University Library

A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Xiong, Xiuqin

Title:
Advancing the Measurement and Valuation of Health-Related Quality of Life in Australian children

Date:
2024-03

Persistent Link:
https://hdl.handle.net/11343/350431

Terms and Conditions:
Terms and Conditions: Copyright in works deposited in Minerva Access is retained by the copyright owner. The work may not be altered without permission from the copyright owner. Readers may only download, print and save electronic copies of whole works for their own personal non-commercial use. Any use that exceeds these limits requires permission from the copyright owner. Attribution is essential when quoting or paraphrasing from these works.

Advancing the Measurement and Valuation of Health-Related Quality of Life in Australian children

Xiuqin Xiong

(ORCID: 0000-0002-1323-4223)

Submitted in total fulfillment for the degree of

Doctor of Philosophy

March 2024

Melbourne Health Economics

Centre for Health Policy

Melbourne School of Population and Global Health

Faculty of Medicine, Dentistry and Health Science

The University of Melbourne

Abstract

In healthcare practice, the diverse array of treatments available often target a wide range of health issues, posing a challenge for decision-makers tasked with comparing their respective benefits. Health-related Quality of Life (HRQoL) offers a solution by providing a unified metric that captures the overall impact of health conditions on an individual's life, irrespective of the type of health conditions or the interventions applied. The use of HRQoL enables the measurement and comparison of the health benefits associated with different treatments, facilitating more informed and efficient resource allocation. Particularly in the realm of child healthcare services, where young individuals are significant users, the need for robust HRQoL measurement instruments is crucial. However, the existing instruments for assessing HRQoL in children are limited in their suitable age range and validation and valuation evidence, especially for younger age groups. This deficiency impedes their effective application in economic evaluations within pediatric healthcare practices.

This PhD thesis endeavors to bridge this critical gap through six original studies focusing on application and improving the measurement and valuation of the HRQoL instruments for Australian children. By addressing these limitations, this research aims to enhance the quality and reliability of HRQoL assessment in pediatric healthcare. Specifically, it seeks to extend HRQoL measurement to encompass young children aged 2-4 years old. Ultimately, these efforts aim to improve decision-making processes and resource allocation in this vital domain.

Declaration

This is to certify that:

i. The thesis comprises only my original work towards the Doctor of Philosophy (PhD), except where indicated in the acknowledgements.

ii. Due acknowledgement has been made in the text to all other material used.

iii. The thesis is fewer than 100, 000 words in length, exclusive of tables, bibliographies and appendices.

Xiuqin Xiong

Preface

I gratefully acknowledge the guidance and work of many others in the five published papers, and one finished study contained in this thesis. All of the six studies were conducted during my PhD candidature. I was the primary author for each paper/manuscript, leading the writing of the manuscripts, from initial drafting to the final revisions. I contributed 90% of the statistical analyses and over 60% of the planning, study design and interpretation of results.

First author publications & manuscripts arising from this thesis:

- Xiong, X., Dalziel, K., Carvalho, N., Xu, R., & Huang, L. (2022). Association between 24-hour movement behaviors and health-related quality of life in children. Quality of life research: an international journal of quality of life aspects of treatment, care and rehabilitation, 31(1), 231–240. https://doi.org/10.1007/s11136-021-02901-6

- Xiong, X., Dalziel, K., Huang, L., & Rivero-Arias, O. (2023). Test-Retest Reliability of EQ-5D-Y-3L Best-Worst Scaling Choices of Adolescents and Adults. Value in health: the journal of the International Society for Pharmacoeconomics and Outcomes Research, 26(1), 50–54. https://doi.org/10.1016/j.jval.2022.07.007

- Xiong, X., Dalziel, K., Huang, L., Mulhern, B., & Carvalho, N. (2023). How do common conditions impact health-related quality of life for children? Providing guidance for validating pediatric preference-based measures. Health and quality of life outcomes, 21(1), 8. https://doi.org/10.1186/s12955-023-02091-4

- Xiong, X., Huang, L., Herd, D. W., Borland, M. L., Davidson, A., Hearps, S., Mackay, M. T., Lee, K. J., Dalziel, S. R., Dalziel, K., Cheek, J. A., & Babl, F. E. (2023). Cost-effectiveness of Prednisolone to Treat Bell Palsy in Children: An Economic Evaluation Alongside a Randomized Controlled Trial. Neurology, 100(24), e2432–e2441. https://doi.org/10.1212/WNL.0000000000207284

- Xiong X, Carvalho N, Huang L, Chen G, Jones R, Devlin N, Mulhern B, Dalziel K. Psychometric Properties of Child Health Utility 9D (CHU9D) Proxy Version Administered to Parents and Caregivers of Children Aged 2-4 Years Compared with Pediatric Quality of Life Inventory™ (PedsQL). Pharmacoeconomics. 2024 Jan 27. https://doi.org/10.1007/s40273-024-01355-1

- Suggested reference (manuscript finished): Xiong X, Dalziel K, Huang L, Carvalho N, Devlin N. Valuing the Child Health Utility 9D (CHU9D) for children under 5 years in Australia.

Other co-authored papers during PhD

- Devine, A., Xiong, X., Gottlieb, S. L., de Mello, M. B., Fairley, C. K., & Ong, J. J. (2022). Health-related quality of life in individuals with genital herpes: a systematic review. Health and quality of life outcomes, 20(1), 25. https://doi.org/10.1186/s12955-022-01934-w

- Jones, R., O'Loughlin, R., Xiong, X., Bahrampour, M., Devlin, N., Hiscock, H., Chen, G., Mulhern, B., Dalziel, K., & Quality of Life in Kids: Key Evidence to Strengthen Decisions in Australia (QUOKKA) Project Team (2023). Comparative Psychometric Performance of Common Generic Paediatric Health-Related Quality of Life Instrument Descriptive Systems: Results from the Australian Paediatric Multi-Instrument Comparison Study. PharmacoEconomics, 10.1007/s40273-023-01330-2. Advance online publication. https://doi.org/10.1007/s40273-023-01330-2

- Jones, R., O'Loughlin, R., Xiong, X., Bahrampour, M., McGregor, K., Yip, S., Devlin, N., Hiscock, H., Mulhern, B., Dalziel, K., & On Behalf Of The Quality Of Life In Kids Key Evidence To Strengthen Decisions In Australia Quokka Project Team (2023). Collecting Paediatric Health-Related Quality of Life Data: Assessing the Feasibility and Acceptability of the Australian Paediatric Multi-Instrument Comparison (P-MIC) Study. Children (Basel, Switzerland), 10(10), 1604. https://doi.org/10.3390/children10101604

- Renee Jones, Brendan Mulhern, Nancy Devlin, Harriet Hiscock, Gang Chen, Rachel O'Louglin, Xiuqin Xiong, Mina Bahrampour, Kristy McGregor, Shilana Yip, and Kim Dalziel on behalf of the Quality Of Life in Kids: Key evidence to strengthen decisions in Australia (QUOKKA) project team. Australian Paediatric Multi-Instrument Comparison (P-MIC) Study: Technical Methods Paper [Online]

Conferences and speaking engagements:

- Xiong, X., Dalziel, K., Carvalho, N., Xu, R., & Huang, L. Association between 24-hour movement behaviors and health-related quality of life in children. Australian Health Economics Doctoral (AHED) Workshop, oral presentation, September 16, 2020

- Xiong, X., Dalziel, K., Huang, L., Mulhern, B., & Carvalho, N. How do common conditions impact health-related quality of life for children? Providing guidance for validating pediatric preference-based measures. Virtual ISPOR 2021, poster presentation, May 17-20, 2021

- Xiong, X., Dalziel, K., Huang, L., & Rivero-Arias, O. Test-Retest Reliability of EQ-5D-Y-3L Best-Worst Scaling Choices of Adolescents and Adults. International health economics association (iHEA) congress, oral presentation, July 15, 2021

- Xiong, X., Dalziel, K., Huang, L., & Rivero-Arias, O. Test-Retest Reliability of EQ-5D-Y-3L Best-Worst Scaling Choices of Adolescents and Adults. Australian Health Economics Society (AHES) annual conference, oral presentation, September 22, 2021

- Xiong, X., Huang, L., Herd, D. W., Borland, M. L., Davidson, A., Hearps, S., Mackay, M. T., Lee, K. J., Dalziel, S. R., Dalziel, K., Cheek, J. A., & Babl, F. E. Cost-effectiveness of Prednisolone to Treat Bell Palsy in Children: An Economic Evaluation Alongside a Randomized Controlled Trial. Health Technology Assessment International (HTAi) Annual meeting, Adelaide, June 24-28th, 2023

Translation of research and other contributions

- Xiuqin Xiong, Dr Li Huang, Dr Natalie Carvalho and Professor Kim Dalziel. Physical activity is critical for children's quality of life. Pursuit. 11 August 2021. https://pursuit.unimelb.edu.au/articles/physical-activity-is-critical-for-children-s-quality-of-life

Source of funding

I am a grateful recipient of the China Scholarship Council - University of Melbourne PhD Scholarship.

Acknowledgements

I wish to express my deepest gratitude to my primary supervisor, Professor Kim Dalziel, whose unwavering professionalism and support have been instrumental throughout my journey. Being her student has been an honor, and I am immensely thankful for her supervision, guidance, and provision of resources without which I would not have achieved so much. Professor Dalziel's expertise, commitment to excellence, meticulous review of my work, and insightful comments and suggestions have greatly contributed to my academic growth. Her belief in my abilities has instilled in me the confidence and independence necessary for future research endeavors. Her constructive feedback, encouragement, and mentorship have been invaluable in overcoming challenges and reaching significant milestones.

I extend my heartfelt appreciation to my co-supervisors, Associate Professor Natalie Carvalho and Dr. Li Huang, whose combined expertise complemented Professor Dalziel's guidance perfectly. Dr. Huang's support in statistical analysis, Stata code, and structural coherence of my work, coupled with Associate Professor Carvalho's assistance in language refinement, literature review, referencing, and study implications, have been indispensable. Their kindness and mentorship have enriched my PhD experience, and I am deeply grateful for their guidance and support.

Special thanks are also due to Associate Professor Helen Jordan, who chaired my advisory committee with dedication and ensured my continuous support throughout my candidature. I am also indebted to Professor Nancy Devlin, whose valuable advice and feedback significantly contributed to the development of my PhD projects.

My heartfelt thanks go to my peers and colleagues for their camaraderie, intellectual exchange, and unwavering support throughout this journey.

Finally, I extend my deepest appreciation to my family members, friends, and mentors, who provided unwavering support during moments of frustration and fatigue. I am especially indebted to my partner, Rongbin Xu, whose steadfast encouragement and support were crucial in navigating the challenges of this arduous journey.

The completion of this thesis would not have been possible without the collective support, guidance, and encouragement of all those mentioned above, for which I am profoundly grateful.

Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1. Background

1.1.1. Importance of economic evaluation

Health expenditure is increasing in Australia and globally. In 2020-21, there was an estimated $220.9 billion spending on health goods and services in Australia, which equated to an average of approximately $8,617 per person and comprised 10.7% of overall economic activity.[1] Australia's health expenditure increased consistently faster than that of the Organization for Economic Cooperation and Development (OECD) median.[2] On the other hand, resources are always scarce. The budget for healthcare is also finite. New health technologies or services will have to compete with existing technologies or services within the healthcare budget or drive up the healthcare service budget. Healthcare expenditure typically rises with national wealth. However, the rate of growth in healthcare expenditure is faster than the Consumer Price Index (CPI) and Gross Domestic Product (GDP), which stresses the importance of decisions on resource allocation.[3] Decision makers need to make choices and priorities about which treatments or services to fund to achieve the best health overall.

Economic evaluation is such a tool to inform resource allocation and achieve the best value within a limited budget. Economic evaluation is a framework to evaluate allocative efficiency by comparing costs and benefits associated with alternative programs or treatments, aiming to aid the decision-making process.[4] There exist four primary forms of economic evaluation: cost-minimization analysis, cost-effectiveness analysis, cost-utility analysis and cost benefit analysis. In Australia, a cost-utility analysis is preferred over a cost-effectiveness analysis as the Pharmaceutical Benefits Advisory Committee (PBAC) guidelines state, where possible, as cost-utility analysis allows comparison across interventions, disease areas and medicines.[5] This thesis only covers cost-effectiveness and cost-utility analysis as these two are the main types informing resource allocation in Australia according to the Pharmaceutical Benefits Advisory Committee Guidelines.[6] It is beyond the scope of this thesis to discuss the details of the other types of economic evaluation.

Cost-effectiveness analysis compares both the costs and health outcomes of an intervention to another intervention (or the status quo) by estimating how much it costs to gain a unit of a health outcome, like a life year gained or a death prevented. Cost-utility analysis can be deemed as a special type of cost-effectiveness analysis. Cost-utility analysis often uses quality-adjusted life years (QALY) as the effectiveness outcome. QALY is a measure that incorporates the impact on a person's length of life and

the impact on their health-related quality of life (HRQoL) during that period.[7] It has the advantage of being able to compare different diseases and programs as a QALY is a united unit, not specific to a disease or program. The disability-adjusted life year (DALY) is an alternative to the QALY and has been favored in the cost-effectiveness work in developing countries.[8] This thesis focuses on informing priority setting and resource allocation decisions in Australian context and thus doesn't go into details of DALY.

1.1.2. Importance of child health

Children are important health service users. In Australia, reports from the Australian Bureau of Statistics (ABS) National Health Survey 2017-2018 showed that around 43% of children had at least 1 long-term condition, while 20% had 2 or more long-term conditions.[9] The most common chronic conditions were respiratory system diseases and mental and behavioral diseases. For example, about 10% of Australian children had asthma, and about 6% of children had anxiety-related problems.[9]

Disease burden in young children is large. Globally, the majority of child and adolescent death occurred during the first five years of life. An estimated 5.0 million children under 5 years died in 2019 (mostly from preventable and treatable diseases),[10] accounting for over 80% child mortality among children and adolescents under 19 years old.[11]

Providing healthcare services for children is crucial, as early interventions that promote childhood health not only yield immediate benefits but also contribute to long-term advantages in adulthood.[12] Across childhood, the first five years where the effects of risks and plasticity are greatest, is essential to health, human capital and wellbeing across the life course.[13]

The medical cost for children is substantial. The expected global pediatric healthcare market is around USD 15,984 million by 2025,[14] which highlights the need to assess value of new technologies for children by economic evaluation. It is therefore critical to correctly measure health benefits to inform health care resource allocation related to childhood diseases and healthcare services or interventions.

1.2. Child Health Measurement

1.2.1. How to measure child health?

According to the World Health Organization (WHO), "Health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity". Traditional clinical and biomedical metrics define health using natural units like blood pressure or respiratory function, yet they fail to capture the subjective experience for the patient.[8] To estimate the overall influence of health conditions

on individual's life, health-related quality of life (HRQoL) was introduced. The benefits of health care can be measured by the improvement of HRQoL.[15] HRQoL can play a role in facilitating the evaluation of the effectiveness of healthcare programs or treatments in such cases.

Two basic approaches are available for the measurement of HRQoL: generic instruments and specific instruments (such as disease specific, condition or symptom specific and population specific).[16] Condition-specific measures typically contain mainly dimensions of health related to that specific condition and are designed to be sensitive to small changes relevant to that condition. However, they do not allow comparison across different health conditions. In contrast, generic measures are designed to include multiple dimensions of health and to be comparable, however they may be less sensitive than condition-specific measures in corresponding health conditions.[17]

Measures of HRQoL can also be classified as with or without a preference-weighted scoring algorithm (previously being called non-preference-based and preference-based measures). The main difference is how they calculate the total scores. Measures without a preference-weighted scoring algorithm (this type of measure also being called health profiles), such as PedsQL and KIDSCREEN-10, usually have a simple summative scoring system, assigning equal importance to different dimensions (i.e., same weight to each dimension/item).[15] In contrast, measures with a preference-weighted scoring algorithm can convert patient-reported values to an index based on quantified population preferences for different health dimensions, which is called the utility value. This utility value is then used to weight length of life, enabling the calculation of quality-adjusted life-years (QALYs) which is the health outcome used in the cost-utility analysis.[16, 18] HRQoL measures without preference-weighted scoring algorithm can be mapped to those with a preference-weighted scoring algorithm for economic evaluation when necessary.[19]

HRQoL can be applied: 1) to describe, monitor and predict levels of health service or population health using aggregate data; 2) to improve clinical practice in individual patient care; 3) to determine the gains of an intervention in randomized controlled trials or other clinical studies; 4) to inform health care resource allocation decisions.[15]

In summary, to inform health care resource allocation, the measure of HRQoL should be a comprehensive reflection of the impacts upon different aspects of health and wellbeing important to patients, comparable between different health conditions/health care programs, and be scored using community/population preferences to produce utility values.

The following generic HRQoL measures accompanied with preference-weighted scoring algorithm are suitable to measure HRQoL in children for economic evaluations according to recent reviews[20, 21], the

Quality of Well-Being Scale (QWB), the Health Utility Index Mark 2 (HUI2), the HUI3, the Sixteen-dimensional measure of HRQoL (16D), the Seventeen-dimension HRQoL (17D), the Assessment of Quality of Life 6-Dimension (AQoL-6D) Adolescent, the Child Health Utility 9D (CHU9D) and CHU9D with guidance notes for children under 5 years old, the EQ-5D Youth version (EQ-5D-Y) and EQ-5D-Y adapted version for 2-4 years old children [22], the Adolescent Health Utility Measure (AHUM), EuroQol Toddler and Infant Populations[23] (EQ-TIPS, formerly known as TANDI) HRQoL measure, the Health Status Classification System for Pre-School Children (HSCS-PS),[24] Health Utilities Preschool (HuPS) and Infant health-related Quality of life Instrument (IQI).[25, 26] Detailed information of these HRQoL instruments is presented in Table 1-3 below.

It is important that the psychometric performance of these HRQoL measures is evaluated before application. Key properties for consideration include practicality/feasibility, reliability, validity, and responsiveness. Feasibility refers to whether it is feasible or practical to use this measure in real life, including such as completion time and difficulty to complete the questionnaire. Reliability means that a measure can reproduce the results with the minimum amount of random error. It usually includes test-retest reliability (consistency across time) and inter-rater reliability (consistency across different raters). Validity refers to the extent to which an instrument can measure its intended construct/purpose. The commonly used validity measures are content validity (appropriateness and comprehensiveness of the items, response options, and instructions) and construct validity (the extent to which the dimension scores correlate with health indicators or other similar measures). Responsiveness refers to the ability of the instrument to measure meaningful changes in HRQoL.[27, 28]

1.2.2. Challenges and research gaps in measuring child health

The selection of instruments

One challenge is the selection of instruments. According to a recent systematic review, none of the instruments demonstrated superior performance than others across all psychometric properties.[26] There exists considerable variability in the evaluated psychometric properties and the performance of psychometric properties across instruments. A recent Australian pediatric multi-instrument comparison study revealed that both the EQ-5D-Y-5L and CHU9D descriptive systems exhibited satisfactory performance, meeting predetermined criteria across ceiling effects, test-retest reliability, known group validity, convergent and divergent validity, as well as responsiveness.[29] It is suggested that stakeholders involved in instrument selection should set a baseline standard of scientific rigor tailored to the specific research context. For example, the considerations could be (1) which properties are relevant; (2) the relative importance for different properties; (3) the threshold of acceptable performance.

Lack of instruments and psychometric property evidence for preschool children

Most childhood generic HRQoL measures designed to be accompanied with preference-based value sets are suitable for children 5 years and above. In comparison to established instruments like EQ-5D-Y, HUI2/3, and CHU9D, the more recent instruments specifically designed for preschool children (CHSCS-PS, IQI, EQ-TIPS), exhibit greater evidence gaps in psychometric performance. The gaps were more significant for reliability (e.g., test-retest) and proxy-child agreement.[26]

Age-appropriateness for instruments targeting children of different ages

Young children under 5 years old may have different health dimensions from older children due to biological and psychosocial development [30]. Applying HRQoL instruments designed for older children to children under 5 years old may present specific issues, such as inappropriate dimensions and lack of sensitivity. However, there is limited empirical evidence to support or refute these concerns. The precision and reliability of reporting and measuring HRQoL in children under 3 years old poses particular challenges due to their distinct different developmental stages and limited ability to effectively express themselves or self-report, particularly in children with developmental delays or neurodevelopmental conditions. Inappropriate health dimensions and the use of caregiver's proxy report can potentially introduce variability and bias.

In addition, there is a lack of available instruments for children under 5 years old. Current approaches to address this include adapting wording from existing measures (e.g., EQ-5D-Y, CHU9D) or developing a new measure to include relevant health dimensions (e.g., EQ-TIPs, IQI) to better represent pre-school children's health status. The adaptation or modification from existing instrument guarantees consistent construct of HRQol measurement across different ages and thus facilitates comparisons of HRQoL between different ages.[31] On the other hand, the development of a completely new instrument may increase the appropriateness, relevance and comprehensiveness of child-specific health dimensions. There is no clear evidence about which is better. However, it is important to evaluate the performance of these new instruments before their application.

Self-report vs proxy report

Self-report involves evaluating children's HRQoL based on their own personal, subjective experiences. In contrast, a proxy report involves assessing the HRQoL of a child by a respondent other than the child, such as a parent, guardian, caregiver, teacher, or health professional. There is evidence that self-report and proxy-report usually provide different responses.[8, 32] Regulatory decision makers are encouraged to consider children's experience reported by children themselves. Children's perspective should be

appropriately considered, taking into account their age and level of maturity.[33] It is recommended by the ISPOR Good Research Practices that children should self-report HRQoL whenever it is possible.[8, 32] However, it may not be feasible all the time, such as when the child is too young or cognitively impaired. The ISPOR Good Research Practices states that "the assessment of health status in children younger than 5 years old must rely on clinical measures and observational reports of parents or other adults as there is no clear evidence of reliability or validity of self-report measures in this age group".[32] For children between 5-7, child-report is possible, but reliability and validity often seem problematic.[32] Therefore, proxy-report may be used where adequate measurement properties have not been demonstrated for a child age group,.[32] Parent proxy responses are useful as parents typically understand the most important experiences of their children. Furthermore, parents are usually the decision makers in their children's health care interventions.[32] One study evaluated differences between children's, parents' and doctors' perceptions of health states in children aged 5-18 years with various chronic conditions.[34] It found that the overall score did not differ between responders but with poorer agreement for subjective domains. In addition, parent–child agreement was higher than doctor–child agreement. In particular, patients experiencing significant pain or emotional distress, as well as those diagnosed with severe cerebral palsy or chronic neurological conditions, were more likely to have their subjective well-being under-reported by doctors and parents.[34] Similarly, Wanni-Arachchige et al. (2024) showed that there is disagreement in self-versus proxy-reporting of HRQoL in children with Attention-Deficit/Hyperactivity Disorder, with parents rating their children's HRQoL lower than the children themselves. A recent systematic review concluded that additional evidence is necessary to develop comprehensive best practice guidance regarding the rationale, when, and methodologies for employing self- and proxy reports in assessing HRQoL among child populations.[35]

Recall period

There are different recall periods for different child health measures for economic evaluation. For example, EQ-5D-Y uses "Today", CHU9D uses "Today/Last night", while HUI2 uses past weeks (1, 2 or 4 weeks) or usual. While recall periods are varied, most instruments used today or right now recall period (refer to Table 1-1). Choosing a recall period also relates to the frequency of assessments. Daily assessments with a 1-day recall period are often impractical, while longer intervals may capture more events. However, longer periods introduce challenges like varied reporting and recall bias. More research or consideration about recall period is valuable for future research to decide which HRQoL measures to use for clinical trials or economic evaluations.[36] A recent systematic review found that quality of life reported tend to be lower when using a seven-day recall period compared to a one-day recall period.[37]

It also showed that participants have mixed preferences for recall periods to measure health, depending on the condition or symptoms.[37] More research is warranted to understand this issue.

Spillover effects and capturing QoL in parents/caregivers

The impact on HRQoL for caregivers and family members of ill patients has been termed "spillover effects". Multiple national guidance bodies recommended the inclusion of spillover effects (related with caregiver and family member) in the measurement of child HRQoL and cost-effectiveness analyses.[38] The spillover effect estimation and incorporation methods are still emerging, which makes the adoption of spillover effects into cost-effectiveness slow.[38] Recently, a new generic measure, the EQ Health and Wellbeing (EQ-HWB) is designed to assess a range of effects including impact on the health and wellbeing of care recipients and caregivers, aiming for evaluating interventions in health, public health, and social care.[39] Further investigation is warranted into the technical methodologies and guidelines regarding the integration of the quality of life impact on parents due to child health issues.

1.3. Child Health Valuation

1.3.1. How to value child health?

What is valuing health?

Valuing health, sometimes referred to as preference elicitation techniques, aims to obtain the utilities (values or preference weights or QALY weights) attached to health states, i.e., putting the 'Q' into QALYs. Health states can be defined by customized health descriptions, or vignettes or by dimensions and levels in HRQoL instruments.[8]

What are the valuation methods? There are direct and indirect approaches to obtain the utilities of health states. Direct approaches include cardinal techniques (mainly including visual analogue scale (VAS), standard gamble (SG) and time trade-off (TTO)) and ordinal or ranking techniques (including discrete choice experiment (DCE), rating/ranking, and best-worst scaling (BWS)). Ordinal techniques do not directly produce utilities that are anchored onto 1-0 full health-dead scale, indicating that a distinct task is necessary to anchor the ordinal responses. Various methods are available for this anchoring purpose, such as extra TTO data, extra VAS data, duration attribute for estimating relative preference with time for DCE, or a rescaling method with or without additional data.[40]

Indirect approaches include the multi-attribute utility instruments (MAUIs) (also known as HRQoL measures accompanied with preference-weighted scoring system),[8] which involve the use of direct valuation approaches as well in the development of their preference-weighted scoring algorithms.

Importance to value child health

A growing number of cost utility studies are carried out in health care programs for children. The trends based on the latest Pediatric Economic Database Evaluation (PEDE) data showed that the number of cost-utility analysis exceeded the number of cost-effectiveness analysis in 2021 (Figure below).[41] This shows that there is increasing interest and empirical studies for measuring and valuing child health in recent years. This highlights the importance of valuing child health.

