers for all papers (DR, AK). Two reviewers independently double extracted the psychometric analyses for 3 papers (DR, AK) and after comparing extractions, undertook single extraction of the psychometric data of the remaining papers (DR, AK). Data were extracted around: the preference-based measure(s) used; whether it was the English version of the measure; preference weights applied (where applicable); whether the paper assessed the index (i.e. the utility scores generated by the measure), dimensions or both index and dimensions;

other health-related quality of life measures or clinical measures used; age of participants (mean age and age range); proportion of females; whether the sample consisted of members of the general population, patients or both; clinical area (where applicable); whether the measure was self-reported and/or proxy -reported by parents/caregivers or both; and sample size.

Psychometric performance of the measures, including both the performance of the utility index and dimensions where this information was available, was assessed using an approach based on a previous review examining the psychometric performance of the adult generic preference-based measures[29], which assessed: known-group validity; convergent validity; responsiveness; reliability; acceptability and feasibility. Data were extracted separately for dimensions and the utility index where this was reported. Some aspects are more relevant to dimensions, for example inter-rater or inter-modal reliability, but most aspects are relevant for both dimensions and index scores. Typically preference-based measures are scored using their value set to generate a utility index score. Whilst preference-based measures can be scored using summative scoring of dimensions and levels this is not typically recommended. Psychometric performance is reported both for the index score and the dimensions since examining the dimension performance is indicative of the performance of the index, and is independent of any country value set that is used to generate the index score.

Where reported, data were extracted for each of the psychometric assessments around: brief summary of analysis undertaken; whether the results were in accordance with clinical expectation (where relevant); and whether the findings were statistically significant. The aspects of psychometric performance that were extracted and assessed are summarised below.

3.2.1. Known-group validity

Known group construct validity assesses the ability to differentiate between groups of different severity, or a less rigorous test of case–control construct validity which examines the ability to differentiate between people with and without the condition. Evidence of known-group validity is determined using the ability to determine a statistically significance difference at the 5% level across known groups is reported,

along with whether the direction of the difference is in accordance with clinical expectation i.e. shows difference in the expected direction e.g. general population with higher index scores than patients. Where studies assess dimensions, it is not typically expected that all dimensions will necessarily capture known-group differences, as not all conditions impact on all dimensions.

3.2.2. Convergent validity

Convergent validity assesses the strength of association between the measure of interest and other measures of health-related quality of life (generic or condition-specific) or disease severity using either correlation coefficients (a more conventional technique) or statistical significance in regression analyses. Evidence of convergent validity is determined by whether moderate (0.41-0.60) or good (0.61-0.8) (or higher and almost perfect) agreement has been observed. It is recognised that these are arbitrary cut-offs, but these are often reported in the papers included in the review (and are based on established criteria, see for example Landis and Koch (1977)). Convergent validity should not be expected between all dimensions of different measures, for example, pain dimensions in two measures would be expected to be correlated, whereas pain in one measure would not be expected to be correlated with mobility in the other measure. Therefore, the convergent validity that is reported focuses upon expected correlations where these are motivated in theory, rather than including poor correlations between dimensions that would not be expected to be correlated. Where studies have reported regression analyses between clinical measures this has not been extracted.

3.2.3. Responsiveness

Responsiveness assesses the ability to capture change over time, where change is expected, for example due to treatment effects. Evidence of responsiveness is determined by the ability to determine a statistically significance change at the 5% level over time. It is also reported whether the direction of the change is in accordance with clinical expectation e.g. higher index scores at the end of treatment than at baseline. Details of the analysis are provided, as these can vary widely across studies depending on the study design. Where dimensions are assessed, it is not necessarily

expected that all dimensions will be responsive since not all conditions or treatments impact on all dimensions.

3.2.4. Reliability

Reliability assesses the degree of change where no change in health is observed using other health-related quality of life or clinical measures. Evidence of reliability is determined by whether the measure is able to reproduce the same value on two separate administrations when there has been no change in health, where this can be over time (test-retest reliability), between methods of administration (inter-modal reliability) or between raters i.e. self-report and parent proxy-report (inter-rater reliability). Reliability can be difficult to summarise, since in some studies reliability may be observed for most but not all dimensions, and hence the level of agreement reported in the studies has been extracted (for example moderate agreement at 0.41 to 0.6, or good agreement at 0.61 to 0.8). However, if reliability is not observed for some dimensions this raises issues around reliability of the whole measure.

3.2.5. Acceptability and feasibility

Acceptability and feasibility assess the practicality of a measure for administration in a specific group of people, and covers aspects such as burden of completion and whether the person completing the measure can meaningfully respond to the questions being asked. Evidence of acceptability and feasibility is indicated where the study demonstrates, for example low missing data or high levels of understanding. A lack of evidence for acceptability and feasibility is concluded where the study demonstrates, for example, high levels of missing data or low levels of understanding. For child and adolescent measures this includes whether the child and adolescent or their proxy can meaningfully complete the measure, since there may be problems of understanding for younger people and problems of knowing the required information (for example how the child feels emotionally) for proxy report. Missing data can also be used to indicate acceptability and feasibility since high levels of missing data indicates that the person completing the measure has not completed some dimensions. Though this can occur for many reasons, it indicates that the measure will not produce useable data for all participants which can impact on the results obtained.

4. RESULTS

4.1. SEARCH RESULTS

An example of the search in MEDLINE is presented below in Table 2.

Table 2: MEDLINE search terms and number of retrieved records for EQ-5D-Y, CHU9D, HUI2 and AQoL-6D

#	Searches	Results
1	(((euroqol or euro qol) adj3 youth) or eq-5d-y or eq 5d y).mp.	55
2	(health utilities index or hui).mp.	1412
3	(aqol or assessment of quality of life).mp.	1654
4	(child health utility or chu9d or chu-9d or chu 9d).mp.	39
5	or/1-4	3117
6	(child* or adolesc* or kid or kids or youngster* or teen* or youth* or infant* or newborn* or perinat* or neonat* or parent proxy).mp.	3892062
7	(pediatri* or paediatri*).mp.	342033
8	6 or 7	3941571
9	5 and 8	802
10	limit 9 to english language	707

The search for HUI3 in Table 3 yielded a further 207 records, 85 were unique from the previous search in March 2019.

Table 3: MEDLINE search terms and number of retrieved records for HUI3 in September 2019

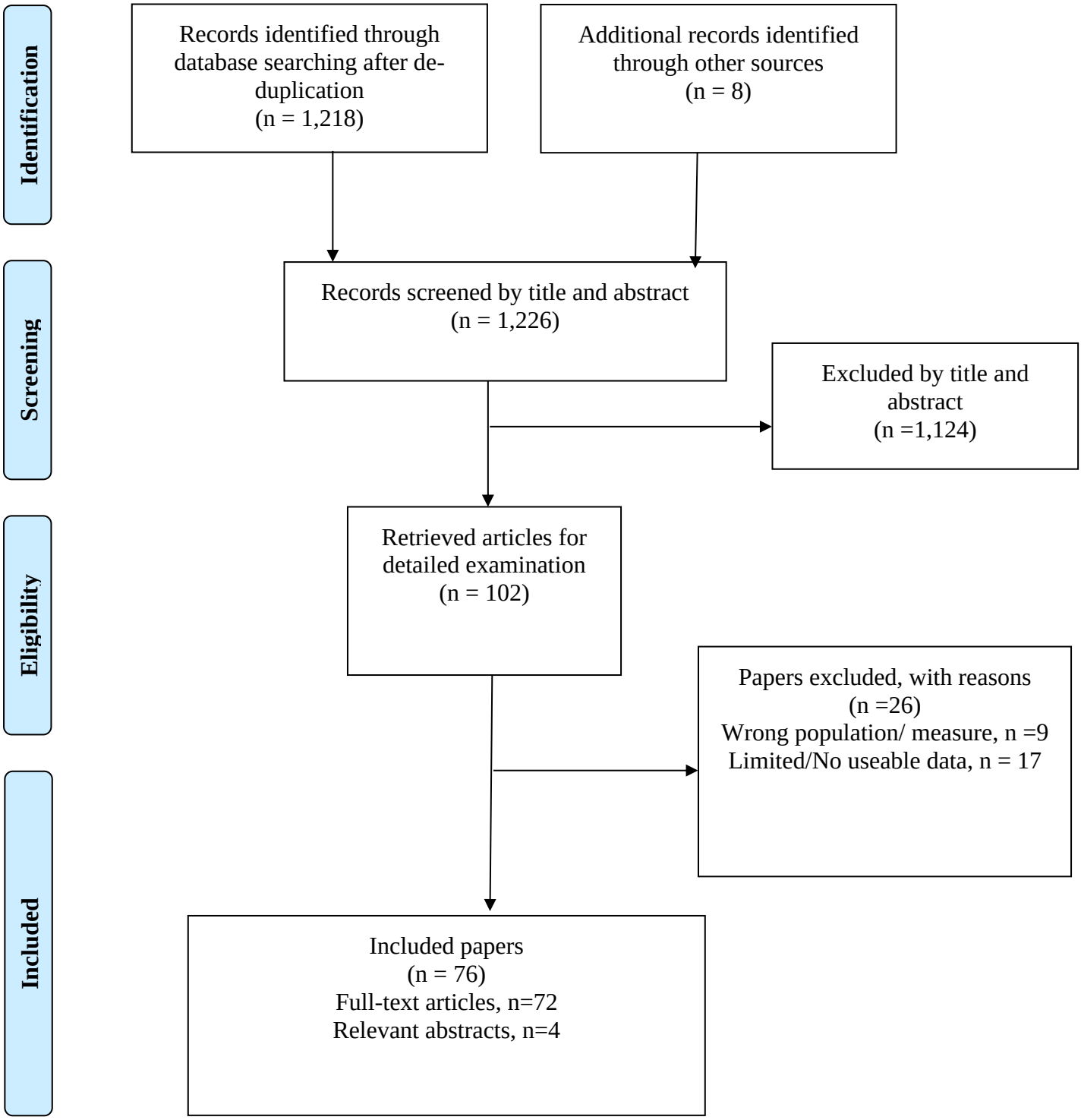
#	Searches	Results
1	(health utilities index mark 3 or health utilities index 3 or hui mark3 or hui mark-3 or hui mark 3 or hui3 or hui-3 or hui 3).mp.	484
2	(child* or adolesc* or kid or kids or youngster* or teen* or youth* or infant* or newborn* or perinat* or neonat* or parent proxy).mp.	4101208
3	(pediatri* or paediatri*).mp.	377859
4	2 or 3	4155614
5	1 and 4	175
6	limit 5 to english language	172

4.2. INCLUDED STUDIES

A total of 1,218 unique records were retrieved, with 8 additional records identified from reference lists. Of these, 102 records were examined in detail. Following the exclusion of 26 papers (see Appendix), 76 papers including 72 full-text articles and 4 conference

abstracts[30-33] were considered suitable for providing evidence for the psychometric assessment of EQ-5D-Y, CHU9D, HUI2, HUI3 and/or AQoL-6D. A summary of included papers is presented in Table 4.

Figure 1: PRISMA diagram outlying selection of studies



4.3. SUMMARY OF STUDIES INCLUDED

Characteristics of included studies are summarised in Table 4 and the psychometric properties and measures assessed per paper are summarised in Table 5. Out of the 76 studies, 52 studies assess only one of the child and adolescent-specific preference-based measures analysed here. Nineteen studies assess both HUI2 and HUI3 [30, 34-50], two studies assess CHU9D and EQ-5D-Y [51, 52], one assesses EQ-5D-Y and HUI2[53], one assesses CHU9D and AQoL-6D[54], and one assesses CHU9D and HUI2[15, 55]. Forty-two studies assess HUI3, 26 studies assess HUI2, 20 studies assess EQ-5D-Y, 12 studies assess CHU9D, and one study assesses AQoL-6D. One study[56] compares the EQ-5D-Y 3 level and 5 level versions.

In total nine studies apply UK value sets (one study also applies the UK EQ-5D value set to EQ-5D-Y). The only study identified for the AQoL-6D uses the Australian adolescent and adult value sets. For the CHU9D seven studies use only the UK value set, one study uses the Australian adolescent value set, one uses both the Australian adolescent and adult value sets, and two studies use both the UK and the Australian adolescent value set. For the EQ-5D-Y there is no accepted value set, and hence 15 studies do not generate utility scores, whereas one study uses UK EQ-5D, one uses Australian EQ-5D, one uses French EQ-5D, one uses Spanish EQ-5D and one uses an unofficial US EQ-5D-Y value set. For the HUI2, twenty studies use the Canadian value set, two use the UK value set, three do not use a value set and one does not report the value set used. For the HUI3 there is only a Canadian value set, though this is not used in four studies due to a focus on dimensions in those studies. Since EQ-5D-Y and CHU9D are recently developed measures, the majority of studies were published from 2010 onwards with only six studies conducted prior to 2000 and another fourteen studies in the review conducted prior to 2010.

The data assessed in the studies are from a variety of countries, with Canada (n=16), UK (n=12) USA (n=9), and Australia (n=8) having the largest number of included studies, followed by Netherlands (n=4), Sweden (n=4), Spain (n=3), China (n=2), Germany (n=2), South Africa (n=2), and many countries with one study (France, Hong Kong, Italy, New Zealand, South Korea, Taiwan, Thailand, and Turkey), two

multinational studies (each included Germany, Italy, South Africa, Spain, Sweden), one study in Australia and New Zealand, one study in UK and Ireland, one study in UK and USA, and one study where country is not reported.

The number of studies using the English language version of the measures are as follows: HUI3 (n = 34); HUI2 (n = 22); CHU9D (n = 11); EQ-5D-Y (n = 6); and AQoL (n =1). Other languages for EQ-5D-Y included: Swedish (n = 5), Spanish (n =4), German (n =2), Chinese (Simplified n = 1, Taiwanese n = 1), Afrikaan (n = 1), Korean (n =1) and Italian (n = 2). One study uses the Chinese version of CHU9D. One study uses both the French and the English version of the HUI2 and one study uses only the French version of HUI2. Other languages for HUI2 included: Chinese, French, Thai, Turkish, (all 1 study each) and Spanish (study also included English version); and for HUI3 included: Dutch (n=3) and Chinese. French, German, Thai, Turkish (all 1 study each) and Spanish (study also included English version). The version of the measure is unknown for 2 studies using EQ-5D-Y and one study using HUI3.

The majority of studies assess a clinical population (n=49), though some studies assess the measure using only a general population sample (n=15) and other studies compare the general population and clinical population samples (n=12). A wide range of conditions are covered in the studies: acute lymphoblastic leukaemia; adolescent or juvenile idiopathic scoliosis; allergic conditions; asthma (n=3); autism spectrum disorders; cancer (n=5); central nervous system tumours survivors; cerebral palsy (n=3); childhood brain tumour survivors; childhood cancer survivors; chronic illness (n=3); chronic kidney disease (n=2); cystic fibrosis; deafness (n=2) and permanent hearing loss (n=1); dental caries, carious surfaces, restored surfaces or missing teeth; depression (n=2); Down syndrome; eczema; foetal alcohol spectrum disorder; functional motor, orthopaedic and medical disabilities; Hodgkin's disease (n=2); Hunter syndrome; idiopathic clubfoot; medulloblastoma and ependymona; neurological disability and preterm births; obstructive hydrocephalus; osteonecrosis secondary to treatment of developmental dysplasia of the hip; overweight and obese (n=2); underweight, healthy weight, overweight or obese; BMI≥85th percentile with Type 2 diabetes, pre-diabetes or insulin resistance; stutter (n=2); Type 1 diabetes mellitus (n=2); vision impairment or blindness; as well as children and adolescents receiving mental health services; adolescents attending well child appointments or

obesity clinic; children participating in an obesity prevention programme; children who when born had extremely low birth weight (n=3) and very preterm born children (n=1 includes both); and children and adolescents who were acutely ill, or with chronic health condition/disability, or in intensive care; and one study included a range of conditions (acute otitis media, bacteraemia, chronic lung disease, hearing loss, epilepsy, meningitis, mild mental retardation, pneumonia).

In total 30 studies administer the measures to the children/adolescents using only self-report, fourteen studies administer the measures using only proxy-report, 27 studies use both self-report and proxy-report for the same children, though for eleven of these studies restrictions are given around when self-complete was administered, for example a minimum age or only where the child was able to self-complete, and one of the studies administered the measures separately and then as a dyad. Three studies use either self or proxy report depending on the age of the child, and two studies do not report who completes the measure.

The age range of children and adolescents included in each study varies. Eleven studies include children aged below five years which is below the recommended age for the measures used in these studies (note the minimum recommended age for CHU9D and EQ-5D-Y is 4 and for HUI2 and HUI3 is 5)[8]. Mean age varies from 6.4 years[51] to 16[57, 58]. The percentage of female subjects in the samples ranges from 14.7%[59] to 80.6%[56].

Sample size varies considerably across the studies, from seven subjects[60] to 9,949[61]. Thirteen studies have sample sizes below 50, fifteen studies have sample sizes between 50 and 99, fifteen studies have sample sizes between 100 and 199, fifteen studies have sample sizes between 200 and 499, eleven studies have sample sizes between 500 and 999, and seven studies have sample sizes greater than or equal to 1000.

The studies assess a range of psychometric properties of the measures, with no study assessing all properties extracted in this review. Across all of the studies, 48 studies assess known-group validity, 33 studies assess convergent validity, fourteen studies

assess responsiveness, 24 studies assess reliability, and 19 studies assess acceptability and feasibility.