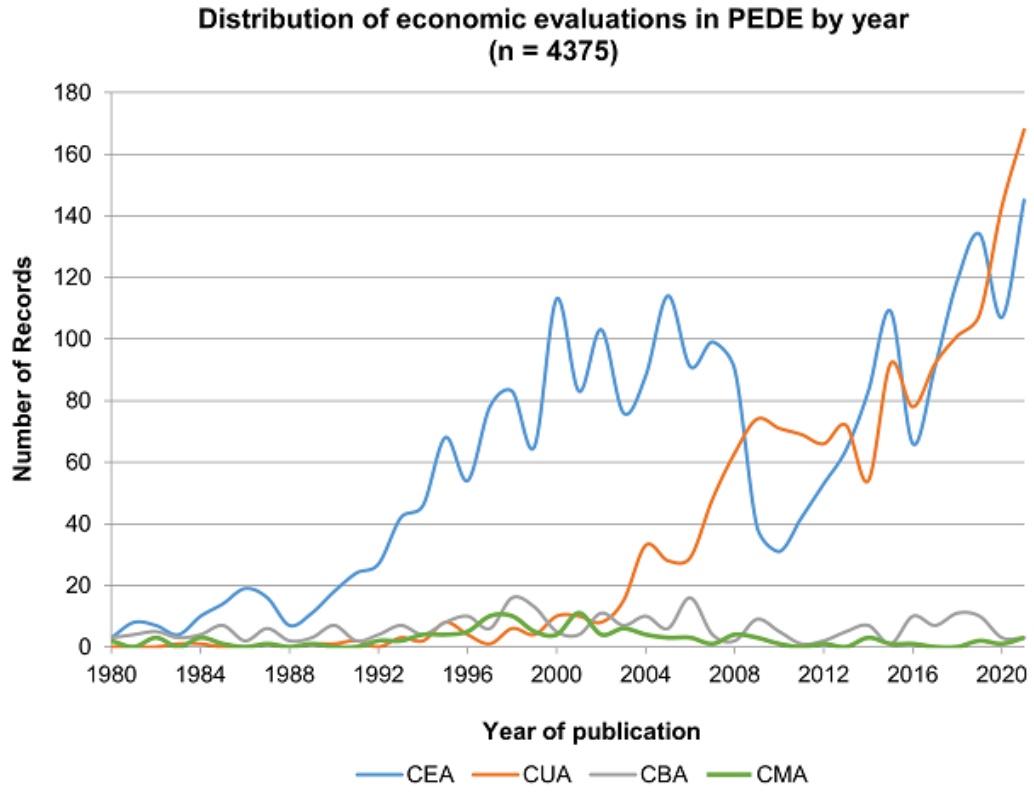


Figure 1-1 Trends in Economic Evaluation

Note: Reference: Pediatric Economic Database Evaluation (PEDE). Trends in Economic Evaluation. Retrieved from http://pede.ccb.sickkids.ca/pede/trends.jsp

Key methodological considerations of valuation of child health

The generally considered four methodological decisions in the valuation of child health include: the source of preferences, the perspective taken, the elicitation methods (including anchoring methods for ordinal valuation tasks) and which mode of administration.[21]

Whose preference/values to use? Currently, the source of values for valuing child health includes the general adult population, adolescents/children, parents, family members, patients, and health professionals.[42] The general adult population are often deemed as the appropriate source to obtain preferences for health states in child specific MAUIs in that they pay tax to fund the health care system. This also provides potential comparability in resource allocation for the whole population since adult MAUIs and child MAUIs use the same source of preferences.[21] Young children (under 6 years old) are too young to understand elicitation techniques and have to use preference obtained from the adult population[43], but whether adolescent population's preference should be considered in decision making raises debate. There are literature indicating that preference from adults differ from those from adolescents.[44] For example, one recent study identified that Australian adults, considering both their own perspective and that of a 10-year-old child, prioritize physical pain or discomfort more highly and assign less importance to being very worried or sadness compared to Australian adolescents.[45] Some researchers suggest combining the preferences of adults and adolescents to form an index, or at least incorporating children's preference as sensitivity analysis in economic evaluations to understand differences.[21]

Which perspective to take? Usually, the person completing the valuation task is asked to imagine someone (themselves or others) to experience health states described in the valuation tasks. Preference elicitation in adolescents could ask them to imagine experiencing the health state themselves or by imagining another adolescent, where the latter is more cognitively challenging. When adults are asked to elicit preferences for pediatric MAUIs, there are multiple possible perspectives, including their own health as adults, their health state during childhood, and the health state of an imagined child (hypothetical or child they know) at a certain age. Each of these perspectives has potential biases. Imagining experiencing the health state themselves will be challenging when it comes to irrelevant situations such as doing homework for example as described in CHU9D instrument. Asking adults to imagine health states being experienced when they were children is prone to recall bias. Asking adults to imagine the health state experienced by another child might depend on their experience with children of that age or the influence of their views on children or child health. Adults may think of a variety of different children when performing a valuation task such as their own, grandchildren or children they work with. There is evidence that many adults will think of a specific child even if asked about a hypothetical child, which is called anchoring bias [46]. Some adults may lack experience with children and find imagining health states for children difficult.[21] When comparing the values between adults taking the perspective of themselves with the perspective of adults thinking of children, evidence were mixed.[42] This again highlights that the perspective and framing matter to the responses received.

What valuation techniques to be used in valuing childhood health? The application of the preference elicitation tasks has changed over time with the emergence of new methods.[42] SG has been used since 1996, with a decrease in frequency since mid-2000. TTO has been consistently used since 2003.[42] DCE first appeared in this field in 2011 and has become more popular recently. BWS was first reported in the same paper as DCE, but has been relatively less used.[42] Presently, TTO is primarily employed to anchor preferences obtained through DCE and BWS methodologies to a utility scale.[42]

Which mode of administration? Traditional utility direct elicitation techniques are usually carried out face to face since the task is difficult and needs the explanation and help from the interviewer. Web-based surveys for obtaining health state values is becoming popular recently, particularly with the growing population of DCE methods.[45, 47] Online administration mode has the advantage of facilitating broad geographical coverage of a population at relatively low cost. However, there is also concern relating to the quality of the responses achieved.[8]

Currently available child specific MAUIS with value sets

Table 1-4 summarized the childhood specific MAUIs with value sets available. Most value sets were for instruments suitable for children 5 years and above. Only the IQI suitable for 0–12-month-old children has an available value set, and HuPS has a scoring algorithm to value health for children aged 2-4 years. The development of value sets for instruments suitable for children under 5 years old has a prominent research gap.

Current use of valuation methodologies in valuing child health

For the valuation techniques used in valuing child health, systematic reviews indicated that direct valuation methods and adult-specific MAUIs were the two most popular choices, while the use of childhood-specific MAUIs was minimal.[48, 49] Kwon et al 2019 reviewed the patterns and trends in the measurement and valuation of childhood health utilities using studies published until June 2017.[49] From Kwon et al 2019, the valuation methods used in child health states valuation were grouped into six key categories: (1) VAS (EQ-5D VAS: 8.8%, EQ-5D-Y VAS: 5.8%; stand-alone VAS: 6.3%); (2) trade-off-based direct valuation methods-TTO (4.3%), SG (5.7%), chained gamble and adjusted SG (3.6%); (3) adult-specific MAUIs (EQ-5D: 10.7%, SF-6D: 0.9%, AQoL-5D: 0.4%, 15D: 0.05%); (4) MAUIs compatible with both childhood and adult populations (QWB: 5.6%, HUI2: 12.1%, HUI3: 20.7%); (5) childhood-specific MAUIs (EQ-5D-Y: 2.7%, CHU9D: 5.8%, 16D: 1.8%, 17D:1.0%, AQoL-6D: 1.3%); (6) mapping non-preference-based clinical measures to utility indices (0.4%).[49] Baily et al. 2022 reviewed Pharmaceutical Benefits Advisory Committee Public Summary Documents in Australia to examine the methods used to value child health in decision making in Australia.[48] Baily et al. found

that out of 62 documents containing information relating to children and utilities, 16 included adult HRQoL measures, 11 included direct elicitation, and only four included child-specific HRQoL measures, with the remaining 31 documents not clear about the HRQoL sources. It was advised to regularly utilize child-specific measures to enhance the quality of evidence for decision makers regarding funding medicines for pediatric use.[48]

For the source of population and perspectives taken in available value sets for childhood specific MAUIs, Kwon et al. 2022 found 21 preference-based value sets for ten generic multidimensional childhood patient-reported outcome measures, with seven based on adolescents' preferences, and 14 based on adults' preferences (seven from the perspective of or on behalf of the child, and seven adopting an adult's perspective).[50]

1.3.2. Challenges and research gaps in valuing child health

Variance in methodologies in valuing health

The preference elicitation techniques including the methods to anchor values on the utility scale of 0-1 varied considerably between studies and value sets. The range and distribution of values of different value sets also varied.[50] It is hard to decide which preference elicitation techniques are most suitable for valuing child health due to differences in reporting.[42] A recent published checklist for studies reporting the valuation of child health is a great help to improve consistency and quality of reporting in future studies valuing child health.[51]

Philosophical debate about the source of preferences and perspectives taken in valuation tasks

Determining whose preferences and which perspective to consider is subject to normative and philosophical deliberation, ultimately resting with policymakers' decisions. However, research in this area would provide evidence and enrich discussions. The identified evidence gaps includes the choice of child age in health state valuation tasks; the suitability and acceptability of valuation tasks for adolescents especially when addressing the concept of a 'dead' health state; the methods to anchor preferences for adolescent/child health states to 0-1 scale; and the generation and use of combined preferences from different sources, for example, adult and adolescent preferences.[52]

Lack of available value sets appropriate for children under 5 years of age

Another prominent gap is valuing health for children under 5 years old. As the development/adaptation of MAUIs for children under 5 years old is more recent, the development of the accompanied value set is even limited.

1.4. Overarching aim

There are many unresolved research issues relating to the measurement, valuation, and application of HRQoL in the pediatric population. Evidence for young children under 5 years old is especially lacking.

This thesis aims to explore the current application of HRQoL in children and to improve the measurement and valuation of HRQoL in children. It includes six individual health economics studies covering various methodologies. These studies demonstrate the application of HRQoL in longitudinal observational studies and economic evaluations, identified evidence gaps, and advanced methods to measure and value HRQoL, finally contributing to better decision-making in health care resource allocation.

1.5. Thesis structure

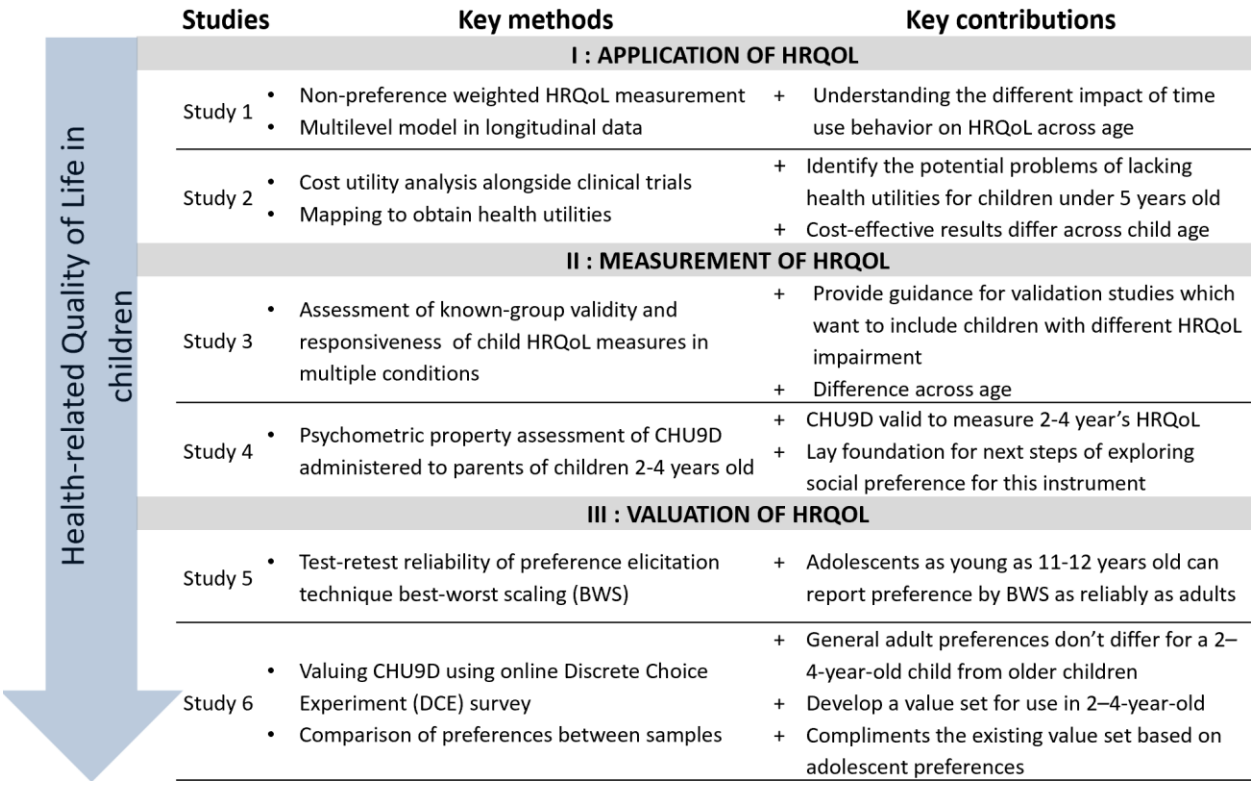
Chapter 1 of this thesis introduces key concepts and the structure of this research. Six individual studies followed and are presented from chapter 2 to chapter 7. The six studies cover various methodologies related to HRQoL including multilevel modeling of HRQoL in longitudinal observational studies, cost-utility analysis, psychometric property assessment for HRQoL measures, evaluating the reliability of preferences elicitation techniques, and scoring HRQoL instruments using discrete choice experiments. These studies together contribute to advancing methods in the measurement and valuation of HRQoL in children. In the final discussion chapter, the thesis concludes with a summary of the main findings, and a discussion of the implications, limitations, and opportunities for future research. With a focus on children, especially young children under 5 years of age, this thesis rigorously contributes to understanding the current application of HRQoL in children and improving the measurement and valuation of HRQoL.

The tables below outline the structure of the thesis, summarizes the key aims of each study along with their publications and summarizes the key methodologies and contributions from each chapter.

Table 1-1 Thesis structure

Studies	Key aims	Publications
I : APPLICATION OF HRQOL		
Study 1	To apply HRQoL measure without preference-weighted scoring system across child ages	Xiong, X., et al., Association between 24-hour movement behaviors and health-related quality of life in children. Qual Life Res, 2022.
Study 2	To apply HRQoL measure with preference-weighted scoring system in a cost-utility analysis across child ages	Xiong, X., et al., Cost-effectiveness of Prednisolone to Treat Bell Palsy in Children: An Economic Evaluation Alongside a Randomized Controlled Trial. Neurology, 2023.
II : MEASUREMENT OF HRQOL		
Study 3	To explore the impact of various conditions on HRQOL across child age, and to inform the recruitment of clinical samples for psychometric property assessment of HRQOL instruments	Xiong, X., et al., How do common conditions impact health-related quality of life for children? Providing guidance for validating pediatric preference-based measures. Health and Quality of Life Outcomes, 2023.
Study 4	To assess psychometric property of CHU9D 2-4 years old version; To provide the foundation for its valuation	Xiong, X., et al., Psychometric Properties of Child Health Utility 9D (CHU9D) Proxy Version Administered to Parents and Caregivers of Children Aged 2-4 Years Compared with Pediatric Quality of Life Inventory™ (PedsQL). Pharmacoeconomics, 2024.
III : VALUATION OF HRQOL		
Study 5	To evaluate preference elicitation technique (best-worst scaling); To evaluate the reliability of adolescents' reported preferences	Xiong, X., et al., Test-Retest Reliability of EQ-5D-Y-3L Best-Worst Scaling Choices of Adolescents and Adults. Value in Health, 2022.
Study 6	To compare general adult population's preference for health between children aged 2–4 years versus To value CHU9D 2-4y version	Manuscript finished; suggested reference: Xiong, X., et al., Valuing the Child Health Utility 9D (CHU9D) for children under 5 years in Australia.

Table 1-2 Summary of key methods and contributions from each study



Abbreviations: health-related quality of life (HRQoL)

1.6. Chapter content

1.6.1. Section I: Application of health-related quality of life

Chapter 2 presents an empirical study using HRQoL data from the Longitudinal Study of Australian Children (LSAC). This study aims to assess the associations between adherence to 24-hour movement behaviors guidelines and child HRQoL across child age. This study is an example demonstrating the application of non-preference-weighted HRQoL data. It highlights the advantage of using HRQoL as an outcome as it captures multiple dimensions of health and thus can show the overall impact on a child's general health and functional status. In addition, this study includes the 24-hour movement guidelines as the exposure instead of single time use behavior (screen use, physical activity, or sleep), which emphasizes the interactions between activities during a day. It made use of the time-use diary in LSAC to enable more accurate estimates of time use behavior. This study covers children with a wide age range (2-15 years) and enables subgroup analysis by age. Linear mixed modeling was used to analyze the longitudinal data. The study found that 1) meeting physical activity guidelines had the strongest

association with HRQoL compared with meeting screen or sleep time guidelines overall, but results differed by age; 2) the association appears strongest for adolescents, where the HRQoL increment of meeting all three guidelines almost reached the minimal clinical importance threshold; 3) the results for young children aged 2-4 years old is unique, of which only meeting all three guidelines showed HRQoL improvement. These results highlight that health dimensions have different importance for 2-4 year olds compared with older children.

Chapter 3 presents another example of the application of HRQoL- the cost-effectiveness of prednisolone to treat bell's palsy in children. This study is a prospectively planned economic evaluation alongside a randomized controlled trial. The time horizon for the economic evaluation was 6 months following randomization. 180 children aged from 6 months to 17 years who presented within 72 hours of onset of bell's palsy were included in the cost-effectiveness analysis. The intervention was oral prednisolone or taste-matched placebo administered for 10 days. The incremental cost-effectiveness ratio comparing prednisolone with placebo was estimated. The health care sector perspective was considered for the cost. Effectiveness was measured using QALYs and recovery. It was found that prednisolone was more costly, however, also led to more QALYs over a 6-month period compared to placebo. The probability that prednisolone is cost-effective is 83% giving conventional willingness-to-pay threshold of A$50,000 per QALY gained and the cost-effectiveness is mainly driven by the older age group (12-17 years old). The HRQoL data was only available for 5-17 years old as CHU9D is only available for children aged 5-17 years old. No other validated HRQoL measures with preference-weighted scoring system for children under 5 years old was available at the time of conducting the study. Health utilities for children under 5 years old were obtained by mapping PedsQL to CHU9D utilities using published algorithm. Children under 5 years old were only a small proportion of the whole study participants. The study explored whether results would differ if using mapped utilities for all participants and found that the probability of being cost-effective was very different from the main results (50% vs 83%). If using all mapped data, the decision would change from being cost-effective to not cost-effective. This suggests that it may be problematic to only use utilities obtained by mapping method and highlights the research gap in the measurement and valuation of HRQoL in children under 5 years old.

1.6.2. Section II: Measurement of health-related quality of life

Chapter 4 presents a study comparing known-group validity and responsiveness of inferred EQ-5D-Y and inferred CHU9D across 27 common chronic child health conditions. The impact on overall HRQoL and different individual health dimensions were explored. It identified the top 10 conditions with the largest HRQoL impact. It used data from LSAC with over 10,000 children at baseline. Child age ranges from 2 to

18 years old. The MAUIs relevant to this study are EQ-5D-Y and CHU9D, with the health dimensions for the two MAUIs mapped from PedsQL items. It is essential for psychometric property assessment studies for MAUIs to include participants with various degrees of HRQoL impairment or expected changes in HRQoL. The study provided valuable information for future studies to select disease groups with various impact on HRQoL. This chapter is a good preparation for chapter 5 which needs recruiting various clinical samples to evaluate the psychometric performance of multiple child specific MAUIs. In addition, this study provided validation evidence for children under 5 years old. This study enables a comparison among a wide range of child chronic health conditions in one study. The impact on HRQoL measurement due to different wording and recall period has also been explored by comparing real CHU9D and inferred CHU9D using relevant PedsQL items (Sensitivity analysis in this study).

Chapter 5 presents a study which assessed the psychometric properties of CHU9D proxy version with guidance notes for children under 5 years old. The data came from a large Australian paediatric multi-instrument comparison (P-MIC) study. Respondents were parents or caregivers of 2-4 years old children in Australia for this study. Completion time, reported difficulty, ceiling and floor effects, test-retest reliability, convergent and divergent validity, known-group validity, and responsiveness were assessed. The study provides evidence that CHU9D with guidance notes is valid and reliable in terms of a series of pre-defined criteria (at least moderate test-retest reliability, convergent validity and known-group validity; significant score changes or at least small effect size of responsiveness) administered to parents/caregivers of children aged 2-4 years old. The findings provided evidence supporting the wide use of CHU9D with guidance notes to measure HRQoL for children aged 2-4 years old. Further studies to explore appropriate preference-weighted scoring for CHU9D with guidance notes to allow application in economic evaluation are important next steps.

1.6.3. Section III: Valuation of health-related quality of life

There are circumstances where children's own preferences are desired, and others where the preferences of the general population of taxpayers are preferred. Methods are less established for obtaining preferences reliably from children. Chapter 6 addresses an important research question in the valuation of HRQoL, which is whether adolescents can report health state preferences reliably using best-worst scaling (BWS), a variant of discrete choice experiments (DCE). BWS and DCE are becoming popular as they are relatively easy in terms of comprehension and administration (e.g., online survey). Profile case BWS is believed to be less cognitively demanding than traditional DCE and has been used to elicit preference from adolescents. Test-retest reliability of a valuation method is important as it determines whether it can provide consistent preferences. However, this evidence for BWS in health state valuation is lacking. The

study aimed to investigate the test-retest reliability of BWS to elicit preference for EQ-5D-Y-3L in adolescents compared to adults using community-based samples. The methods used to assess the test-retest reliability includes simple agreement, kappa statistics, comparison of BWS marginal frequencies and comparison of relative attribute importance between baseline and follow up. This study added to the evidence that generally healthy adolescents as young as 11-12 years old can complete BWS tasks to report preferences for health reliably, with best choices slightly more reliable than worst choices.

Chapter 7 focuses on valuing children's HRQoL using preferences derived from the adult population. It involves the valuation of the CHU9D proxy version with guidance notes for Australian children aged 2-4 years old. This study fulfills the last piece of evidence to extend health utilities measurement for economic evaluation to children as young as 2 years old. For CHU9D, there are available value sets for its original version for children aged 5-17 years old. It is not known if the value sets for older children could be used in 2–4-year-old children as it is not clear whether the general population adults' preferences differ between these two age groups. In addition, CHU9D only has a value set based on adolescent preferences in Australia while it is common to have preferences from general adult population from the tax-payer perspective. Each is desirable for different reasons and understanding differences between the two is important. Therefore, for CHU9D 5–17-year version, a value set developed from general population adults is lacking. This study first investigated whether general population adults' preferences for children aged 2-4 years old differ from preferences for children aged 5-17 years. Secondly it explored the preference-weighted scoring appropriate for use in children aged 2-4 years old. It used an online survey of DCE to obtain preferences from the general adult population. The sample was randomly allocated to two arms, one arm taken from the perspective of a 2–4-year-old and the other arm of a 10-year-old child. It used VAS to anchor the latent preference values to 0-1 where 1=full health and 0=dead. It found that general adults' preferences for health did not differ between a 2–4-year-old and a 10-year-old. The pooled data of the two arms were used to generate the value set appropriate for use for both age groups since there is no appreciable difference in preferences found. The developed value set enables accurate calculation of QALYs for cost-utility analysis including children aged 2-4 years old. The value set also compliments the existing CHU9D value set based on adolescent's preferences for policy makers who may want to include difference sources of preferences in Australia.

Chapter 8 concluded the thesis by summarizing the main findings, discussion of implications, strengths and limitations, and identified unresolved issues and future research opportunities.

1.7. References

[1] (July 5, 2023). *Health expenditure Australia 2020-21*. Available: https://www.aihw.gov.au/reports/health-welfare-expenditure/health-expenditure-australia-2020-21/contents/summary

[2] "Australia's health expenditure: an international comparison," Australian Institute of Health and Welfare, Australian Institute of Health and Welfare2019, Available: https://www.aihw.gov.au/reports/health-welfare-expenditure/health-expenditure-international-comparison/summary.

[3] R. Calder, R. Dunkin, C. Rochford, and T. Nichols, "Australian health services: too complex to navigate," 2019.

[4] M. F. Drummond, M. J. Sculpher, K. Claxton, G. L. Stoddart, and G. W. Torrance, *Methods for the economic evaluation of health care programmes*. Oxford university press, 2015.

[5] T. P. B. A. Committee. (2021, Nov 8). *PBAC guidelines; Section 3A 1.2, PBAC guidelines, 2016*.

[6] A. G. Department of Health and Aged Care. (2016, 21/02/2024). *The Pharmaceutical Benefits Advisory Committee Guidelines: section 3A.1.2 Type of economic evaluation*.

[7] S. J. Whitehead and S. Ali, "Health outcomes in economic evaluation: the QALY and utilities," *British Medical Bulletin,* vol. 96, no. 1, pp. 5-21, 2010.

[8] J. Brazier, J. Ratcliffe, J. Saloman, and A. Tsuchiya, *Measuring and valuing health benefits for economic evaluation*. OXFORD university press, 2016.

[9] A. I. o. H. a. waelfare, "Chronic conditions and burden of diseases," Australian Institute of Health and Welfare, Australian Institute of Health and Welfare2022.

[10] World Health Organization. (2019, July 16th, 2022). *Children: reducing mortality. Newsroom, Factsheets.* Available: https://www.who.int/news-room/factsheets/detail/children-reducing-mortality.

[11] T. Vos *et al.*, "Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019," *The Lancet,* vol. 396, no. 10258, pp. 1204-1222, 2020/10/17/ 2020.

[12] Q. Wang, H. Zhang, J. A. Rizzo, and H. Fang, "The Effect of Childhood Health Status on Adult Health in China," vol. 15, no. 2, p. 212, 2018.

[13] L. M. Richter *et al.*, "Investing in the foundation of sustainable development: pathways to scale up for early childhood development," (in eng), *Lancet,* vol. 389, no. 10064, pp. 103-118, Jan 7 2017.

[14] P. Dimitri *et al.*, "Medical Device Development for Children and Young People-Reviewing the Challenges and Opportunities," (in eng), *Pharmaceutics,* vol. 13, no. 12, Dec 17 2021.

[15] P. Dolan, "The measurement of health-related quality of life for use in resource allocation decisions in health care," *Handbook of health economics,* vol. 1, pp. 1723-1760, 2000.

[16] "Measuring Health-Related Quality of Life," *Annals of Internal Medicine,* vol. 118, no. 8, pp. 622-629, 1993.

[17] J. Brazier and S. Dixon, "The use of condition specific outcome measures in economic appraisal," (in eng), *Health Econ,* vol. 4, no. 4, pp. 255-64, Jul-Aug 1995.

[18] G. Calaminus and R. Barr, "Economic evaluation and health-related quality of life," (in eng), *Pediatr Blood Cancer,* vol. 50, no. 5 Suppl, pp. 1112-5, May 2008.

[19] T. Lambe *et al.*, "Mapping the Paediatric Quality of Life Inventory (PedsQL™) Generic Core Scales onto the Child Health Utility Index–9 Dimension (CHU-9D) Score for Economic Evaluation in Children," *PharmacoEconomics,* vol. 36, no. 4, pp. 451-465, 2018/04/01 2018.

[20] G. Chen and J. Ratcliffe, "A Review of the Development and Application of Generic Multi-Attribute Utility Instruments for Paediatric Populations," (in eng), *Pharmacoeconomics,* vol. 33, no. 10, pp. 1013-28, Oct 2015.

[21] D. Rowen, O. Rivero-Arias, N. Devlin, and J. Ratcliffe, "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?," (in eng), *Pharmacoeconomics,* Jan 6 2020.

[22] K. Dalziel *et al.*, "A Qualitative Investigation to Develop an Adapted Version of the EQ-5D-Y-3L for Use in Children Aged 2-4 Years," (in eng), *Value Health,* vol. 26, no. 10, pp. 1525-1534, Oct 2023.

[23] J. Verstraete and R. Amien, "Cross-Cultural Adaptation and Validation of the EuroQoL Toddler and Infant Populations Instrument Into Afrikaans for South Africa," *Value in Health Regional Issues,* vol. 35, pp. 78-86, 2023/05/01/ 2023.

[24] S. Saigal *et al.*, "Development, reliability and validity of a new measure of overall health for pre-school children," (in eng), *Qual Life Res,* vol. 14, no. 1, pp. 243-57, Feb 2005.

[25] K. M. Beusterien, J.-E. Yeung, F. Pang, and J. Brazier, "Development of the multi-attribute adolescent health utility measure (AHUM)," *Health and Quality of life outcomes,* vol. 10, no. 1, p. 102, 2012.

[26] J. Kwon *et al.*, "Systematic Review of the Psychometric Performance of Generic Childhood Multi-attribute Utility Instruments," *Applied Health Economics and Health Policy,* 2023/05/03 2023.

[27] J. Brazier and M. J. H. e. Deverill, "A checklist for judging preference-based measures of health related quality of life: learning from psychometrics," vol. 8, no. 1, pp. 41-51, 1999.

[28] L. B. Mokkink *et al.*, "COSMIN Study Design checklist for Patient-reported outcome measurement instruments," ed, 2019.

[29] R. Jones *et al.*, "Comparative Psychometric Performance of Common Generic Paediatric Health-Related Quality of Life Instrument Descriptive Systems: Results from the Australian Paediatric Multi-Instrument Comparison Study," *PharmacoEconomics,* 2023/11/13 2023.

[30] J. Ball, R. M. Bindler, and K. J. J. Cowen, "Child health nursing: Partnering with children & families," 2010.

[31] N. Wille *et al.*, "Development of the EQ-5D-Y: a child-friendly version of the EQ-5D," *Quality of Life Research,* vol. 19, no. 6, pp. 875-886, 2010/08/01 2010.

[32] L. S. Matza *et al.*, "Pediatric patient-reported outcome instruments for research to support medical product labeling: report of the ISPOR PRO good research practices for the assessment of children and adolescents task force," (in eng), *Value Health,* vol. 16, no. 4, pp. 461-79, Jun 2013.

[33] U. Nations. (1989, 21/02/2024). *Convention on the rights of the child-Article 12.1*.

[34] A. M. Morrow, A. Hayen, S. Quine, A. Scheinberg, and J. C. Craig, "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 and Development,* vol. 38, no. 2, pp. 186-195, 2012.

[35] C. Mpundu-Kaambwa *et al.*, "A Systematic Review of International Guidance for Self-Report and Proxy Completion of Child-Specific Utility Instruments," *Value in Health,* vol. 25, no. 10, pp. 1791-1804, 2022/10/01/ 2022.

[36] N. Bansback *et al.*, "Impact of the recall period on measuring health utilities for acute events," vol. 17, no. 12, pp. 1413-1419, 2008.

[37] T. Peasgood, J. M. Caruana, and C. Mukuria, "Systematic Review of the Effect of a One-Day Versus Seven-Day Recall Duration on Patient Reported Outcome Measures (PROMs)," *The Patient - Patient-Centered Outcomes Research,* vol. 16, no. 3, pp. 201-221, 2023/05/01 2023.

[38] E. Wittenberg, L. P. James, and L. A. Prosser, "Spillover Effects on Caregivers' and Family Members' Utility: A Systematic Review of the Literature," *PharmacoEconomics,* vol. 37, no. 4, pp. 475-499, 2019/04/01 2019.

[39] J. Brazier *et al.*, "The EQ-HWB: Overview of the Development of a Measure of Health and Wellbeing and Key Results," *Value in Health,* vol. 25, no. 4, pp. 482-491, 2022/04/01/ 2022.

[40] H. Wang, D. L. Rowen, J. E. Brazier, and L. Jiang, "Discrete Choice Experiments in Health State Valuation: A Systematic Review of Progress and New Trends," (in eng), *Appl Health Econ Health Policy,* vol. 21, no. 3, pp. 405-418, May 2023.

[41] (2023, Nov 22, 2023). *Trends in Economic Evaluation*.

[42] C. Bailey *et al.*, "Preference Elicitation Techniques Used in Valuing Children's Health-Related Quality-of-Life: A Systematic Review," *PharmacoEconomics,* vol. 40, no. 7, pp. 663-698, 2022/07/01 2022.

[43] W. J. Ungar, "Challenges in health state valuation in paediatric economic evaluation: are QALYs contraindicated?," (in eng), *Pharmacoeconomics,* vol. 29, no. 8, pp. 641-52, Aug 2011.

[44] J. Ratcliffe, K. Stevens, T. Flynn, J. Brazier, and M. G. Sawyer, "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," (in eng), *Value Health,* vol. 15, no. 5, pp. 730-6, Jul-Aug 2012.

[45] K. Dalziel, M. Catchpool, B. Garcia-Lorenzo, I. Gorostiza, R. Norman, and O. Rivero-Arias, "Feasibility, Validity and Differences in Adolescent and Adult EQ-5D-Y Health State Valuation in Australia and Spain: An Application of Best-Worst Scaling," (in eng), *Pharmacoeconomics,* Jan 24 2020.

[46] F. Lieder, T. L. Griffiths, M. H. QJ, and N. D. Goodman, "The anchoring bias reflects rational use of cognitive resources," (in eng), *Psychon Bull Rev,* vol. 25, no. 1, pp. 322-349, Feb 2018.

[47] B. M. Craig, W. Greiner, D. S. Brown, and B. B. Reeve, "Valuation of Child Health-Related Quality of Life in the United States," vol. 25, no. 6, pp. 768-777, 2016.

[48] C. Bailey, K. Dalziel, P. Cronin, N. Devlin, and R. Viney, "How are Child-Specific Utility Instruments Used in Decision Making in Australia? A Review of Pharmaceutical Benefits Advisory Committee Public Summary Documents," (in eng), *Pharmacoeconomics,* vol. 40, no. 2, pp. 157-182, Feb 2022.

[49] J. Kwon, S. W. Kim, W. J. Ungar, K. Tsiplova, J. Madan, and S. Petrou, "Patterns, trends and methodological associations in the measurement and valuation of childhood health utilities," (in eng), *Quality of Life Research,* vol. 28, no. 7, pp. 1705-1724, Jul 2019.

[50] J. Kwon *et al.*, "Systematic Review of Conceptual, Age, Measurement and Valuation Considerations for Generic Multidimensional Childhood Patient-Reported Outcome Measures," *PharmacoEconomics,* vol. 40, no. 4, pp. 379-431, 2022/04/01 2022.

[51] C. Bailey *et al.*, "The RETRIEVE Checklist for Studies Reporting the Elicitation of Stated Preferences for Child Health-Related Quality of Life," *PharmacoEconomics,* 2024/01/13 2024.

[52] D. Rowen, O. Rivero-Arias, N. Devlin, and J. Ratcliffe, "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?," (in eng), *Pharmacoeconomics,* vol. 38, no. 4, pp. 325-340, Apr 2020.

[53] Wanni Arachchige Dona S, Badloe N, Sciberras E, Gold L, Coghill D, Le HND. The Impact of Childhood Attention-Deficit/Hyperactivity Disorder (ADHD) on Children's Health-Related Quality of

Life: A Systematic Review and Meta-Analysis. J Atten Disord. 2023;27(6):598-611. doi:10.1177/10870547231155438

1.8. Tables and Figures

Table 1-3 Summary of the classification systems of child- and adolescent-specific generic preference-based measure

Measure	Age appropriate to measure health for (years unless otherwise specified)	Classification system content	Country of origin	Self/proxy-report	Recall period
AHUM	12–18	Self-care; pain; limitations walking around (mobility); perceptions of strenuous activities; self-image; health perceptions	UK	self-assess	N/A
AQoL-6D	Adolescent	Independent living; relationships; mental health; coping; pain; senses	Australia	self-assess	Unspecified
CHU9D	4–17 (with a proxy version with guidance notes for children under 5 years old)	Worry; sadness; pain; tiredness; annoyance; school; sleep; daily routine; activities	UK	Self-assess/proxy-assess	Today/last night
EQ-5D-Y	4–15 (with adapted version for under 5 years old children)	Mobility; looking after myself; doing usual activities; having pain or discomfort; feeling worried, sad or unhappy	UK	Self-assess/proxy-assess	Today
HUI2	5 upwards	Sensation; mobility; emotion; cognition; self-care; pain; fertility	Canada	Self-assess/interviewer-	Current (past 1 week, past 2

HUI3	5 upwards	Vision; hearing; speech; ambulation; dexterity; emotion; cognition; pain	Canada	administered/proxy-assess	weeks, past 4 weeks) or usual
QWB-SA	Unclear	Chronic symptoms or problems; acute physical problems; mental health; mobility; physical activity; social activity	USA		Past 3 days, not including today
16D	12–15	Mobility; vision; hearing; breathing; sleeping; eating; elimination; speech; mental function; discomfort and symptoms; school and hobbies; friends; physical appearance; depression; distress; vitality	Finland		Right now/today
17D	8–11	Mobility; breathing; school and hobbies; friends; hearing; vision; eating; elimination; vitality; sleeping; anxiety; discomfort and symptoms; learning and memory; ability to concentrate; depression; speech; physical appearance	Finland	Interviewer-administered/proxy-assess	Right now/today
HSCS-PS	2.5-5	Vision, hearing, speech, mobility, dexterity, self-care, emotion, learn/remember, think/problem-solve, pain, general health, behavior	Canada	Proxy report	Past week
EQ-TIPS	0-36 months	Movement, play, pain, relationships, communication, and eating	South Africa	Proxy report	Today
HuPS	2-4	vision, hearing, speech, ambulation, dexterity, emotion, cognition, and pain and discomfort (n=12)	Canada	Proxy report	Past week

IQI	0-1	Sleeping; feeding; breathing; stooling/poo; mood; skin; interaction (n=7)	UK, New Zealand, Singapore	Proxy report	Today

Note: *Adapted from Table 1, in Rowen et al 2020[1] and Table 2 in Joseph Kwon 2022[2]

Table 1-4 Summary of the value set methodologies of child- and adolescent-specific generic measures accompanied with preference-weights

Measure (suitable age range)	**Country**	**Whose values**	**Perspective**	**Elicitation technique**	**Mode of administration**	**Method of anchoring onto the 1–0 full health–dead scale**	**Year of publication**	**Sample age range (years)**
AHUM (12-18y)	UK	Adult	Own health	TTO	Face-to-face interview with props	TTO utility values are directly generated onto the 1–0 scale	2012	18 upwards
AQoL-6D (Adolescent)	Australia, Fiji, New Zealand, Tonga	Adolescents	Own health	TTO	Class test of 10–15 participants with 2 facilitators	TTO utility values are directly generated onto the 1–0 scale	2010	Not specified
CHU9D (4-17y)	Australia[3]	Adolescents	Own health	Best–worst scaling	Online	TTO utility values elicited from a sample of young adults	2011–2016	11–17
	China[4]	Adolescents	Own health	Best–worst scaling	Classroom	TTO utility values elicited from a sample of young adults	2019	BWS: 9–17, TTO: 18–19

	Netherlands[5]	Adult general population	Own health	Discrete choice experiment with duration	Online survey	Modelled latent scale values anchored using duration coefficient	2018	18 upwards
	UK[6]	Adult general population	Own health	Standard gamble	Face-to-face interview with props	Standard gamble utility values are directly generated onto the 1–0 scale	2012	18 upwards
EQ-5D-Y-3L (4-15y)	US	Adult general population	7- or 10-year-old child	Discrete choice experiment involving problems with one attribute for *x* years, followed by full health for *y* years	Online survey	Modelled latent scale values, argued are directly on the 1–0 scale	2016	18 upwards
	Japan[7]	Adult general population	10-year-old child	Discrete choice experiment	computer-assisted personal interview	composite time tradeoff (cTTO)	2021	
	Slovenia[8]	Adult general population	10-year-old child	Discrete choice experiment	online DCE survey and face-to-face interviews for anchoring study	composite time tradeoff (cTTO)	2021	

	Germany[9]	Adult general population	10-year-old child	Discrete choice experiment	online DCE survey and face-to-face interviews for anchoring study	composite time tradeoff (cTTO)	2022	
	China[10]	Adult general population	10-year-old child	Discrete choice experiment	face-to-face or one-on-one computer-assisted personal interview	composite time tradeoff (cTTO)	2022	
HUI2 (5 upwards)	Canada	Parents of school-aged children (subsample of parents of childhood cancer patients)	Child aged 10 years	Standard gamble and VAS	Face-to-face interview with props	Standard gamble utility values are directly generated onto the 1–0 scale	1996	Not specified
	UK	Adult general population	Child aged 10 years	Standard gamble and VAS	Face-to-face interview with props	Standard gamble utility values are directly generated onto the 1–0 scale	2005	18 upwards
HUI3 (5 upwards)	Canada	Adult general population	Own health	Standard gamble and VAS	Face-to-face interview with props	Standard gamble utility values are directly generated onto the 1–0 scale	2002	16 upwards
QWB (unclear)	US	Adult general population	Own health	VAS	Unclear	VAS values elicited, assuming 0 = dead/worst state	2008	18 upwards

16D (12-15y)	Finland	Adolescents aged 12–15 years	Own health	VAS	Classrooms after oral instruction	Value of dead elicited on VAS	1996	12–15
17D (8-11y)	Finland	Parents	8- to 11-year-old child	VAS	Unclear	Value of dead elicited on VAS	1996	8–11
HuPS (2-4y)	Canada	Parent of children at 2 to 6 years of age and clinician	Unclear	Continuity with HUI3 scoring algorithm	Unclear	NA	2023	
IQI (0-1y)	China-HK, UK, US	General adult population and primary caregivers of infants and toddlers (0–3 years)	Adult choose for infant	DCE	Online survey	Rescaling	2020	

Note: adapted from Table 3, in Rowen et al 2020[1] and added studies after 2020.

Reference of tables and figures:

[1] D. Rowen, O. Rivero-Arias, N. Devlin, and J. Ratcliffe, "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?," (in eng), *Pharmacoeconomics,* Jan 6 2020.

[2] J. Kwon *et al.*, "Systematic Review of Conceptual, Age, Measurement and Valuation Considerations for Generic Multidimensional Childhood Patient-Reported Outcome Measures," *PharmacoEconomics,* vol. 40, no. 4, pp. 379-431, 2022/04/01 2022.

[3] J. Ratcliffe *et al.*, "Valuing the Child Health Utility 9D: Using profile case best worst scaling methods to develop a new adolescent specific scoring algorithm," (in eng), *Soc Sci Med,* vol. 157, pp. 48-59, May 2016.

[4] G. Chen, F. Xu, E. Huynh, Z. Wang, K. Stevens, and J. Ratcliffe, "Scoring the Child Health Utility 9D instrument: estimation of a Chinese child and

adolescent-specific tariff," (in eng), *Qual Life Res,* vol. 28, no. 1, pp. 163-176, Jan 2019.
[5] D. Rowen, B. Mulhern, K. Stevens, and J. H. Vermaire, "Estimating a Dutch Value Set for the Pediatric Preference-Based CHU9D Using a Discrete Choice Experiment with Duration," (in eng), *Value Health,* vol. 21, no. 10, pp. 1234-1242, Oct 2018.
[6] K. Stevens, "Valuation of the Child Health Utility 9D Index," (in eng), *Pharmacoeconomics,* vol. 30, no. 8, pp. 729-47, Aug 1 2012.
[7] T. Shiroiwa, S. Ikeda, S. Noto, T. Fukuda, and E. Stolk, "Valuation Survey of EQ-5D-Y Based on the International Common Protocol: Development of a Value Set in Japan," *Medical Decision Making,* vol. 41, no. 5, pp. 597-606, 2021/07/01 2021.
[8] V. Prevolnik Rupel and M. Ogorevc, "EQ-5D-Y Value Set for Slovenia," (in eng), *Pharmacoeconomics,* vol. 39, no. 4, pp. 463-471, Apr 2021.
[9] S. Kreimeier *et al.*, "EQ-5D-Y Value Set for Germany," *PharmacoEconomics,* vol. 40, no. 2, pp. 217-229, 2022/12/01 2022.
[10] Z. Yang *et al.*, "Estimating an EQ-5D-Y-3L Value Set for China," *PharmacoEconomics,* vol. 40, no. 2, pp. 147-155, 2022/12/01 2022.

SECTION I: Application of health-related quality of life

Chapter 2: Association between 24-hour movement behaviors and health-related quality of life in children

Published in Quality of Life Research (2022) with Dalziel, K., Carvalho, N., Xu, R., & Huang, L

Citation: Xiong, X., Dalziel, K., Carvalho, N., Xu, R., & Huang, L. (2022). Association between 24-hour movement behaviors and health-related quality of life in children. Quality of life research: an international journal of quality of life aspects of treatment, care and rehabilitation, 31(1), 231–240. https://doi.org/10.1007/s11136-021-02901-6

2.1. Study impact

This paper has received wide attention and has contributed to policy making. A media article was written for this paper, and was published on *Pursuit*, the official media platform of the University of Melbourne. Based on the findings, the media report[1] emphasized the critical importance of physical activity for children's quality of life: "*Increasing the daily level of physical activity could be the easiest 'win' for parents to improve their child's wellbeing, benefiting adolescents and children from low-income families most*". In addition, this paper was featured in 9 NEWS (TV broadcast), on 18 Aug 2021 for its relevance to playground closures during the (nationwide) Covid19 lockdown. At the same time, this research caught the attention of the Policy and Government Relations Officer in the University of Melbourne and was forwarded to the Minister for Education and Minister for Mental Health. A meeting was held with staff from the Federal Department of Education, Skills and Employment (DESE) to discuss outcomes from the research and suggestions as to how to promote the importance of activity, particularly in Covid lockdown. Key messages to different stakeholders were summarized. For example, to parents, the key message is "*more attention on exercise rather than guilt about screens*"; to policy makers, "*reopening activities such as basketball courts for high school children may be as important as playgrounds for younger children given the strong impact of physical activity on HRQoL in older children*"; to educational leaders: "*relationship with health-related QoL on school days may be larger than non-school days; formal/informal opportunities on school days*". This research has received formal acknowledgement and commendation from the Minister for Education and Minister for Mental Health: "*I acknowledge the importance of the findings*", "*fantastic to be provided with evidence that points to an effective way for parents to improve their child's wellbeing through increasing levels of physical activity*", "*I am thrilled to have announced the Active*

[1] https://pursuit.unimelb.edu.au/articles/physical-activity-is-critical-for-children-s-quality-of-life

Schools initiative at a time when our students need it most". Brochures or fliers for their internal use to broadcast the key information from this article has also been made and shared across all schools in Australia.

2.2. Abstract

PURPOSE To assess the associations between adherence to 24-hour movement behaviours guidelines and child general health and functional status measured by health-related quality of life.

METHODS The Longitudinal Study of Australian Children (2004-2016), a nationally representative sample with data available for children aged 2-15 years was used. Physical activity time, recreational screen time and sleep time were calculated from time use diaries, and classified as 'meeting guidelines' or 'not' based on the age-specific 24-hour movement guidelines. Child general health and functional status was measured using the multidimensional Pediatric Quality of Life Inventory (PedsQL). Associations between meeting guidelines and PedsQL were assessed using linear mixed effects models.

RESULTS 8,919 children were included. Each additional guideline met was associated with a 0·52 (95% confidence interval [CI]: 0·39-0·65) increase in PedsQL total score. Compared with meeting no guidelines, the effect of meeting physical activity guidelines alone (β=0·93, 95% CI: 0·42-1·44) was larger compared to meeting screen (β=0·66, 95% CI: 0·06-1·27) or sleep time (β=0·47, 95% CI: 0·04-0·89) guidelines alone. The highest increment was observed in meeting both screen time and physical activity guidelines (β=1·89, 95% CI: 1·36-2·43). Associations were stronger in children from lower-income families (β for meeting all versus none =2·88, 95% CI: 1·77-3·99) and children aged 14-15 years (β=4·44, 95% CI: 2·49-6·40).

CONCLUSIONS The integration of screen time and physical activity guidelines is associated with the highest PedsQL improvement. The association between guidelines adherence and PedsQL appears stronger for adolescents, and those from low-income families.

2.3. Introduction

During 2010 to 2012, national recommendations for physical activity and sedentary behavior in children were released in Australia, the United Kingdom, and Canada [1-4]. The recommendations in the movement behaviors are backed by evidence from physical inactivity [5-7], excessive digital media use [8, 9], insufficient sleep [10, 11] and their associated unfavorably physical, psychological, social, and cognitive health outcomes. In 2016, Canada was the first country to release integrated 24-hour movement guidelines (a mix of physical activity, screen use and sleep in each 24-hour period, please see Appendix 1 for details) for school-aged children and youth [12], followed by the

Australian Government in 2017, and New Zealand, South Africa, the United Kingdom, the United States, and by the World Health Organization (WHO) [13]. The 24-hour movement paradigm shifts the focus from individual activity components to the whole period and emphasizes the interactions between activities during a day. For example, time spent on physical activity has its own health benefit but it also has positive impacts on sleep [14] and screen time [15], whilst sleep and screen time interact with each other too [16]. The guidelines from all jurisdictions are broadly equivalent regarding the recommended time per activity, with only minor variations in wording and use of 4 or 5 years as the age cutoff for early years [13]. Taking physical activity as an example, the WHO recommends that children and adolescents aged 5-17 years should do at least an average of 60 minutes of moderate to vigorous physical activity per day across the week, mostly aerobic (see Appendix 1 for the detailed age specific guidelines**)**. The 24-hour guideline was evidence-based and supported by systematic reviews examining the relationships between and among movement behaviors, stakeholder survey and focus groups/stakeholder interviews.

Nevertheless, evidence on the association between adherence to 24-hour guideline and child general health outcomes in a population representative sample is lacking. Traditionally, health impact was primarily measured using mortality and morbidity. At the present time, the importance of health-related quality of life (HRQOL) beyond survival is widely recognized by clinicians, researchers and policy makers to assist care management and policy decisions. [17-19] Advantages of using HRQOL include evaluating the overall influence of health conditions on individual's life and capturing clinical and non-clinical benefits of care such as symptom relief.[20, 21] The multidimensional construct of HRQOL often includes physical, emotional, and social dimensions of health [17], assessing the self-reported health status and impact of health as opposed to objective biological indicators and diagnoses. HRQOL instruments can be generic or condition specific,[19] with some classified as non-preference based (e.g. PedsQL, KIDSCREEN) while others preference-based (e.g. EQ-5D-Y, CHU9D). Non-preference based instruments are mainly applied in clinical studies, population health studies, etc., while preference-based measurements are mainly used in economic evaluations.[20]

Children's HRQOL could be affected by many factors, such as socio-economic status[22], special health care needs[23], or physical activity[5]. There are a number of contributors that are also related to children's physical activity, such as socioeconomic status[24], parental physical activity[25], and family environment[26]. Several studies have investigated the relationship between HRQOL and physical activity, screen use and sleep separately[5, 9, 27] or a combination of two of them in selected age groups of children [28, 29]. To the best of our knowledge, only two previous studies directly investigated the association between 24-hour movement behaviors and HRQOL [30, 31]. In a multinational, observational study of children aged 9-11 years, children had significantly higher HRQOL when they met the screen time recommendation only, the screen plus sleep

recommendations, and all three recommendations compared with those who met none [30]. Another study followed children aged 3-5 until they were aged 9-11 years and found no significant associations between baseline 24-hour movement behaviors (capturing physical activity, screen use and sleep) and later HRQOL among 471 children [31]. The two studies had either a narrow age range or a relatively small sample with the majority of the participants from urban or relatively high socio-economic background families. Also, children's recreational screen time or sleep time in these studies was obtained from survey questionnaires rather than a time use diary, which may be subject to recall or social desirability bias [32, 33].

This study aims to investigate how adherence to the integrated 24-hour movement guidelines was associated with HRQOL in a population representative sample with children aged 2-15 years We used a nationally representative sample with a comprehensive age range and time use diaries which allowed comparison between different age groups and more accurate estimates to complement the existing evidence.

2.4. Methods

2.4.1. Study design and participants

Data from all seven waves of the Longitudinal Study of Australian Children (LSAC) were used. The LSAC data was a deidentified, publicly available, existing dataset provided by the Department of Social Services, the Australian Institute of Family Studies and the Australian Bureau of Statistics. The LSAC, which commenced in 2004, involves repeated biennial assessment ('waves') of over 10,000 children across two age cohorts (a birth cohort of 5,107 children aged 0-1 year in 2003-2004, and a kindergarten cohort of 4,983 children aged 4-5 years in 2003-2004). The LSAC used a two-stage cluster randomized design with stratification by state and then by major metropolitan center to obtain a geographically representative sample of children and their families. The LSAC sampling design and field methods are detailed elsewhere [34]. Each wave was approved by The Australian Institute of Family Studies Ethics Committee, and families provided written informed consent.

2.4.2. Time use data

Physical activity, recreational screen use, and sleep time were derived from the time use diaries which documented 24-hour use of time for each child. We have focused on children aged 2-15 years to align to the period when HRQOL and time use data were available.

Two types of time use diaries were available: a parent-completed diary for children aged 2-9 years, and a child-completed diary for those aged 10-15 years. For the parent-completed diaries, the primary caregiver, usually the mother, completed two diaries with one for a weekday and one for a weekend day. The primary carer was asked to complete a diary the weekday immediately following an

interview, as well as one weekend diary randomly selected by the interviewer to achieve a random allocation of weekdays and a random allocation of weekend days [35]. They were asked to record what the child was doing in 15-minute blocks of time, from a list of pre-coded activities in 96 blocks throughout the day [35]. The average daily recreational screen use, physical activity and sleep time were calculated as a weighted average of the week. The average missingness of the 96 time blocks was 6.56% for children aged 2-9 years. Individual's missing entries in the 96 blocks were imputed by age, sex, socioeconomic status, day of the week, wave, cohort and time block of the day using random forest technique [36]. Based on the complete dataset without time use missing values, the accuracy of the imputation models compared with the observed time use ranged from 82% to 99%.

Children aged 10-15 years old were required to self-record their own time use the day before the scheduled interview. The diary included the start time of each activity, and sleep and awake time [37]. The time use was electronically recorded by an interviewer the next day during the interview so that any uncertain entries could be clarified. Due to the concerns of survey burden, each child was required to complete only one diary on a randomly selected day by the interviewer [37]. Missingness of the time use data for sleep, screen, and physical activities ranged from 0.06% to 1.52%. This was considered to be minimal and we did not impute the missing data for children aged 10-15 years.

Based on the time use data, we defined adherence to guidelines on physical activity, recreational screen use and sleep according to the age-specific 24-hour movement guidelines (see **Appendix 1** for details). The time use diaries were not collected from the aged 6-9 years (wave 4 and 5) of the birth cohort, where data from the kindergarten cohort (6-9 years from wave 2 and 3) alone was used.

2.4.3. Health-related quality of life

We used the Pediatric Quality of Life Inventory (PedsQL) Version 4.0, an established, standardized, generic instrument[17] for HRQOL assessment in children and adolescents available across the age ranges of 2-18 years. The PedsQL measures four health dimensions: physical, emotional, social and school functioning [17]. There are 23 items (21 items for 2-4 years) in the PedsQL Inventory. In the LSAC, PedsQL was filled out by the study child's primary caregiver, who rated the frequency of each item in the past month with a 5-point Likert scale from 0 (Never) to 4 (Almost always). Items are reversed scored and linearly transformed to a 0-100 scale (0=100, 1=75, 2=50, 3=25, 4=0) in the PedsQL, with higher scores indicating better HRQOL [17]. The total score was calculated as the sum of the score of each item divided by the number of items answered. If more than 50% of the items are missing, the total score should not be computed. A 4.5 points change in the total score for parent proxy-report is considered to be clinically meaningful [17].

2.4.4. Statistical analysis

Covariates considered to be associated with HRQOL were controlled for, including age, sex, indigenous status, language spoken at home, number of siblings, household income and parental education [38-40]. Age was grouped into 2-4, 5-13 and 14-15 years corresponding to the 24-hour guideline categories. Language spoken at home was English or otherwise. Number of siblings was categorized as single child, one sibling, two or more. Household income was categorized as lowest 25%, middle 50% and highest 25% using quartiles in each wave-cohort. Highest parental education was whether any parent has bachelor's degree or above. We also adjusted for children's general health using a 2-question sequence Children With Special Health Care Needs screener [41]. The screener identified children having more than average health care needs that is expected to last more than 12 months, implicating an estimated 16% of the Australian children under 18 years of age [41].

Linear mixed modeling (or multi-level model) with random intercept for individuals was used to evaluate the association between meeting 24-hour movement guidelines and HRQOL [42]. The random intercept accounts for repeated measurements of individuals by defining the child identifier as the cluster variable [43]. The coefficients of the linear mixed effects model can be interpreted as the average difference in HRQOL in response to both within- and the inter-individual changes.[44-46]

Two regressions were run in the primary analysis, one treating guideline adherence as a continuous variable (number of guidelines met: 0,1,2,3) and the other as a categorical variable (meeting different combinations of three guidelines: none, sleep, screen, physical activity, sleep and screen, sleep and physical activity, screen and physical activity, all three). To explore potential heterogeneity, subgroup analyses were conducted by age, sex and household income for all children included, and by whether it was a school day (school day for children aged 4-15 years). All covariates in the primary analysis except for the grouping variable were included in the subgroup analyses. We also performed a sensitivity analysis by repeating our main analysis using the data without imputation of the missing values.

The Stata statistical software package (version 16.0, College Station, Texas) was used for data cleaning and analysis, and R (version 3.5.3, "caret" package) was used for data imputation.

2.5. Results

Observations with the outcome variable PedsQL missing were dropped, leaving 33,168 person-wave observations from 8,874 children aged 2-15 years old. The demographic characteristics, average time per activity, 24-hour guideline adherence, and average PedsQL scores are described by age group in Table 2-1. Overall, physical activity and sleep time were lower in older children, and recreational screen time were higher in older children. Of the three guidelines, an average of 1·6 (SD 0·9)

guidelines were met by children aged 2-15 years, 16·6% of children met all three guidelines, and 10·2% met none. The mean of the PedsQL total score was slightly lower in older age groups.

Single guideline adherence for all children and by subgroups are presented in Figure 2-1. Overall, adherence to recreational screen time recommendation was the lowest. Boys had relatively low adherence to screen time guidelines, while girls had lower adherence to physical activity guidelines. Children from the lowest household income families had the lowest percentages of adherence to all three guidelines. On non-schooldays, screen time guideline adherence was much lower compared to schooldays.