Table 4: Characteristics of included studies

Study reference	English version of measure	Country	Value set	General population	Condition, where relevant	Self-report	Proxy report	Age range of children (years)	Mean age	% female	N
AQoL-6D											
Ratcliffe, 2012b[54]	Yes	Australia	Australia adolescent and adult	Yes	No	Yes	No	11 to 17	15 (1.7)	51	500
CHU9D											
Canaway, 2013[51]	Yes	UK	UK	Yes	No	Yes	No	6 to 7	6.4	43	160
Chen, 2015[52]	Yes	Australia	Australia adult	Yes	No	Yes	No	11 to 17	14 (2)	51	2020
Foster Page, 2015[62]	Yes	New Zealand	UK	No	Dental caries, carious surfaces, restored surfaces or missing teeth	Yes	No	6 to 9	8.3 (0.7)	56	87
Frew, 2015[63]	Yes	UK	UK	Yes	Underweight, healthy weight, overweight or obese	Yes	No	5 to 6	6.3 (0.31)	48.3	1344
Furber, 2015[64]	Yes	Australia	UK and Australian adolescent	No	Receiving mental health services	No	Yes	5 to 17	11.7 (5.8)	47.5	200
Oluboyede, 2019[65]	Yes	UK	UK	Yes	No	Yes	No	11 to 18	15.4	50.6	975

Study reference	English version of measure	Country	Value set	General population	Condition, where relevant	Self-report	Proxy report	Age range of children (years)	Mean age	% female	N
Petersen, 2018[66]	Yes	Australia	Australia adolescent	Yes	No	Yes	No	15 to 17	15.8 (0.8)	53	775
Ratcliffe, 2012a[55]	Yes	Australia	UK	Yes	No	Yes	No	11 to 17	15 (1.9)	48	500
Ratcliffe, 2012b[54]	Yes	Australia	Australia adolescent and adult	Yes	No	Yes	No	11 to 17	15 (1.7)	51	500
Sach, 2017[32]	Yes	UK	UK	No	Eczema	Unknown	Unknown	5 and above	Not reported	Not reported	137
Stevens, 2012a[67]	Yes	Australia	UK	Yes	No	Yes	No	11 to 17	14.5 (2.0)	45.3	961
Xu, 2014[68]	No, Chinese	China	UK and Australian adolescent	Yes	No	Yes	No	9 to 19	14.1 (2.5)	45.5	815
EQ-5D-Y											
Åström, 2018[69]	No, Swedish	Sweden	No	Yes	No	Yes	No	13 to 18	15.9 (1.6)	49.4	6574
Bergfors, 2015[70]	No, Swedish	Sweden	No	No	Asthma	Yes	No	8 to 16	12.1 (2.4)	41.5	94
Burstrom, 2014[71]	No, Swedish	Sweden	No	Yes	Functional motor, orthopaedic and medical disabilities	Yes	No	Clinical population 7 to 17, general population 8 to 16	Clinical population 12.0 (3.1), general population 13.3 (2.7)	Clinical population 60.6, general population 48.9	478, Clinical population n=71, general populati

Study reference	English version of measure	Country	Value set	General population	Condition, where relevant	Self-report	Proxy report	Age range of children (years)	Mean age	% female	N
											on n=407
Canaway, 2013[51]	Yes	UK	UK EQ-5D	Yes	No	Yes	No	6 to 7	6.4	43	160
Chen, 2015[52]	Yes	Australia	Australia EQ-5D	Yes	No	Yes	No	11 to 17	14 (2)	5100%	2020
Eidt-Koch, 2009[72]	No, German	Germany	No	No	Cystic fibrosis	Yes	Yes	8 to 17	8 to 13 years (n=55) 10.8 (1.7), 14 to 17 years (n=41) 15.9 (1.8)	8 to 13 years (n = 55) 56.4, 14 to 17 years (n = 41) 41.5	96
Hernandez, 2015[31]	Unclear	Spain	France EQ-5D	No	Asthma	Yes	No	6 to 11	Not reported	Not reported	69
Hsu, 2018[73]	No, Taiwanese	Taiwan	No	No	Chronic kidney disease	Yes	No	7 to 18	11.96 (4.08)	35	68
Jelsma, 2010[74]	Yes	South Africa	No	Yes	No	Yes	No	Not reported	15.5 (1.3)	50	522
Kim, 2018[61]	No, Korean	South Korea	No	No	Allergic conditions	Yes	No	7 to 13	10.2 (1.8)	48.6	9949
Loof, 2019[75]	No, Swedish	Sweden	No	Yes	Idiopathic clubfoot	Yes	Yes	8 to 10	Idiopathic clubfoot 9.4 (0.6), General population 9.5 (0.6)	Idiopathic clubfoot 29, General population 30	215, Idiopathic clubfoot n=106, General population n=109

Study reference	English version of measure	Country	Value set	General population	Condition, where relevant	Self-report	Proxy report	Age range of children (years)	Mean age	% female	N
Mayoral, 2017[33]	Unknown	Unknown	Spain EQ-5D	No	Type I diabetes mellitus	Unknown	Unknown	Unknown	Unknown	Unknown	136
Oluboyede, 2013[53]	Yes	UK	No	Yes	No	Yes	No	11 to 18	12.7	53	49
Perez-Sousa, 2018[76]	No, Spanish	Spain	No	No	Overweight and obese	Yes	Yes	6 to 14	Intervention group 9.6 (2.1), control group 8.7 (1.6)	47	151
Ravens-Sieberer, 2010[21]	Yes in South Africa only, Spanish, German, Italian	Germany, Italy, South Africa, Spain, Sweden	No	Yes	No	Yes	No	8 and above	Germany 13.8 (1.9), Italy 11.8 (2.2), South Africa 15.5 (1.3), Spain 13.0 (2.7), Sweden 13.2 (2.7)	Germany 49.1, Italy 52.0, South Africa 49.6, Spain 49.2, Sweden 48.9	2809, Germany n=756, Italy n=415, South Africa n=258, Spain n=973, Sweden n=407
Robles, 2015[77]	No, Spanish	Spain	No	Yes	No	Yes	No	8 to 18	11.7 (2.8)	54	923
Scalone, 2011[78]	No, Italian	Italy	No	Yes	Acute lymphoblastic leukaemia	Yes	No	8 to 15	Not reported. Median age 9.4 years	28	440, Clinical population n=25,

Study reference	English version of measure	Country	Value set	General population	Condition, where relevant	Self-report	Proxy report	Age range of children (years)	Mean age	% female	N
											general population n=415
Scott, 2017[79]	Yes and Afrikaan	South Africa	US	Yes	Acutely ill; or chronic health condition/disability	Yes	No	8 to 12	10.5 (1.45)	Not reported	329
Wille, 2010[22]	Yes in South Africa only	Germany, Italy, South Africa Spain, Sweden,	No	Yes	No	Yes	No	8 to 18	Germany 13.9 (1.8), South Africa 15.5 (1.3), Spain 13.0 (2.7)	Not reported	1987 Germany n=756, South Africa n=258, Spain n=973
Wong, 2019[a][56]	No, Chinese	China	No	No	Adolescent or juvenile idiopathic scoliosis	Yes	No	8 to 17	14.0 (1.9)	80.6	129
HUI2											
Banks, 2008[34]	Yes	Canada	Canada	No	Cancer - undergoing chemotherapy	Yes, children aged 10 and over	Yes	2 to 18	9.5 (SD not reported)	35	29
Barr, 1997[35]	Yes	Canada	Canada	No	Cancer	Yes but due to low numbers were excluded	Yes - nurse-investigator, parents	Not reported	Not reported	67%	18

Study reference	English version of measure	Country	Value set	General population	Condition, where relevant	Self-report	Proxy report	Age range of children (years)	Mean age	% female	N
						from the analysis					
Belfort, 2011[36]	Yes	USA	Canada	No	Attending well-child appointments or obesity clinic	Yes	Yes	5 to 18	10.8	47%	76
Boran, 2011[37]	No, Turkish	Turkey	Canada	No	Cancer during neutropenia (adverse effect associated with cytotoxic therapy)	No	Yes	11 mths to 14 years	7.7 (3.4)	48%	50
Dickerson, 2018[80]	Yes	USA	Canada	No	Depression	Yes	No	13 to 17	15.3 (1.34)	65.2	392
Feeny, 2004[38]	Yes	Canada	Canada	Yes	Extremely low weight at birth	Yes	No	12 to 16	Extremely low weight at birth 14(1.6), born at term 14.4 (1.3)	Not reported	Extremely low birth weight 150, controls 125
Furlong, 2012[39]	Yes	Canada	Canada	Yes	Acute lymphoblastic leukaemia	Yes	Yes	5 to 18	Not reported for sample with HUI2/HUI3	Not reported for sample with HUI2/HUI3	Patients - 196
Glaser, 1999[81]	Yes	UK	Canada	No	Central nervous system tumours survivors	Yes	Yes	6 to 16	10.5	66.7	30

Study reference	English version of measure	Country	Value set	General population	Condition, where relevant	Self-report	Proxy report	Age range of children (years)	Mean age	% female	N
Kennedy, 1999[82]	Yes	UK	Canada	No	Childhood brain tumour survivors	Yes aged 16 and over	Yes for ages below 16	2 to 11	Not reported, median 5	Not reported	32
Klaassen, 2010a[83]	Yes	Canada	Canada	No	Hodgkin's disease	Yes	No	8 to 17	14.7	55.1	51
Klaassen, 2010b[40]	Yes	Canada	Canada	No	Hodgkin disease	Yes	Yes	8.9 to 18	14.7	55%	49
Kulpeng, 2013[41]	No, Thai	Thailand	Canada	N	Meningitis, bacteremia, pneumonia, acute otitis media, hearing loss, chronic lung disease, epilelps, mild mental retardation	Yes, age 7 and above who were able to communicate	Yes	5 to 14	10 (3)	38%	173
Le Gales, 1999[42]	No, French	France	N/A	No	Medulloblastoma and ependymona	Yes, with assistance by parent if child aged below 10	Yes	5 to 19	12 (4)	34.90%	43
Lynch, 2016[43]	Yes	USA	Canada	Yes	Depression	Yes	No	13 to 17	Nondepressed 15.2 (1.39), Subthreshold depression 15.4 (1.27), Full	Nondepressed 51.3%, Subthreshold depression 58.5%, Full depression 79.9%	392

Study reference	English version of measure	Country	Value set	General population	Condition, where relevant	Self-report	Proxy report	Age range of children (years)	Mean age	% female	N
									depression 15.4 (1.34)		
Mok, 2014[44]	No, Chinese	Hong Kong	Canada	No	Down Syndrome	No	Yes	5 to 18	Not reported	44%	30
Morrow, 2012[45]	Yes	Australia	N/A	No	Chronic Illness	Yes, age 12 and over and able to complete	Yes	5 to 18	12.2 (SD not reported)	45.80%	131
Nixon Speechley, 1999[46]	Yes	Canada	Canada	No	Childhood cancer survivors	No	Yes	7 to 16	12	79.5	250
Oluboyede, 2013[53]	Yes	UK	No	Yes	No	Yes	No	11 to 18	12.7	53	49
Petrou, 2013[47]	Yes	UK and Republic of Ireland	HUI2 - UK, HUI3 – Canada	Yes	Neurological disability and preterm births	No	Yes	Patients: 10 years 1 month to 11 years and 1 month, Controls: 9 years 9 months to 12 years 3 months	Median age for each sample: 10 years 11 months	Patients: 44.3%, Controls 59.9%	Patients 79, Controls 252
Ratcliffe, 2012a[55]	Yes	Australia	UK	Yes	No	Yes	No	11 to 17	15 (1.9)	48	500

Study reference	English version of measure	Country	Value set	General population	Condition, where relevant	Self-report	Proxy report	Age range of children (years)	Mean age	% female	N
Stevens, 2012b[30]	Yes	UK	UK	No	Intensive care	Yes aged over 11	Yes	5 and above	Not reported	Not reported	685
Sung, 2003[48]	Yes	Canada	Canada	No	Cancer	No	Yes	1 to 18	7.2 (4.0)	Not reported	36
Sung, 2004[49]	Yes	Canada	Canada	No	Chronic Illness	Yes	Yes	12 to 17	13.7 (1.7)	45%	19
Trevino, 2013[50]	Yes and Spanish	USA	Canada	Yes	Obesity	Yes	No	10 to 12	Not reported	53.10%	4979
Trudel, 1998[84]	Yes	Canada	Canada	No	Cancer	No	Yes	4 to 18	9.1 (3.8)	31.1	61
Ungar, 2012[85]	Yes	Canada	Canada	No	Asthma	Yes, solo then as dyad	Yes, solo then as dyad	8 to 17	10.9 (2.4)	45	91
HUI3											
Banks, 2008[34]	Yes	Canada	Canada	No	Cancer - undergoing chemotherapy	Yes, children aged 10 and over	Yes	2 to 18	9.5 (SD not reported)	35	29
Barr, 1997[35]	Yes	Canada	Canada	No	Cancer	Yes but due to low numbers were excluded from the analysis	Yes - nurse-investigator, parents	Not reported	Not reported	67%	18
Belfort, 2011[36]	Yes	USA	Canada	No	Attending well-child appointments or obesity clinic	Yes	Yes	5 to 18	10.8	47%	76

Study reference	English version of measure	Country	Value set	General population	Condition, where relevant	Self-report	Proxy report	Age range of children (years)	Mean age	% female	N
Boran, 2011[37]	No, Turkish	Turkey	Canada	No	Cancer during neutropenia (adverse effect associated with cytotoxic therapy)	No	Yes	11 months to 14 years	7.7 (3.4)	48%	50
Boulton, 2006[86]	Yes	England	Canada	No	Vision impairment or blindness	Yes	Yes - parents	3 to 8	6 years 2 months (1y 6 months)	41%	79
Cheng, 2000[87]	Yes	USA	Canada	No	Deafness	Yes	Yes	Not reported	10 (4.9)	40%	22
de Sonneville-Koedoot, 2014[88]	No, Dutch	Netherlands	Canada	Yes	Stutter	Yes	Yes	3 to 6.3	Not reported	30%	197
de Sonneville-Koedoot, 2015[89]	No, Dutch	Netherlands	Canada	No	Stutter	Yes	Yes	3 to 6.3	Not reported	30%	198
Dickerson, 2018[80]	Yes	USA	Canada	No	Depression	Yes	No	13 to 17	15.3 (1.34)	65.2	392
Feeny, 2004[38]	Yes	Canada	Canada	Yes	Extremely low weight at birth	Yes	No	12 to 16	Extremely low weight at birth 14(1.6), born at term 14.4 (1.3)	Not reported	Extremely low birth weight 150, controls 125
Francis, 2019[90]	Yes	Australia New Zealand	Canada	No	Chronic kidney disease	Yes - age 13 and over	Yes - age 12 and under	6 to 18	median 12.6	0.38	375

Study reference	English version of measure	Country	Value set	General population	Condition, where relevant	Self-report	Proxy report	Age range of children (years)	Mean age	% female	N
Furlong, 2012[39]	Yes	Canada	Canada	Yes	Acute lymphoblastic leukaemia	Yes	Yes	5 to 18	Not reported for sample with HUI2/HUI3	Not reported for sample with HUI2/HUI3	Patients - 196
Janse, 2008[91]	No, Dutch	Netherlands	Canada	No	Chronic Illness - cystic fibrosis admitted for pneumonia, newly diagnosed acute lymphatic leukaemia, juvenile idiopathic arthritis, or asthma	Yes	Yes	10 to 17	13 (1.7)	0.35	60
Kennes, 2002[92]	Yes	Canada	N/A	No	Cerebral palsy	Yes	Yes	5 to 13	8 years and 5 mths (SD 1 year 11 mths)	45.8%	408
Klaassen, 2010a[83]	Yes	Canada	Canada	No	Hodgkin disease	Yes	No	8 to 17	14.7	55.1	51
Klaassen, 2010b[40]	Yes	Canada	Canada	No	Hodgkin disease	Yes	Yes	8.9 to 18	14.7	55%	49
Kulkarni, 2010[93]	Yes	Canada	Canada	No	Obstructive hydrocephalus	No	Yes	5 to 18	treatment 12.3 (4.0), Shunt as first treatment 12.0 (4.0)	Not reported	47

Study reference	English version of measure	Country	Value set	General population	Condition, where relevant	Self-report	Proxy report	Age range of children (years)	Mean age	% female	N
Kulpeng, 2013[41]	No, Thai	Thailand	Canada	No	Meningitis, bacteremia, pneumonia, acute otitis media, hearing loss, chronic lung disease, epilepsy, mild mental retardation	Yes, age 7 and above who were able to communicate	Yes	5 to 14	10 (3)	38%	173
Le Gales, 1999[42]	No, French	France	N/A	No	Medulloblastoma and ependymona	Yes, with assistance by parent if child aged below 10	Yes	5 to 19	12 (4)	34.90%	43
Lee, 2011[94]	Yes	USA	Canada	No	Type 1 Diabetes	Yes	Yes	8 to 18	13.7 (3.1)	51.70%	238
Livingston, 2008[57]	Yes	Canada	Canada	No	Cerebral palsy	No	Yes	13 to 20	16 (1 year, 9 months)	46.50%	185
Lovett, 2010[43, 95]	Yes	UK	Canada	No	Deafness	No	Yes	18 months to 16 years	Unilateral 7.2 (3.7), Bilateral 7.3 (3.9)	Unilateral 60%, Bilateral 46.7%	Unilateral 20, Bilateral 30
Lynch, 2016[43]	Yes	USA	Canada	Yes	Depression	Yes	No	13 to 17	Nondepressed 15.2 (1.39), Subthreshold depression 15.4 (1.27), Full	Nondepressed 51.3%, Subthreshold depression 58.5%, Full depression 79.9%	392

Study reference	English version of measure	Country	Value set	General population	Condition, where relevant	Self-report	Proxy report	Age range of children (years)	Mean age	% female	N
									depression 15.4 (1.34)		
Mattera, 2018[60]	Yes	UK and USA	N/A	No	Hunter Syndrome	Yes, age 12 and over and able to complete	Yes, aged under 12 or unable to self-report, but this data cannot be extracted as is merged with caregiver report up to aged 26	12 to 17	Not reported	Not reported	Self report 7
Mok, 2014[44]	No, Chinese	Hong Kong	Canada	No	Down Syndrome	No	Yes	5 to 18	Not reported	44%	30
Morrow, 2012[45]	Yes	Australia	N/A	No	Chronic Illness	Yes, age 12 and over and able to complete	Yes	5 to 18	12.2 (SD not reported)	45.80%	131
Nixon Speechley, 1999[46]	Yes	Canada	Canada	No	Childhood cancer survivors	No	Yes	7 to 16	12 (SD not reported)	79.5	250
Penn, 2011[96]	Yes	UK	Canada	Yes	Childhood brain tumours	Yes aged 8 and over	Yes	3 to 16	10.5 (SD not reported)	Patients 51.7%,	29 patients,