The regression analysis showed that HRQOL was positively associated with the number of guidelines met, and the PedsQL total score were 0·52 (95%CI: 0·39-0·65) higher with each additional guideline met (Table 2-2). When 'guideline adherence' was treated as a categorical variable, the PedsQL total score were 1·6 1(95%CI: 1·16-2·07) higher for children meeting all three guidelines compared with those meeting none. Meeting physical activity guidelines alone had similar effect with meeting both screen and sleep guidelines (β=0·93, 95%CI: 0·42-1·44 versus β=0·83, 95%CI: 0·34-1·32). Meeting screen time and physical activity guidelines was associated with greater QOL improvements (β=1.89, 95% CI: 1.36-2.43) compared with meeting screen time and sleep guidelines (β=0.83, 95% CI: 0.34-1.32). Results using data without imputation of the missing values were consistent with the main results (Appendix 2).

In Figure 2, guideline adherence was estimated using the total number of guidelines met whilst in Figure 3, adherence was whether each individual guideline was met. The scales of the associations were different by subgroups (Figure 2-2**,** Figure 2-3). Meeting guidelines was associated with greatest HRQOL improvements for the 14-15 years old (β for meeting all guidelines versus meeting none =4·44, 95% CI: 2·49-6·40) and the lowest 25% income group (β=2·88, 95% CI: 1·77-3·99). The associated improvements with guideline adherence were lowest in the 2-4 years old and the highest income group. We did not find notable difference by gender and by school-day. When meeting all three guidelines compared with meeting none of the guidelines, the HRQOL difference in 14-15 years old children (4.44 points) is approaching the minimal clinically important difference, 4.50 points.

For different combinations of guidelines, the patterns of the associations also differ based on age groups. Figure 3 showed that for 2-4 years old, only meeting all three guidelines were associated with HRQOL improvement (meeting one or two the guidelines are not associated with an improved HRQOL compared to meeting none). This is in contrast with 5-12 years old where meeting any of the guidelines was associated with improvement. For 14-15 years old, meeting sleep time or recreational screen time guideline alone was not associated with HRQOL improvement.

2.6. Discussion

Understanding the importance of time use for child health is now more important than ever. In 2021, the latest National Child Health Poll in Australia revealed that the number one health concern for parents is excessive screen time with more than 90% of parents reporting it as a big problem or somewhat of a problem in the community. 'Not enough exercise' was also ranked as the 7th in the top ten health concerns for children. [47] In this study, we found that meeting the recreational screen time plus physical activity guidelines has the strongest positive association with child HRQOL. Meeting the physical activity guidelines alone appears to be more important than meeting other guidelines alone overall and has a greater positive association with HRQOL when in combination with meeting other guidelines. This is consistent with previous systematic reviews which also found that physical activity was most consistently associated with positive health indicators in children aged 5-17 years [48]. One previous study including 3,040 students aged 11-18 years also found that those who were physically active every school day and low screen-based media users had higher HRQOL (PedsQL total score: boys, by 6.6 points; girls, by 7.8 points) compared with the physical inactive and high screen-based media users [28], One explanation for the stronger association with meeting physical activity guidelines would be its associated benefit in reducing screen time, building social connectedness, enhancing self-esteem and increasing sleep quality, which interact to result in better psychosocial health [7]. Teenagers aged 14-15 years and children from low household income families showed stronger relationship between time use and HRQOL. The results provided additional insight on the possible HRQOL impact for children in environment that restricts physical activities such as the unexpected lockdown during the COVID-19 pandemic.

Associations between time use behavior and HRQOL appear to vary according to different development stages. For 2-4 years old, it appears that none of the three guidelines can be neglected, namely that only when all three guidelines were met can a HRQOL improvement be expected. In children aged 14-15-years, meeting all movement guidelines has the highest HRQOL benefit comparing to meeting none, with the HRQOL difference clinically meaningful (4.44 points differences which approaches the 4.50 points minimal clinically important difference) [17]. Zhu (2019) also found that failing to meet these guidelines had stronger associations with depression and anxiety in 12-17 years compared to 6-11 years [49].

Guidelines adherence was poorer in children from low-income families, and the HRQOL improvement associated with guidelines adherence was higher for this group. Although guideline adherence is likely to contribute to the HRQOL, an caveat is that economically disadvantaged children could also have poorer physical health [40] which might also have a direct impact on HRQOL. The finding thus needs to be interpreted with caution where it is important to understand the interactions of socioeconomic status and health.

Strengths of our study include employing a large nationally representative dataset with a wide age range and measuring guideline adherence by time use diaries, which tends to provide 'the most accurate and comprehensive information' [50]. It has a potential advantage over simple recall questions in accuracy (intentionally over- or under-reporting or memory error in simple recall questions)[32, 33]. Its objectivity also enables comparability across countries and over time[51]. We have carried out a rich set of subgroup analyses, which may provide more detailed information for targeted groups.

Several limitations have been identified. While we had access to time use dairy data on physical activity, it is known that accelerometry data is more accurate, especially at capturing low intensity exercise [52], which could be important particularly for the 2-4 year age group. However, it is often not practical for population studies with large sample sizes and repeated survey design to adopt accelerometry measurement due to excessive cost [52, 53]. Due to data availability and the aim of matching to 24-hour guidelines, we tried to choose the best measure and take advantage of the unique time use diary data to answer the research question. We acknowledge that there are other tools to measure time use behavior and it would be interesting to compare the results when data are available. Most of the effect sizes in our study did not reach minimal clinically important difference. However smaller effect sizes may still be important at a population level or to groups with greater vulnerability [54]. The causal relationship between guideline adherence and HRQOL could not be determined in this study. Aiming to explore the age-specific effects of guideline adherence for a wider range of child ages, we did not explore whether previous adherence impacts future outcomes. Future research on this would be valuable. The screen time use guideline in our study focused on the recreational screen time, which is consistent with the current guideline. However, non-recreational screen time such as online learning, is increasingly used in education. Future studies on non-recreational screen use may be warranted especially considering the fast-changing movement behavior norms.

2.7. Conclusions

While meeting physical activity guidelines alone has the strongest association with HRQOL compared with meeting screen and sleep time guidelines alone, the integration of screen time and physical activity guidelines is associated with the highest HRQOL increment. The association between guidelines adherence and HRQOL appears strongest for adolescents, and those from low-income families.

2.8. Reference

[1] M. S. Tremblay *et al.*, "Canadian sedentary behaviour guidelines for the early years (aged 0–4 years)," vol. 37, no. 2, pp. 370-380, 2012.

[2] M. S. Tremblay *et al.*, "Canadian physical activity guidelines for the early years (aged 0–4 years)," *Applied Physiology, Nutrition, and Metabolism,* vol. 37, no. 2, pp. 345-356, 2012.

[3] "Start Active, Stay Active: A Report on Physical Activity for Health from the Four Home Countries' Chief Medical Officers," ed. UK Department of Health: Crown Copyright London, England, 2011.

[4] "Move and play every day national physical activity recommendations for children 0–5 years," in *Australian Governmental, Department of Health*, ed. Australian Governmental Department of Health, 2010.

[5] A. M. Marker, R. G. Steele, and A. E. Noser, "Physical activity and health-related quality of life in children and adolescents: A systematic review and meta-analysis," (in eng), *Health Psychol,* vol. 37, no. 10, pp. 893-903, Oct 2018.

[6] V. J. Poitras *et al.*, "Systematic review of the relationships between objectively measured physical activity and health indicators in school-aged children and youth," *Applied Physiology, Nutrition, and Metabolism,* vol. 41, no. 6, pp. S197-S239, 2016.

[7] D. Lubans *et al.*, "Physical Activity for Cognitive and Mental Health in Youth: A Systematic Review of Mechanisms," (in eng), *Pediatrics,* vol. 138, no. 3, Sep 2016.

[8] G. Lissak, "Adverse physiological and psychological effects of screen time on children and adolescents: Literature review and case study," (in eng), *Environ Res,* vol. 164, pp. 149-157, Jul 2018.

[9] N. Stiglic and R. M. Viner, "Effects of screentime on the health and well-being of children and adolescents: a systematic review of reviews," *BMJ open,* vol. 9, no. 1, p. e023191, 2019.

[10] J. P. Chaput *et al.*, "Systematic review of the relationships between sleep duration and health indicators in school-aged children and youth," (in eng), *Appl Physiol Nutr Metab,* vol. 41, no. 6 Suppl 3, pp. S266-82, Jun 2016.

[11] J. P. Chaput *et al.*, "Systematic review of the relationships between sleep duration and health indicators in the early years (0-4 years)," (in eng), *BMC Public Health,* vol. 17, no. Suppl 5, p. 855, Nov 20 2017.

[12] M. S. Tremblay *et al.*, "Canadian 24-Hour Movement Guidelines for Children and Youth: An Integration of Physical Activity, Sedentary Behaviour, and Sleep," (in eng), *Appl Physiol Nutr Metab,* vol. 41, no. 6 Suppl 3, pp. S311-27, Jun 2016.

[13] M. S. Tremblay, "Introducing 24-Hour Movement Guidelines for the Early Years: A New Paradigm Gaining Momentum," *Journal of Physical Activity & Health,* vol. 17, no. 1, pp. 92-95, 2020.

[14] C. Lang, N. Kalak, S. Brand, E. Holsboer-Trachsler, U. Pühse, and M. Gerber, "The relationship between physical activity and sleep from mid adolescence to early adulthood. A systematic review of methodological approaches and meta-analysis," *Sleep Medicine Reviews,* vol. 28, pp. 32-45, 2016/08/01/ 2016.

[15] S. Park, "Associations of physical activity with sleep satisfaction, perceived stress, and problematic Internet use in Korean adolescents," *BMC Public Health,* vol. 14, no. 1, p. 1143, 2014/11/05 2014.

[16] C. A. Magee, J. K. Lee, and S. A. Vella, "Bidirectional Relationships Between Sleep Duration and Screen Time in Early Childhood," *JAMA Pediatrics,* vol. 168, no. 5, pp. 465-470, 2014.

[17] J. W. Varni, T. M. Burwinkle, M. Seid, and D. Skarr, "The PedsQL™* 4.0 as a Pediatric Population Health Measure: Feasibility, Reliability, and Validity," *Ambulatory Pediatrics,* vol. 3, no. 6, pp. 329-341, 2003/11/01/ 2003.

[18] P. M. Fayers and D. Machin, *Quality of life: the assessment, analysis and interpretation of patient-reported outcomes*. John Wiley & Sons, 2013.

[19] G. H. Guyatt, D. H. Feeny, and D. L. Patrick, "Measuring health-related quality of life," (in eng), *Ann Intern Med,* vol. 118, no. 8, pp. 622-9, Apr 15 1993.

[20] P. Dolan, "The measurement of health-related quality of life for use in resource allocation decisions in health care," *Handbook of health economics,* vol. 1, pp. 1723-1760, 2000.

[21] I. B. Wilson and P. D. Cleary, "Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes," (in eng), *Jama,* vol. 273, no. 1, pp. 59-65, Jan 4 1995.

[22] M. S. Didsbury *et al.*, "Socio-economic status and quality of life in children with chronic disease: A systematic review," (in eng), *J Paediatr Child Health,* vol. 52, no. 12, pp. 1062-1069, Dec 2016.

[23] M. Mohler-Kuo and M. Dey, "A comparison of health-related quality of life between children with versus without special health care needs, and children requiring versus not requiring psychiatric services," (in eng), *Qual Life Res,* vol. 21, no. 9, pp. 1577-86, Nov 2012.

[24] R. Stalsberg and A. V. Pedersen, "Effects of socioeconomic status on the physical activity in adolescents: a systematic review of the evidence," *Scandinavian Journal of Medicine & Science in Sports,* https://doi.org/10.1111/j.1600-0838.2009.01047.x vol. 20, no. 3, pp. 368-383, 2010/06/01 2010.

[25] G. J. Welk, K. Wood, and G. Morss, "Parental influences on physical activity in children: An exploration of potential mechanisms," *Pediatric exercise science,* vol. 15, no. 1, pp. 19-33, 2003.

[26] V. Cleland, A. Timperio, J. Salmon, C. Hume, A. Telford, and D. Crawford, "A longitudinal study of the family physical activity environment and physical activity among youth," ed: SAGE Publications Sage CA: Los Angeles, CA, 2011.

[27] Q. Xiao *et al.*, "Sleep characteristics and health-related quality of life in 9- to 11-year-old children from 12 countries," *Sleep Health,* vol. 6, no. 1, pp. 4-14, Feb 2020.

[28] K. E. Lacy *et al.*, "Screen time and physical activity behaviours are associated with health-related quality of life in Australian adolescents," *Quality of Life Research,* vol. 21, no. 6, pp. 1085-1099, 2012/08/01 2012.

[29] N. Motamed-Gorji *et al.*, "Association of screen time and physical activity with health-related quality of life in Iranian children and adolescents," *Health Qual Life Outcomes,* vol. 17, no. 1, p. 2, Jan 5 2019.

[30] H. Sampasa-Kanyinga *et al.*, "Associations between meeting combinations of 24-h movement guidelines and health-related quality of life in children from 12 countries," *Public Health,* vol. 153, pp. 16-24, 2017/12/01/ 2017.

[31] T. Hinkley *et al.*, "Prospective associations with physiological, psychosocial and educational outcomes of meeting Australian 24-Hour Movement Guidelines for the Early Years," (in eng), *Int J Behav Nutr Phys Act,* vol. 17, no. 1, p. 36, Mar 10 2020.

[32] J. Boase and R. Ling, "Measuring Mobile Phone Use: Self-Report versus Log Data," *Journal of Computer-Mediated Communication,* vol. 18, no. 4, pp. 508-519, 2013.

[33] R. J. Fisher, "Social desirability bias and the validity of indirect questioning," *Journal of consumer research,* vol. 20, no. 2, pp. 303-315, 1993.

[34] C. Soloff, Lawrence, D., & Johnstone, R., "LSAC sample design (Technical Paper No. 1)," *Melbourne: Australian Institute of Family Studies. ,* 2005.

[35] J. Baxter, *Children's time use in the Longitudinal Study of Australian Children: Data quality and analytical issues in the 4-year cohort*. Australian Institute of Family Studies, 2007.

[36] F. Tang, *Random Forest Missing Data Approaches*. University of Miami, 2017.

[37] J. Corey, J. Gallagher, E. Davis, and M. Marquardt, "The Times of Their Lives: Collecting time use data from children in the Longitudinal Study of Australian Children (LSAC)," *LSAC technical paper,* vol. 13, 2014.

[38] J. Liu, M. Sekine, T. Tatsuse, Y. Fujimura, S. Hamanishi, and X. Zheng, "Association among number, order and type of siblings and adolescent mental health at age 12," (in eng), *Pediatr Int,* vol. 57, no. 5, pp. 849-55, Oct 2015.

[39] T. Sanders, P. D. Parker, B. Del Pozo-Cruz, M. Noetel, and C. Lonsdale, "Type of screen time moderates effects on outcomes in 4013 children: evidence from the Longitudinal Study of Australian Children," (in eng), *Int J Behav Nutr Phys Act,* vol. 16, no. 1, p. 117, Nov 29 2019.

[40] N. J. Spurrier, M. G. Sawyer, J. J. Clark, and P. Baghurst, "Socio-economic differentials in the health-related quality of life of Australian children: results of a national study," *Australian and New Zealand Journal of Public Health,* vol. 27, no. 1, pp. 27-33, 2003.

[41] L. Huang, G. L. Freed, and K. Dalziel, "Children with special health care needs: how special are their health care needs?," *Academic Pediatrics,* 2020.

[42] G. M. Fitzmaurice and C. Ravichandran, "A Primer in Longitudinal Data Analysis," *Circulation,* vol. 118, no. 19, pp. 2005-2010, 2008.

[43] S. E. Humphrey and J. M. LeBreton, *The handbook of multilevel theory, measurement, and analysis*. American Psychological Association, 2019.

[44] A. Bell and K. Jones, "Explaining fixed effects: Random effects modeling of time-series cross-sectional and panel data," *Political Science Research and Methods,* vol. 3, no. 1, pp. 133-153, 2015.

[45] O. Torres-Reyna, "Panel data analysis fixed and random effects using Stata (v. 4.2)," *Data & Statistical Services, Priceton University,* vol. 112, 2007.

[46] J. M. Neuhaus and J. D. Kalbfleisch, "Between-and within-cluster covariate effects in the analysis of clustered data," *Biometrics,* pp. 638-645, 1998.

[47] M. Danchin, "Top 10 child health problems: what Australian parents think," in "RCH National Child Health Poll," The Royal Children's Hospital Melbourne, MelbourneMarch 2021, Available: https://www.rchpoll.org.au/wp-content/uploads/2021/03/NCHP20-Poll-report-A4_FA.pdf, Accessed on: 26 March 2021.

[48] T. J. Saunders *et al.*, "Combinations of physical activity, sedentary behaviour and sleep: relationships with health indicators in school-aged children and youth," (in eng), *Appl Physiol Nutr Metab,* vol. 41, no. 6 Suppl 3, pp. S283-93, Jun 2016.

[49] X. Zhu, J. A. Haegele, and S. Healy, "Movement and mental health: Behavioral correlates of anxiety and depression among children of 6–17 years old in the US," *Mental Health and Physical Activity,* vol. 16, pp. 60-65, 2019.

[50] D. J. Harding, "Measuring children's time use: A review of methodologies and findings," *Center for Research on Child Wellbeing. Working paper,* pp. 97-1, 1997.

[51] A. Bauman, M. Bittman, and J. Gershuny, "A short history of time use research; implications for public health," *BMC Public Health,* vol. 19, no. 2, p. 607, 2019/06/03 2019.

[52] T. Hinkley, J. Salmon, D. Crawford, A. D. Okely, and K. D. Hesketh, "Preschool and childcare center characteristics associated with children's physical activity during care hours: an observational study," (in eng), *Int J Behav Nutr Phys Act,* vol. 13, no. 1, p. 117, Nov 11 2016.

[53] Ž. Pedišić and A. Bauman, "Accelerometer-based measures in physical activity surveillance: current practices and issues," *British Journal of Sports Medicine,* vol. 49, no. 4, p. 219, 2015.

[54] D. Osoba, G. Rodrigues, J. Myles, B. Zee, and J. Pater, "Interpreting the significance of changes in health-related quality-of-life scores," *Journal of clinical oncology,* vol. 16, no. 1, pp. 139-144, 1998.

2.9. Tables and Figure

Table 2-1 Key characteristics of the sample at person-year response level

	2-4 years (N=6749)	5-13 years (N=8246)	14-15 years (N=3075)	Total (N=8919)
Demographics				
Special health care needs, yes	12.8	15.8	19.6	15.4
Female, yes	48.4	49.0	49.7	48.9
Indigenous, yes	2.3	2.4	2.0	2.4
Speaking English at home, yes	91.5	90.7	90.2	90.9
Two parent family, yes	90.7	86.2	83.4	87.1
Parental education bachelor or above, yes	46.8	46.4	46.8	46.6
Number of siblings				
Single child	14.0	8.8	11.6	10.4
1 sibling	51.4	45.7	46.7	47.3
>=2 siblings	34.7	45.5	41.6	42.3
Household income (AU$1,000 per week)	1.3(0.9)	2.0(1.5)	2.5(1.7)	1.9(1.5)
Outcomes				
Time use				
Screen time (hours per day)	2.0(1.4)	2.7(2.2)	3.7(3.0)	-
Physical activity (hours per day)	2.1(1.5)	1.8(1.6)	1.1(1.4)	-
Sleep time (hours per day)	11.3(1.9)	10.0(1.4)	9.1(1.6)	-
24-hour guidelines adherence				
Number of guidelines met	1.3(0.7)	1.8(0.9)	1.3(0.9)	1.6(0.9)
Percentage meeting three guidelines	5.4	22.1	10.7	16.6
Percentage meeting no guideline	11.2	8.5	19.1	10.2
PedsQL total score	81.9(10.1)	79.3(13.1)	78.5(14.6)	79.9(12.6)

Notes: Data are % or mean (SD). The percentages are calculated based on person-wave observations. N is the number of unique children. Screen time was recreational screen time. According to 24-hour movement guidelines, total sleep time was used for 0-4 years, and sleep time at night was used for 5-15 years. Household income prior to 2016 was inflated to 2016 Australian dollars. PedsQL total score ranges from 0-100 points.

Table 2-2 Association between meeting 24-hour movement guidelines and HRQOL

	Adherence to guidelines as continuous variable (model 1)		Adherence to guidelines as categorical variable (model 2)	
	β coefficient (95%CI)	p value	β coefficient (95%CI)	p value
Special health care needs (reference: no)	-5.34 (-5.70, -4.97)	<0.001	-5.33 (-5.69, -4.97)	<0.001
Female (reference: male)	0.28 (-0.13, 0.69)	0.182	0.30 (-0.11, 0.71)	0.155

Indigenous (reference: no)	-2.81 (-4.07, -1.56)	<0.001	-2.82 (-4.08, -1.57)	<0.001
Speak English at home (reference: no)	3.01 (2.33, 3.70)	<0.001	3.00 (2.32, 3.68)	<0.001
Two parent family (reference: no)	2.02 (1.53, 2.50)	<0.001	2.01 (1.52, 2.49)	<0.001
Parental education bachelor or above (reference: no)	0.34 (-0.05, 0.72)	0.086	0.34 (-0.04, 0.73)	0.081
Number of siblings (reference: single child)				
One sibling	0.38 (-0.15, 0.92)	0.160	0.38 (-0.16, 0.91)	0.166
>=2 siblings	1.01 (0.44, 1.57)	<0.001	1.00 (0.44, 1.57)	0.001
Age group (reference: 2-4 years)				
5-13 years	-2.18 (-2.44, -1.93)	<0.001	-2.35 (-2.62, -2.07)	<0.001
14-15 years	-2.40 (-2.81, -1.98)	<0.001	-2.49 (-2.92, -2.07)	<0.001
Income group (reference: lowest 25%)				
Middle 50%	0.72 (0.36, 1.07)	<0.001	0.71 (0.36, 1.07)	<0.001
Highest 25%	1.27 (0.82, 1.72)	<0.001	1.26 (0.81, 1.71)	<0.001
Number of guidelines met	0.52 (0.39, 0.65)	<0.001	-	-
Guidelines met (reference: none)	-	-		
Sleep only	-	-	0.47 (0.04, 0.89)	0.032
Screen time only	-	-	0.66 (0.06, 1.27)	0.031
Physical activity only	-	-	0.93 (0.42, 1.44)	<0.001
Screen time+sleep	-	-	0.83 (0.34, 1.32)	0.001
Physical activity+sleep	-	-	1.15 (0.71, 1.59)	<0.001
Screen time+physical activity	-	-	1.89 (1.36, 2.43)	<0.001
All met	-	-	1.61 (1.16, 2.07)	<0.001

Note: HRQOL means health-related quality of life. Linear mixed effects model is used. Model 1 treated guidelines adherence as a continuous variable (x=0,1,2,3). Model 2 treated guidelines adherence as a categorical variable (x= none, sleep, screen, physical, sleep+screen, sleep+physical, screen+physical, all three).

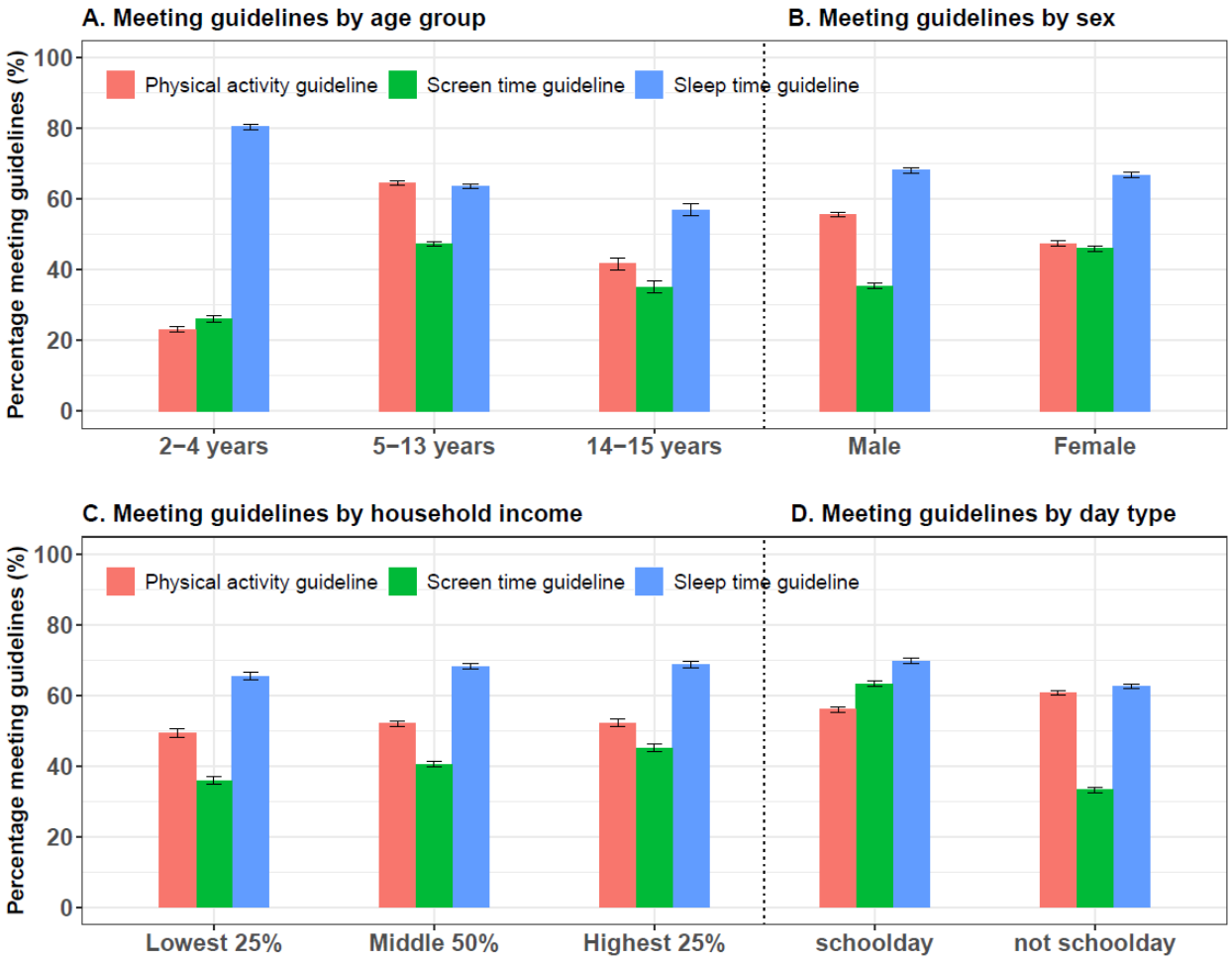


Figure 2-1 Single movement guideline adherence

Note: 'Whether a school-day' data is available for 4-15 years old. All other data are available for 2-15 years old. The error bars represent the 95% confidence intervals of the proportions of guidelines adherence.