Study reference	English version of measure	Country	Value set	General population	Condition, where relevant	Self-report	Proxy report	Age range of children (years)	Mean age	% female	N
										Controls 50%	32 controls
Petrou, 2013[47]	Yes	UK and Republic of Ireland	HUI2 - UK, HUI3 - Canada	Yes	Neurological disability and preterm births	No	Yes	Patients: 10 years 1 month to 11 years and 1 month, Controls: 9 years 9 months to 12 years 3 months	Median age for each sample: 10 years 11 months	Patients: 44.3%, Controls 59.9%	Patients 79, Controls 252
Rhodes, 2012[97]	Yes, English or Spanish	US	Canada	No	Adolescents with BMI≥85th percentile with Type 2 diabetes, pre-diabetes or insulin resistance	Yes	Yes	12 to 18	15.5 (2.0)	Not reported	107
Roposch, 2011[98]	Yes	UK	Canada	No	Osteonecrosis Secondary to Treatment of Developmental Dysplasia of the Hip	Yes	No	4 to 18	14 (2.5)	83%	72

Study reference	English version of measure	Country	Value set	General population	Condition, where relevant	Self-report	Proxy report	Age range of children (years)	Mean age	% female	N
Rosenbaum, 2007[58]	Yes	Canada	Canada	No	Cerebral palsy	No	Yes	13 to 20	16 (1 year, 9 months)	45.30%	203
Smith-Olinde, 2008[99]	Yes	USA	Canada	No	Permanent hearing loss	No	Yes	5 to 10	7.3 (1.9)	48.50%	103
Stade, 2006[100]	Yes	Canada	Canada	No	Children and youth prenatally exposed to alcohol, Foetal Alcohol Spectrum Disorder (FASD)	Yes, where feasible and possible	Yes	8 to 21	14.5 (SD not reported)	57.10%	126
Stevens, 2012b[30]	Yes	UK	Canada	No	Intensive care	Yes aged over 11	Yes	5 and above	Not reported	Not reported	685
Sung, 2003[48]	Yes	Canada	Canada	No	Cancer	No	Yes	1 to 18	7.2 (4.0)	Not reported	36
Sung, 2004[49]	Yes	Canada	Canada	No	Chronic Illness	Yes	Yes	12 to 17	13.7 (1.7)	45%	19
Tan, 2018[101]	Yes	Australia	Canada	No	Part of an obesity prevention intervention	No	Yes	2 to 5 years (not clearly reported)	Not reported	51%	368
Tilford, 2012[59]	Yes	USA	Canada	Yes	Autism spectrum disorders	No	Yes	4 to 17	8.6 (3.3)	14.70%	150

Study reference	English version of measure	Country	Value set	General population	Condition, where relevant	Self-report	Proxy report	Age range of children (years)	Mean age	% female	N
Trevino, 2013[50]	Yes, English and Spanish	USA	Canada	Yes	Obesity	Yes	No	10 to 12	Not reported	53.10%	4979
Ungar, 2012[85]	Yes	Canada	Canada	No	Asthma	Yes - solo then as dyad	Yes - solo then as dyad	8 to 17	10.9 (2.4)	45	91
Verrips, 2001[102]	Not reported	Netherlands	Canada	No	Very low birth weight children	Yes	Yes	14	Phone and postal 14.3 (0.2), face-to-face and postal 14.3 (0.1), postal only 14.2 (0.2)	Phone and postal 46%, face-to-face and postal 49%, postal only 51%	684 (Phone and postal 100, face-to-face and postal 103, postal only 481)
Wolke, 2013[103]	No, German	Germany	Canada	Yes	Very low birth weight (VLBW) and very preterm (VP) born children	Yes, with exception of adolescents with moderate to severe disability	Yes	13	Not reported	Controls 50.5%, VLBW/VP I (no or mild disability) 44.8%, VLBW/VP II (moderate to severe disability) 75.0%	Controls 282, VLBW/VP I 260, VLBW/VP II 12

Notes: [a] Wong et al (2019) compare the EQ-5D-Y 3 level and 5 level versions.

VLBW/VP = Very low birth weight/very pre-term.

Table 5: Measures of interest and psychometric properties assessed in included studies

Study reference	EQ-5D-Y	CHU9D	HUI2	HUI3	AQoL-6D	Known group validity	Convergent validity	Responsiveness	Reliability	Acceptability and feasibility	General population, clinical population or both
Åström, 2018[69]	✓					✓					General
Banks, 2008[34]			✓	✓			✓	✓			Clinical
Barr, 1997[35]				✓				✓		✓	Clinical
Belfort, 2011[36]			✓	✓		✓			✓		Clinical
Bergfors, 2015[70]	✓						✓				Clinical
Boran, 2011[37]			✓	✓		✓		✓			Clinical
Boulton, 2006[86]				✓		✓					Clinical
Burstrom, 2014[71]	✓					✓	✓				Both
Canaway, 2013[51]	✓	✓				✓	✓		✓	✓	General
Chen, 2015[52]	✓	✓				✓	✓				General
Cheng, 2000[87]			✓	✓				✓			Clinical
de Sonneville-Koedoot, 2014[88]				✓		✓					Both
de Sonneville-Koedoot, 2015[89]				✓				✓			Clinical
Dickerson, 2018[80]			✓	✓			✓	✓			Clinical
Eidt-Koch, 2009[72]	✓						✓				Clinical
Feeny, 2004[38]			✓	✓		✓	✓				Both
Foster Page, 2015[62]		✓					✓	✓			Clinical

Study reference	EQ-5D-Y	CHU9D	HUI2	HUI3	AQoL-6D	Known group validity	Convergent validity	Responsiveness	Reliability	Acceptability and feasibility	General population, clinical population or both
Francis, 2019[90]				✓		✓					Clinical
Frew, 2015[63]		✓				✓	✓				Both
Furber, 2015[64]		✓				✓	✓				Clinical
Furlong, 2012[39]			✓	✓		✓				✓	Both
Glaser, 1999[81]			✓						✓	✓	Clinical
Hernandez, 2015[31]	✓					✓					Clinical
Hsu, 2018[73]	✓								✓		Clinical
Janse, 2008[91]				✓					✓		Clinical
Jelsma, 2010[74]	✓									✓	General
Kennedy, 1999[82]			✓			✓					Clinical
Kennes, 2002[92]				✓			✓				Clinical
Kim, 2018[61]	✓					✓				✓	Clinical
Klaassen, 2010a[83]			✓	✓			✓	✓			Clinical
Klaassen, 2010b[40]			✓	✓					✓		Clinical
Kulkarni, 2010[93]				✓			✓				Clinical
Kulpeng, 2013[41]			✓	✓			✓		✓		Clinical
Le Gales, 1999[42]			✓	✓		✓			✓	✓	Clinical
Lee, 2011[94]				✓					✓	✓	Clinical
Livingston, 2008[57]				✓			✓				Clinical

Study reference	EQ-5D-Y	CHU9D	HUI2	HUI3	AQoL-6D	Known group validity	Convergent validity	Responsiveness	Reliability	Acceptability and feasibility	General population, clinical population or both
Loof, 2019[75]		✓				✓					Both
Lovett, 2010[95]				✓		✓					Clinical
Lynch, 2016[43]			✓	✓		✓					Clinical
Mattera, 2018[60]				✓		✓	✓				Clinical
Mayoral, 2017 [33]	✓					✓	✓	✓			Clinical
Mok, 2014[44]			✓	✓		✓				✓	Clinical
Morrow, 2012[45]			✓	✓					✓		Clinical
Nixon Speechley, 1999[46]			✓	✓			✓				Clinical
Oluboyede, 2013[53]	✓		✓							✓	General
Oluboyede, 2019[65]		✓				✓	✓				General
Penn, 2011[96]				✓		✓			✓		Both
Perez-Sousa, 2018[76]	✓							✓	✓		Clinical
Petersen, 2018[66]		✓				✓	✓				General
Petrou, 2013[47]			✓	✓		✓					Both
Ratcliffe, 2012a[55]		✓	✓		✓	✓	✓				General
Ratcliffe, 2012b[54]		✓			✓	✓					General
Ravens-Sieberer, 2010[21]	✓					✓	✓		✓	✓	General
Rhodes, 2012[97]				✓		✓	✓		✓		Clinical

Study reference	EQ-5D-Y	CHU9D	HUI2	HUI3	AQoL-6D	Known group validity	Convergent validity	Responsiveness	Reliability	Acceptability and feasibility	General population, clinical population or both
Robles, 2015[77]	✓					✓			✓	✓	General
Roposch, 2011[98]				✓		✓				✓	Clinical
Rosenbaum, 2007[58]				✓		✓	✓				Clinical
Sach, 2017[32]		✓				✓	✓	✓			Clinical
Scalone, 2011[78]	✓					✓	✓		✓	✓	Both
Scott, 2017[79]	✓					✓	✓	✓	✓	✓	Both
Smith-Olinde, 2008[99]				✓		✓				✓	Clinical
Stade, 2006[100]				✓		✓			✓		Clinical
Stevens, 2012a[67]		✓				✓	✓				General
Stevens, 2012b[30]			✓	✓		✓		✓	✓	✓	Clinical
Sung, 2003[48]			✓	✓			✓			✓	Clinical
Sung, 2004[49]			✓	✓					✓		Clinical
Tan, 2018[101]				✓		✓					Clinical
Tilford, 2012[59]				✓		✓	✓				Both
Trevino, 2013[50]			✓	✓		✓					General
Trudel, 1998[84]			✓			✓	✓		✓		Clinical
Ungar, 2012[85]			✓	✓		✓	✓	✓	✓		Clinical
Verrips, 2001[102]				✓					✓		Clinical
Wille, 2010[22]	✓									✓	General
Wolke, 2013[103]				✓		✓					Both
Wong, 2019[56]	✓								✓		Clinical

Study reference	EQ-5D-Y	CHU9D	HUI2	HUI3	AQoL-6D	Known group validity	Convergent validity	Responsiveness	Reliability	Acceptability and feasibility	General population, clinical population or both
Xu, 2014[68]		✓				✓					General

4.4. KNOWN-GROUP VALIDITY

Table 6 presents the results of all studies assessing known group validity (n=48).

4.4.1. AQoL-6D

One study assesses known-group validity for the AQoL-6D index[54], finding that AQoL-6D significantly captured known-group differences for general health and long-standing illness.

4.4.2. CHU9D

Eleven studies assess known-group validity for CHU9D, with ten finding CHU9D significantly captured known-group differences. Of these two studies assess both the CHU9D index and the dimensions, and all other studies only assess the index. The known-group differences assessed are: healthy/less healthy (derived using PedsQL); general health category (excellent, very good, good, fair or poor); long-term illness; clinical banding (derived using Strengths and Difficulties Questionnaire (SDQ); eczema severity (derived using Patient Oriented Eczema Measure (POEM)); self-assessed weight; illness or disability; physical activity level; and sleep hours (above or below median level). One study did not find evidence that the CHU9D was able to find a statistically significant difference between two groups of weight status (healthy and underweight; overweight and obese).

4.4.3. EQ-5D-Y

Twelve studies assess known-group validity for EQ-5D-Y, with all finding EQ-5D-Y significantly captured known-group differences. However for four studies evidence of known-group validity was not found for all five dimensions (note that for EQ-5D-Y dimensions are assessed for eight studies whereas the index is assessed for four studies). In addition for one study, evidence of known-group validity was not found for all known-groups examined. The known-group differences assessed are: self-reported condition; patients/general population; healthy/less healthy (derived using PedsQL); general health; long-term illness; allergic symptoms; chronic condition; clinical banding (derived using SDQ); groups reflecting severity (mainstream school, chronic disability, chronically ill, acutely ill); well-controlled/not well-controlled asthma; idiopathic clubfoot with/without neurodevelopmental difficulties; (probably have) mental disorder; metabolic control (this is not statistically significant). One paper[72] assesses whether

responses differ for clinical subgroups by Cystic Fibrosis Questionnaire (CFQ) subscales for patients reporting any problem on EQ-5D-Y, but as this analysis is not straightforward to interpret as being indicative of performance of EQ-5D-Y (since it excludes all those reporting no problems), this has not been included in the table.

4.4.4. HUI2

Fourteen studies assess known-group validity for the HUI2, with eleven finding HUI2 significantly captured known-group differences for: neutropenic/non-neutropenic; low birth weight/birth at term; acute lymphoblastic leukaemia patients/general population; SSEN (Statement of Special Educational Needs); nondepressed/depressed and depression severity; behavioural problems; speech problems; learning problems; hearing problems; vision problems; degree of impairment; disability; general health; and undergoing active treatment/follow-up (for a subset of HUI2 dimensions) but that it does not significantly capture known-group differences for: radiation dose of treatment; high risk of behavioural problems or emotional problems (though note small sample size of 32); long-standing illness; endocrine problems; fasting glucose; and fasting insulin. Weight categories are assessed in two studies with one finding significant differences and the other did for the index but when assessing dimensions only the mobility dimension was significant. One study does not report the findings, and one study states only that HUI2 did not significantly capture known-group differences for persistent asthma severity (though note small sample size of 91). Ten studies assess the HUI2 index, one study assesses the dimensions only and three studies assess both the index and dimensions.

4.4.5. HUI3

Twenty-four studies assess known group validity for HUI3, with seventeen finding HUI3 significantly captured known-group differences for: neutropenic/non-neutropenic; ophthalmic conditions; stutter; low birth weight at birth/birth at term; chronic kidney disease stages; dialysis/transplant; acute lymphoblastic leukaemia patients/general population; nondepressed/depressed; depression severity; endocrine problems; behavioural problems; speech problems; learning problems; hearing problems; vision problems; patients with childhood brain tumours/controls; disability; impairment severity; Bucholz-Ogden Grades; gross motor function levels; foetal alcohol syndrome disorder/general population; Asperger's syndrome/autism

disorder; severity level of autism symptoms; language use; weight categories; fasting glucose; fasting insulin; very low birth weight/very pre-term with/without disability vs full term. Significant differences were not found for: radiation dose of treatment; bilateral/unilateral cochlear implants (though note small sample size of 50 in the study); degree of hearing loss; foetal alcohol symptoms vs foetal alcohol effects. Two studies do not report the findings, and one study states only that HUI2 did not significantly capture known-group differences for persistent asthma severity (though note small sample size of 91). Weight categories are assessed in three studies, with one finding significant differences, and the other two finding significant differences for the index but when assessing dimensions only the mobility dimension was significant.

Table 6: Known group validity (48 studies)

Measure	Study reference	Index or dimensions or both assessed	Groups defined as	Mean differences across groups in direction consistent with clinical expectation	Difference between groups statistically significant
AQoL-6D	Ratcliffe, 2012b[54]	Index	General health (excellent, very good, good, fair/poor); long-standing illness/disability/medical condition	Yes	Yes
CHU9D	Canaway, 2013[51]	Index	Healthy group and less healthy group categorised using PedsQL score	Yes	Yes
CHU9D	Chen, 2015[52]	Index	Self-assessed general health (excellent, very good, good, fair or poor); long-term disability/illness/medical condition	Yes	Yes
CHU9D	Frew, 2015[63]	Both	Weight status (healthy and underweight; overweight and obese)	No	No
CHU9D	Furber, 2015[64]	Index	SDQ clinical band (normal, borderline, abnormal)	Yes	Yes
CHU9D	Oluboyede, 2019[65]	Index	Weight status (normal/overweight/obese); self-assessed weight (very overweight/moderately overweight/slightly overweight/about the right weight/slightly underweight/moderately underweight/very underweight); General health (Excellent/very good/good/fair/poor); Illness or disability (yes/no)	Yes	Yes
CHU9D	Petersen, 2018[66]	Index	General health (excellent, very good, good, fair/poor); long term disability/illness/medical condition	Yes	Yes
CHU9D	Ratcliffe, 2012a[55]	Index	General health (excellent, very good, good, fair/poor); long standing disability/illness	Yes	Yes for general health, no for long-standing disability/illness
CHU9D	Ratcliffe, 2012b[54]	Index	General health (excellent, very good, good, fair/poor); long-standing illness/disability/medical condition	Yes	Yes
CHU9D	Sach, 2017[32]	Index	POEM group, at baseline and follow-up	Yes	Yes

Measure	Study reference	Index or dimensions or both assessed	Groups defined as	Mean differences across groups in direction consistent with clinical expectation	Difference between groups statistically significant
CHU9D	Stevens, 2012a[55]	Both	General health (excellent, very good, good, fair, poor); long-standing disability/illness/medical condition	Yes	Yes
CHU9D	Xu, 2014[68]	Both	General health (assessed for index only); physical activity level; sleep hours in the last 7 days above and below median time	Yes for index and for relevant dimensions	Yes for index; Yes for physical activity for most dimensions; Yes for sleep for all dimensions
EQ-5D-Y	Åström, 2018[69]	Dimensions	Self-reported condition	Yes	Yes
EQ-5D-Y	Burstrom, 2014[71]	Dimensions	Patient and general population samples	Yes	Yes
EQ-5D-Y	Canaway, 2013[51]	Index	Healthy group and less healthy group categorised using PedsQL score	Yes	Yes
EQ-5D-Y	Chen, 2015[52]	Index	Self-assessed general health (excellent, very good, good, fair or poor); long-term disability/illness/medical condition	Yes	Yes
EQ-5D-Y	Kim, 2018[61]	Dimensions	Allergic symptoms (wheezing or whistling in the chest, runny or blocked nose, itchy rash)	Yes	Yes
EQ-5D-Y	Loof, 2019[75]	Dimensions	Patient and general population samples; Idiopathic clubfoot and neurodevelopmental difficulties (NDD)/ idiopathic clubfoot and no NDD	Yes	Yes with exception of self-care and worried/sad unhappy for patient/general population, and exception of mobility and self-care for patient subsamples
EQ-5D-Y	Mayoral, 2017[33]	Index	Probable/not probable mental disorders; good/poor metabolic control	Yes for mental disorders; unknown for metabolic control	Yes for mental disorders; no for metabolic control
EQ-5D-Y	Ravens-Sieberer, 2010[21]	Dimensions	SDQ scores (normal, borderline/abnormal) in German and Spain samples; chronic condition	Yes in general, though not for self-care with exception of Sweden	Yes for some dimensions for some countries, not for self-care with exception of Sweden
EQ-5D-Y	Robles, 2015[77]	Dimensions	General health (excellent/very good/good, fair/poor); SDQ	Yes	Yes with exception of self-care for general health
EQ-5D-Y	Scalone, 2011[78]	Dimensions	General population and patient samples using matching	Yes	No with exception of worried/sad/unhappy