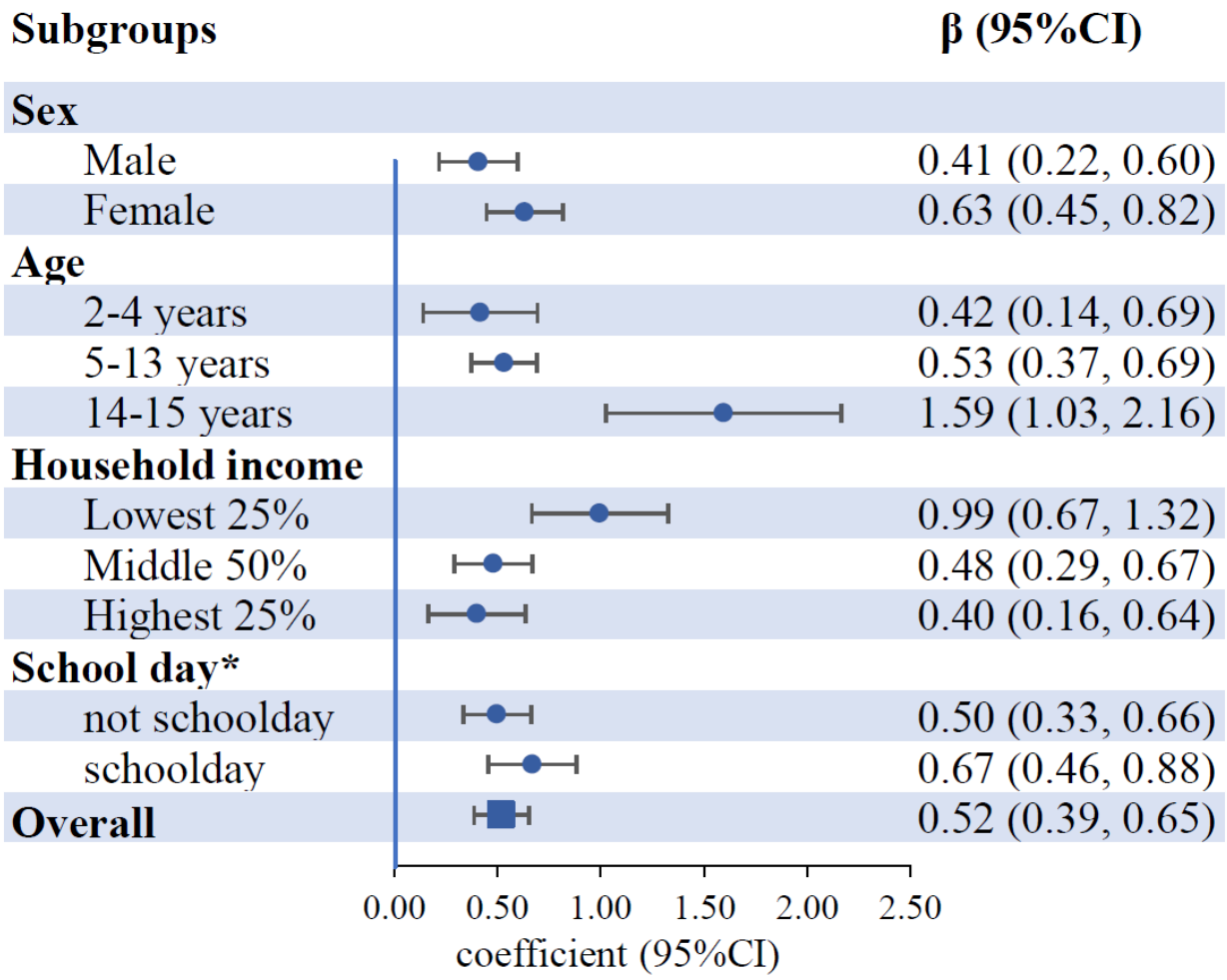


Figure 2-2 Association between meeting individual movement guidelines and HRQOL in subgroups

Note: * 'Whether a school-day' data is available for 4-15 years old. All other data are available for 2-15 years old. In this figure the model treated guideline adherence as a continuous variable: the total number of guidelines met. All the covariates in the primary analysis except the grouping variable were included in the subgroup analyses. The covariates include special health care needs, female, Indigenous status, speak English at home, two parent family, parental education, number of siblings, age group, and income group.

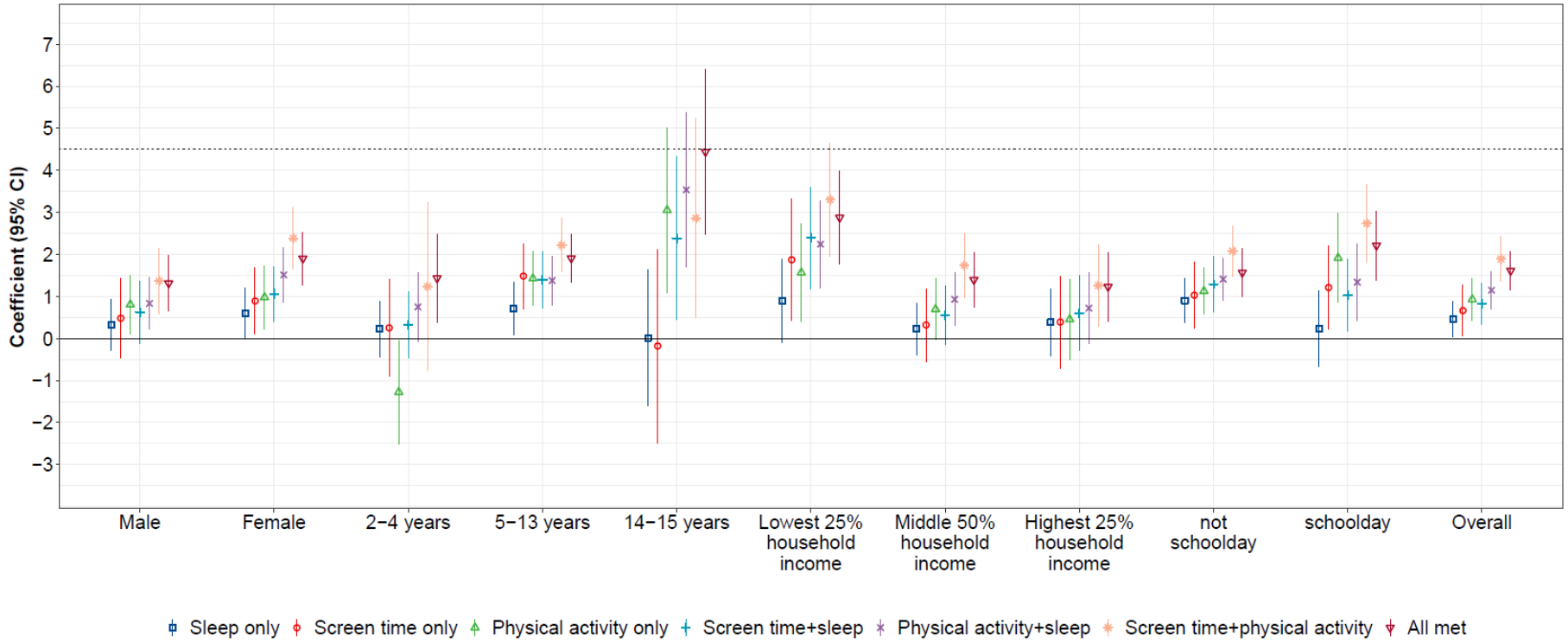


Figure 2-3 Association between meeting combinations of movement guidelines and HRQOL in subgroups

Note: In this figure the model treated guideline adherence as a categorical variable: meeting different combinations of three guidelines: none, sleep, screen, physical activity, sleep and screen, sleep and physical activity, screen and physical activity, all three. Reference group is "none". 'Whether a school-day' data is available for 4-15 years old. All the covariates in the primary analysis except the grouping variable were included in the subgroup analyses. The covariates include special health care needs, female, Indigenous status, speak English at home, two parent family, parental education, number of siblings, age group, and income group.

2.10. Supplementary materials

2.10.1. Appendix 1 The time limits in the guidelines used to define adherence

Table S 1. The time limitations of the guidelines used to define adherence in this study

<table>
<tr><th></th><th>2-4 years</th><th>5-13 years</th><th>14-15 years</th></tr>
<tr><td>Physical activity</td><td>at least 180 minutes spent in a variety of physical activities</td><td colspan="2">accumulating 60 minutes of moderate to vigorous physical activity</td></tr>
<tr><td>Screen time</td><td>no more than 1 hour sedentary screen time per day</td><td colspan="2">no more than 2 hours recreational sedentary screen time per day</td></tr>
<tr><td rowspan="2">Sleep</td><td>aged 2 years: 11 to 14 hours total sleep</td><td rowspan="2">an uninterrupted 9 to 11 hours of sleep per night</td><td rowspan="2">an uninterrupted 8 to 10 hours of sleep per night</td></tr>
<tr><td>aged 3-4 years: 10-13 hours total sleep</td></tr>
</table>

Note: 1. Age 5 is included in 'under 5 years' and '5 to 17 years' in Australian 24-hour movement guidelines. We choose age 5 to belong to the older group to compare with guidelines in other countries (e.g. Canada) and World Health Organizations. 2.The activities coded to each behaviour category are detailed in Table S3.

Detailed 24-hour movement guidelines in Australia *(extraction for 2-15 years)*

1.1. 24-hour movement guidelines in Australia for the early years (under 5 years)

Available from https://www1.health.gov.au/internet/main/publishing.nsf/Content/ti-0-5years

Toddlers (aged 2 years)

• Physical activity: At least 180 minutes spent in a variety of physical activities including energetic play, spread throughout the day; more is better;

• Sedentary Behaviour: Not being restrained for more than 1 hour at a time (e.g., in a stroller, car seat or high chair) or sitting for extended periods. Sedentary screen time should be no more than 1 hour; less is better. When sedentary, engaging in pursuits such as reading and storytelling with a caregiver is encouraged; and

• Sleep: 11 to 14 hours of good quality sleep, including naps, with consistent sleep and wake-

up times.

Pre-schoolers (aged 3–5 years)

• Physical activity: At least 180 minutes spent in a variety of physical activities, of which at least 60 minutes is energetic play, spread throughout the day; more is better;

• Sedentary behaviour: Not being restrained for more than 1 hour at a time (e.g., in a stroller or car seat) or sitting for extended periods. Sedentary screen time should be no more than 1 hour; less is better. When sedentary, engaging in pursuits such as reading and storytelling with a caregiver is encouraged; and

• Sleep: 10 to 13 hours of good quality sleep, which may include a nap, with consistent sleep and wake-up times.

1.2. 24-hour movement guidelines in Australia for Children and Young People (5-17 years)

Available from https://www1.health.gov.au/internet/main/publishing.nsf/Content/health-24-hours-phys-act-guidelines

• Physical activity: Accumulating 60 minutes or more of moderate to vigorous physical activity per day involving mainly aerobic activities; several hours of a variety of light physical activities;

• Sedentary behaviour: Limiting sedentary recreational screen time to no more than 2 hours per day; breaking up long periods of sitting as often as possible;

• Sleep: An uninterrupted 9 to 11 hours of sleep per night for those aged 5–13 years and 8 to 10 hours per night for those aged 14–17 years; and consistent bed and wake-up times.

2.10.2. Appendix 2 Sensitive analysis using data before imputation

Table S 2. Sensitive analysis before imputation.

	Adherence to guidelines as continuous variable (model 1)		Adherence to guidelines as categorical variable (model 2)	
	β coefficient	95%CI	β coefficient	95%CI

Special health care needs (reference: no)	-5.35***	-5.71, -4.99	-5.34***	-5.70, -4.98
Female (reference: male)	0.27	-0.14, 0.68	0.29	-0.12, 0.70
Indigenous (reference: no)	-2.81***	-4.06, -1.55	-2.81***	-4.07, -1.56
Speak English at home (reference: no)	2.98***	2.29, 3.66	2.98***	2.30, 3.66
Two parent family (reference: no)	2.04***	1.56, 2.52	2.03***	1.54, 2.51
Parent with bachelor degree or above (reference: no)	0.33	-0.06, 0.71	0.33	-0.05, 0.72
Number of siblings (reference: single child)				
One sibling	0.38	-0.16, 0.92	0.38	-0.16, 0.92
>=2 siblings	1.00***	0.43, 1.56	0.99***	0.43, 1.56
Age group (reference: 2-4 years)				
5-13 years	-2.16***	-2.42, -1.91	-2.30***	-2.58, -2.02
14-15 years	-2.45***	-2.87, -2.04	-2.50***	-2.92, -2.08
Income group (reference: lowest 25%)				
Middle 50%	0.71***	0.35, 1.07	0.71***	0.35, 1.06
Highest 25%	1.27***	0.82, 1.72	1.27***	0.82, 1.72
Number of guidelines met	0.59***	0.46, 0.73	-	-
Guidelines met (reference: none)	-	-		
Sleep only	-	-	0.52*	0.12, 0.93
Screen time only	-	-	0.69*	0.16, 1.21
Physical activity only	-	-	0.79***	0.33, 1.24
Screen time+sleep	-	-	0.89***	0.42, 1.37
Physical activity+sleep	-	-	1.28***	0.84, 1.72
Screen time+physical activity	-	-	1.61***	1.15, 2.08
All met	-	-	1.80***	1.34, 2.25

Note: * for p<0.05, ** for p<0.01, and *** for p<0.001.

2.10.3. Appendix 3 Allocation of pre-determined LSAC time-use categories to physical activity and screen time

Table S 3. Allocation of pre-determined LSAC time-use categories to physical activity and screen time

Wave 1 of K (4-5 years)	Wave 2, 3 of K (6-9 years)	Wave 2, 3 of B (2-5 years)
Physical activity	**Physical activity**	**Physical activity**
Walk for travel or for fun	Walking (for travel or fun)	Walking (for travel or fun)
Ride bicycle, trike etc (travel or fun)	Riding bicycle, scooter, roller blades etc (for travel or fun)	Riding bicycle, scooter, roller blades etc. (for travel or fun)
Other play, Other Activities*0.5[a]	Active free play (e.g. running, climbing, ball game)	Active free play (e.g. running, climbing, ball game)
Other exercise - swim / dance/ run about	Organised sport/physical activity (e.g. swim, dance, Auskick)	
Screen time	**Screen time**	**Screen time**
Watching TV, video, DVD, movie	Watching TV, video, DVD, movie	Watching TV, video, DVD, movie
Use computer / computer games	Using computer/computer game	Using computer, computer game
Other activities	**Other activities**	**Other activities**
Sleeping, napping	Sleeping, napping	Sleeping, napping
Not sure what child was doing	Not sure what child was doing	Not sure what child was doing
Awake in bed	Awake in bed	Awake in bed
Eating, drinking, being fed	Eating and drinking	Eating, drinking, being fed
Bathe, dress, hair care, health care	Bathing, dressing, hair care, health care	Bathing, dressing, hair care, health care
Do nothing, bored / restless	Doing nothing, bored/restless	Doing nothing, bored/restless

Crying, upset, tantrum	Sulking, upset	Crying, upset, tantrum
Destroy things, create mess	Arguing, fighting	Arguing, fighting
Held, cuddled, comforted, soothed	Being hugged, comforted, helped to calm down	Destroying things, creating mess
Being reprimanded, corrected	Being reprimanded, corrected	Being reprimanded
Listening to tapes, CD's, radio, music	Listening to tapes, CDs, radio, music	Being held, (cuddled), comforted, soothed
Read a story, talk / sing, talked / sung to	Being read to or told a story	Listening to tapes, CDs, radio, music
Colour, look at book, educational game	Reading or looking at book by self	Read a story, talk/sing, talked/sung to
Being taught to do chores, read, etc	Quiet free play (e.g. board game, craft, dress-ups)	Drawing/colouring, looking at book, etc.
Visiting people, special event, party	Helping with chores, jobs	Quiet free play (e.g. board game, craft, dress-ups)
Organised lessons / activities	Visiting people, special event or outing	Being taught to do chores
Travel in pusher or on bicycle seat	Other organised lesson/activity (e.g. music, drama)	Visiting people, special event, outing
Travel in car / another household vehicle	Travel in car	Other organised lesson/activity (e.g. music, drama)
Travel on public transport, ferry, plane	Travel on public transport	Travel in pusher/bicycle seat
Taken places with adult (eg shopping)	Being taken places with adult (e.g. shopping)	Travel in car
		Travel on public transport
		Being taken places with adult (e.g. shopping)
Wave 4 of K (age 10-11 years)	**Wave 5 of K (12-13 years)**	**Wave 6,7 of B&K (age 10-15 years)**
Physical activity[b]	**Physical activity**	**Physical activity**
Organised team sports and training i.e--football, bascket ball, netball, cricket	Organised team sports and training	Archery / Shooting sports

Organised individual sport i.e. swimming, dancing, tennis, marital arts, gymnastics

Taking Pet for a walk

By foot

By bike, scooter, skateboard etc.

Ball games, riding a bike, scooter, skateboard, skipping, running, chasing

Organised individual sport and training

Walking pets / playing with pets

By foot

By bike, scooter, skateboard etc.

Unstructured active play

Active club activities

Athletics / Gymnastics

Fitness / Gym / Exercise

Ball Sports

Martial arts / Dancing

Motor Sports / Roller Sports / Cycling

Water/Ice/Snow Sports

Organised team sports and training other

Archery / Shooting sports (individual)

Athletics / Gymnastics (individual)

Fitness / Gym / Exercise (individual)

Martial arts / Dancing (individual)

Motor Sports / Roller Sports / Cycling (individual)

Ball Sports (individual)

Water/Ice/Snow Sports (individual)

Organised individual sport and training other

Archery / Shooting sports (unstructured)

Athletics / Gymnastics (unstructured)

Fitness / Gym / Exercise (unstructured)

Ball Sports (unstructured)

Martial arts / Dancing (unstructured)

Motor Sports / Roller Sports / Cycling (unstructured)

Water/Ice/Snow Sports (unstructured)

Unstructured active play Other

Walking pets/playing with pets

Active club activities

By foot

By bike, scooter, skateboard etc

Recreational Screen time

Electronic media, games, computer use

Computer games - internet

Computer games - not internet

Xbox, Playstation, Nintendo, WII etc

Internet not covered elsewhere

TV/DVD

Texting, email, social networking such--facebook/twitter

Skype or Webcam

Recreational Screen time

Playing games

Watching TV programs or movies/videos

Spending time on social networking sites

Downloading/posting media (e.g. music, videos, applications)

Internet shopping (excluding downloading/posting media)

General Internet browsing (excluding homework)

General application use (e.g. Microsoft Office; excluding homework)

Electronic device use nec.

Video chatting (e.g. Skype)

Online chatting / Instant messaging

Recreational Screen time

Playing games (electronic device)

Playing games (Electronic device) nfd

Watching TV programs or movies/videos

Spending time on social networking sites

Downloading/posting media

Internet shopping

General Internet browsing

General application use

Electronic device use nec

Video chatting

Online chatting / Instant messaging

Non-recreational Screen time

Computer for homework - internet

Computer for homework - not internet

Other activities

Eating/drinking

Personal/Health care

Bathing, dressing, toileting, teeth bru

Dentist, Doctor, Chiropractor, Physio e

Chores

Making own bed, tidying own room

Making, preparing own food

Getting self-ready, packing/unpacking own school/sports bag

Cleaning, tidying other rooms

Cooking, meal preparation, making lunch, setting table for others

Washing dishes, stacking and emptying dishwasher

Gardening, putting out the bin

Taking care of siblings, other children

Taking care of pets

Non-recreational Screen time

Creating/maintaining websites (excluding social networking profile)

Texting/emailing

Other activities

Retailing (including fast food)

Pamphlet delivering

Umpiring/refereeing

Car washing

Gardening / lawn mowing

Babysitting

Animal care

Working in a family business or farm

Work nec.

Volunteering

Eating/drinking

Cleaning teeth

Showering/bathing

Getting dressed / getting ready

Non-recreational Screen time

Doing homework (electronic device)

Creating/maintaining websites

Texting/emailing

Other activities

Retailing

Hospitality (including fast food)

Clerical/office

Labourers and related workers

Gardening / lawn mowing

Babysitting

Apprenticeships/trades persons

Working in a family business or farm

Work Other

Umpiring (work)

Car washing (work)

Animal care (work)

Volunteering (work)

Eating/drinking

Active Activities

Scouts, girl guides, cadets, youth groups etc.

Shopping

Going out to museums, cultural events, fairs, community events

Cinema

going to Live Sporting Events

Non-Active Activities

Private music, language, religion, tutoring

Listening to music, CDs, playing music for lesiure

Reading or being read to for leisure

Board or card games, puzzles, toys, art and craft

Non-Active Club Activities i.e. Chess Club, art/craft groups

Doing nothing

Sleeping/napping

Homework (not on computer) including music practice

School Lessons

Communication

Personal care nec.

Doctor

Dentist

Physiotherapist / Chiropractor

Medical/Health care nec.

Cleaning/tidying

Laundry/clothes care

Food/drink preparation

Food/drink clean up

Gardening / lawn mowing

Animal care (excluding active play)

Home maintenance

Taking care of siblings

Chores nec.

Shopping

Going out to a concert, play, museum, art gallery, community or school event, an amusement park etc.

Religious activities / ritual ceremonies

Cleaning teeth

Showering/bathing

Getting dressed / getting ready

Personal care nec

Doctor

Dentist/Orthodontist

Physiotherapist / Chiropractor

Medical/Health care

Personal care/Medical/Health Care nec.

Cleaning/tidying

Laundry/clothes care

Clothes making

Food/drink preparation

Food/drink clean up

Gardening (maintenance chores)

Cleaning grounds/garage/shed/outside of house (chore

Pool care (chores)

Talking face to face

Talking on a landline phone

Talking on a mobile phone

Travel

By private car

Travel by public transport such as bus,train, tram, ferry, taxi, plane

Other

Attending live sporting events

Active activities nec

Private music lessons/practice, academic tutoring

Listening to music

Playing musical instruments or singing for leisure

Reading or being read to for leisure

Unstructured non-active play

Non-active club activities

Doing nothing

Sleeping/napping (not end of day bed-time)

Doing homework (not via electronic devices)

Non-active activities nec.

Doing homework

School lessons

Talking face-to-face (in person not via electronic devices)

Talking on a landline phone (not video chat)

Talking on a mobile phone (not video chat)

Non-verbal interaction (e.g. cuddles)

Negative face-to-face communication

Animal care

Home maintenance

Design/Home Improvement

Heat/water/power upkeep

Car/boat/bike care

Selling/disposing of household assets

Rubbish/Recycling

Packing

Household management Other

Taking care of siblings (chores)

Chores nec

Shopping

Shopping

Purchasing consumer goods

Purchasing durable goods

Window shopping

Purchasing repair services

Purchasing administrative services

Purchasing personal care services

Communication nec.

By private motor vehicle/bike

By public/chartered transport such as bus, taxi or aeroplane

Travel nec.

Filling out the diary

Other

Purchasing other services

Attendance at movies / cinema

Attendance at concert/theatre

Attendance at museum / exhibition / art gallery

Attendance at zoo / animal park / botanic garden

Attendance at other mass events

Going out nec

Religious practice

Weddings, funerals, rites of passage

Religious activities / ritual ceremonies nec

Attending live sporting events

Active activities nec

Private music lessons/practice, academic tutoring

Listening to music

Playing musical instruments or singing for leisure

Reading or being read to for leisure

Chess, card, paper and board games / crosswords

Games of chance / gambling

Hobbies, collections

Handwork crafts (excl. clothes making)

Arts

Unstructured non-active play nec

Attend courses (excluding school /university)

Clubs

Religious groups

Doing nothing

Sleeping/napping (not end of day bed-time)

Doing homework (not via electronic devices)

Non-active activities nec

School lessons

Talking face-to-face

Talking on a landline phone

Talking on a mobile phone

Non-verbal interaction

Negative face-to-face communication

Communication nec

By private motor vehicle/bike

By public/chartered transport

Travel nec

Illegal activities

Filling out the diary

	Other
	Uncodeable activity;

Note: [a] The absence of the category 'Active free play' in Wave 1 of the K-cohort (4/5 years), which was available in all other waves (2-9 years), produced marked inconsistencies in physical-activity time. This issue was resolved by allocating 50% of the time in the 'Other play, other activities' category to *Physical Activity* in Wave 1 of the K-cohort.

[b] For 5-15 years old, the intensity of physical activity is classified according to *The Compendium of Physical Activities* (https://sites.google.com/site/compendiumofphysicalactivities/Activity-Categories). When it was not clear the activity involved physical activity to a moderate level, we took a conservative approach and excluded such activity from the computation of physical activity time.

Chapter 3: Cost-effectiveness of Prednisolone to Treat Bell Palsy in Children: An Economic Evaluation Alongside a Randomized Controlled Trial.

Published in Neurology (2023) with Huang, L., Herd, D. W., Borland, M. L., Davidson, A., Hearps, S., Mackay, M. T., Lee, K. J., Dalziel, S. R., Dalziel, K., Cheek, J. A., & Babl, F. E.

Citation: Xiong, X., Huang, L., Herd, D. W., Borland, M. L., Davidson, A., Hearps, S., Mackay, M. T., Lee, K. J., Dalziel, S. R., Dalziel, K., Cheek, J. A., & Babl, F. E. (2023). Cost-effectiveness of Prednisolone to Treat Bell Palsy in Children: An Economic Evaluation Alongside a Randomized Controlled Trial. Neurology, 100(24), e2432–e2441. https://doi.org/10.1212/WNL.0000000000207284

3.1. Abstract

Background and Objectives: Bell's palsy is the third most frequent diagnosis in children with sudden onset neurological dysfunction. The cost-effectiveness of treating Bell's palsy with prednisolone in children is unknown. We aimed to assess the cost-effectiveness of prednisolone in treating Bell's Palsy in children compared with placebo.

Methods: This economic evaluation was a prospectively planned secondary analysis of a double-blinded, randomized, placebo-controlled superiority trial (BellPIC) conducted from 2015 to 2020. Time horizon was 6 months since randomization. Children aged 6 months to <18 years who presented within 72 hours of onset of clinician diagnosed Bell's palsy and who completed the trial were included (N=180). Interventions were oral prednisolone, or taste matched placebo administered for 10 days. Incremental cost-effectiveness ratio comparing prednisolone with placebo was estimated. Costs were considered from a healthcare sector perspective and included Bell's palsy related medication cost, doctor visits and medical tests. Effectiveness was measured using quality-adjusted life-years (QALYs) based on Child Health Utility 9D. Nonparametric bootstrapping was performed to capture uncertainties. Pre-specified sub-group analysis by age 12-18 years versus <12 years was conducted.

Results: The mean cost per patient was A$760 in the prednisolone group and A$693 in the placebo group over the 6-month period (difference A$66, 95% confidence interval [CI]: -A$47, A$179). QALYs over 6-months was 0.45 in the prednisolone group and 0.44 in the placebo group (difference 0.01, 95%CI: -0.01,

0.03). The incremental cost to achieve one additional recovery was estimated to be A$1577 using prednisolone compared with placebo, and cost per additional QALY gained was A$6625 using prednisolone compared with placebo. Given a conventional willingness-to-pay threshold of A$50,000 per QALY gained (equivalent to US$35,000 or £28,000), prednisolone is very likely cost-effective (probability is 83%), . Sub-group analysis suggests that this was primarily driven by the high probability of prednisolone being cost-effective in children aged 12-18 years (probability is 98%) and much less so for those <12 years (probability is 51%).

Discussion: This provides new evidence to stakeholders and policy makers when considering whether to make prednisolone available in treating Bell's palsy in children aged 12-18 years.

Trial Registration: Australian New Zealand Clinical Trials Registry ACTRN12615000563561.

Key word Bell's palsy, steroid, prednisolone, children, cost-effectiveness, economic evaluation

3.2. Introduction

Bell's palsy is the third most frequent diagnosis in children with sudden onset neurological dysfunction[1], and the most common acute facial paralysis for people of all ages.[2] In the United Kingdom, the incidence rate is more than 6 per 100 000 person-years in children aged <14 years and more than 20 per 100 000 person-years in people aged 15-29 years.[3] The symptoms of Bell's palsy can impact the functioning of face, mouth and eyes resulting in impaired verbal communication and social interaction.[4] Children with Bell's palsy have been reported to be distressed, embarrassed and treated differently because of the facial condition.[4]

Steroids are inexpensive, and may reduce the inflammatory process, neural edema and compression of the nerve in the facial canal.[5] In adults there is high level evidence that steroids improve recovery in Bell's palsy[6, 7] and the American Academy of Otolaryngology-Head and Neck Surgery Foundation published a clinical practice guideline recommending the use of oral steroids for Bell's palsy patients 16 years and older within 72 hours of symptom onset.[8] Whether steroids may also be appropriate for treating Bell's palsy in children is unknown.[9] One systematic review of six studies reported that the role of steroid treatment for Bell's palsy in children is inconclusive.[10]

In line with the effectiveness evidence and in terms of cost-effectiveness, treating Bell's palsy with steroids has been suggested to be cost-effective in adults. One economic evaluation based on a National Institute for Health Research commissioned trial conducted in 2004-2006 found that compared with no prednisolone, prednisolone was on average less costly and more effective (77% probability of being cost effective at £30,000 willingness-to-pay threshold).[11] However, there is no evidence regarding the cost-effectiveness of using steroids to treat Bell's palsy in children.

Prednisolone is a commonly used steroid. We recently conducted a double-blinded, randomized, placebo-controlled superiority trial (BellPIC) comparing prednisolone with placebo for the treatment of Bell's palsy in children.[12] At 1 month post randomization, no statistically significant difference was found in complete recovery (49% versus 57% for prednisolone compared with placebo), which is the primary outcome of the trial. Complete recovery rates at 3 and 6 months were not significantly different either (90% versus 85%, and 99% versus 93% respectively). In the current paper, we focused on comparing the cost and effectiveness of using prednisolone versus placebo to facilitate healthcare planning decisions and comparison across health conditions, where effectiveness is mainly estimated using a generic health-related quality of life instrument collected out to six months post randomization.

3.3. Methods

An economic evaluation was prospectively planned alongside the randomized clinical trial,[9, 12] conducted following the Second Panel on Cost-Effectiveness in Health and Medicine guidelines.[13] The reporting of the economic evaluation followed the updated Consolidated Health Economic Evaluation Reporting Standards (CHEERS).[14] The cost-effectiveness analysis was conducted from a healthcare sector perspective, and the time horizon is the 6-month period since randomization.