Measure	Study reference	Index or dimensions or both assessed	Groups defined as	Mean differences across groups in direction consistent with clinical expectation	Difference between groups statistically significant
EQ-5D-Y	Scott, 2017[79]	Dimensions	Four groups reflecting severity (mainstream school, chronic disability, chronically ill, acutely ill)	Yes	Yes
EQ-5D-Y	Hernandez, 2015[31]	Index	Well-controlled and not well-controlled asthma	Yes	Yes
HUI2	Belfort	Index	Healthy weight vs overweight or obese	Yes	No
HUI2	Boran	Both	Neutropenic vs non neutropenic (note this is examined over time so also reported under responsiveness)	Yes	Index - Yes; Dimensions - mobility, emotion, self-care
HUI2	Feeny, 2004[38]	Index	Low weight at birth vs birth at term	Yes	Yes
HUI2	Furlong, 2012[39]	Index	Patients vs general population	Yes	Yes
HUI2	Kennedy, 1999[82]	Index	SSEN; high risk of behavioural problems; emotional problems	Yes for SSEN and high risk of behavioural problems	Yes for SSEN; No otherwise
HUI2	Le Gales, 1999[42]	Dimensions	Radiation dose of treatment	No	No
HUI2	Lynch, 2016[43]	Index	Nondepressed vs depressed; depression severity (nondepressed vs subthreshold; subthreshold vs full depression; nondepressed vs full depression; subthreshold vs moderate depression; moderate depression vs severe depression; subthreshold vs severe depression)	Yes	Yes with exception of moderate depression vs severe depression
HUI2	Mok, 2014[44]	Index	Endocrine problems (yes/no); behavioural problems; speech problems; learning problems; hearing problems; vision problems	Yes	Yes with exception of endocrine problems. Multiple regression was also reported but not explained
HUI2	Petrou, 2013[47]	Index	Impairment (no impairment vs mild impairment; no impairment vs moderate impairment; no impairment vs severe impairment) and disability (yes/no)	Yes	Yes

Measure	Study reference	Index or dimensions or both assessed	Groups defined as	Mean differences across groups in direction consistent with clinical expectation	Difference between groups statistically significant
HUI2	Ratcliffe, 2012a[55]	Index	General health (excellent, very good, good, fair/poor); long standing disability/illness	Yes	Yes for general health, no for long-standing disability/illness
HUI2	Stevens, 2012b[30]	Index	Different degrees of in-hospital severity of illness, at 6 months and 12 months	Not reported	Not reported
HUI2	Trevino, 2013[50]	Both	Index - Normal weight vs overweight; normal weight vs obese; normal weight vs severely obese; fasting glucose; fasting insulin; Dimensions - weight (normal weight, overweight, obese, severely obese)	Index - Yes, Dimensions – mobility	Index - Yes for normal weight vs obese; normal weight vs severely obese, Dimensions - mobility
HUI2	Trudel, 1998[84]	Both	Patients undergoing active treatment and those on follow up	Yes for some dimensions	Yes for emotion, pain, self-care dimensions and index
HUI2	Ungar, 2012[85]	Index	Mild, moderate or severe persistent asthma	Not reported	No
HUI3	Belfort, 2011[36]	Index	Healthy weight vs overweight or obese	Yes	No
HUI3	Boran, 2011[37]	Both	Neutropenic vs non neutropenic (note this is examined over time so also reported under responsiveness)	Yes	Index - Yes; Dimensions – emotion
HUI3	Boulton, 2006[86]	Both	Ophtalmic conditions (visual pathway condition, condition of the eye and nystagmus alone)	Yes	Yes
HUI3	De Sonneville, 2014[88]	Both	Children who stutter vs general population	Yes	Yes for speech, emotion, cognition and for the index
HUI3	Feeny, 2004[38]	Index	Low weight at birth vs birth at term	Yes	Yes
HUI3	Francis, 2019[90]	Both	Different chronic kidney disease stages; dialysis vs transplant	Yes	Yes
HUI3	Furlong, 2012[39]	Index	Patients vs general population	Yes	Yes
HUI3	Le Gales, 1999[42]	Dimensions	Radiation dose of treatment	No	No

Measure	Study reference	Index or dimensions or both assessed	Groups defined as	Mean differences across groups in direction consistent with clinical expectation	Difference between groups statistically significant
HUI3	Lovett, 2010[95]	Index	Bilateral vs unilateral cochlear implants	Yes	No
HUI3	Lynch, 2016[43]	Index	Nondepressed vs depressed; depression severity (nondepressed vs subthreshold; subthreshold vs full depression; nondepressed vs full depression; subthreshold vs moderate depression; moderate depression vs severe depression; subthreshold vs severe depression)	Yes	Yes
HUI3	Mok, 2014[44]	Index	Endocrine problems (yes/no); behavioural problems; speech problems; learning problems; hearing problems; vision problems	Yes	Yes. Multiple regression was also reported but not explained.
HUI3	Penn, 2011[96]	Both	Patients vs controls	Yes	Yes for index for parent report at 3 timepoints (T1, T6, T12) and self-report for the only timepoint reported (T12), Yes for all timepoints parent report for ambulation, emotion, cognition and pain, Yes for cognition only for self-report
HUI3	Petrou, 2013[47]	Both	Dimensions - children with disability vs children without disability; Index - impairment (no impairment vs mild impairment; no impairment vs moderate impairment; no impairment vs severe impairment) and disability (yes/no)	Yes	Dimensions - Yes with exception of emotion, Index – Yes
HUI3	Rhodes, 2012[97]	Index	Diagnosis (Type 2 diabetes, prediabetes, insulin resistance)	No	No
HUI3	Roposch, 2011[98]	Both	Bucholz-Ogden Grades I, II, III, IV	Index - Yes for grades I,II,III vs grade IV but not within grades I,II,III. Dimensions - median scores reported with	Index - Yes for grades I and II vs grades III and IV. Dimensions - no

Measure	Study reference	Index or dimensions or both assessed	Groups defined as	Mean differences across groups in direction consistent with clinical expectation	Difference between groups statistically significant
				no difference except pain grade IV	
HUI3	Rosenbaum, 2007[58]	Dimensions	Gross Motor Function Classification System - Level I, Level II, Level III, Level IV, Level V	Yes for ambulation and cognition, mainly for vision, speech, dexterity	Yes
HUI3	Smith-Olinde, 2008[99]	Index	Degree of hearing loss (mild vs moderate; moderate vs severe; severe vs profound no cochlear implant; severe vs profound with cochlear implant	Yes	No - not significant for severe vs profound no cochlear implant; severe vs profound with cochlear implant; significance not otherwise reported
HUI3	Stade, 2006[100]	Both for FAS vs FAE, Index for FASD vs general population	Foetal Alcohol Syndrome vs Foetal Alcohol Effects; FASD (all sample) vs general population	Yes	No for FAS vs FAE, Yes for FASD vs general population
HUI3	Stevens, 2012b[30]	index	Different degrees of in-hospital severity of illness, at 6 months and 12 months	Not reported	Not reported
HUI3	Tan, 2018[101]	Both	Weight (healthy weight, overweight, obese)	No	No
HUI3	Tilford, 2012[59]	Index	Asperger's disorder vs autism disorder; severity level of autism symptoms; no problems with language use and understanding vs severe problems with language use and understanding	Yes; Yes for most symptoms though not always for moderate vs severe problems; Yes	Yes; Yes for many symptoms (compulsive behaviours, anxiety, sensory issues, sleep disturbance, hyperactivity, attention span, eating habits, social interactions, self-stimulatory and repetitive behaviours, self-injurious behaviour, lost or losing skills previously had); Yes
HUI3	Trevino, 2013[50]	Both	Index - Normal weight vs overweight; normal weight vs obese; normal weight vs severely obese; fasting glucose; fasting insulin; Dimensions - weight (normal weight, overweight, obese, severely obese)	Index - Yes, Dimensions - speech	Index - Yes for normal weight vs obese; normal weight vs severely obese, Dimensions - speech (though not significant for severely obese)

Measure	Study reference	Index or dimensions or both assessed	Groups defined as	Mean differences across groups in direction consistent with clinical expectation	Difference between groups statistically significant
HUI3	Ungar, 2012[85]	Index	Mild, moderate or severe persistent asthma	Not reported	No
HUI3	Wolke, 2013[103]	Dimensions	Parent report: VLBW/VP I vs full-term, VLBW/VP II vs full term, VLBW/VP I vs VLBW/VP II; For self-report VLBW/VP I vs Full-term	Yes for vision, speech, dexterity, ambulation, cognition, no for hearing, emotion (with exception of parental report VLBW/VP I vs VLBW/VP II) and pain	Yes always for dexterity and ambulation, mostly for vision, speech, cognition, sometimes for emotion and pain (though not in expected clinical direction)

Notes: FAE = Foetal alcohol effects; FAS=Foetal alcohol syndrome; FASD= Foetal alcohol syndrome disorder; NDD=Neurodevelopmental difficulties; POEM=Patient Oriented Eczema Measure; SDQ=Strengths and Difficulties Questionnaire; SSEN=Statement of Special Educational Needs; VLBW/VP = Very low birth weight/very pre-term.

4.5. CONVERGENT VALIDITY

Table 7 presents the results of all studies assessing convergent validity (n=33).

4.5.1. AQoL-6D

No studies assess the convergent validity of AQoL-6D.

4.5.2. CHU9D

Ten studies assess convergent validity of CHU9D, with all finding some significant correlations and one finding a significant relationship using regression analysis. Significant correlations are found between CHU9D dimensions and: similar EQ-5D-Y dimensions; similar PedsQL domain scores; SDQ items; similar HUI2 dimensions; and KIDSCREEN-10 scores (where domains are similar). Significant correlations are found between CHU9D utilities and: similar EQ-5D-Y dimensions; global measure of health; PedsQL total score; SDQ score; HUI2 utility; ADQOL (Atopic dermatitis-specific preference-based measure) utility; and WAItE (Weight-specific Adolescent Instrument for Economic evaluation) index.

4.5.3. EQ-5D-Y

Nine studies assess convergent validity of EQ-5D-Y, with all finding some significant correlations, and no studies use regression analysis. Significant correlations are found between EQ-5D-Y dimensions and: similar PAQLQ (Paediatric Asthma Quality of Life Questionnaire) domains; KIDSCREEN domains; KIDSCREEN index; general health; life satisfaction; similar CHU9D dimensions; CHU9D utility; PedsQL domain summary scores; PedsQL items; CFQ scales (Cystic Fibrosis Questionnaire); WeeFim dimensions (researcher-reported measure capturing functional independence); Faces Pain Scale. Typically significant correlations are not found between all EQ-5D-Y dimensions and the dimensions or domains of other measures, but where the domains/dimensions conceptually overlap, which would be expected.

One study examines the correlation between EQ-5D-Y and EQ-VAS[69] with no other correlations examined, finding significant correlations for mobility, looking after myself and usual activities. Another study[61] also examines the relationship between EQ-5D-Y and EQ-VAS by looking at problem reporting rates by dimension and VAS score according to allergic symptoms reported. Differences in responses for EQ-5D and EQ-

5D-Y are higher for the pain/discomfort and worried/sad/unhappy dimensions. These studies have not been reported within the tables since EQ-VAS is a self-complete score of the individual's view of their own health on a 0-100 scale and can be considered a component of the EQ-5D-Y. Another study[73] examines the correlations between EQ-5D-Y dimensions, a clinical measure (eGFR, estimated glomerular filtration rate) as well as comorbidities and clinical conditions, finding some significant correlations, though none for pain/discomfort and few for worried/sad/unhappy.

4.5.4. HUI2

Ten studies assess convergent validity of HUI2, with all finding some significant correlations, and one study also finding significant relationships with other measures using regression analysis. Significant correlations are found between HUI2 dimensions and: similar CHU9D dimensions; similar CHQ summary component scores (Child Health Questionnaire); CHQ domain scores; POQOLS (Paediatric Oncology Quality of Life Scale); and CBCL (Child Behaviour Checklist). Significant correlations are found between HUI2 utilities and: CDRS-R (Child Depression Rating Scale-Revised); CHQ summary component; CHQ physical; CHQ psychosocial; EQ-5D; HUI3, Lansky play-performance scale; PedsQL; PedsQL cancer; and CHU9D utility. One study found significant correlations using parent proxy-report but not when assessing self-report responses.

4.5.5. HUI3

Fifteen studies assess convergent validity of HUI3, with fourteen studies finding significant correlations. Significant correlations are found between HUI3 dimensions and: CHQ domains; CHQ summary component scores; HS-FOCUS; PedsQL domains. Significant correlations are found between HUI3 index and: CDRS-R; CHQ physical; CHQ psychosocial; cognitive functioning; GMFCS; EQ-5D; HOQ; HUI2, Lansky play-performance scale; PedsQL; PedsQL cancer; Vineland-II adaptive behaviour scales. Two studies do not find significant correlations between HUI3 and Quality of Life Instrument for People With Developmental Disabilities. One study uses both regression analysis and correlations.

Table 7: Convergent validity (33 studies)

Measure	Study reference	Correlation examined	Other measures examined for correlation	Significant correlation(s) (0.41 and above)	Regression analysis undertaken	Regression details	Regression analysis shows significant relationship
CHU9D	Canaway, 2013[51]	Yes	EQ-5D-Y dimensions and CHU9D dimensions, PedsQL dimension summary scores	Yes - amongst similar dimensions/domains	No		
CHU9D	Chen, 2015[52]	Yes	EQ-5D-Y and CHU9D dimensions, EQ-5D-Y and CHU9D utilities	Yes - amongst similar dimensions/domains. Agreement highest for higher utilities	No		
CHU9D	Foster Page, 2015[62]	Yes	CHU9D utility and Global measure of health	Yes	No		
CHU9D	Frew, 2015[63]	Yes	Dimensions and PedsQL domains; CHU9D utility and PedsQL total score	Yes for utilities, no across dimensions/domains	No		
CHU9D	Furber, 2015[64]	Yes	CHU9D utility and SDQ total score; CHU9D dimensions and SDQ items	Yes for utility/scores and amongst similar dimensions/domains	Yes	OLS regression of SDQ total on CHU9D dimensions	Yes for 4 dimensions
CHU9D	Oluboyede, 2019[65]	Yes	CHU9D utility and WAItE index	Yes	No		
CHU9D	Petersen, 2018[66]	Yes	CHU9D dimensions and PedsQL dimensions; CHU9D utility PedsQL total score	Yes – between utility score and total score, amongst similar dimensions/domains though to a lesser extent for pain, tired, sleep, social functioning	No		
CHU9D	Ratcliffe, 2012a[55]	Yes	CHU9D and HUI2 dimensions, CHU9D and HUI2 utilities	Yes	No		
CHU9D	Sach, 2017[32]	Yes	Correlation of CHU9D and ADQOL utility scores at baseline and follow-up	Yes – fair at baseline and moderate at follow-up	No		

Measure	Study reference	Correlation examined	Other measures examined for correlation	Significant correlation(s) (0.41 and above)	Regression analysis undertaken	Regression details	Regression analysis shows significant relationship
CHU9D	Stevens, 2012a[67]	Yes	CHU9D dimensions and KIDSCREEN-10 score	Yes – though only for some of the similar dimensions/domains	No		
EQ-5D-Y	Bergfors, 2015[70]	Yes	Dimensions with PAQLQ domains, PAQLQ total score, SRH	Yes – for usual activities and pain/discomfort with PAQLQ, pain/discomfort with SRH	No		
EQ-5D-Y	Burstrom, 2014[71]	Yes	Dimensions with Kidscreen domains, general health, life satisfaction. Undertaken separately for patient and general population samples	Yes – higher correlations in general population sample and with KIDSCREEN over general health and life satisfaction	No		
EQ-5D-Y	Canaway, 2013[51]	Yes	EQ-5D-Y dimensions, CHU9D dimensions, PedsQL domain summary scores	Yes - amongst similar dimensions/domains	No		
EQ-5D-Y	Chen, 2015[52]	Yes	EQ-5D-Y dimensions and CHU9D dimensions, EQ-5D-Y and CHU9D utilities	Yes - amongst similar dimensions/domains. Agreement highest for higher utilities	No		
EQ-5D-Y	Eidt-Koch, 2009[72]	Yes	Dimensions with CFQ scales	Yes - amongst similar dimensions/domains	No		
EQ-5D-Y	Mayoral, 2017[33]	Yes	EQ-5D-Y utility, mobility, anxiety/depression and KIDSCREEN	Yes - anxiety/depression and EQ-5D-Y utility with KIDSCREEN	No		
EQ-5D-Y	Ravens-Sieberer, 2010[21]	Yes	EQ-5D-Y dimensions and Kidscreen (index, Physical wellbeing and psychological wellbeing scores), general health	Yes – though rarely for self-care	No		

Measure	Study reference	Correlation examined	Other measures examined for correlation	Significant correlation(s) (0.41 and above)	Regression analysis undertaken	Regression details	Regression analysis shows significant relationship
			and life satisfaction scores				
EQ-5D-Y	Scalone, 2011[78]	Yes	EQ-5D-Y dimensions and PedsQL items	Yes - amongst similar dimensions/domains	No		
EQ-5D-Y	Scott, 2017[79]	Yes	EQ-5D-Y dimensions and PedsQL dimensions, WeeFIM dimensions and Faces Pain Scale, analyses undertaken separately for different groups	Yes – for some similar dimensions/domains, though only significant for all similar dimensions/domains for the acutely ill group	No		
HUI2	Banks, 2008[34]	Yes	For self-report - HUI2 and HUI3 utilities, PedsQL, PedsQL cancer; for proxy-report - HUI2 and HUI3 utilities, PedsQL, PedsQL cancer, CHQ physical, CHQ psychosocial	No for self-report; Yes for parent-report for all with exception of CHQ psychosocial	No		
HUI2	Dickerson, 2018[80]	Yes	Utility with CDRS-R, reported separately for full sample/sample depressed at baseline	Yes	Yes	Change from baseline to 12 week follow-up regressed on whether have >20% improvement in CDRS-R score and age, ethinic minority, gender, baseline value	Yes
HUI2	Feeny, 2014[38]	Yes	HUI2 and HUI3 utilities	Yes	No		