3.3.1. Standard Protocol Approvals, Registrations, and Patient Consents

The trial was approved by the institutional ethics committee at the RCH (HREC/15/RCHM/V4) and received governance approval by the institutional ethics offices at each participating site. Written informed consent was obtained for each participant from a parent or legal guardian and the child if deemed competent. The study was registered with the Australian New Zealand Clinical Trials Registry ACTRN12615000563561.

3.3.2. Data

The full trial protocol[9] and main study findings[12] have been published elsewhere. Briefly, the Bell's Palsy in Children (BellPIC) trial was a randomized, double-blinded, placebo-controlled trial of the use of prednisolone to improve recovery from Bell's palsy. Study sites were 10 hospitals in Australia and 1 in New Zealand in the PREDICT (Paediatric Research in Emergency Department International Collaborative) research network.[15] Patients aged 6 months to <18 years, diagnosed by an emergency department (ED) clinician with Bell's palsy and onset of symptoms less than 72 hours prior to randomization were recruited. A 10-day treatment without dosage taper was administered. Participants received either oral prednisolone 1 mg/kg/day (based on weight categories) to a maximum of 50mg/day or taste matched placebo for 10 days. The primary outcome of the trial was complete recovery defined by House Brackmann facial paralysis scales at 1 month, with grades 1, 2, 3, 4, 5 and 6 indicating normal (complete recovery), mild dysfunction, moderate dysfunction, moderately severe dysfunction, severe dysfunction and total paralysis respectively.[16] Health-related quality of life as a secondary outcome was collected for the economic evaluation as specified in the published protocol at 1, 3 and 6 months post randomization or until the participant fully recovered.

3.3.3. Costs

Prednisolone was supplied as Redipred oral liquid, which contains the active ingredient prednisolone (equivalent to prednisolone 5 mg/mL). Redipred, as well as the taste matched placebo was supplied by

Aspen Pharmacare Pty Ltd (St Leonards, NSW, Australia). Bell's palsy related health service costs were assessed via a self-reported survey administered at month 1, 3 and 6 after randomization, and unrelated service use was not asked. The health service use categories included general practitioner (GP), hospital in-patient, hospital out-patient, or ED visits and other health services (patients were asked to describe in text), medical tests including blood tests, neuroimaging (head computed tomography (CT), head magnetic resonance imaging (MRI)) and lumbar puncture. Costs for the initial non-admitted ED visit that led to the diagnosis of Bell's palsy was also included.

Costs were estimated using the physical units of health care items used multiplied by unit costs, expressed in 2020 Australian currency price. Unit cost for GP visits was obtained from the Medical Benefits Schedule.[17] Unit cost for ED visits was obtained from the Independent Hospital Pricing Authority.[18] There were no hospital inpatient admissions reported by patients during the study period, thus costing for inpatient visits was not relevant. Detailed unit costs including those for other care items were summarized in Appendix eTable 1. The total healthcare cost over a 6-month period was the sum of all medication and health service costs during that time.

3.3.4. Effectiveness

Effectiveness outcome focused on quality-adjusted life-years (QALYs) over 6 months as QALYs are comparable across health conditions and routinely used to facilitate healthcare planning decisions. Complete recovery (defined as a House Brackmann facial paralysis scale=1) at 6 months was also considered. The QALYs were calculated based on health utilities estimated using the Child Health Utility 9D (CHU9D)[19] and the Pediatric Quality of life Inventory (PedsQL)[20].

The CHU9D is a paediatric generic preference-based measure of health-related quality of life, consisting of a descriptive system and a set of preference weights.[21] It gives a single value ranging from 0 (equivalent to dead) to 1 (equivalent to perfect health) for a health state defined by its descriptive system.[22] CHU9D is available for children aged 5-18 years (5-7 years parent-reported and 8-18 years self-reported). For children aged 2-4 years, utilities were obtained using PedsQL scores mapped to CHU9D using established mapping algorithms.[23] PedsQL is a non-preference-based quality of life instrument and is available for children aged 2-18 years (2-4 years parent-reported and 5-18 years self-reported).

Participants were followed until full recovery or 6 months post randomization, whichever occurred first. For those who recovered earlier than 6 months, it is assumed that they remained recovered for the rest of the follow-up period sustaining their last observable quality-of-life. Utility for patients at the time of

recruitment (baseline) was estimated using the mean from unrecovered patients at 1 month, assuming that baseline utilities are the same for prednisolone and placebo groups due to randomization. The QALYs over the 6-month period was estimated by calculating the area under the curve using the trapezium rule (Appendix eFigure 1).[24]

3.3.5. Cost-effectiveness

The incremental cost-effectiveness ratio, defined as the difference in cost divided by the difference in effectiveness, was estimated and interpreted as the cost per additional QALY gained at 6 months, or cost to achieve one additional complete recovery at 6 months.

A prespecified cost-effectiveness subgroup analysis by age (6 months to <12 years vs. 12-18 years) was also conducted. Age 12 was chosen as the cut-point to be consistent with the pre-planned sub-group analysis of the primary outcome.[12]

3.3.6. Missing data

Missingness including CHU9D utilities (15.2% missing) and facial paralysis scales (7.8% missing) were imputed using multiple imputation with predictive mean matching within chained equations, assuming that data were missing at random (missing at random assumption was tested using the Little Missing Completely at Random test [25, 26]). The imputation model included baseline characteristics of age, gender, weight, treatment group, hospital, length of illness, side of palsy, face pain, and facial paralysis severity. Imputations were drawn from a pool of 5 donors (nearest neighbors).[27, 28] Missing costs were limited (1.5% missing) and were assumed to be zero as it is most common to have no healthcare use.

3.3.7. Uncertainty and sensitivity analysis

To capture sampling uncertainty, probabilistic sensitivity analysis was conducted using bootstrapping with 1000 replications drawing from cost and effectiveness data observed at the patient level. The bootstrapping results are graphically presented on a cost-effectiveness plane, with each of the 1000 dots representing one mean cost and effectiveness difference between treatment and placebo group. We avoided presenting confidence intervals (CIs) directly for the cost-effectiveness ratios as the CIs may not be interpretable when the lower and upper bounds locate in different quadrants,[29] nevertheless the cost-effectiveness plane can be used to locate the middle 95% of all bootstrapped dots which correspond to the conventional 95% CI. The acceptability curves were also constructed by counting the proportion of bootstrap replicates that are acceptable under various willingness-to-pay levels, summarized as the probability that the therapy is cost-effective.[30] We used A$ 28,000 to A$ 76,000 per QALY as the range

of willingness to pay threshold according to published literature, reporting on a A$50,000/QALY threshold which corresponds to approximately US$35,000 or £28,000.[31, 32] [29]

One-way sensitivity analyses were performed to assess the robustness of the findings. Firstly, cost inputs were varied. This includes varying the unit cost of medication and medical health service use, and removing extreme cost outliers (3 patients with costs over A$1500 and 1 patient with cost over A$2000 respectively). Secondly, alternative approaches for dealing with missing data including complete case analysis, mean imputation, median imputation, and regression imputation were used. Thirdly, we considered an alternate mapping approach using PedsQL to estimate CHU9D utilities.[23]

Student's t test for continuous outcomes and Chi-squared test for dichotomous outcomes were used. All analyses were conducted using STATA version 16 (Statacorp, College Station, Texas, USA).

3.3.8. Data availability

The authors support data sharing. Data from the BellPIC trial have used identifiable individual patient data that are subject to restriction due to ethics, consent, and privacy issues. Anonymized participant data and data dictionary will be available on request from the corresponding author where possible within these constraints for use.

3.4. Results

187 participants aged 6 months to <18 years were recruited, 93 randomized to prednisolone and 94 to placebo. The baseline sample characteristics of each group are presented in Table 3-1. Seven participants who withdrew or were lost to follow up and had no quality-of-life or cost data (5 from prednisolone, 2 from placebo) were dropped. A total of 180 participants were included.

Total costs disaggregated by type of health care used are presented in Table 3-2 (costs by different age groups presented in Appendix eTable 2). The cost-effectiveness outcomes are summarized in Table 3-3. Participants in the prednisolone group had a mean incremental cost of A$66 (95% CI: -A$47, A$179) compared to patients in the placebo group. Participants in the prednisolone group had a mean incremental QALY gain of 0.01 (95% CI: -0.01, 0.03), resulting in an ICER of A$6625 per QALY gained. This is lower than conventional willingness-to-pay thresholds of A$50,000 per QALY,[31, 32] suggesting that prednisolone is likely cost-effective relative to placebo. The acceptability curve (Figure 3-1**)** indicated that prednisolone was cost-effective compared with placebo in 83.0% of bootstrapped replicates if willingness-to-pay for a QALY gain is $50,000 (Table 3-3). The probability that prednisolone is cost-effective rises from 79.0% to 84.1% when willingness-to-pay increases from A$28,000 to A$76,000,

implying that the conclusion remains unchanged given common willingness-to-pay thresholds (Table 3-3),

The recovery rate at 6 months for the prednisolone group was 0.97 and for the placebo group was 0.92 (difference in proportions 0.04 (95%CI: -0.03, 0.11), p=0.221, Table 3-3). The incremental cost to achieve one additional recovery was estimated to be A$1577 for prednisolone compared to placebo. The uncertainty when using recovery rate as effectiveness outcome was also demonstrated with the cost-effectiveness plane and acceptability curve (Appendix eFigure 2).

The subgroup analysis by age revealed that prednisolone is more cost-effective in children aged 12-18 years than those <12 years (Figure 3-2). This is primarily driven by the much greater incremental QALY gain of 0.042 (95%CI: 0.002,0.081) in children aged 12-18 years comparing prednisolone to placebo, whereas the QALY gain was 0.001 (95%CI: -0.016,0.018) in those <12 years. There was also a larger gain in recovery of 0.16 (95%CI: -0.01,0.30; p=0.059) comparing prednisolone to placebo in children aged 12-18 years, compared with 0.004 (95%CI: -0.07,0.08; p=0.922) in the younger age group at 6 months. The cost difference between prednisolone and placebo was similar in different age groups. The probability of prednisolone being cost-effective is 98% in children aged 12-18 years and 51% in those <12 years at willingness-to-pay threshold of A$50,000 per QALY. One-way sensitivity analyses suggest that the conclusion remains under various scenarios (Appendix eFigure 3).

3.5. Discussion

We evaluated the cost-effectiveness of prednisolone treatment compared with placebo for Bell's palsy in children. Prednisolone was more costly, however, also led to more QALYs gained over a 6-month period. The additional cost is likely worthwhile given the additional QALYs gained using conventional value judgement thresholds, whereas the gains are primarily driven by children aged 12-18 years. Given a willingness-to-pay of A$50,000 per QALY gained, the probability that prednisolone is good value is 98% in children aged 12-18 years. For children <12 years on the other hand, the probability that prednisolone is cost effective is 51%, meaning that it is uncertain whether prednisolone is good value for this age group. In addition, the average incremental cost to achieve one additional recovery appears to be quite reasonable considering the social impact of Bell's Palsy in children aged 12-18 years (A$429), for children <12 years it appears much more expensive (A$11,188). This added evidence to the conclusion we had above. To the best of our knowledge, this is the first study evaluating the cost-effectiveness of using steroids to treat Bell's palsy in children.

The results that prednisolone is highly likely cost-effective and good value for children aged 12-18 years but may not be for under 12 years are not surprising. In adults, oral steroids are recommended by the American Academy of Otolaryngology-Head and Neck Surgery Foundation to help improve recovery in Bell's palsy in patients 16 years and older.[8] As age 16 years is within our trial population (up to <18 years) and there is little biological reason that treatment with prednisolone would suddenly stop being efficacious for children who are slightly younger. It is reasonable to find that steroids have offered good value in those children aged 12-18 years. Our results are thus in line with the only previous cost-effectiveness analysis in adults of prednisolone treatment for Bell's palsy,[11] which reported that prednisolone is likely cost-effective with a 77% probability compared to placebo at willingness-to-pay of £30,000 (2006/07 currency price). For children younger than 12 years on the other hand, our results suggest that evidence supporting the use of steroids is not strong, consistent with the overall clinical findings.

It is worth noticing that no significant difference in the primary outcome (recovery defined by facial palsy score at 1 month) was found between the two treatment groups, nor for the pre-specified subgroup of children aged 12-18 years, in the main clinical paper.[12] The authors concluded that the study, although underpowered, does not provide evidence that early treatment with prednisolone improves complete recovery.[12] The clinical paper and the current economic evaluation agree in conclusion for under 12 years old and differ for 12-18 years old. The difference may be due to the following reasons. The economic evaluation is centered around patient quality of life outcomes, where a generic instrument instead of a condition specific measure is used to capture general quality of life impact. Further, it is possible that prednisolone improves HRQoL independent of its effect on Bell's Palsy recovery. In this study, we found that prednisolone has resulted significantly better QALYs in older children over the 6 months period. This is possible because Bell's palsy had a larger impact on quality of life for adolescents than for younger children, which is supported by our supplementary results which showed a larger quality of life difference between recovered and unrecovered patients in the older age group (Appendix eTable 3). A previous systematic review also reported an increase in social difficulties with age in children with facial palsy.[33] QALY reflected the accumulative quality-of-life difference over the whole 6-month period and thus may drive the high probability of prednisolone being cost-effective in 12-18 years old.

3.6. Strengths and limitations

Our study has several strengths including using individual level data from a gold standard randomized controlled trial, with a wide range of child age (6 months to <18 years). We used quality-of-life observations collected as part of the trial instead of assumed utilities based on published literature.

Comprehensive sensitivity analyses were performed to capture uncertainty. A prespecified subgroup analysis was also conducted to capture heterogeneity of patient population.

Several limitations were also identified. The time horizon was 6 months and longer-term impacts were not measured. Nevertheless, Bell's palsy is an acute disease with 97% and 93% children recovered at 6 months in the treatment and placebo groups in this current trial and literature shows all children recovered within 12 months[34]. Recurrent Bell's palsy is uncommon,[35] hence long term evaluation is unlikely to have a meaningful impact on our results and conclusion. In actual clinical practice, a placebo would not be administered in patients not receiving a medication (prednisolone). The use of a placebo, although in line with the guidelines for clinical trials of no known or available alternative therapy, may underestimate the benefit of prednisolone compared with usual practice due to its psychosomatic effects. Another limitation is that CHU9D was not designed for children under 4 years, and thus the CHU9D utilities for children aged 2-4 years old in this study were mapped from an established algorithm. Further evidence based on directly collected health utilities for 2-4 years may be valuable. We primarily reported on a A$50,000/QALY threshold corresponding to approximately US$35,000 or £28,000, and appropriate willingness-to-pay levels may vary across countries and over time. Nevertheless, we used A$28,000 to A$76,000 per QALY as the range of willingness-to-pay and the results suggested that this has little meaningful impact on the probability that prednisolone is cost-effective (probability rises from 79.0% to 84.1%) and thus the conclusion remains. Finally, baseline utility was not collected due to concerns for patients and parents stress when at diagnosis and recruitment and was assumed to be the same for both patient groups due to randomization. We used the mean utility from unrecovered patients at 1 month as baseline for both groups. Nevertheless, the QALY difference between the two groups would remain unchanged even if other baseline utility values were applied as long as same level of baseline utility can be expected for both groups given randomization.

3.7. Conclusions

Our results suggest that using prednisolone to treat Bell's palsy in children is likely cost-effective compared with placebo over a 6-month period in children aged 12-18 years. The benefit primarily comes from improvement in children's quality of life and QALYs, which is increasingly valued by clinicians in practice. In children aged <12 years, there is no strong evidence supporting the use of prednisolone to treat Bell's palsy. Prednisolone for treating Bell's palsy in 12 to 18 years old may be considered by stakeholders and policy makers.

3.8. Disclosure

The authors report no relevant disclosures beyond the funding information listed above. Go to Neurology.org/N for full disclosures.

3.9. References

[1] M. T. Mackay *et al.*, "Stroke and nonstroke brain attacks in children," (in eng), *Neurology,* vol. 82, no. 16, pp. 1434-40, Apr 22 2014.

[2] N. Hato, S. Murakami, and K. Gyo, "Steroid and antiviral treatment for Bell's palsy," *The Lancet,* vol. 371, no. 9627, pp. 1818-1820, 2008/05/31/ 2008.

[3] S. Rowlands, R. Hooper, R. Hughes, and P. Burney, "The epidemiology and treatment of Bell's palsy in the UK," (in eng), *Eur J Neurol,* vol. 9, no. 1, pp. 63-7, Jan 2002.

[4] M. Lee, M. Mackay, L. Blackbourn, and F. E. Babl, "Emotional impact of Bell's palsy in children," (in eng), *J Paediatr Child Health,* vol. 50, no. 3, pp. 245-6, Mar 2014.

[5] D. Cope and R. Bova, "Steroids in Otolaryngology," *The Laryngoscope,* https://doi.org/10.1097/MLG.0b013e31817c0b4d vol. 118, no. 9, pp. 1556-1560, 2008/09/01 2008.

[6] F. M. Sullivan *et al.*, "Early Treatment with Prednisolone or Acyclovir in Bell's Palsy," *New England Journal of Medicine,* vol. 357, no. 16, pp. 1598-1607, 2007.

[7] M. Engström *et al.*, "Prednisolone and valaciclovir in Bell's palsy: a randomised, double-blind, placebo-controlled, multicentre trial," (in eng), *Lancet Neurol,* vol. 7, no. 11, pp. 993-1000, Nov 2008.

[8] R. F. Baugh *et al.*, "Clinical practice guideline: Bell's palsy," (in eng), *Otolaryngol Head Neck Surg,* vol. 149, no. 3 Suppl, pp. S1-27, Nov 2013.

[9] F. E. Babl *et al.*, "Bell's Palsy in Children (BellPIC): protocol for a multicentre, placebo-controlled randomized trial," (in eng), *BMC Pediatr,* vol. 17, no. 1, p. 53, Feb 13 2017.

[10] J. Pitaro, S. Waissbluth, and S. J. Daniel, "Do children with Bell's palsy benefit from steroid treatment? A systematic review," *International Journal of Pediatric Otorhinolaryngology,* vol. 76, no. 7, pp. 921-926, 2012/07/01/ 2012.

[11] R. A. Hernández, F. Sullivan, P. Donnan, I. Swan, and L. Vale, "Economic evaluation of early administration of prednisolone and/or aciclovir for the treatment of Bell's palsy," (in eng), *Fam Pract,* vol. 26, no. 2, pp. 137-44, Apr 2009.

[12] F. E. Babl *et al.*, "Efficacy of Prednisolone for Bell Palsy in Children: A Randomized, Double-Blind, Placebo-Controlled, Multicenter Trial," (in eng), *Neurology,* Aug 25 2022.

[13] G. D. Sanders *et al.*, "Recommendations for Conduct, Methodological Practices, and Reporting of Cost-effectiveness Analyses: Second Panel on Cost-Effectiveness in Health and Medicine," (in eng), *Jama,* vol. 316, no. 10, pp. 1093-103, Sep 13 2016.

[14] D. Husereau *et al.*, "Consolidated Health Economic Evaluation Reporting Standards 2022 (CHEERS 2022) statement: updated reporting guidance for health economic evaluations," (in eng), *Bmj,* vol. 376, p. e067975, Jan 11 2022.

[15] F. Babl *et al.*, "Paediatric Research in Emergency Departments International Collaborative (PREDICT): first steps towards the development of an Australian and New Zealand research network," (in eng), *Emerg Med Australas,* vol. 18, no. 2, pp. 143-7, Apr 2006.

[16] J. W. House and D. E. Brackmann, "Facial nerve grading system," (in eng), *Otolaryngol Head Neck Surg,* vol. 93, no. 2, pp. 146-7, Apr 1985.

[17] MBS. (December 1st). *Unit cost for GP*. Available: http://www9.health.gov.au/mbs/search.cfm?q=23&sopt=I

[18] *The Round 23 National Hospital Cost Data Collection (NHCDC) collected public hospital cost information for the 2018–19 financial year.*

[19] K. J. Stevens, "Working with children to develop dimensions for a preference-based, generic, pediatric, health-related quality-of-life measure," (in eng), *Qual Health Res,* vol. 20, no. 3, pp. 340-51, Mar 2010.

[20] J. W. Varni, M. Seid, and P. S. Kurtin, "PedsQL 4.0: reliability and validity of the Pediatric Quality of Life Inventory version 4.0 generic core scales in healthy and patient populations," (in eng), *Med Care,* vol. 39, no. 8, pp. 800-12, Aug 2001.

[21] *Measuring & Valuing Health. A brief overview of the Child Health Utility 9D (CHU9D).* Available: https://licensing.sheffield.ac.uk/product/CHU-9D

[22] J. Ratcliffe, T. Flynn, F. Terlich, K. Stevens, J. Brazier, and M. Sawyer, "Developing Adolescent-Specific Health State Values for Economic Evaluation," *PharmacoEconomics,* vol. 30, no. 8, pp. 713-727, 2012/08/01 2012.

[23] R. Sweeney, G. Chen, L. Gold, F. Mensah, and M. Wake, "Mapping PedsQL(TM) scores onto CHU9D utility scores: estimation, validation and a comparison of alternative instrument versions," (in eng), *Qual Life Res,* vol. 29, no. 3, pp. 639-652, Mar 2020.

[24] J. N. Matthews, D. G. Altman, M. J. Campbell, and P. Royston, "Analysis of serial measurements in medical research," (in eng), *Bmj,* vol. 300, no. 6719, pp. 230-5, Jan 27 1990.

[25] R. J. A. Little, "A Test of Missing Completely at Random for Multivariate Data with Missing Values," *Journal of the American Statistical Association,* vol. 83, no. 404, pp. 1198-1202, 1988/12/01 1988.

[26] D. I. Rhon, M. Kim, C. V. Asche, S. C. Allison, C. S. Allen, and G. D. Deyle, "Cost-effectiveness of Physical Therapy vs Intra-articular Glucocorticoid Injection for Knee Osteoarthritis: A Secondary Analysis From a Randomized Clinical Trial," *JAMA Network Open,* vol. 5, no. 1, pp. e2142709-e2142709, 2022.

[27] A. Manca and S. Palmer, "Handling missing data in patient-level cost-effectiveness analysis alongside randomised clinical trials," *Applied Health Economics and Health Policy,* vol. 4, no. 2, pp. 65-75, 2005/06/01 2005.

[28] M. Franklin, R. M. Hunter, A. Enrique, J. Palacios, and D. Richards, "Estimating Cost-Effectiveness Using Alternative Preference-Based Scores and Within-Trial Methods: Exploring the Dynamics of the Quality-Adjusted Life-Year Using the EQ-5D 5-Level Version and Recovering Quality of Life Utility Index," *Value in Health,* 2022/01/13/ 2022.

[29] H. A. Glick, J. A. Doshi, S. S. Sonnad, and D. Polsky, *Economic evaluation in clinical trials*. OUP Oxford, 2014.

[30] E. Fenwick and S. Byford, "A guide to cost-effectiveness acceptability curves," *British Journal of Psychiatry,* vol. 187, no. 2, pp. 106-108, 2005.

[31] B. George, A. Harris, and A. Mitchell, "Cost-effectiveness analysis and the consistency of decision making: evidence from pharmaceutical reimbursement in australia (1991 to 1996)," (in eng), *Pharmacoeconomics,* vol. 19, no. 11, pp. 1103-9, 2001.

[32] L. C. Edney, H. Haji Ali Afzali, T. C. Cheng, and J. Karnon, "Estimating the Reference Incremental Cost-Effectiveness Ratio for the Australian Health System," *PharmacoEconomics,* vol. 36, no. 2, pp. 239-252, 2018/02/01 2018.

[33] M. Hotton *et al.*, "A Systematic Review of the Psychosocial Adjustment of Children and Adolescents with Facial Palsy: The Impact of Moebius Syndrome," *International Journal of Environmental Research and Public Health,* vol. 17, no. 15, p. 5528, 2020.

[34] E. Ünüvar, F. Oğuz, M. Sıdal, and A. Kılıç, "Corticosteroid treatment of childhood Bell's palsy," *Pediatric Neurology,* vol. 21, no. 5, pp. 814-816, 1999/11/01/ 1999.

[35] D. B. Pitts, K. K. Adour, and R. L. Hilsinger, Jr., "Recurrent Bell's palsy: analysis of 140 patients," (in eng), *Laryngoscope,* vol. 98, no. 5, pp. 535-40, May 1988.

3.10. Tables and figures

Table 3-1 Baseline characteristics

Variables	Prednisolone (N=93)	Placebo (N=94)
Age		
Mean (SD)	10.06 (4.68)	8.66(4.36)
Median (p25-p75)	11.13 (5.17-14.01)	9.44(5.09-11.84)
Age group, No. (%)		
6 months to <12 years	54 (58.06)	73 (77.66)
12 to <18 years	39 (41.94)	21 (22.34)
Sex, No. (%)		
Male	45 (48.39)	45(47.87)
Female	48 (51.61)	49(52.13)
House Brackmann (clinician) category		
Non-severe (II to IV), No. (%)	85 (91.40)	81 (86.17)
Severe (V and VI), No. (%)	8 (8.60)	13 (13.83)
House Brackmann (clinician) score		
Mean (SD)	3.48(0.76)	3.64(0.80)
Median (p25-p75)	3.00(3.00-4.00)	4.00(3.00-4.00)
House Brackmann- Parent Perception		
Mean (SD)	3.72(1.27)	3.81(0.95)
Median (p25-p75)	4.00(3.00-4.00)	4.00(3.00-4.00)

Abbreviations: No.=Number of observations; SD = Standard deviation; p25 = percentile 25, p75 =percentile 75. House Brackmann scale is treated both as categories (non-severe vs severe) and continuous scores (ranging from 1 to 6, with higher score indicating more severe symptoms; 1=normal, 2=mild dysfunction, 3=moderate dysfunction, 4=moderately severe dysfunction,5=severe dysfunction, 6=total paralysis).

Table 3-2 Total cost presented by cost category

	Prednisolone (N=88)		Placebo (N=92)		Mean difference (95%CI)	*P* value
	Freq	Mean (SD)	Freq	Mean (SD)		
Initial ED visit		572		572	0	NA
Medication cost		35.03(13.63)		0	0	NA
Follow up cost (6 months)		152.48(458.72)		121.26(303.14)	31.22(-81.90,144.35)	0.589
GP	13	8.89(27.63)	17	7.65(16.64)	1.24(-5.39,7.87)	0.715
ED	5	78.00(379.84)	7	49.74(182.90)	28.26(-58.25,114.77)	0.523
Inpatient	0	0	0	0	0	NA
Outpatient	9	18.70(77.82)	7	22.21(120.05)	-3.51(-33.21,26.19)	0.817

Other health service use	3	8.02(47.30)	6	18.33(86.93)	-10.32(-30.89,10.26)	0.327
Medical tests	14	38.87(117.56)	10	23.33(93.37)	15.55(-15.40,46.49)	0.326
Total cost (medication and follow up)		759.51(458.64)		693.26(303.14)	66.25(-46.86,179.36)	0.252

Abbreviations: ED = Emergency Department; SD = standard deviation; Freq = Frequency; CI = Confidence interval; NA = not applicable.

There was no hospital admission reported during study period, thus no inpatient cost.