Measure	Study reference	Correlation examined	Other measures examined for correlation	Significant correlation(s) (0.41 and above)	Regression analysis undertaken	Regression details	Regression analysis shows significant relationship
HUI2	Klaassen, 2010a[83]	Yes	HUI2 utilities with PedsQL; PedsQL cancer; Lansky Play-Performance Scale; HUI3 utility	Yes	No		
HUI2	Kulpeng, 2013[41]	Yes	HUI2, HUI3, EQ-5D utilities - assessed separately for self-report and caregiver proxy-report	Yes	No		
HUI2	Nixon Speechley, 1999[46]	Yes	CHQ summary component scores with HUI2 utility scores; CHQ domain scores and HUI2 dimensions	Yes – for the index and amongst similar dimensions/domains	No		
HUI2	Ratcliffe, 2012a[55]	Yes	CHU9D and HUI2 dimensions, CHU9D and HUI2 utilities	Yes for index and dimensions	No		
HUI2	Sung, 2003[48]	Yes	HUI2 dimensions and utility with CHQ summary component scores	Yes – amongst similar dimensions/domains, not for utility	No		
HUI2	Trudel, 1998[84]	Yes	HUI2 dimensions and POQOLS, CBCL and TRF	Yes, for some HUI2 dimensions and POQOLS and CBCL	No		
HUI2	Ungar, 2012[85]	Yes	HUI2 utility and dimensions with PedsQL domains	Yes – for utility and amongst similar dimensions/domains (stronger for HUI2 than HUI3)	No		
HUI3	Banks, 2008[34]	Yes	For self-report - HUI2 and HUI3 utilities, PedsQL, PedsQL cancer; for	No for self-report; Yes for parent-report for all with	No		

Measure	Study reference	Correlation examined	Other measures examined for correlation	Significant correlation(s) (0.41 and above)	Regression analysis undertaken	Regression details	Regression analysis shows significant relationship
			proxy-report – HUI2 and HUI3 utilities, PedsQL, PedsQL cancer, CHQ physical, CHQ psychosocial	exception of CHQ psychosocial			
HUI3	Dickerson, 2018[80]	Yes	Utility with CDRS-R, reported separately for full sample/sample depressed at baseline	Yes	No		
HUI3	Feeny, 2004[38]	Yes	HUI2 and HUI3 utilities	Yes	No		
HUI3	Kennes, 2002[92]	Yes	HUI3 dimensions and GMFCS	Yes	No		
HUI3	Klaassen, 2010a[83]	Yes	HUI3 utilities with PedsQL; PedsQL cancer; Lansky Play-Performance Scale; HUI2 utility	Yes	No		
HUI3	Kulkarni, 2010[93]	Yes	HUI3 utilities with HOQ	Yes	No		
HUI3	Kulpeng, 2013[41]	Yes	HUI3, HUI2, EQ-5D utilities - assessed separately for self-report and caregiver proxy-report	Yes	No		
HUI3	Livingston, 2008[57]	Yes	HUI3 dimensions with Short Version of the Quality of Life Instrument for People with Developmental Disabilities	No	No		
HUI3	Mattera, 2018[60]	Yes	HS-FOCUS total scores with HUI3 dimensions	Yes for speech, dexterity and cognition	No		

Measure	Study reference	Correlation examined	Other measures examined for correlation	Significant correlation(s) (0.41 and above)	Regression analysis undertaken	Regression details	Regression analysis shows significant relationship
HUI3	Nixon Speechley, 1999[46]	Yes	CHQ summary component scores with HUI3 utility scores; CHQ domain scores and HUI3 dimensions	Yes for the index; Yes amongst similar dimensions/domains	No		
HUI3	Rhodes, 2012[97]	Yes	HUI3 utility with PedsQL self-report and proxy-report	Yes HUI3 and PedsQL both self-report, HUI3 and PedsQL both proxy report, PedsQL self-report and HUI3 proxy report, but not for HUI3 self-report and PedsQL proxy report	No		
HUI3	Rosenbaum, 2007[58]	Yes	HUI3 utility with Quality of Life Instrument for People With Developmental Disabilities	No	No		
HUI3	Sung, 2003[48]	Yes	HUI3 dimensions and utility with CHQ summary component scores	Dimensions - Yes for pain only, Index - Yes for CHQ physical summary score but not for the psychosocial score	No		
HUI3	Tilford, 2012[59]	Yes	ADOS, Vineland-II Adaptive Behavior Scales, Cognitive functioning with HUI3 utilities	Yes for Vineland-II Adaptive Behavior Scales and cognitive functioning	Yes	OLS regression of HUI3 with ADOS, Vineland-II Adaptive Behavior Scales, Cognitive functioning	Yes for Vineland-II Adaptive Behavior Scales and cognitive functioning
HUI3	Ungar, 2012[85]	Yes	HUI3 utility and dimensions with PedsQL domains	Yes – for utility and amongst similar dimensions/domains	No		

Measure	Study reference	Correlation examined	Other measures examined for correlation	Significant correlation(s) (0.41 and above)	Regression analysis undertaken	Regression details	Regression analysis shows significant relationship
				(stronger for HUI2 than HUI3)			

Notes: ADOS = Autism Diagnostic Observation Schedule; ADQOL=Atopic dermatitis-specific preference-based measure; CBCL=Child Behaviour Checklist; CDRSR=Child Depression Rating Scale-Revised; CHQ=Child Health Questionnaire; CFQ=Cystic Fibrosis Questionnaire; GMFCS = Gross Function Motor Classification System; HOQ = Hydrocephalus Outcome Questionnaire; HS-FOCUS = Hunter Syndrome-Functional Outcomes for Clinical Understanding Scale; PAQLQ=Paediatric Asthma Quality of Life Questionnaire; PedsQL=Paediatric Quality of Life Inventory; POQOLS=Paediatric Oncology Quality of Life Scale; SDQ=Strengths and Difficulties Questionnaire (note here proxy report version was used); SRH=Self-rated health questionnaire; TRF=Teacher's report form, teachers' version of the CBCL; WAItE=Weight-specific Adolescent Instrument for Economic evaluation; Wee-FIM=researcher-reported measure capturing functional independence.

4.6. RESPONSIVENESS

Table 8 presents the results of all studies assessing responsiveness (n=14).

4.6.1. AQoL-6D

No studies assess the responsiveness of AQoL-6D.

4.6.2. CHU9D

Two studies assess the responsiveness of the CHU9D index, with one study finding CHU9D to be significantly responsive when assessed for different severity groups for eczema (categorised using POEM). The other study finds that the measure is not significantly responsive for capturing different dental classifications (though note small study sample size of 87).

4.6.3. EQ-5D-Y

Three studies assess the responsiveness of EQ-5D-Y, with all studies finding EQ-5D-Y is significantly responsive. One study assesses responsiveness for control and intervention groups, one study assesses responsiveness for acutely ill and chronically ill children, and one study assesses responsiveness for patients with improvement in the intervention group. There is some variation in what is assessed, where one study assesses the dimensions, one study assesses an index and one study assesses a composite score (summed score rather than a utility score).

4.6.4. HUI2

Seven studies assess the responsiveness of HUI2, with four studies finding that HUI2 was significantly responsive, two studies not detecting a significant change (though note small sample sizes of seven and 91), and one study does not report whether the findings are significant. One study finds that HUI2 detects change over time for steroid treatment in cancer patients, one study finds HUI2 is significantly responsive to capture being neutropenic to non neutropenic, and one study finds that HUI2 is significantly responsive for patients with improved depression. In addition, one study finds significant responsiveness for children whose health has changed for only a subset of the time points tested (though note small study sample size of 51). Out of the seven studies, four studies assess the index and three assess both the index and the dimensions.

4.6.5. HUI3

Nine studies assess the responsiveness of HUI3, with five studies finding that HUI3 was significantly responsive, three studies not detecting a significant change (though note small sample sizes of 29 and 91 for two of the studies) and one study does not report whether the findings are significant. HUI3 was found to be significantly responsive to steroid treatment in cancer patients, change from neutropenic to non neutropenic, change between pre and post cochlear implantation, and patients with improved depression. In addition, one study finds significant responsiveness for children whose health has changed for only a subset of the time points tested (though note small study sample size of 51). Out of the nine studies, five studies assess the index and four studies assess both the dimensions and the index.

Table 8: Responsiveness (14 studies)

Measure	Study reference	Index or dimensions or both assessed	Comparison e.g. change over time	Analysis details	Comparison in direction consistent with clinical/expected expectation	Responsiveness of measure is statistically significant
CHU9D	Foster Page, 2015[62]	Index	Change over time from baseline to 1 year follow-up	Analysed separately across carious, restored, missing surfaces and highest caries	Yes	No
CHU9D	Sach, 2017[32]	Index	Change over time from baseline to follow-up (time difference not reported)	Across different POEM (Patient Oriented Eczema Measure) groups	Yes	Yes
EQ-5D-Y	Mayoral, 2017[33]	Index	Change over time from baseline to 9 month follow-up	Patients with improvement in intervention group	Yes	Yes
EQ-5D-Y	Perez-Sousa, 2018[76]	Dimensions	Change over time from baseline to 6 month post-intervention	Analysed by control and intervention groups	Yes	Yes
EQ-5D-Y	Scott, 2017[79]	Composite score	Change over time from baseline to 3 month follow-up	Acutely ill and chronically ill children	Yes	Yes
HUI2	Banks, 2008[34]	Index	Change over time in 1 week intervals over 4 week period beginning on 3rd day of new chemotherapy cycle	Patients rated as improved using a Global Parental Rating of Change using proxy-report responses	Yes	No but sample size was 4 to 9
HUI2	Barr, 1997[35]	Index and reported (and significant) for mobility, emotion and pain	Change over time every week for 3 weeks of steroid treatment	All patients (small sample)	Yes	Yes - but unclear whether is observed for parent report or for nurse and clinical report only
HUI2	Boran, 2011	Both	Change over time from being neutropenic to non neutropenic	Patients undergoing chemotherapy were assessed during a neutropenic phase and	Yes	Index - yes; Dimensions - mobility, emotion, self-care

Measure	Study reference	Index or dimensions or both assessed	Comparison e.g. change over time	Analysis details	Comparison in direction consistent with clinical/expected expectation	Responsiveness of measure is statistically significant
				later non-neutropenic phase		
HUI2	Dickerson, 2018[80]	Index	Change over time from baseline to 12 week follow-up	Patients with >20% improvement in CDRS-R (Children's Depression Rating Scale–Revised) score	Yes	Yes
HUI2	Klaassen, 2010a[83]	Index	Changes over time between 4 timepoints	Those whose health changed	Yes	Yes (between timepoints 1 and 2 only)
HUI2	Stevens, 2012b[30]	Index	Change over time (details not reported)	Not reported	Not reported	Not reported
HUI2	Ungar, 2012[85]	Both	Change over time from baseline to follow-up between 3 and 6 months	Children who demonstrated clinical change between visits	Yes	No
HUI3	Banks, 2008[34]	Index	Change over time in 1 week intervals over 4 week period beginning on 3rd day of new chemotherapy cycle	Patients rated as improved using a Global Parental Rating of Change using proxy-report responses	Yes	No but sample size was 4 to 9
HUI3	Barr, 1997[35]	Index and reported (and significant) for mobility, emotion and pain	Change over time every week for 3 weeks of steroid treatment	All patients (small sample)	Yes	Yes - but unclear whether is observed for parent report or for nurse and clinical report only
HUI3	Boran, 2011[37]	Both	Change over time from being neutropenic to non neutropenic	Patients undergoing chemotherapy were assessed during a neutropenic phase and later non-neutropenic phase	Yes	Index - yes for HUI2 and HUI3; Dimensions - HUI2 mobility, emotion, self-care and HUI3 emotion
HUI3	Cheng, 2000[87]	Both	Change between pre and post cochlear implantation		Yes	Yes

Measure	Study reference	Index or dimensions or both assessed	Comparison e.g. change over time	Analysis details	Comparison in direction consistent with clinical/expected expectation	Responsiveness of measure is statistically significant
HUI3	Dickerson, 2018[80]	Index	Change over time from baseline to 12 week follow-up	Patients with >20% improvement in CDRS-R (Children's Depression Rating Scale–Revised) score	Yes	Yes
HUI3	Feeny, 2004[38]	Index	Change over time from baseline, 3, 6, 12 and 18 months, QALYs from baseline to follow-up	Compared for two different interventions, and difference in QALYs assessed	Yes	No (EQ-VAS found significant difference across interventions)
HUI3	Klaassen, 2010a[83]	Index	Change over time between 4 timepoints	Those whose health changed	Yes	Yes (between timepoints 1 and 2 only)
HUI3	Stevens 2012b	Index	Change over time (details not reported)	Not reported	Not reported	Not reported
HUI3	Ungar, 2012[85]	Both	Change over time from baseline to follow-up between 3 and 6 months	Children who demonstrated clinical change between visits	Yes	No

4.7. RELIABILITY

Table 9 presents the results of all studies assessing reliability (n=24).

4.7.1. AQoL-6D

No studies assess the reliability of AQoL-6D.

4.7.2. CHU9D

One study[51] assesses the reliability of CHU9D examining test-retest reliability. This study finds that arguably reliability is not achieved since test-retest administration of the measure in the morning and the afternoon generated only fair to moderate agreement in the dimension responses (note this is the same study that also did not find test-retest reliability for EQ-5D-Y, though found CHU9D had higher reliability than EQ-5D-Y).

Two studies assess reliability using internal consistency[62, 64], but as this is something not typically examined in preference-based measures, since we would not typically expect the dimensions to be internally consistent, this has not been reported here.

4.7.3. EQ-5D-Y

Eight studies assess reliability of EQ-5D-Y, where six studies assess test-retest reliability, one study assesses inter-rater reliability, and one study assesses agreement between online and paper versions. Out of the studies assessing test-retest reliability, four studies find test-retest reliability, one study finds test-retest reliability with the exception of the usual activities and pain/discomfort dimension, and one study does not find test-retest reliability (note this is the same study that also does not find test-retest reliability for CHU9D[51]). The study assessing inter-rater child/parent-proxy reliability[76] does not find evidence of reliability, as though there is moderate or fair agreement for some dimensions only poor agreement is found for the other dimensions. The study assessing agreement between online and paper versions[77] finds acceptable agreement for the different versions, though agreement was lowest for the worried/sad/unhappy dimension.

4.7.4. HUI2

Eight studies assess reliability of HUI2, with two studies assessing test-retest reliability and seven studies assessing inter-rater reliability. Both studies assessing test-retest reliability find that the HUI2 is reliable. Four studies find inter-rater child/parent-proxy reliability, whereas three studies do not find evidence of inter-rater child/parent-proxy reliability (note that two had small sample sizes of 19 and 91).

4.7.5. HUI3

Fourteen studies assess reliability of HUI3, with all testing inter-rater reliability and one study also testing test-retest reliability and one study also testing inter-modality agreement. Eight of the studies found evidence of inter-rater reliability whereas 6 studies did not find evidence of inter-rater reliability in particular for the dimensions of cognition, emotion and pain. There was no evidence of test-retest reliability or inter-modality agreement with concerns raised around the impact of modality on reporting of cognition, emotion and pain.

Table 9: Reliability (24 studies)

Measure	Study reference	Analysis	Reliability observed	Relevant information
CHU9D	Canaway, 2013[51]	Test-retest; Pupils completed the measures in the morning and in the afternoon	No	Dimensions assessed, fair to moderate agreement. CHU9D higher reliability than EQ-5D-Y
EQ-5D-Y	Canaway, 2013[51]	Test-retest; Pupils completed the measures in the morning and in the afternoon	No	Dimensions assessed, fair to moderate agreement. CHU9D higher reliability than EQ-5D-Y
EQ-5D-Y	Hsu, 2018[73]	Test-retest; 6 months later	Yes	Fair to almost perfect agreement in dimensions of mobility, usual activities, looking after myself, but slight to no agreement for pain/discomfort and worried/sad/unhappy
EQ-5D-Y	Perez-Sousa, 2018[76]	Inter-rater; Child/parent proxy	No	At baseline moderate agreement for mobility and worried/sad/unhappy, poor agreement for other dimensions. After 6 months of intervention fair agreement for pain/discomfort and worried/sad/unhappy dimensions, poor agreement for other dimensions
EQ-5D-Y	Ravens-Sieberer, 2010[21]	Test-retest; Third of Italy and Spain samples administered 7-10 days later	Yes	Significant or identical agreement in dimensions, only exception Italy sample for mobility
EQ-5D-Y	Robles, 2015[77]	Agreement between online and paper versions	Yes	Acceptable agreement, lowest for worried/sad/unhappy
EQ-5D-Y	Scalone, 2011[78]	Test-retest; Sub-sample completed measure again 10 days later	Yes	Significant agreement for all dimensions, though Bland-Altman plot below standard threshold for repeatability
EQ-5D-Y	Scott, 2017[79]	Test-retest; Sub-sample completed measure again 24 hours later	Yes - but not for usual activities or pain/discomfort	Significant agreement with exception of usual activities with poor agreement and pain/discomfort with fair agreement
EQ-5D-Y	Wong, 2019[56]	Test-retest; Measure administered 2-3 weeks after baseline	Yes	Good agreement, except for sad/unhappy in 3L version
HUI-2	Glaser, 1999[81]	Inter-rater; Child/parent proxy/physician/physiotherapist	Yes	Fair, moderate to high agreement, of both dimensions and utility scores, with exception of child/physician where agreement is poor
HUI2	Klaassen, 2010b[40]	Inter-rater reliability at 4 time points; child/parent	Yes	Significant agreement at T1 and T3 but not T2 and T4

Measure	Study reference	Analysis	Reliability observed	Relevant information
HUI2	Kulpeng, 2013[41]	Inter-rater reliability; child/parent	Yes	Significant difference in utility for patients/parents in hearing loss, no other significant differences
HUI2	Morrow, 2012[45]	Inter-rater reliability; child/parent	No	No significant inter-rater reliability for any dimension, moderate reliability for sensation, mobility, pain (could not be assessed for self-care as high proportion of responses in a single severity level)
HUI2	Stevens, 2012b[30]	Inter-rater; Child/proxy	Yes	No details reported
HUI2	Sung, 2004[49]	Inter-rater reliability; child/parent	No	
HUI2	Trudel, 1998[84]	Test-retest; Children off treatment completed measure 2-4 weeks later	Yes	Significant or identical agreement for each dimension and utility score
HUI2	Ungar, 2012[85]	Inter-rater child/parent proxy when undertaken independently; Test-retest from baseline to follow-up for children who remained stable	No for inter-rater reliability (though high agreement for child and dyad report); Yes for test-retest reliability	Inter-rater: no significant agreement (note high agreement between child and dyad report). Test-retest: significant agreement for utility score
HUI3	Belfort, 2011[36]	Inter-rater reliability; child/parent	Yes	Inter-reliability for HUI3 index but not for all dimensions (not reported)
HUI3	Janse, 2008[91]	Inter-rater reliability; child/parent	No	Very good agreement for hearing, good agreement for vision, moderate agreement for dexterity and poor agreement for speech, ambulation, emotion, cognition, pain
HUI3	Klaassen, 2010b[40]	Inter-rater reliability at 4 time points; child/parent	Yes	Significant agreement at T1 and T3 but not T2 and T4
HUI3	Kulpeng, 2013[41]	Inter-rater reliability; child/parent	Yes	Significant difference in utility for patients/parents in hearing loss, no other significant differences
HUI3	Le Gales, 1999[42]	Inter-rater reliability; child/parent	No	High agreement between raters for hearing, vision, speech, ambulation, dexterity, but low agreement for emotion, cognition and pain
HUI3	Lee, 2011[94]	Inter-rater reliability child/parent; test-retest reliability for control group	Yes	Yes inter-rater reliability; Moderate test-retest reliability
HUI3	Morrow, 2012[45]	Inter-rater reliability; child/parent	No	No significant inter-rater reliability for any dimension, moderate reliability for vision, ambulation and pain (could not be assessed for hearing and speech as high proportion of responses in a single severity level)

Measure	Study reference	Analysis	Reliability observed	Relevant information
HUI3	Penn, 2011[96]	Inter-rater reliability; child/parent	Yes	Patients - Correlation between parent-report and self-report for HUI3 attributes of vision, hearing, speech, ambulation, dexterity, and cognition was good, moderate for emotion, poor for pain. Controls - moderate inter-rater reliability
HUI3	Rhodes, 2012[97]	Inter-rater reliability; child/parent	Yes	No significant difference in index scores, significant correlation in index scores, significant differences for pain dimension only
HUI3	Stade, 2006[100]	Inter-rater reliability; child/parent	Yes	Yes inter-rater reliability
HUI3	Stevens 2012b	Inter-rater; Child/proxy	Yes	No details reported
HUI3	Sung, 2004[49]	Inter-rater reliability; child/parent	No	Differences observed but non-significant (potentially due to low sample size)
HUI3	Ungar, 2012[85]	Inter-rater reliability child/parent when undertaken independently; Test-retest from baseline to follow-up for children who remained stable	No	Inter-rater: no significant agreement (note high agreement between child and dyad report). Test-retest: no significant agreement (note significant for HUI2)
HUI3	Verrips, 2001[102]	Inter-rater reliability child/parent; inter-modality reliability	No	Inter-rater reliability high for vision, hearing, ambulation, dexterity, moderate for speech, low for cognition, emotion, pain; reliability by mode of administration - face-to-face interview/telephone interview/postal survey - high reliability for vision, hearing, speech, ambulation, dexterity, low for cognition, emotion, pain - with more psychological dysfunction reported in interviews

4.8. ACCEPTABILITY AND FEASIBILITY

Table 10 presents the results of all studies assessing accessibility and feasibility (n=17).

4.8.1. AQoL-6D

No studies assess the acceptability and feasibility of AQoL-6D.

4.8.2. CHU9D

One study assesses the acceptability and feasibility of CHU9D, using missing data, time to complete the measure and interviewer ratings of respondent understanding. The study finds that the CHU9D is acceptable and feasible.

4.8.3. EQ-5D-Y

Nine studies assess the acceptability and feasibility of EQ-5D-Y, where eight studies assess missing data, one study assesses whether assistance is required to complete the measure, one study assesses whether respondents agreed to complete the measure, one study uses therapist feedback and one study uses cognitive interviews. All except one study find that the EQ-5D-Y is acceptable and feasible, where the other study has 10.2% missing EQ-5D-Y data (note that this is the same study that finds that HUI2 is not acceptable and feasible, and that also finds lower missing data for EQ-5D-Y in comparison to HUI2 and verbal clarification is required for some respondents for HUI2 but not for EQ-5D-Y[53]).

4.8.4. HUI2

Seven studies assess the acceptability and feasibility of HUI2, where three studies assess missing data, one study assesses whether assistance is required to complete the measure, one study assesses time to complete the measure, one study assesses completion rates, one study assesses difficulty to understand and complete, and one study assesses the acceptability and consistency of the Chinese translation. Four studies find that the HUI2 is acceptable and feasible, however one study has 26.5% missing HUI2 data, and verbal clarification is required for some respondents (note that this is the same study that found that EQ-5D-Y was not acceptable and feasible[53]), one had completion rates varying from 72% to 85%, and one study did not report their findings.

4.8.1. HUI3

Nine studies assess the acceptability and feasibility of HUI3, where two studies assess missing data, two studies assess ceiling effects, one study assesses completion rates, one study assesses difficulty to understand and complete, one study assesses the acceptability and consistency of the Chinese translation, one assesses time to complete, and one assesses ease of completion. Six studies finding evidence of acceptability and feasibility, one study finds ceiling effects in osteonecrosis secondary to treatment of developmental dysplasia of the hip, one study had completion rates varying from 72% to 85%, and one study did not report their findings.

Table 10: Acceptability and feasibility (19 studies)

Measure	Study reference	Analysis	Acceptability and feasibility observed	Issues raised, where relevant
CHU9D	Canaway, 2013[51]	Missing data; time to complete; interviewer ratings of respondent understanding	Yes	Interviewers rated 7.1% of children as having poor/very poor understanding
EQ-5D-Y	Canaway, 2013[51]	Missing data; time to complete; interviewer ratings of respondent understanding	Yes	Interviewers rated 7.1% of children as having poor/very poor understanding
EQ-5D-Y	Jelsma, 2010[74]	Missing data	Yes	
EQ-5D-Y	Kim, 2018[61]	Missing data	Yes	
EQ-5D-Y	Oluboyede, 2013[53]	Missing data; whether assistance required to complete	No	10.2% missing
EQ-5D-Y	Ravens-Sieberer, 2010[21]	Missing data	Yes	
EQ-5D-Y	Robles, 2015[77]	Missing data	Yes	
EQ-5D-Y	Scalone, 2011[78]	Whether respondents agreed to self-complete	Yes	
EQ-5D-Y	Scott, 2017[79]	Missing data; therapist feedback	Yes	Some 8-9 year-olds had difficulty understanding usual activities dimension
EQ-5D-Y	Wille, 2010[22]	Missing data; cognitive interviews	Yes	
HUI2	Furlong, 2012[39]	Completion rates	No	Completion rates varied among treatment phases from 72% to 85% at baseline. Missing assessment rate varied from 16% to 62% for the 2 year post treatment point.
HUI2	Glaser, 1999[81]	Missing data	Yes	
HUI2	Le Gales, 1999[42]	Difficulty in understanding and competing	Yes	Most of the children (86.5%) and the parents (97.4%) said they had no difficulty in understanding the questionnaire; 81.1% of the children and 89.7% of the parents said they had no difficulty in answering the questions (both HUI2 and HUI3)

Measure	Study reference	Analysis	Acceptability and feasibility observed	Issues raised, where relevant
HUI2	Mok, 2014[44]	Acceptability and consistency of translation	Not reported	Testing on 5 healthy Chinese adults was also conducted to ensure consistency between English and Chinese version of HUI
HUI2	Oluboyede, 2013[53]	Missing data; whether assistance required to complete	No	26.5% missing; verbal clarification required for approximately 10 respondents
HUI2	Stevens, 2012b[30]	Missing data; time to complete	Yes	
HUI2	Sung, 2003[48]	Ease of completion	Yes	
HUI3	Barr, 1997[35]	Independent completion	Yes	No children under 8 could complete the measures independently
HUI3	Furlong, 2012[39]	Completion rates	No	Completion rates varied among treatment phases from 72% to 85% at baseline. Missing assessment rate varied from 16% to 62% for the 2 year post treatment point.
HUI3	Le Gales, 1999[42]	Difficulty in understanding and competing	Yes	Most of the children (86.5%) and the parents (97.4%) said they had no difficulty in understanding the questionnaire; 81.1% of the children and 89.7% of the parents said they had no difficulty in answering the questions (both HUI2 and HUI3)
HUI3	Lee, 2011[94]	Missing data	Yes	Missing data - 3% for self-report, 6% for parent-report
HUI3	Mok, 2014[44]	Acceptability and consistency of translation	Not reported	Testing on 5 healthy Chinese adults was also conducted to ensure consistency between English and Chinese version of HUI
HUI3	Roposch, 2011[98]	Ceiling effects	No	51% sample in full health, varied from 67% (pain) to 100% (hearing, speech, dexterity) for each dimension
HUI3	Smith-Olinde, 2008[99]	Ceiling effects	Yes	Examined ceiling effects, where vision, ambulation, dexterity, emotion and pain suffered from ceiling effects with over 80% at highest level, but impact was observed as expected in hearing and speech
HUI3	Stevens 2012b	Missing data; time to complete	Yes	
HUI3	Sung, 2003[48]	Ease of completion	Yes	

4.9. OTHER PSYCHOMETRIC ANALYSES

Other psychometric analyses are undertaken in some studies, but these typically involve plots where the same finding is reported statistically in the sections above. For example, studies often report Bland-Altman plots alongside tests of convergent validity, to assess agreement between different measures. Some studies plot the distribution of utility values in the sample using different value sets for the CHU9D[54, 68], or distribution of responses across the different countries included in the sample[21]. One study[22] undertakes cognitive interviews to assess comprehensibility, possible misinterpretations, and acceptance of EQ-5D-Y. The study finds some general difficulties interpreting 'looking after myself' but as the item was understood by the majority of respondents it was left unchanged. One study[84] assesses the content validity of HUI2 using a literature review, expert opinion and informal discussions with parents, finding that the HUI2 dimensions are adequate for children with cancer.

4.10. RESULTS SUMMARY

Table 11 summarises the results of all analyses. The number of entries reflect the number of studies where each psychometric property is assessed. EQ-5D-Y has the largest amount of evidence of good psychometric performance in proportion to the number of studies that have examined its psychometric performance (note this is for the dimensions). The CHU9D is assessed in fewer studies, but the majority of studies find evidence of good psychometric performance. The evidence for HUI2 and HUI3 are more mixed, and for AQoL-6D the evidence is based on only one study.

Table 11: Summary of psychometric performance by measure and utility index (i.e. country value set)

	Dimensions or utility index i.e. country value set	Known group validity	Convergent validity	Responsiveness	Inter-rater reliability	Test-retest reliability	Inter-modality reliability	Acceptability and feasibility
AQoL-6D	Australian adolescent utilities	✓						
	Australian adult utilities	✓						
CHU9D	Dimensions	✓✓×	✓✓✓✓✓✓×			×		✓×
	Australian adolescent utilities	✓✓✓✓	✓✓					
	Australian adult utilities	✓✓	✓					
	UK utilities	✓✓✓✓✓✓±	✓✓✓✓✓✓	✓×				
EQ-5D-Y	Dimensions	✓✓✓✓✓±±×	✓✓✓✓✓✓✓±	✓✓	×	✓✓✓✓±×	✓	✓✓✓✓✓✓✓✓
	UK EQ-5D utilities	✓	✓					
	Australian EQ-5D utilities	✓						
	French EQ-5D utilities	✓						
	Spanish EQ-5D utilities	±	✓	✓				
	US EQ-5D-Y utilities							
HUI2	Dimensions	±±✓×	✓✓✓✓	✓✓×	✓××	✓		✓✓✓✓××
	Canadian utilities	✓✓✓✓✓✓✓✓±××	✓✓✓✓✓✓✓✓±×	✓✓✓✓××	✓✓✓××	✓✓		
	UK utilities	±	✓✓					
HUI3	Dimensions	✓✓✓✓✓✓±±×××××	✓✓✓×	✓✓×	✓✓±×××××		×	
	Canadian utilities	✓✓✓✓✓✓✓✓✓✓✓✓ ±×××××	✓✓✓✓✓✓✓✓✓✓±±× ×	✓✓✓±×××	✓✓✓✓×××	✓×		✓✓✓✓✓✓××

Notes: ✓ Evidence demonstrating significant performance × Property is examined but no significant evidence is found ± Evidence is mixed or inconclusive evidence found. Each symbol represents the findings of one study assessing that psychometric property. Where studies assess multiple psychometric properties a symbol is recorded for each psychometric property assessed.

5. DISCUSSION

The review has outlined the evidence around the psychometric performance of the child and adolescent-specific measures of AQoL-6D, CHU9D, EQ-5D-Y, HUI2 and HUI3. Overall the published evidence is limited, since there are few studies comparing measures, studies with small sample sizes that may not be powered to detect statistical significance, and only a relatively small number of studies within the same condition. Relatively few studies use UK value sets to generate utility values. There is both a limited number and heterogeneity of published studies, as the evidence is based on a relatively small number of studies across a range of countries, a range of different populations and conditions, using different study designs, different languages, different value sets and many different statistical techniques. The wide variation in studies makes it difficult to synthesise the evidence to generate a consistent picture of the overall performance of each measure. In particular, evidence is limited assessing responsiveness, with only fourteen studies assessing responsiveness. There is a concern raised across all measures around their reliability. Only HUI2 performs strongly for test-retest reliability. None of the measures perform strongly for inter-rater reliability between child self-report and parent proxy-report (though AQoL-6D and CHU9D are not assessed). The findings suggest that there is reason for concern around the comparability of self-report and proxy responses to measure HRQOL of children and adolescents.

For CHU9D the review found evidence of known-group validity and convergent validity, mixed evidence of responsiveness and acceptability and feasibility, but the only study assessing test-retest reliability did not find evidence of reliability. For EQ-5D-Y the review found evidence for its dimensions of known group validity, convergent validity, responsiveness, test-retest reliability, acceptability and feasibility, but the only study assessing inter-rater reliability did not find evidence of reliability. There is no evidence available around the psychometric performance of potential UK utility values since there is no UK value set, nor any official value set for any country, for the EQ-5D-Y. For HUI2 the review found evidence of test-retest reliability and mixed evidence of known-group validity, convergent validity, responsiveness, inter-rater reliability, acceptability and feasibility, as good performance was not found unanimously across these aspects of psychometric performance. For HUI3 the review found mixed

evidence of known-group validity, convergent validity, responsiveness, inter-rater reliability, test-retest reliability and acceptability and feasibility, with a proportion of studies not demonstrating evidence of known group validity, responsiveness or reliability. Only one study assessed the psychometric performance of AQoL-6D.

There is a large amount of evidence of good performance for the EQ-5D-Y dimensions, however good psychometric performance is not reported unanimously in all studies assessing the measure, and there is more mixed evidence around reliability. More studies assess the psychometric performance of HUI3 than the other measures, but the evidence of HUI3 is more mixed. This means that for HUI3 there are a larger number of studies finding evidence of good psychometric performance, but the proportion of studies who do not find evidence of good psychometric performance is larger than for the other measures. HUI2 is also assessed in a large number of studies, though the performance is mixed. In contrast, EQ-5D-Y and CHU9D are assessed in fewer studies but the proportion of studies that find evidence of good psychometric performance is larger.

In particular, there are two studies that have some findings contrary to other studies. One study[51] assesses test-retest reliability of CHU9D and EQ-5D-Y, finding no evidence for either measure. As this is the only study assessing reliability of CHU9D this can lead to a larger perceived impact that the CHU9D is not reliable, yet the finding should be validated by other studies, in particular since the lack of evidence of test-retest reliability of EQ-5D-Y is contrary to some of the other studies assessing test-retest reliability of EQ-5D-Y where evidence of reliability is found (though consistent with others). The study administered the measures in the morning and afternoon of the same day. The authors of the study stated that there were no clear directional changes between the morning and afternoon responses, and suggested further research to better understand this finding, for example using a think-aloud study [51].

Another study[53] found that neither EQ-5D-Y nor HUI2 were acceptable and feasible as they had high levels of missing data, but this is contrary to most other studies assessing acceptability and feasibility for these measures. This suggests that the higher levels of missing data may have been study specific (note also the small study sample size of 49).

For EQ-5D-Y there is no official value set, and the good psychometric performance that is observed is based mainly on the performance on the dimensions. Whilst it could be anticipated that a UK utility index would have the same psychometric performance, this can only be confirmed through data analyses. The value set may not have sufficiently large differences in utility decrements for different severity levels of each dimension.

Few studies assessed measures within the same clinical area. However, even where there were multiple studies within a clinical area the evidence is limited. For example, three studies assessed the performance of measures in patients with asthma, where two assessed EQ-5D-Y [31, 70] and one assessed HUI2[85]. EQ-5D-Y was found to have known-group validity and convergent validity, with no assessment of responsiveness, reliability, acceptability or feasibility. HUI2 was found to have convergent validity, responsiveness, test-retest reliability, but the study assessed and found no evidence for known-group validity or inter-rater reliability. On the basis of these findings it is difficult to recommend usage of either measure over the other, since for EQ-5D-Y there is limited evidence available but the evidence that is available suggests good performance, whereas for HUI2 there is wider evidence available but the evidence is mixed. Equally, whilst the evidence is mixed it is difficult to determine whether known-group validity would be expected since the sample size was 91. Differences in samples may also potentially impact on results. Six studies assessed the performance of measures for overweight and obese people or obesity prevention programmes, though two studies involved a general population sample[50, 65] and four studies involved patient samples[36, 63, 76] [101]. For these studies assessing weight, there was not evidence of good psychometric performance for HUI2 and HUI3, though there was evidence of good psychometric performance for EQ-5D-Y and CHU9D (though this was not unanimous for CHU9D).

Some studies had small sample sizes, with 28 out of the 76 studies having a sample size below 100. Sample size has not been used to assess the studies, but it should be taken into consideration that some studies may not have found significant evidence of the psychometric performance due to the sample size, meaning that the result may not be indicative of the performance of the measure. In particular for HUI2 and HUI3

this may have impacted on the results, where for HUI2 15 of 26 studies assessing performance had sample sizes below 100 and for HUI3 18 of 42 studies had sample sizes below 100. In the literature there are no clear guidelines or accepted practice around how to generate sample sizes for studies assessing psychometric performance of patient-reported outcome measures[104], nor to our knowledge preference-based measures.

Appropriateness of the statistical analyses undertaken was not assessed, though the data were extracted according to the authors of this reports viewpoints of what was regarded as assessments for each of the psychometric properties, not whether the study claimed the psychometric properties were assessed (for some studies this differed).