Table 3-3 Cost-effectiveness analysis results for prednisolone versus placebo over 6 months

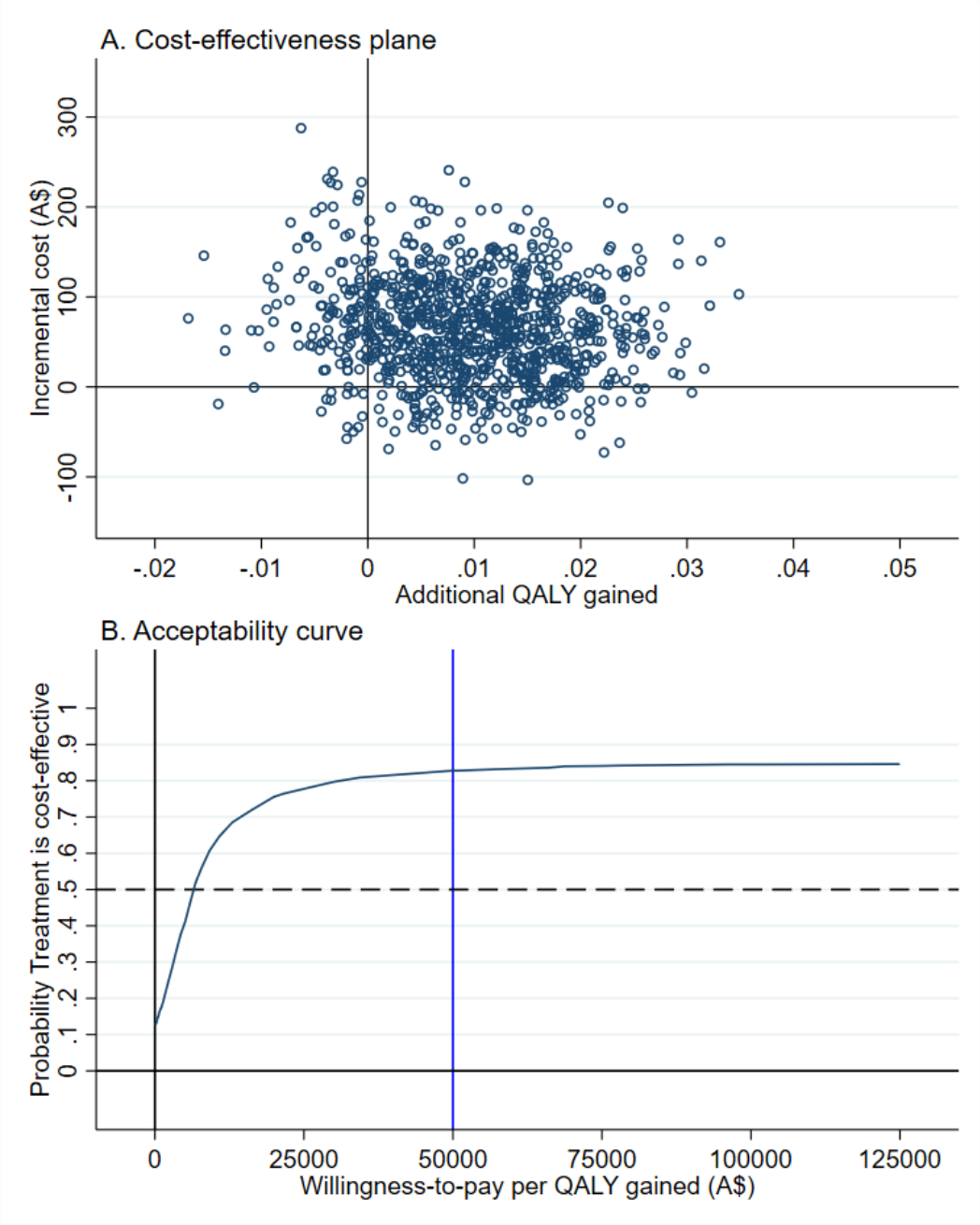
Variable	Prednisolone	Placebo	Mean Difference (95%CI), *p* values[a]	ICER[b]	Acceptability, %			
					WTP= 28000	WTP= 42000	WTP= 50000	WT= 76000
Total sample (6months to < 18years)	N=88	N=92						
Cost (95%CI), A$	759.51(662.33,856.69)	693.26(630.48,756.04)	66.25 (-46.86,179.36), *p*=0.252					
Effectiveness: QALY (95%CI)	0.45(0.44,0.46)	0.44(0.43,0.45)	0.01 (-0.01,0.03), *p*=0.296	A$6625 per QALY gained	0.790	0.819	0.828	0.841
Effectiveness: Recovery (rate)	0.966	0.924	0.042 (-0.03, 0.11), *p*=0.221	A$1577 per recovery achieved				
Older children (12 to <18years)	N=37	N= 22						
Total medical cost (95%CI)	811.61(604.36,1018.85)	745.08(576.67,3917.49)	66.53 (-221.13,354.19), *p*=0.652					
QALY (95%CI)	0.44(0.42,0.46)	0.40(0.36,0.44)	0.042 (0.002,0.081), *p*=0.040	A$1584 per QALY gained	0.974	0.979	0.983	0.986
Recovery (rate)	0.973	0.818	0.155[c] (-0.007,0.302), *p*=0.059	A$429 per recovery achieved				
Younger children (6months to <12years)	N=51	N=70						
Total medical cost (95%CI)	721.71(640.02,803.40)	676.97(611.16,742.77)	44.75 (-57.06,146.56), *p*=0.391					
QALY (95%CI)	0.46(0.44,0.47)	0.45(0.44,0.47)	0.0014 (-0.0155,0.0184), *p*=0.867	A$31,964 per QALY gained	0.468	0.496	0.505	0.522
Recovery (rate)	0.961	0.957	0.004 (-0.07, 0.08), *p*=0.922	A$11,188 per recovery achieved				

Abbreviations: ICER, incremental cost-effectiveness ratio; CI = Confidence interval; QALY, quality-adjusted life-years; WTP, willingness to pay per QALY gained.

Note: Conventional willingness to pay threshold is not available for cost per additional recovery achieved and thus the probability is not calculated. More decimals were used for ICERs for clearer results. Recovery rate was reported for 180 observations. The recovery rate reported in the main clinical paper might be slightly different (N=187), however the conclusion remains (no significant difference between treatment groups).

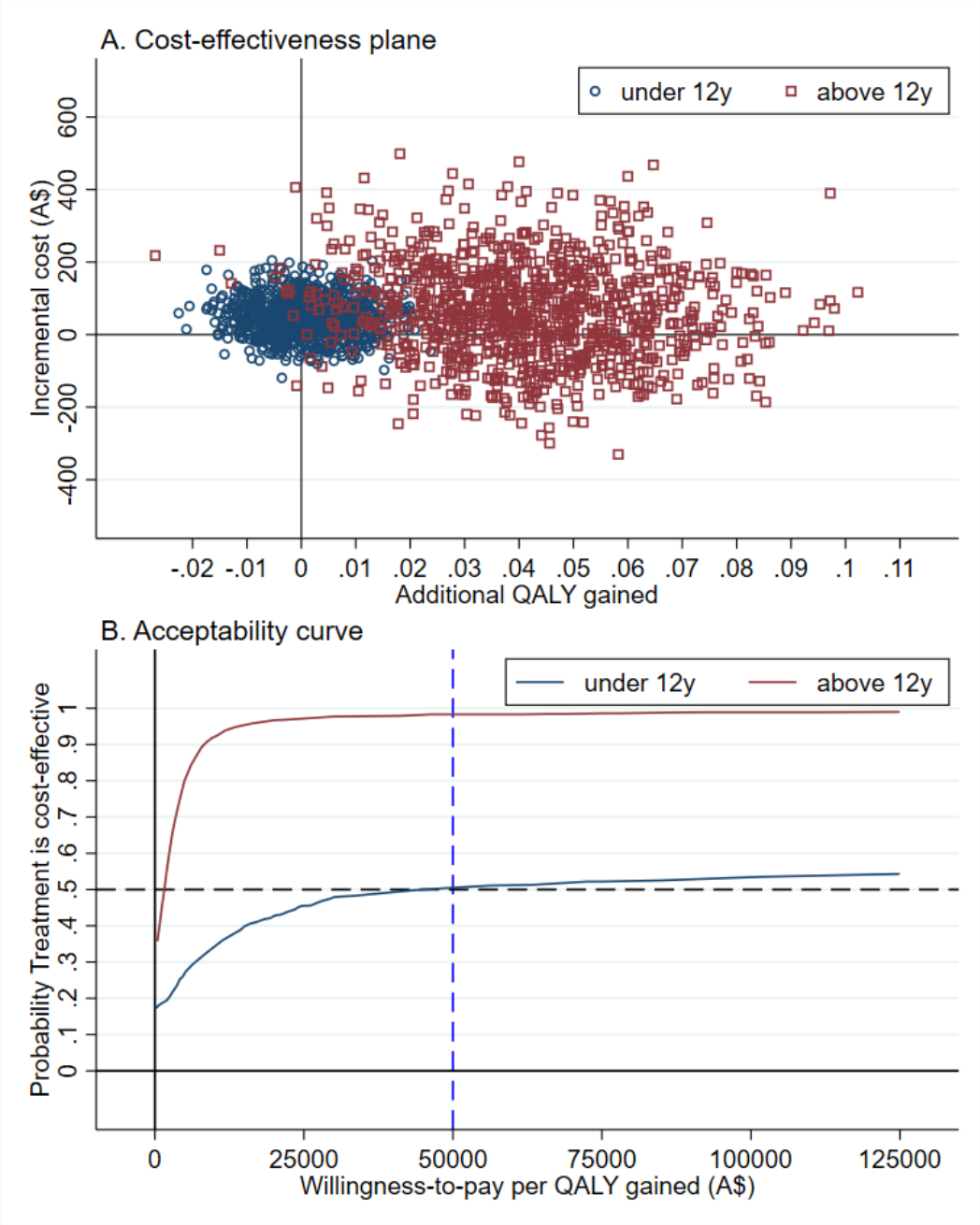
[a] P value obtained by Chi-squared test (for recovery rate) and 2-tailed t test (for QALY and cost).

[b] For QALY outcome, ICER= cost difference/ QALY difference. For recovery outcome, ICER=cost difference/recovery rate difference.



Abbreviations: QALY, quality-adjusted life-year.

Figure 3-1 Cost-effectiveness plane and acceptability curve comparing prednisone with placebo, total sample



Abbreviations: QALY, quality-adjusted life-year.

Figure 3-2 Cost-effectiveness planes and acceptability curves comparing prednisolone with placebo, by 12 years old

3.11. Supplemental materials

Cost input table

eTable 1 Unit cost prices and sources

Type of health services	Base case	Low	High	Source
Medication (30 mL bottle of Redipred)	13.99	9.50	27.55	Chemist, hospital price, PBS price (general patient charge) [a]
General Practitioner Attendances	39.10	17.90	75.75	MBS items 23, 3, 36 (Level B, A, C) [b]
Emergency Departments (non-admitted presentation) [a]	572.00	435.00	675.00	IHPA report round 23 [c]
Outpatient				
Ophthalmology	203.65			MBS item 109
Neurology	159.35			MBS item 110 (consultant physician services including neurology)
Physiotherapy	91.50			MBS item 82035
Speech pathology	91.50			MBS item 82020, 93036, 93041
ENT	160.20			MBS item 82300
Pediatrician	159.35	NA	278.75	MBS item 110 (consultant physician services including pediatrician), 135
Medical tests				
Blood test	32.55	7.85	NA	MBS item 66596, 65060
MRI	403.20	336.00	NA	MBS item 63040, 63043, 63046
CT	230.40			MBS item 56022
Lumbar puncture	78.35			MBS item 39000
Doppler ultrasound	173.60			MBS item 55238, 55244
X-ray	106.55			MBS item 12306
Acupuncture	74.60			MBS item 197

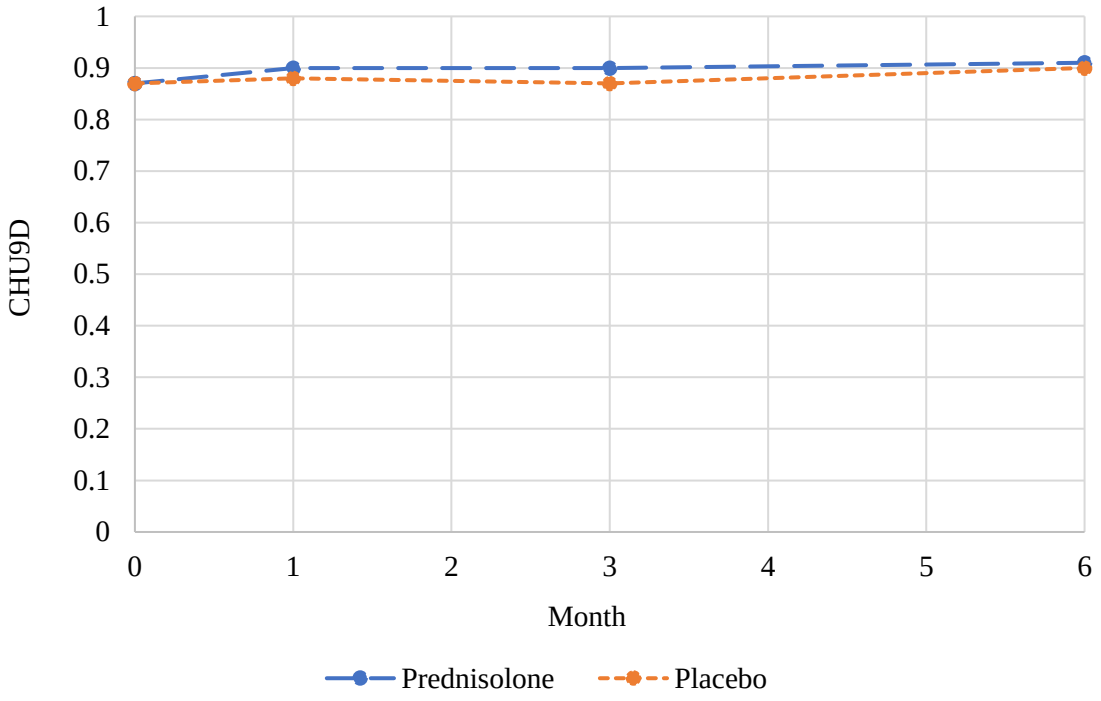
Note: The frequency of doctor visits was obtained as a free text variable, asking "Please specify how many times they used each service". As the number of children having health services was small, we calculated the cost as close to the original description as possible. Some observations for outpatient visits not only described the frequency but also the detailed type of doctor they visited. The frequency variable was checked one by one, and we tried to use the specific type of doctor visits and relevant frequencies wherever possible. For example, one observation reported "neurology 1, physiotherapy 1, speech pathology 1", we then used the specific unit cost of each type of doctor visit multiplied by its frequency and summed them up. For those only with simple frequency information, we used the unit cost of the most frequently reported type of doctor visit as its unit cost. There is no data about the frequency of medical tests used (see details in above example survey questions). One medical test was assumed considering the short period.

[a]Market price for "Redipred" obtained on Dec 19, 2020, from https://www.chemistwarehouse.com.au/buy/7377/redipred-5mg-ml-oral-liquid-30ml---prednisolone-sodium-phosphate; hospital price was provided by the staff of the Royal Children's Hospital, Melbourne, Australia; Pharmaceutical Benefits Scheme (PBS) price: https://www.pbs.gov.au/medicine/item/8285C, the general patient charge, searched on Dec 19, 2020.

[b]MBS prices were obtained from Medicare Benefits Schedule (MBS) website, http://www.mbsonline.gov.au/internet/mbsonline/publishing.nsf/Content/Home, searching each item for each health service, searched on Dec 19, 2020.

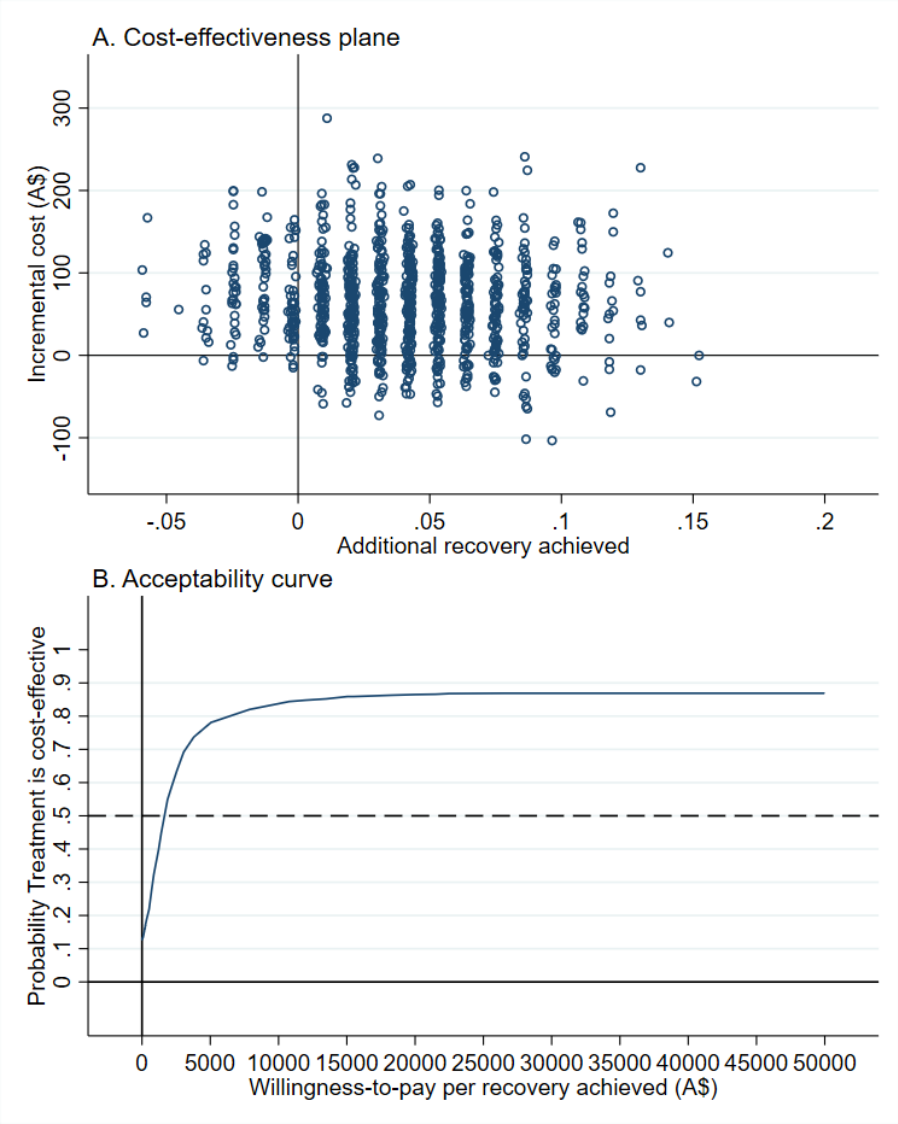
[c] Base case for unit cost for Emergency Department (ED) visits were the average of the ED unit cost of 8 states across Australia (lowest 435, highest 675).

Quality-of-life over 6 months

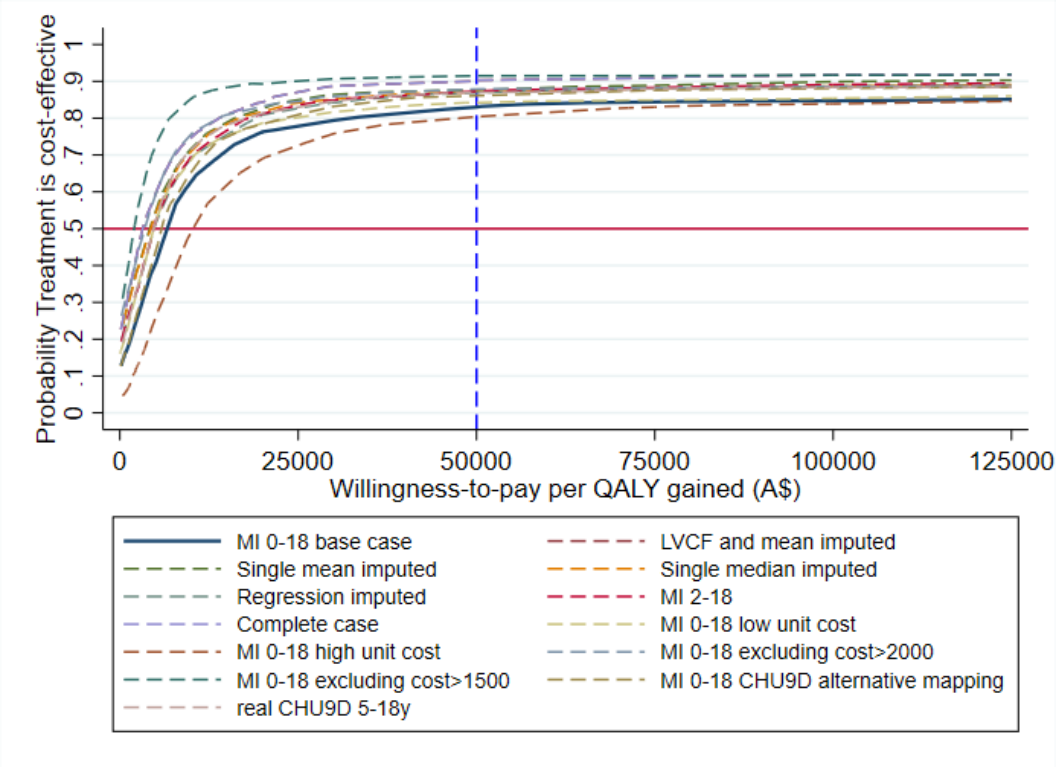


eFigure 1 Quality-of-life over the 6-month period, measured using the CHU9D. The areas under the curves are used to estimate the total QALYs for prednisolone and placebo respectively.

Sensitivity analyses



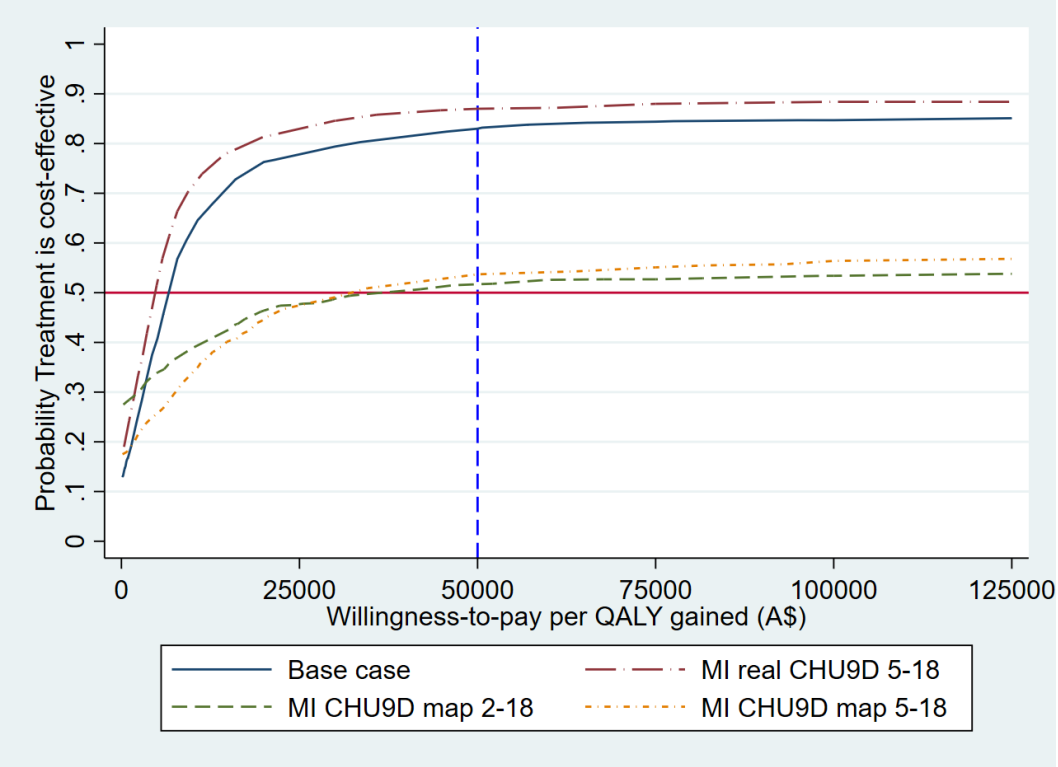
eFigure 2(A) Cost-effectiveness plane for recovery outcome comparing prednisolone with placebo using bootstrapping and (B), acceptability curve showing the probability that prednisolone is cost-effective compared with placebo, given potential willingness to pay thresholds.



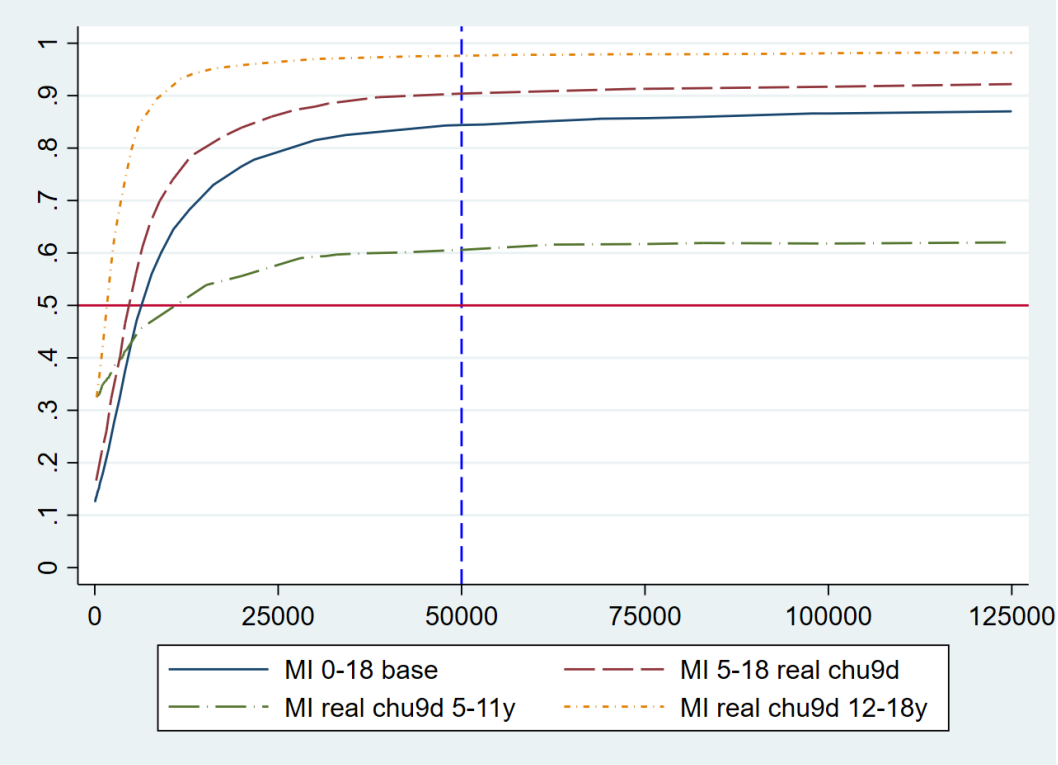
eFigure 3 Acceptability curves of the probability that prednisolone is cost-effective compared with placebo (using QALY as effectiveness outcome, combining all sensitivity analyses)

Note: Base case used total score based algorithm, sensitivity analysis used alternative dimension-score based algorithm. Both algorithms are from the same literature, Sweeney, R., et al. (2020). There were 5 children under age 2 years with complete missing CHU9D utilities and the base case MI included these children.

eFigure4 below explores the impact of using mapped CHU9D on the cost-effectiveness result. Analyses in different scenarios were conducted to see the difference: (1) using mapped CHU9D from 2-4 years and real CHU9D for 5-18 years (base case), (2) using all real CHU9D data for 5-18 years, (3) using all mapped CHU9D for 5-18 years, (4) using all mapped CHU9D data for 2-18 years old. The Figure shows that the result of using all real CHU9D in 5-18 is very different from the result of using all mapped CHU9D in 5-18. This also shows that using all real CHU9D has similar results with our base case, which confirms that the small part of children under 5 years old has a small impact on our main results. However, this means that the results of 6 month to <12y sub age group is not as reliable as 12-18y age group and future analysis using all real CHU9D is valuable. eFigure5 confirms that the difference by age group still exists when using all real CHU9D data.



eFigure4 Sensitivity analysis: comparing real CHU9D and mapped CHU9D



eFigure5 Sensitivity analysis: real CHU9D by age

Details of costs in different types by age

eTable 2 Costs by category in different age groups

	Prednisolone		Placebo		Difference
	Freq	Mean (SD)	Freq	Mean (SD)	Mean(95%CI)
Aged 6m-12y		N=51		N=70	
Medication cost		27.17(13.00)		NA	
Follow up cost total until 6 months		122.55(293.25)		104.97(275.98)	17.58(-84.67,119.83)
GP	6	6.90(20.25)	14	8.38(17.48)	-1.48(-8.22,5.27)
Emergency Department	3	44.86(192.90)	7	65.37(207.56)	-20.51(-93.23,52.21)
Outpatient	7	24.37(94.46)	4	7.47(47.17)	16.89(-8.72,42.51)
Other health service use	2	6.25(31.24)	4	10.36(53.59)	-4.12(-20.55,12.32)
Tests (e.g., blood tests, neuroimaging)	7	40.17(120.96)	6	13.38(70.52)	26.79(-7.50,61.08)
Total cost (medication and follow up cost)		149.71(290.45)		104.97(275.98)	44.75(-57.06,146.56)
Aged 12y-18y		N=37		N= 22	
Medication cost		45.88(2.37)		NA	
Follow up cost total until 6 months		193.73(621.35)		173.08(379.83)	20.65(-266.92,308.22)
GP	7	11.62(35.53)	3	5.33(13.73)	6.29(-9.24,21.83)
Emergency Department	2	123.68(541.56)	0	0.00(0.00)	123.68(-103.43,350.78)
Outpatient	2	10.89(46.19)	3	69.09(228.30)	-58.20(-133.84,17.45)
Other health service use	1	10.45(63.56)	2	43.68(149.75)	-33.23(-88.10,21.65)
Tests (e.g., blood tests, neuroimaging)	7	37.09(114.32)	4	54.98(141.62)	-17.89(-83.89,48.11)
Total cost (medication and follow up cost)		239.61(621.58)		173.08(379.83)	66.53(-221.13,354.19)

Abbreviations: SD = standard deviation; Freq = Frequency; CI = Confidence interval; NA = not applicable.