Methodological limitations of the review include missing studies of child and adolescent preference-based measures in mixed adolescent and adult populations due to the paediatrics filter applied in the database search. It is also possible that some relevant studies were incorrectly excluded at the title and abstract sift stage as each citation was sifted by one reviewer and there may have been reviewer error. Statistical mapping analyses have not been included in the review since mapping assessments are undertaken to generate predictions rather than assess association *per se*, though it is recognised that mapping analyses can provide some evidence of associations between measures.

The review has also not extracted the comparative performance of adult measures where these are also used, but convergent validity using correlations and/or regression analyses has been extracted and reported where this has been undertaken. Comparisons of EQ-5D and EQ-5D-Y were beyond the remit of this review, though there are published studies available where both measures are administered to the same people at the same time. Studies that administered one or more measures and summarised their results were not included in the review unless they assessed psychometric properties. Therefore it is possible that there are clinical studies that may not have been captured in our search of the literature that report whether the child and adolescent-specific preference-based measures found a

statistically significant change over time or difference across treatments if they have not reported that they have assessed responsiveness or known-group validity.

6. CONCLUSIONS

The review of published evidence on the psychometric performance of a selection of child and adolescent-specific generic preference-based measures has found that the evidence is limited.

From the current evidence, EQ-5D-Y has the largest amount of evidence of good psychometric performance in proportion to the number of studies that have examined its psychometric performance. The majority of the evidence related to EQ-5D-Y is based on dimensions. The CHU9D is assessed in fewer studies, but the majority of studies find evidence of good psychometric performance. There are a larger number of studies assessing the psychometric performance of HUI2 and HUI3, but the evidence of good psychometric performance is more mixed, with a larger proportion not finding evidence of good psychometric performance. However, for HUI2 and HUI3 the studies are more limited in their sample sizes and statistical power and this is likely to have impacted on their performance. For AQoL-6D the evidence is based on only one study. The review is informative in indicating patient populations where the psychometric performance of one or measures has been assessed, and providing an overview of the evidence found.

6.1. SUGGESTED POINTS FOR CONSIDERATION BY NICE

The review has highlighted that there is limited published evidence around the psychometric performance of EQ-5D-Y, CHU9D, HUI2, HUI3 and AQoL-6D. The evidence is further limited in particular for NICE in that:

1) the AQoL-6D and EQ-5D-Y studies do not involve use of a UK value set, since there are no UK value sets currently available;

2) Only eight CHU9D studies use the UK value set;

3) Only two HUI2 studies use the UK value set.

Different value sets can have different psychometric properties, and drawing conclusions about the performance of an instrument based on the classification system alone may be misleading.

The following points are suggested for consideration:

- Given the paucity of evidence comparing measures, and the limitations relating much of the evidence that does exist, NICE must consider whether it is appropriate to recommend a specific instrument at this time.
- This review does not cover all available child and adolescent-specific generic preference-based measures, as the following also are potential candidates for use: AHUM; QWB; 16D; 17D. However, the review included the currently available measures the authors consider as most appropriate for use to inform UK policy using criteria around: intended and worded appropriately for use in children and adolescents; applicability across conditions using a generic classification system; development (or validation) with an English-speaking population; potential availability and feasibility of inclusion in datasets used to inform UK policy.
- Overall given the evidence available examining the psychometric performance of EQ-5D-Y, CHU9D, HUI2, HUI3 and AQoL-6D, the EQ-5D-Y has the largest amount of evidence of good psychometric performance in proportion to the number of studies that have examined its psychometric performance, followed by CHU9D. Any choice of measure for recommendation for use to inform policy would require additional considerations including but not limited to: content validity of the dimensions and severity levels in the measure; the appropriateness of the methods used to generate the value set; projected usage in trials and other relevant studies used to inform health technology assessment; relationship to adult EQ-5D since models often require utility values into adulthood.
- Though a large number of conditions are assessed in studies included in the review, not all conditions are assessed and many are only assessed in one study. New evidence may be needed to demonstrate the performance of a measure when it is applied in a patient population where it has not previously been validated.

6.2. RECOMMENDATIONS FOR FUTURE RESEARCH

The following are potential research questions that would be informative around the psychometric performance of the main generic child and adolescent-specific preference-based measures:

- What is the comparative psychometric performance of the main generic child and adolescent-specific preference-based measures, when administered to the same patients? Answering this research question could involve:
    - Primary data collection of the main child and adolescent-specific preference-based measures of interest administered to patients, preferably with a range of conditions across different ICD classifications. This would enable psychometric analyses to be undertaken across different measures using the same sample and applying the same statistical methods. In particular data collection could focus upon reliability where the evidence is mixed for EQ-5D-Y and limited for CHU9D. In addition, data collection could be linked to an intervention, and/or clinical measures, to determine responsiveness.
    - Accessing existing datasets of one or more of the main child and adolescent-specific preference-based measures of interest administered to patients to conduct independent analyses on these datasets, particularly where some of these datasets may not have had psychometric analyses published.
- Do the main generic child and adolescent-specific preference-based measures have content validity of dimensions and severity levels across the age range of respondents that they are recommended for?
- What is the impact of using self-report EQ-5D-Y versus proxy-report EQ-5D? Since many economic evaluations in children and adolescents use adult EQ-5D values in their economic model, this would be informative around the impact of using child and adolescent EQ-5D-Y over adult EQ-5D. This could include a review of studies comparing both the results and psychometric performance of EQ-5D and EQ-5D-Y. This could be extended to other adult preference-based measures and/or other child and adolescent preference-based measures (for example CHU9D).
- When, and at what ages, should self-report and proxy-report administrations of a measure be used to generate utility values to inform the economic model?
- Do any new UK value sets have good psychometric performance (note that CHU9D and EQ-5D-Y are expected to have new value sets in the next few years)? This could be assessed using either new or existing datasets.

- Does new evidence around the psychometric performance of the main child and adolescent-specific preference-based measures confirm the findings of this review? This could involve regular annual updates to the excel spreadsheet associated with the review that summarises all studies assessing the psychometric performance of selected child and adolescent preference-based measures (for example EQ-5D-Y and CHU9D).
- Do the findings of the review differ if a quality assessment is undertaken of the studies included in the review that assess psychometric performance of the main child and adolescent-specific preference-based measures?

7. REFERENCES

1. Finch, A., C. Mukuria, and J. Brazier, *What is the evidence for the performance of generic preference-based measures? A systematic overview of reviews.* European Journal of Health Economics, 2018. **19**(4): p. 557-570.
2. Davis, S. and A. Wailoo, *A review of the psychometric performance of the EQ-5D in people with urinary incontinence.* Health & Quality of Life Outcomes, 2013. **11**(20).
3. Yang, Y., J. Brazier, and L. Longworth, *EQ5D in skin conditions: an assessment of validity and responsiveness.* European Journal of Health Economics, 2014. **16**(9): p. 927-939.
4. Papaioannou, D., J. Brazier, and G. Parry, *How valid and responsive are generic health status measures, such as EQ-5D and SF-36, in schizophrenia? A systematic review.* Value in Health, 2011. **14**(6): p. 907-920.
5. Yang, Y., L. Longworth, and J. Brazier, *An assessment of validity and responsiveness of generic measures of health-related quality of life in hearing impairment.* Quality of Life Research, 2013. **22**: p. 2813-28.
6. Chen, G.R., J., *A Review of the Development and Application of Generic Multi-Attribute Utility Instruments for Paediatric Populations.* Pharmacoeconomics, 2015. **33**(10): p. 1013-28.
7. Kwon, J., et al., *A Systematic Review and Meta-analysis of Childhood Health Utilities.* Medical Decision Making, 2018. **38**(3): p. 277-305.
8. Rowen, D., et al., *Review of valuation methods of preference-based measures of health for economic evaluation in child and adolescent populations: Where are we now and where are we going?* Pharmacoeconomics, 2020: p. Forthcoming.
9. National Institute for Health Clinical Excellence, *Guide to the methods of technology appraisal 2013.* NICE, 2013.
10. Hill JH, R.D., Pennington, Wong R, Wailoo A, *A review of the methods used to estimate and model utility values in nice technology appraisals for paediatric populations. DSU Report.* 2019.
11. Richardson, J., et al., *Measurement of the Quality of Life for Economic Evaluation and the Assessment of Quality of Life (AQoL) Mark 2 Instrument.* Australian Economic Review, 2004. **37**(1): p. 62-88.
12. Moodie, M.R., 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 in Health, 2010. **13**(8): p. 1014-27.
13. Stevens, K., *Valuation of the Child Health Utility 9D Index.* Pharmacoeconomics, 2012. **30**(8): p. 729-47.
14. Ratcliffe, J.C., L.; Flynn, T.; Sawyer, M.; Stevens, K.; Brazier, J.; Burgess, L., *Valuing Child Health Utility 9D health states with a young adolescent sample: a feasibility study to compare best-worst scaling discrete-choice experiment, standard gamble and time trade-off methods.* Applied Health Economics & Health Policy, 2011. **9**(1): p. 15-27.
15. Ratcliffe, J.F., T.; Terlich, F.; Stevens, K.; Brazier, J.; Sawyer, M., *Developing adolescent-specific health state values for economic evaluation: an application of profile case best-worst scaling to the Child Health Utility 9D.* Pharmacoeconomics, 2012. **30**(8): p. 713-27.

16. Ratcliffe, J.C., G.; Stevens, K.; Bradley, S.; Couzner, L.; Brazier, J.; Sawyer, M.; Roberts, R.; Huynh, E.; Flynn, T., *Valuing Child Health Utility 9D Health States with Young Adults: Insights from a Time Trade Off Study.* Applied Health Economics & Health Policy, 2015. **13**(5): p. 485-92.
17. Ratcliffe, J.H., E.; Chen, G.; Stevens, K.; Swait, J.; Brazier, J.; Sawyer, M.; Roberts, R.; Flynn, T., *Valuing the Child Health Utility 9D: Using profile case best worst scaling methods to develop a new adolescent specific scoring algorithm.* Social Science & Medicine, 2016. **157**: p. 48-59.
18. Rowen, D., et al., *Estimating a Dutch Value Set for the Pediatric Preference-Based CHU9D Using a Discrete Choice Experiment with Duration.* Value in Health, 2018. **21**(10): p. 1234-1242.
19. Chen, G., et al., *Scoring the Child Health Utility 9D instrument: estimation of a Chinese child and adolescent-specific tariff.* Quality of Life Research, 2019. **28**(1): p. 163-176.
20. Devlin, N.J. and R. Brooks, *EQ-5D and the EuroQol Group: Past, Present and Future.* Applied Health Economics & Health Policy, 2017. **15**(2): p. 127-137.
21. Ravens-Sieberer, U.W., N.; Badia, X.; Bonsel, G.; Burstrom, K.; Cavrini, G.; Devlin, N.; Egmar, A. C.; Gusi, N.; Herdman, M.; Jelsma, J.; Kind, P.; Olivares, P. R.; Scalone, L.; Greiner, W., *Feasibility, reliability, and validity of the EQ-5D-Y: results from a multinational study.* Quality of Life Research, 2010. **19**(6): p. 887-97.
22. Wille, N.B., X.; Bonsel, G.; Burstrom, K.; Cavrini, G.; Devlin, N.; Egmar, A. C.; Greiner, W.; Gusi, N.; Herdman, M.; Jelsma, J.; Kind, P.; Scalone, L.; Ravens-Sieberer, U., *Development of the EQ-5D-Y: a child-friendly version of the EQ-5D.* Quality of Life Research, 2010. **19**(6): p. 875-86.
23. Craig, B.M.G., W.; Brown, D. S.; Reeve, B. B., *Valuation of Child Health-Related Quality of Life in the United States.* Health Economics, 2016. **25**(6): p. 768-77.
24. Kind, P., et al., *Can adult weights be used to value child health states? Testing the influence of perspective in valuing EQ-5D-Y.* Quality of Life Research, 2015. **24**(10): p. 2519-39.
25. Kreimeier, S., et al., *Valuation of EuroQol Five-Dimensional Questionnaire, Youth Version (EQ-5D-Y) and EuroQol Five-Dimensional Questionnaire, Three-Level Version (EQ-5D-3L) Health States: The Impact of Wording and Perspective.* Value in Health, 2018. **21**(11): p. 1291-1298.
26. Torrance, G.W., et al., *Multiattribute utility function for a comprehensive health status classification system. Health Utilities Index Mark 2.* Medical Care, 1996. **34**(7): p. 702-22.
27. McCabe C, S.K., Roberts J, Brazier J, *Health state values for the HUI 2 descriptive system: results from a UK survey.* Health economics, 2005; 14:231-244, 2005.
28. Feeny, D., et al., *Multiattribute and single-attribute utility functions for the health utilities index mark 3 system.* Medical Care, 2002. **40**(2): p. 113-28.
29. Longworth, L., et al., *Use of generic and condition-specific measures of health-related quality of life in NICE decision-making: A systematic review, statistical modelling and survey.* Health Technology Assessment, 2014. **18**: p. 1-224.
30. Stevens, K.J.F., J. V., *An assessment of the psychometric performance of the Health Utilities Index 2 and 3 in children following discharge from a U.K. pediatric intensive care unit.* Pediatric Critical Care Medicine, 2012. **13**(4): p. 387-92.

31. Hernandez, G., et al., *Validity of the EQ-5D-Y in a cohort of asthmatic children in Europe.* Quality of Life Research, 2015. **24**: p. 174-174.
32. Sach, T.H.M., E.; Thomas, K.; Montgomery, A.; Harrison, E.; Williams, H.; Clothes Trial Team, *The comparative performance of the CHU-9D and the ADQoL amongst children aged 5 years ir more with eczema: evidence from the CLOTHES randomised controlled trial.* Value in Health, 2017. **20**(9): p. A406-A406.
33. Mayoral, K., et al., *Validity of the EQ-5D-Y in children and adolescents with diabetes.* Quality of Life Research, 2017. **26**: p. 71.
34. Banks, B.A., N.J. Barrowman, and R. Klaassen, *Health-related quality of life: changes in children undergoing chemotherapy.* Journal of Pediatric Hematology/Oncology, 2008. **30**(4): p. 292-7.
35. Barr, R., et al., *Health-related quality of life during post-induction chemotherapy in children with acute lymphoblastic leukemia in remission.* International Journal of Oncology, 1997. **11**(2): p. 333-9.
36. Belfort, M.B., et al., *Health state preferences associated with weight status in children and adolescents.* BMC Pediatrics, 2011. **11**: p. 12.
37. Boran, P., et al., *Translation and cultural adaptation of health utilities index with application to pediatric oncology patients during neutropenia and recovery in Turkey.* Pediatric Blood & Cancer, 2011. **56**(5): p. 812-7.
38. Feeny, D., et al., *Comparing directly measured standard gamble scores to HUI2 and HUI3 utility scores: group- and individual-level comparisons.* Social Science & Medicine, 2004. **58**(4): p. 799-809.
39. Furlong, W., et al., *Health-related quality of life among children with acute lymphoblastic leukemia.* Pediatric Blood & Cancer, 2012. **59**(4): p. 717-24.
40. Klaassen, R.J.B., Ronald D.; Hughes, Joanna; Rogers, Paul; Anderson, Ronald; Grundy, Paul; Ali, S.; Yanofsky, Rochelle; Abla, Oussama; Silva, Mariana; Carret, Anne-Sophie; Cappelli, Mario, *Nurses provide valuable proxy assessment of the health-related quality of life of children with Hodgkin Disease.* Cancer, 2010. **116**(6): p. 1602-1607.
41. Kulpeng, W., et al., *Variation of health-related quality of life assessed by caregivers and patients affected by severe childhood infections.* BMC Pediatrics, 2013. **13**: p. 122.
42. Le Gales, 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.* International Journal of Cancer - Supplement, 1999. **12**: p. 112-8.
43. Lynch, F.L., et al., *Measuring Health-related Quality of Life in Teens With and Without Depression.* Medical Care, 2016. **54**(12): p. 1089-1097.
44. Mok, W.K., et al., *Validation and application of health utilities index in Chinese subjects with down syndrome.* Health & Quality of Life Outcomes, 2014. **12**: p. 144.
45. Morrow, A.M., et al., *A comparison of doctors', parents' and children's reports of health states and health-related quality of life in children with chronic conditions.* Child: Care, Health & Development, 2012. **38**(2): p. 186-95.
46. Nixon Speechley, K.M., E.; Desmeules, M.; Schanzer, D.; Landgraf, J. M.; Feeny, D. H.; Barrera, M. E., *Mutual concurrent validity of the child health questionnaire and the health utilities index: an exploratory analysis using survivors of childhood cancer.* International Journal of Cancer - Supplement, 1999. **12**: p. 95-105.

47. Petrou, S., et al., *The association between neurodevelopmental disability and economic outcomes during mid-childhood.* Child: Care, Health & Development, 2013. **39**(3): p. 345-57.
48. Sung, L.G., M. L.; Doyle, J. J.; Young, N. L.; Ingber, S.; Rubenstein, J.; Wong, J.; Samanta, T.; McLimont, M.; Feldman, B. M., *Construct validation of the Health Utilities Index and the Child Health Questionnaire in children undergoing cancer chemotherapy.* British Journal of Cancer, 2003. **88**(8): p. 1185-90.
49. Sung, L., et al., *Health-related quality of life (HRQL) scores reported from parents and their children with chronic illness differed depending on utility elicitation method.* Journal of Clinical Epidemiology, 2004. **57**(11): p. 1161-6.
50. Trevino, R.P., T.H. Pham, and S.L. Edelstein, *Obesity and preference-weighted quality of life of ethnically diverse middle school children: the HEALTHY study.* Journal of Obesity, 2013. **2013**: p. 206074.
51. Canaway, A.G.F., E. J., *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.* Quality of Life Research, 2013. **22**(1): p. 173-83.
52. Chen, G.F., T.; Stevens, K.; Brazier, J.; Huynh, E.; Sawyer, M.; Roberts, R.; Ratcliffe, J., *Assessing the Health-Related Quality of Life of Australian Adolescents: An Empirical Comparison of the Child Health Utility 9D and EQ-5D-Y Instruments.* Value in Health, 2015. **18**(4): p. 432-8.
53. Oluboyede, Y.T., S.; McCabe, C., *Measuring health outcomes of adolescents: report from a pilot study.* European Journal of Health Economics, 2013. **14**(1): p. 11-9.
54. Ratcliffe, J.S., K.; Flynn, T.; Brazier, J.; Sawyer, M. G., *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): p. 730-6.
55. Ratcliffe, J.S., K.; Flynn, T.; Brazier, J.; Sawyer, M., *An assessment of the construct validity of the CHU9D in the Australian adolescent general population.* Quality of Life Research, 2012. **21**(4): p. 717-25.
56. Wong, C.K.H., et al., *A head-to-head comparison of five-level (EQ-5D-5L-Y) and three-level EQ-5D-Y questionnaires in paediatric patients.* European Journal of Health Economics, 2019. **02**: p. 02.
57. Livingston, M.H. and P.L. Rosenbaum, *Adolescents with cerebral palsy: stability in measurement of quality of life and health-related quality of life over 1 year.* Developmental Medicine & Child Neurology, 2008. **50**(9): p. 696-701.
58. Rosenbaum, P.L., et al., *Quality of life and health-related quality of life of adolescents with cerebral palsy.* Developmental Medicine & Child Neurology, 2007. **49**(7): p. 516-21.
59. Tilford, J.M., et al., *Preference-based health-related quality-of-life outcomes in children with autism spectrum disorders: a comparison of generic instruments.* Pharmacoeconomics, 2012. **30**(8): p. 661-79.
60. Mattera, M., et al., *Validation of the shortened Hunter Syndrome-Functional Outcomes for Clinical Understanding Scale (HS-FOCUS).* Health and Quality of Life Outcomes, 2018. **16**.
61. Kim, S.K., M.W. Jo, and S.H. Kim, *Health-related quality of life by allergy symptoms in elementary school students.* Health & Quality of Life Outcomes, 2018. **16**(1): p. 93.

62. Foster Page, L.A.B., D. M.; Cameron, C. M.; Thomson, W. M., *Can the Child Health Utility 9D measure be useful in oral health research?* International Journal of Paediatric Dentistry, 2015. **25**(5): p. 349-57.
63. Frew, E.J.P., M.; Lancashire, E.; Hemming, K.; Adab, P.; Waves Study co-investigators, *Is utility-based quality of life associated with overweight in children? Evidence from the UK WAVES randomised controlled study.* BMC Pediatrics, 2015. **15**: p. 211.
64. Furber, G.S., 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 & Quality of Life Outcomes, 2015. **13**: p. 22.
65. Oluboyede, Y. and R. Tomos, *Measuring Weight-Specific Quality of Life in Adolescents: An Examination of the Concurrent Validity and Test-Retest Reliability of the WAItE.* Value in Health, 2019. **22**(3): p. 348-354.
66. Petersen, K.D.C., G.; Mpundu-Kaambwa, C.; Stevens, K.; Brazier, J.; Ratcliffe, J., *Measuring Health-Related Quality of Life in Adolescent Populations: An Empirical Comparison of the CHU9D and the PedsQLTM 4.0 Short Form 15.* The Patient: Patient-Centered Outcomes Research, 2018. **11**(1): p. 29-37.
67. Stevens, K.R., 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 in Health, 2012. **15**(8): p. 1092-9.
68. Xu, F., et al., *Measuring and Valuing Health-Related Quality of Life among Children and Adolescents in Mainland China - A Pilot Study.* Plos One, 2014. **9**(2).
69. Astrom, M., et al., *Population health status based on the EQ-5D-Y-3L among adolescents in Sweden: Results by sociodemographic factors and self-reported comorbidity.* Quality of Life Research, 2018. **27**(11): p. 2859-2871.
70. Bergfors, S.A., M.; Burstrom, K.; Egmar, A. C., *Measuring health-related quality of life with the EQ-5D-Y instrument in children and adolescents with asthma.* Acta Paediatrica, 2015. **104**(2): p. 167-73.
71. Burstrom, K.B., A.; Brostrom, E. W.; Sun, S.; Egmar, A. C., *EQ-5D-Y as a health-related quality of life measure in children and adolescents with functional disability in Sweden: testing feasibility and validity.* Acta Paediatrica, 2014. **103**(4): p. 426-35.
72. Eidt-Koch, D.M., T.; Greiner, W., *Cross-sectional validity of the EQ-5D-Y as a generic health outcome instrument in children and adolescents with cystic fibrosis in Germany.* BMC Pediatrics, 2009. **9**: p. 55.
73. Hsu, C.N.L., H. W.; Pickard, A. S.; 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.* International Journal for Quality in Health Care, 2018. **13**: p. 13.
74. Jelsma, J., *A comparison of the performance of the EQ-5D and the EQ-5D-Y health-related quality of life instruments in South African children.* International Journal of Rehabilitation Research, 2010. **33**(2): p. 172-7.
75. Loof, E., et al., *Neurodevelopmental difficulties negatively affect health-related quality of life in children with idiopathic clubfoot.* Acta Paediatrica, 2018. **27**: p. 27.
76. Perez-Sousa, M.A., et al., *Does anthropometric and fitness parameters mediate the effect of exercise on the HRQoL of overweight and obese children/adolescents?* Quality of Life Research, 2018. **27**(9): p. 2305-2312.

77. Robles, N.R., L.; Rodriguez-Arjona, D.; Azuara, M.; Codina, F.; Raat, H.; Ravens-Sieberer, U.; Herdman, M., *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 & Quality of Life Outcomes, 2015. **13**: p. 72.
78. Scalone, L., et al., *Assessing quality of life in children and adolescents: Development and validation of the Italian version of the EQ-5D-Y*. Vol. 8. 2011. 331-341.
79. Scott, D.F., G. D.; Jelsma, J., *The use of the EQ-5D-Y health related quality of life outcome measure in children in the Western Cape, South Africa: psychometric properties, feasibility and usefulness - a longitudinal, analytical study.* Health & Quality of Life Outcomes, 2017. **15**(1): p. 12.
80. Dickerson, J.F.F., D. H.; Clarke, G. N.; MacMillan, A. L.; Lynch, F. L., *Evidence on the longitudinal construct validity of major generic and utility measures of health-related quality of life in teens with depression.* Quality of Life Research, 2018. **27**(2): p. 447-454.
81. Glaser, A.W.F., W.; Walker, D. A.; Fielding, K.; Davies, K.; Feeny, D. H.; Barr, R. D., *Applicability of the Health Utilities Index to a population of childhood survivors of central nervous system tumours in the U.K.* European Journal of Cancer, 1999. **35**(2): p. 256-61.
82. Kennedy, C.R.L., K., *Comparison of screening instruments for disability and emotional/behavioral disorders with a generic measure of health-related quality of life in survivors of childhood brain tumors.* International Journal of Cancer - Supplement, 1999. **12**: p. 106-11.
83. Klaassen, R.J.K., M.; Gaboury, I.; Hughes, J.; Anderson, R.; Grundy, P.; Ali, S. K.; Jardine, L.; Abla, O.; Silva, M.; Barnard, D.; Cappelli, M., *Evaluating the ability to detect change of health-related quality of life in children with Hodgkin disease.* Cancer, 2010. **116**(6): p. 1608-14.
84. Trudel, J.G.R., M.; Dobkin, P. L.; Leclerc, J. M.; Robaey, P., *Psychometric properties of the Health Utilities Index Mark 2 system in paediatric oncology patients.* Quality of Life Research, 1998. **7**(5): p. 421-32.
85. Ungar, W.J.B., K.; Dell, S.; Feldman, B. M.; Marshall, D.; Willan, A.; Wright, J. G., *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): p. 697-712.
86. Boulton, M., et al., *Health-related quality of life of children with vision impairment or blindness.* Developmental Medicine & Child Neurology, 2006. **48**(8): p. 656-61.
87. Cheng, A.K., et al., *Cost-utility analysis of the cochlear implant in children.* JAMA, 2000. **284**(7): p. 850-6.
88. de Sonneville-Koedoot, C., et al., *Health-related quality of life of preschool children who stutter.* Journal of Fluency Disorders, 2014. **42**: p. 1-12.
89. de Sonneville-Koedoot, C., et al., *Economic evaluation of stuttering treatment in preschool children: The RESTART-study.* Journal of Communication Disorders, 2015. **58**: p. 106-18.
90. Francis, A., et al., *Quality of life of children and adolescents with chronic kidney disease: a cross-sectional study.* Archives of Disease in Childhood, 2019. **104**(2): p. 134-140.
91. Janse, A.J., et al., *Quality of life in chronic illness: children, parents and paediatricians have different, but stable perceptions.* Acta Paediatrica, 2008. **97**(8): p. 1118-24.

92. Kennes, J., et al., *Health status of school-aged children with cerebral palsy: information from a population-based sample.* Developmental Medicine & Child Neurology, 2002. **44**(4): p. 240-7.
93. Kulkarni, A.V., et al., *Quality of life after endoscopic third ventriculostomy and cerebrospinal fluid shunting: an adjusted multivariable analysis in a large cohort.* Journal of Neurosurgery. Pediatrics., 2010. **6**(1): p. 11-6.
94. Lee, J.M., et al., *Health utilities for children and adults with type 1 diabetes.* Medical Care, 2011. **49**(10): p. 924-31.
95. Lovett, R.E., et al., *Bilateral or unilateral cochlear implantation for deaf children: an observational study.* Archives of Disease in Childhood, 2010. **95**(2): p. 107-12.
96. Penn, A., et al., *A detailed prospective longitudinal assessment of health status in children with brain tumors in the first year after diagnosis.* Journal of Pediatric Hematology/Oncology, 2011. **33**(8): p. 592-9.
97. Rhodes, E.T., et al., *Health-related quality of life in adolescents with or at risk for type 2 diabetes mellitus.* Journal of Pediatrics, 2012. **160**(6): p. 911-7.
98. Roposch, A., et al., *Functional outcomes in children with osteonecrosis secondary to treatment of developmental dysplasia of the hip.* Journal of Bone & Joint Surgery - American Volume, 2011. **93**(24): p. e145.
99. Smith-Olinde, L., et al., *Health state preference scores for children with permanent childhood hearing loss: a comparative analysis of the QWB and HUI3.* Quality of Life Research, 2008. **17**(6): p. 943-53.
100. Stade, B.C., et al., *Health-related quality of life of Canadian children and youth prenatally exposed to alcohol.* Health & Quality of Life Outcomes, 2006. **4**: p. 81.
101. Tan, E.J., et al., *Is there an association between early weight status and utility-based health-related quality of life in young children?* Quality of Life Research, 2018. **27**(11): p. 2851-2858.
102. Verrips, G.H., et al., *Measuring health status using the Health Utilities Index: agreement between raters and between modalities of administration.* Journal of Clinical Epidemiology, 2001. **54**(5): p. 475-81.
103. Wolke, D., et al., *Self and parent perspectives on health-related quality of life of adolescents born very preterm.* Journal of Pediatrics, 2013. **163**(4): p. 1020-6.e2.
104. Anthoine, E., et al., *Sample size used to validate a scale: a review of publications on newly-developed patient reported outcomes measures.* Health & Quality of Life Outcomes, 2014. **12**(176).
105. Allen, J.I., K. J.; Lewin, T. J.; Attia, J. R.; Kelly, B. J., *Construct validity of the Assessment of Quality of Life - 6D (AQoL-6D) in community samples.* Health & Quality of Life Outcomes, 2013. **11**: p. 61.
106. Barr, R.D., et al., *Health-related quality of life in survivors of tumours of the central nervous system in childhood--a preference-based approach to measurement in a cross-sectional study.* European Journal of Cancer, 1999. **35**(2): p. 248-55.
107. Buysse, C.M., et al., *Long-term health status in childhood survivors of meningococcal septic shock.* Archives of Pediatrics & Adolescent Medicine, 2008. **162**(11): p. 1036-41.
108. Christensen, R., et al., *Change in pain status in children with cerebral palsy.* Developmental Medicine & Child Neurology, 2017. **59**(4): p. 374-379.

109. Cox, C.L., et al., *Proxy assessment of quality of life in pediatric clinical trials: application of the Health Utilities Index 3.* Quality of Life Research, 2005. **14**(4): p. 1045-56.
110. Fu, L., et al., *Measurement of health-related quality of life in survivors of cancer in childhood in Central America: feasibility, reliability, and validity.* Journal of Pediatric Hematology/Oncology, 2006. **28**(6): p. 331-41.
111. Furlong, W.B., R. D.; Feeny, D.; Yandow, S., *Patient-focused measures of functional health status and health-related quality of life in pediatric orthopedics: a case study in measurement selection.* Health & Quality of Life Outcomes, 2005. **3**: p. 3.
112. Gomersall, T., et al., *Measuring quality of life in children with speech and language difficulties: a systematic review of existing approaches.* International Journal of Language & Communication Disorders, 2015. **50**(4): p. 416-35.
113. Hinds, P.S., et al., *The Health Utilities Index 3 invalidated when completed by nurses for pediatric oncology patients.* Cancer Nursing, 2007. **30**(3): p. 169-77.
114. Hoey, H.M., H. M.; Fitzgerald, M.; Mortensen, H. B.; Hougaard, P.; Lynggaard, H.; Skovlund, S. E.; Aanstoot, H. J.; Chiarelli, F.; Daneman, D.; Danne, T.; Dorchy, H.; Garandeau, P.; Greene, S.; Holl, R.; Kaprio, E.; Kocova, M.; Martul, P.; Matsuura, N.; Robertson, K.; Schoenle, E.; Sovik, O.; Swift, P.; Tsou, R. M.; Vanelli, M.; Aman, J.; Hvidore Study Group on Childhood, Diabetes, *Parent and health professional perspectives in the management of adolescents with diabetes: development of assessment instruments for international studies.* Quality of Life Research, 2006. **15**(6): p. 1033-42.
115. Horsman, J.R., et al., *Disability and health-related quality of life in long-term survivors of cancer in childhood in Brazil: An assessment of the construct validity of the health utilities index (HUI3).* Value in Health, 2008. **11**(3): p. A75-A76.
116. Janse, A.J., et al., *Quality of life in chronic illness: perceptions of parents and paediatricians.* Archives of Disease in Childhood, 2005. **90**(5): p. 486-91.
117. Klaassen, R.J., et al., *Validation and reliability of a disease-specific quality of life measure (the TranQol) in adults and children with thalassaemia major.* British Journal of Haematology, 2014. **164**(3): p. 431-7.
118. Mpundu-Kaambwa, C., et al., *A review of preference-based measures for the assessment of quality of life in children and adolescents with cerebral palsy.* Quality of Life Research, 2018. **27**(7): p. 1781-1799.
119. Otto, C., 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.* Quality of Life Research, 2018. **27**: p. S61-S61.
120. Petersson, C., et al., *Comparing children's self-report instruments for health-related quality of life using the International Classification of Functioning, Disability and Health for Children and Youth (ICF-CY).* Health and quality of life outcomes, 2013. **11**: p. 75-75.
121. Petrou, S. and E. Kupek, *Estimating preference-based health utilities index mark 3 utility scores for childhood conditions in England and Scotland.* Medical Decision Making, 2009. **29**(3): p. 291-303.
122. Richardson, J.I., A.; Peacock, S.; Sinha, K.; Khan, M.; Misajon, R.; Keeffe, J., *Utility weights for the vision-related Assessment of Quality of Life (AQoL)-7D instrument.* Ophthalmic Epidemiology, 2012. **19**(3): p. 172-82.

123. Richardson, J.R.P., S. J.; Hawthorne, G.; Iezzi, A.; Elsworth, G.; Day, N. A., *Construction of the descriptive system for the Assessment of Quality of Life AQoL-6D utility instrument.* Health & Quality of Life Outcomes, 2012. **10**: p. 38.
124. Richardson, J.I., A.; Khan, M. A.; Maxwell, A., *Validity and reliability of the Assessment of Quality of Life (AQoL)-8D multi-attribute utility instrument.* The Patient: Patient-Centered Outcomes Research, 2014. **7**(1): p. 85-96.
125. Redouane, B.C., E.; Stephens, D.; Keilty, K.; Mouzaki, M.; Narayanan, U.; Moraes, T.; Amin, R., *Parental Perceptions of Quality of Life in Children on Long-Term Ventilation at Home as Compared to Enterostomy Tubes.* PLoS ONE [Electronic Resource], 2016. **11**(2): p. e0149999.
126. Roncada, C., et al., *Specific instruments to assess quality of life in children and adolescents with asthma.* Jornal de Pediatria, 2013. **89**(3): p. 217-225.
127. Schiariti, V.F., Nora; Cieza, Alarcos; Klassen, Anne; O'Donnell, Maureen, *Content comparison of health-related quality of life measures for cerebral palsy based on the International Classification of Functioning.* Disability and Rehabilitation: An International, Multidisciplinary Journal, 2011. **33**(15-16): p. 1330-1339.
128. 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.* Applied Health Economics and Health Policy, 2011. **9**(3): p. 157-169.
129. Tonmukayakul, U., et al., *A systematic review of utility values in children with cerebral palsy.* Quality of Life Research, 2019. **28**(1): p. 1-12.

APPENDIX

A.1 RETRIEVED ARTICLES EXCLUDED UPON DETAILED EXAMINATION

Table A1: Retrieved articles that were excluded upon detailed examination (n=26)

Allen 2013[105]	Wrong population
Barr 1999[106]	Limited/ no useable data
Buysse 2008[107]	Wrong population
Christensen,2017[108]	Limited/ no useable data
Cox 2005[109]	Limited/ no useable data
Fu 2006[110]	Wrong population
Furlong 2005[111]	Limited/ no useable data
Gomersall 2015[112]	Limited/ no useable data
Hinds 2007[113]	Limited/ no useable data
Hoey 2006[114]	Wrong measure
Horsman 2008[115]	Wrong population
Janse 2005[116]	Limited/ no useable data
Klaassen ,2014[117]	Wrong population
Mpundu-Kaambwa 2018[118]	Limited/ no useable data
Otto 2018[119]	Limited/ no useable data
Petersson 2013[120]	Limited/ no useable data
Petrou 2009[121]	Limited/ no useable data
Ratcliffe 2012c[15]	Limited/ no useable data
Richardson 2012a[122]	Wrong population/measure
Richardson 2012b[123]	Wrong population/measure
Richardson 2014[124]	Wrong population/measure
Redouane 2016 [125]	Limited/ no useable data
Roncada 2013[126]	Limited/ no useable data
Schiariti 2011[127]	Limited/ no useable data
Stevens 2011[128]	Limited/ no useable data
Tonmukayakul 2019[129]	Limited/ no useable data