Comparison of quality of life by recovery

eTable 3 Comparison of quality of life by recovery

	Recovered		Not recovered		Difference (95%CI)	*P* value [a]
	N	Utility Mean (95%CI)	N	Utility Mean (95%CI)		
Base case MI 0-18						
Month 1	94	0.91(0.88,0.93)	86	0.87(0.84,0.90)	0.04(0.00,0.08)	0.055
Month 3	157	0.88(0.86,0.91)	23	0.92(0.88,0.96)	-0.04(-0.11,0.03)	0.301
Month 6	170	0.91(0.88,0.93)	10	0.88(0.80,0.96)	0.03(-0.07,0.12)	0.569
Age: 6 months to <12 years						
Month 1	62	0.91(0.88,0.93)	59	0.89(0.85,0.92)	0.02(-0.03,0.06)	0.396
Month 3	105	0.91(0.88,0.93)	16	0.92(0.87,0.98)	-0.02(-0.09,0.05)	0.595
Month 6	116	0.93(0.91,0.95)	5	0.90(0.78,1.02)	0.03(-0.08,0.14)	0.574
Age: 12 -18 years						
Month 1	32	0.91(0.85,0.97)	27	0.83(0.77,0.89)	0.08(0.00,0.17)	0.054
Month 3	52	0.84(0.78,0.90)	7	0.91(0.82,0.99)	-0.07(-0.23,0.09)	0.374
Month 6	54	0.85(0.80,0.90)	5	0.86(0.75,0.96)	-0.01(-0.18,0.17)	0.950

Abbreviations: CI = Confidence interval; MI: multiple imputation.

[a] P value calculated from t test.

[b] For quality of life comparison, we mainly refer the results at month 1 when the sample between recovered and unrecovered is similar. In month 3 and month 6, the sample size in the unrecovered group is too small. The quality of life of those recovered is higher than those unrecovered at month 1 (quality of life difference 0.04 (95%CI: 0.00, 0.08), p=0.055). In the sub age groups, we again mainly compare the quality of life by recovery in month 1 as the sample size is too small for month 3 and month 6. When comparing the quality of life between recovered and unrecovered patients the impairment on quality of life is larger in the older group (quality of life difference between recovered and unrecovered 0.08 (95%CI: 0.00, 0.17), p=0.054) than in the younger group (quality of life difference 0.02 (95%CI: -0.03, 0.06), p=0.396).

Reference

1. Mpundu-Kaambwa C, Chen G, Russo R, Stevens K, Petersen KD, Ratcliffe J. Mapping CHU9D Utility Scores from the PedsQLTM 4.0 SF-15. PharmacoEconomics 2017; 35(4): 453-67.

2. Lambe T, Frew E, Ives NJ, et al. Mapping the Paediatric Quality of Life Inventory (PedsQL™) Generic Core Scales onto the Child Health Utility Index–9 Dimension (CHU-9D) Score for Economic Evaluation in Children. PharmacoEconomics 2018; 36(4): 451-65.

3. Sweeney R, Chen G, Gold L, Mensah F, Wake M. Mapping PedsQL(TM) scores onto CHU9D utility scores: estimation, validation and a comparison of alternative instrument versions. Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation 2020; 29(3): 639-52.

SECTION II: Measurement of health-related quality of life

Chapter 4: How do common conditions impact health-related quality of life for children? Providing guidance for validating pediatric preference-based measures

Published in Health and quality of life outcomes (2023) with Dalziel, K., Huang, L., Mulhern, B., & Carvalho, N.

Citation: Xiong, X., Dalziel, K., Huang, L., Mulhern, B., & Carvalho, N. (2023). How do common conditions impact health-related quality of life for children? Providing guidance for validating pediatric preference-based measures. Health and quality of life outcomes, 21(1), 8. https://doi.org/10.1186/s12955-023-02091-4

4.1. Abstract

Background: There is increasing interest in the validation of pediatric preference-based health-related quality of life measurement instruments. It is critical that children with various degrees of health-related quality of life (HRQoL) impact are included in validation studies. To inform patient sample selection for validation studies from a pragmatic perspective, this study explored HRQoL impairments between known-groups and HRQoL changes over time across 27 common chronic child health conditions and identified conditions with the largest impact on HRQoL.

Methods: The health dimensions of two common preference-based HRQoL measures, the EQ-5D-Y and CHU9D, were constructed using Pediatric Quality of Life Inventory (PedsQL) items that overlap conceptually. Data was from the Longitudinal Study of Australian Children, a nationally representative sample with over 10,000 children at baseline. Seven waves of data were included for the analysis, with child age ranging from 2 to18 years. Impacts to specific health dimensions and overall HRQoL between those having a specific condition versus not were compared using linear mixed effects models. HRQoL changes over time were obtained

by calculating the HRQoL differences between two consecutive time points, grouped by "Improved" and "Worsened" health status. Comparison among various health conditions and different age groups (2-4 years, 5-12 years and 13-18 years) were made.

Results: Conditions with the largest statistically significant total HRQoL impairments of having a specific condition compared with not having the condition were recurrent chest pain, autism, epilepsy, anxiety/depression, irritable bowel, recurrent back pain, recurrent abdominal pain, and attention deficit hyperactivity disorder (ADHD) for the total sample (2-18 years). Conditions with largest HRQoL improvement over time were anxiety/depression, ADHD, autism, bone/joint/muscle problem, recurrent abdominal pain, recurrent pain in other part, frequent headache, diarrhea and day-wetting. The dimensions included in EQ-5D-Y and CHU9D can generally reflect HRQoL differences and changes. The HRQoL impacts to specific health dimensions differed by condition in the expected direction. The conditions with largest HRQoL impacts differed by age group.

Conclusions: The conditions with largest HRQoL impact were identified. This information is likely to be valuable for recruiting patient samples when validating pediatric preference-based HRQoL instruments pragmatically.

Keywords: HRQoL, preference-based measures, CHU9D, EQ-5D-Y, PedsQL, validation, dimension, children, known-group validity, responsiveness

4.2. Background

Economic evaluation is increasingly used by policy makers to inform healthcare resource allocation decisions due to limited health budgets [1]. The preferred form of economic evaluation presents results as incremental cost per quality-adjusted life-years (QALYs) gained for one intervention relative to another. To calculate QALYs, preference-based health-related quality of life (HRQoL) data are required. Preference-based measures (PBMs) of HRQoL consist of two elements (descriptive systems and value sets) [2]. The focus of this paper is the validation of the descriptive system, which is important before the application of PBMs. Known-groups validity and responsiveness are the two key properties in validation studies.

Known-groups validity is demonstrated when a questionnaire can discriminate between two groups known to differ on the variable of interest [3]. Responsiveness is the ability of an instrument to depict a meaningful change in an indicator between baseline and sometime later [4]. They are the properties this study trying to provide supportive evidence to.

Two popular preference-based HRQoL instruments for children and adolescents are EQ-5D-Y-3L and the Child Health Utility 9 Dimensions (CHU9D) [5]. They are the top two most applied child-specific preference-based measures [6]. EQ-5D-Y-3L is appropriate for 4-15 years old, and CHU9D is appropriate for 4-17 years old [5]. There are some studies underway trialing proxy version of CHU9D and EQ-5D-Y (both 3L and 5L versions) with guidance notes for children 4 years old or younger [7, 8].

Pediatric Quality of Life Inventory (PedsQL) is an established, non-preference based profile generic instrument for HRQoL assessment in children and adolescents available across the age ranges of 2-18 years [9]. It is possible to use PedsQL as a proxy for the dimensions included in EQ-5D-Y and CHU9D instruments with the following reasons. PedsQL is well validated and has been used widely internationally to assess HRQoL across various heath conditions [10-12]. PedsQL has items that widely overlap with those used in pediatric PBMs [13]. Many studies have found correlations between PedsQL and EQ-5D-Y-3L and CHU9D and often use PedsQL as a gold standard for convergent validity in validation studies [4, 13-17]. There are mapping algorithms to infer EQ-5D and CHU9D utility scores from PedsQL [18-20]. Although our study focuses on providing evidence for validation for descriptive system for PBMs instead of utility (i.e., a single index which gives limited information on different health dimensions), the mapping algorithm added evidence to the common concept between PedsQL and EQ-5D-Y and CHU9D.

There are limitations with existing validation studies for EQ-5D-Y-3L and CHU9D. A frequently mentioned methodological limitation is that some dimensions of the measures are less validated [17, 21, 22]. For example, previous validation studies have found that very small proportions of children had any problems in “mobility” or “looking after myself” in the EQ-5D-Y and “daily routine” in the CHU9D, and thus had little HRQoL impairment in these

dimensions [17, 21, 22]. This limits both the strength and generalizability of the validation evidence. Another issue is that there is a lack of validation studies for younger children with a range of health conditions. For example, most validation studies include children above 7 years of age for EQ-5D-Y-3L [15], and children above 11 years of age for CHU9D [13, 14, 23]. Most study samples are of school children with limited or unspecified health conditions. In addition, responsiveness is less validated [24]. There is also a lack of development of appropriate instruments for young children under 5 years [25] (i.e., PBMs for children under 5 years old are either under development or lacking validation evidence). These all contribute to the increasing interest in carrying out future validation studies [8].

It is critical that children with various degree of impairment in HRQoL or expected changes in HRQoL are included in those validation studies. Otherwise, results may be misleading and lack generalizability. Validated instruments are in turn critical for evaluation and priority setting of programs, services, treatments and supports for children and their families. Therefore, it is important for us to know more about expected HRQoL decrements and changes across a range of common childhood conditions.

The aim of this study was to investigate the HRQoL impairment and HRQoL changes over time based on dimensions captured in CHU9D and EQ-5D-Y descriptive system inferred from PedsQL across a wide range of common pediatric conditions and child age. The results can help inform the recruitment of samples in future validation studies.

4.3. Methods

4.3.1. Sample

Data were from the Longitudinal Study of Australian Children (LSAC), a geographically representative sample of Australian children and their families. The LSAC commenced in 2004, followed by repeated biennial assessment (‘waves’) of over 10,000 children across two age cohorts (a birth cohort of children aged 0-1 year and a kindergarten cohort of children aged 4-5 years in 2003-2004). The LSAC sampling design and field methods are detailed elsewhere [26]. LSAC was approved by The Australian Institute of Family Studies Ethics

Committee, and families provided written informed consent. Seven waves of data (from 2004 to 2016) of both cohorts were used except for the first wave of the birth cohort because the children were under 2 years old and did not have HRQoL data collected.

4.3.2. HRQoL measurement

HRQoL data were available in the form of the PedsQL (Version 4.0), which measures four health dimensions: (1) physical, (2) emotional, (3) social, and (4) school functioning and contains 23 items (21 items for 2-4 years) [27]. The PedsQL was completed by the study child's primary caregiver, who rated the frequency of each item in the past month with a 5-point Likert scale from 0 (Never) to 4 (Almost always). Items were reversed scored and linearly transformed to a 0-100 scale (0=100, 1=75, 2=50, 3=25, 4=0) using recommended methods, with higher scores indicating better HRQoL[27]. In waves 6 and 7 of LSAC (children aged 10-18 years), CHU9D data (self-reported and asking about today) was also available.

4.3.3. Using PedsQL as a proxy for preference based instruments

We used PedsQL items to construct scores that closely mirror the dimensions in EQ-5D-Y and CHU9D. EQ-5D-Y is the youth version of the commonly used EQ-5D, containing five dimensions: (1) mobility, (2) looking after myself, (3) usual activities, (4) pain or discomfort, and (5) worried, sad or unhappy. EQ-5D-Y has two versions, 3L (3 response levels for each dimension) and 5L (5 response levels for each dimension). The inferred EQ-5D-Y scores are based on PedsQL items and PedsQL scoring algorithm and cannot tell between EQ-5D-Y-3L and EQ-5D-Y-5L. As EQ-5D-Y 3L and 5L share the same 5 dimensions, our results apply to both version of EQ-5D-Y and thus we only use the term of "EQ-5D-Y" in the following texts.

CHU9D was developed for children from its inception [28], containing nine dimensions: (1) worried, (2) sad, (3) pain, (4) tired, (5) annoyed, (6) schoolwork, (7) sleep, (8) daily routine, and (9) join in activities. Each CHU9D dimension contains five levels of severity. The selection of PedsQL items to represent CHU9D dimensions was straightforward due to use of the same or very similar wording. The selection of items for EQ-5D-Y was based on the description of the dimensions and referring to Scalone, L., et al. (2011) which proposed items

expected to be correlated and correlation coefficients, supplemented with team discussion where decisions were not clear [29]. We identified PedsQL items with overlapping conceptualization to all dimensions in CHU9D and EQ-5D-Y (Appendix 1). The PBMs' dimensions that had multiple PedsQL items identified (two CHU9D dimension and three EQ-5D-Y dimensions) were checked to ensure that the items within each dimension showed significant and mostly moderate correlations (correlation interpretation criteria: weak: <0.3; moderate: 0.3-0.6; strong: >0.6)[13, 30] (Appendix 2). For each CHU9D and EQ-5D-Y dimension, all relevant PedsQL items identified were averaged to calculate a corresponding dimension score. The total score was calculated as the average of the dimension scores. Both dimension and total scores for inferred CHU9D and EQ-5D-Y range from 0 to 100, with higher scores indicating better HRQoL.

4.3.4. Health conditions included

Within the LSAC survey parents reported whether their child had any ongoing health conditions, defined as a health problem that 'exists for some period of time (weeks, months, years) or re-occurs regularly'. If the answer was 'yes', parents were directed to select from a group of different health conditions which varied by age. In total, 33 ongoing conditions were identified for children from 2 to 18 years old. We excluded diseases with a sample size smaller than 30 in order to keep this task manageable and focused on common health conditions as opposed to those rarely reported (leading to the exclusion of palpitation, congenital heart disease and bedwetting). We also excluded diseases that were not specific, such as "other illness" because they can provide little instructive information (leading to the exclusion of other illness, other infection, and other physical disability). There were finally 27 health conditions included (details in **Appendix 3**).

4.3.5. Statistical analyses

We described the demographic information of the study sample including gender, indigenous status, parental education, the Socio-Economic Indexes for Areas (SEIFA), and whether the child had special health care needs [31]. Children were categorized into age groups according to the age cut off points of the PedsQL: 2-4 years, 5-12 years, 13-18 years.

To estimate the HRQoL impairment of having a specific condition compared with not having that condition, we used linear mixed effects models with random intercepts accounting for the hierarchical data structure due to repeated measurement of each child in the longitudinal dataset [32]. The child identifier was used as the cluster or random intercept variable. The dependent variable is HRQoL (total or dimension scores) and the independent variable is a binary variable 1/0, where 1 indicates having a specific health condition and 0 not having the condition. The coefficient of the independent variable indicates the HRQoL impairment, i.e., HRQoL difference between groups with and without the condition. The model simply adjusted for gender and age (continuous variable in years). 15.48% of the observations had multiple conditions. Only one condition was considered for each regression model with results ranked from largest total HRQoL impairment to smallest.

To assess the responsiveness to health changes over time, the HRQoL change between two consecutive time points was calculated using the HRQoL in the current wave minus the HRQoL in the previous wave (two-year interval between waves). Three groups were defined: "Worse", "Improved", and "Unchanged" (detailed definition for health change in **Appendix 4**). It is hypothesized that "Unchanged" group should have minor HRQoL changes over time, with the "Worse" group having decreased HRQoL, and the "Improved" group having increased HRQoL. The size of HRQoL change over time was assessed using the standardized response mean (SRM), which was calculated by dividing the mean change by the standard deviation of the change. The SRM can further facilitate comparison with other studies and interpretation, with SRM < 0.2 being considered small, 0.5 moderate, and 0.8 large [33].

4.3.6. Sensitivity analysis

We included narrower sets of relevant item(s) for dimensions with multiple PedsQL items to see if there is any difference in the result. The detailed items included in each version of sensitivity analysis and its corresponding results are at **Appendix 5**. We also used real CHU9D data (only available in 6th and 7th waves for 10-18 years old) to compare with the HRQoL impairment estimated using inferred CHU9D (**Appendix 6**).

4.4. Results

4.4.1. Participants

Table 1 shows the patient characteristics of the baseline sample and the observations included in this analysis. The combined data used for the analysis were generally similar with the baseline LSAC sample. There were 52,339 observations from all eligible waves, with 8.84% of the observations missing inferred EQ-5D-Y and CHU9D measures.

4.4.2. HRQoL impairment: regression results

The top 10 conditions with significant coefficients in HRQoL total scores are presented. Across all age groups, having a health condition was negatively associated with HRQoL. The shared eight conditions among the top 10 for inferred EQ-5D-Y and CHU9D are recurrent chest pain, autism, epilepsy, anxiety/depression, irritable bowel, recurrent back pain, recurrent abdominal pain, and attention deficit hyperactivity disorder (ADHD) for 2-18 year olds (Figure 4-1, Figure 4-2).

The top 10 conditions were different across age groups, with some common conditions (e.g. ADHD, recurrent abdominal pain). For children aged 2-4 years the top conditions with highest HRQoL impairment include ADHD, frequent headaches, soiling and diarrhoea. For children aged 5-18 years, anxiety/depression, autism, recurrent pain and epilepsy were amongst conditions with relatively high HRQoL impairment.

In general, ‘mobility’ in EQ-5D-Y and ‘daily routine’ in CHU9D had relatively small HRQoL impairment associated with having health conditions. The influence of a health condition on various health dimensions were as expected for typical conditions. For example, for anxiety/depression, the dimensions ‘worried, sad or unhappy’, ‘worried’ and ‘sad’ were associated with the largest HRQoL impairment, while for frequent headache, recurrent pain and bone/joint/muscle problem, the dimensions ‘pain or discomfort’ and ‘pain’ showed the largest HRQoL decrement. To be noted, anxiety/depression also showed significant HRQoL impairment in the dimensions ‘pain or discomfort’ and ‘pain’.

4.4.3. HRQoL change over time

As hypothesized HRQoL changes were generally positive in the "Improved" group and negative in the "Worse" group, except among the 2-4 year olds (Figure 4-3, Figure 4-4). The HRQoL change over time was trivial among children with health condition status unchanged (Appendix 4). HRQoL changes were generally larger when conditions worsened compared to when conditions improved, particularly for the inferred CHU9D. The shared nine conditions for inferred EQ-5D-Y and CHU9D among the top 10 largest SRM in the "Improved" group were anxiety/depression, ADHD, autism, bone/joint/muscle problem, recurrent abdominal pain, recurrent pain in other part, frequent headache, diarrhea and day-wetting in 2-18 years.

Around half of the top 10 conditions in all age groups except 2-4 years had small to moderate changes in overall HRQoL improvement (SRM: 0.2-0.5) for both instruments, with another half having small changes (SRM<0.2). The 2-4 years group all had small changes (SRM<0.2) except ear infection. Inferred CHU9D had larger HRQoL changes than inferred EQ-5D-Y, with anxiety/depression demonstrating moderate to large effects sizes (SRM: 0.5-0.8) when conditions worsened in 5-18 years old.

In 2-4 year olds, both "Improved" and "Worse" groups showed HRQoL improvement in the 'looking after myself', 'pain/discomfort' (EQ-5D-Y), 'sleep', 'daily routine' and 'pain' (CHU9D) dimensions. Additionally, the HRQoL change over time could be different by age group for the same condition. For example, 5-12 years old had larger HRQoL loss in 'sleep' and 'annoyed' when developing autism than 13-18 years old.

Generally, 'pain/discomfort' and 'worried/sad/unhappy' in EQ-5D-Y, and 'worried', 'sad' and 'pain' in CHU9D had relatively large HRQoL change over time. Again, the HRQoL changes across dimensions for typical health conditions were as expected. It is worth noting that depression had a larger HRQoL improvement than anxiety when conditions improved in 13-18 years old, with 'mobility' (EQ-5D-Y) and 'join in activities' (CHU9D) contributing most to this HRQoL increase.

4.4.4. Sensitivity analyses

The analyses using narrower sets of PedsQL items to represent the EQ-5D-Y and CHU9D dimensions were consistent with the main result (Appendix 5), confirming that using slightly different description and number of items had little impact on the main results. The inferred CHU9D and real CHU9D shared 5 common conditions (depression, recurrent chest pain, autism, anxiety, and soiling) among the top 10 conditions with significant HRQoL impairment. The real CHU9D data showed relatively large HRQoL impairment in recurrent conditions with consistent symptoms such as recurrent pain (back/abdominal/other part) and frequent headache. However, the inferred CHU9D demonstrated larger impact for conditions with less frequency but more severe outcomes, such as epilepsy, or with long term mental impact such as ADHD. To compare in the same condition, for example depression, real CHU9D had similar overall HRQoL impairment with inferred CHU9D. However, real CHU9D had much less impairment in 'worried', 'sad' and 'sleep' that are easily influenced by the mood of that particular day but had larger impairment in 'daily routine' which is more stable and also with more detailed description than the inferred CHU9D (Appendix 6).

4.5. Discussion

The shared conditions among the top 10 with significant coefficients on inferred EQ-5D-Y and CHU9D total scores for 2–18-year-olds were recurrent chest pain, ADHD, recurrent abdominal pain, recurrent back pain, epilepsy, anxiety/depression, irritable bowel, and autism. The shared conditions among the top 10 with largest changes in inferred EQ-5D-Y and CHU9D total score over time for 2-18 years were anxiety/depression, ADHD, autism, bone/joint/muscle problem, recurrent abdominal pain, recurrent pain in other part, frequent headache, diarrhea and day-wetting. The impacts to specific health dimensions differed by health conditions in the expected direction.

Identification of conditions with the largest HRQoL impact and which dimensions contribute to the impact may help inform the recruitment of patients in validation studies, especially for studies with limited budget and not being able to include a wide range of conditions. It is most

difficult to recruit patients with large HRQoL impairment in real life. Validation studies frequently reported limitations of samples lacking severe conditions and with high ceiling effects which limited the ability to validate the instruments and suggested further research in a range of clinical conditions.[13, 15, 23, 34, 35] Future studies could consider recruiting some pediatric patients from the top 10 conditions to guarantee the effectiveness and efficiency of validation on known-group validity and responsiveness. It can also help the recruitment for multiple-instrument comparison studies, where conditions with too small HRQoL impact may limit the comparison between instruments. For a 'real world' example, early findings from this work informed the study design of a large validation study for multiple pediatric PBMs comparisons [8]. The results from the 2-4 years old may be especially valuable since few validation studies have included this very young population. However, the results of 2-4 years old should be interpreted with caution as EQ-5D-Y and CHU9D themselves only have experimental versions which are under evaluation and our results are based on inferred EQ-5D-Y and CHU9D from PedsQL items.

It is worth noting that 2-4 years olds had HRQoL improvements over time observed in 'sleep', 'daily routine' and 'looking after myself' dimensions even in the "Worse" group. One reason for this phenomenon may be that HRQoL improvements due to natural developments with age outweighs the decrease due to newly developed conditions. This might suggest that more appropriate dimensions might be needed for this young group to effectively reflect relevant HRQoL changes. Previous studies echo this suggestion [36-38]. More studies are warranted for this very young population in PBMs development or adaptation from existing measures.

Another interesting finding is that the HRQoL changes over time were generally larger when conditions were newly developed compared to when conditions resolved. One explanation for this may be that people are more sensitive to loss than gains, known as loss aversion—that is, changes for the worse (losses) seem larger than equivalent changes for the better [39, 40]. These results support the consideration of loss aversion in economic evaluations.

The consistent results from sensitivity analysis using narrower sets of PedsQL items for EQ-5D-Y and CHU9D dimensions confirmed the robustness of our results considering slightly different wording and number of items for one dimension. The common results between inferred CHU9D and real CHU9D confirmed that they had basically the same concept, while the differences may be due to difference in the degree of detailed description, recall period and proxy report or self-report (inferred CHU9D score used PedsQL items that are parent-reported over a time period of 'the past month', while real CHU9D is child self-reported and asked about 'today'). The sensitivity analysis using real CHU9D data provided some support for our conceptual mapping method using PedsQL data, but also indicated the importance to consider the influence of the other factors on HRQoL measurement. The sensitivity analysis comparing inferred CHU9D and real CHU9D provided valuable information on the impact of parent proxy versus self-completion and recall period. The inferred CHU9D generally had larger or similar overall HRQoL impact than the real CHU9D (Appendix Figure 11), which may suggest that parents tend to worry more about health conditions than children themselves. The real CHU9D had much less impairment in 'worried', 'sad' and 'sleep' dimensions that are easily influenced by random factors of the particular day of being investigated, indicating that a short recall period may reduce the ability of instruments to detect meaningful HRQoL impact on these dimensions.

Our study has several strengths. A large sample with nationally representative children in Australia (over 10, 000 children at baseline) was used. We included 27 common chronic pediatric conditions and allowed for comparison of their impact on HRQoL based on PBM constructs within a single study, which has not been reported previously. A wide childhood ages (2-18 years) enables comparison between age groups, which is valuable. We included dimensions that mirror both EQ-5D-Y and CHU9D, providing useful comparisons across instruments. Furthermore, taking advantage of the longitudinal dataset, the exploration of the HRQoL changes over time can inform responsiveness testing across a wide range of conditions which has been relatively less studied [24].

There are also some limitations. First, we used PedsQL items to mirror EQ-5D-Y and CHU9D dimensions due to not having direct PBM measurement across ages. Although the

conceptualization of EQ-5D-Y and CHU9D dimensions overlapped with PedsQL items, there exist some differences, such as the exact wording of questions, the recall period and the different number of PedsQL items informing different dimensions, that may lead to different HRQoL scores. However, PedsQL has the same number of levels for each response as the CHU9D and the increasingly used EQ-5D-Y 5L, which added to its suitability to reflect the descriptive system of CHU9D and EQ-5D-Y 5L. Sensitivity analysis including narrower set of items for EQ-5D-Y dimension showed consistent results. Comparison between inferred CHU9D and real CHU9D measurement in a small subset of children showed the potential difference for appropriate interpretation and use of our results. Second, only three conditions were included in the analysis of HRQoL changes over time in 2-4 year olds due to the lack of data available for 0-1 years and small sample sizes for some health conditions. Thus, the conditions with largest HRQoL changes over time in 2-4 years old need to be interpreted with caution. Third, for 2-4 year olds, the two year interval may be too long to capture meaningful HRQoL changes over time because this young age is associated with rapid natural development whereby dimensions such as 'daily routine' and 'sleep' have a strong natural history of improvement linked to development/growing. Further HRQoL responsiveness research is needed for this age group. Finally, our analyses have focused on the descriptive systems of two common generic preference-based measures EQ-5D-Y and CHU9D, and the results may not apply to instruments with some very different health dimensions.

4.6. Conclusion

The relationship between childhood health conditions and HRQoL varies by health dimensions and age groups. Validation studies for children should include various conditions with a range of expected HRQoL impacts where relevant and possible. When there is difficulty including patients from some disease areas, top candidates from this study may be considered based on resources and aim of the validation study.

List of abbreviations

QALYs: quality-adjusted life-years; HRQoL: health-related quality of life; PBMs: Preference-based measures; CHU9D: Child Health Utility 9 Dimensions; PedsQL: Pediatric Quality of Life Inventory; LSAC: Longitudinal Study of Australian Children, SEIFA: Socio-Economic Indexes for Areas; SRM: standardized response mean; ADHD: attention deficit hyperactivity disorder.

4.7. Reference

[1] M. Krahn, S. Bryan, K. Lee, and P. J. Neumann, "Embracing the science of value in health," (in eng), *Canadian Medical Association Journal* vol. 191, no. 26, pp. E733-E736, Jul 2 2019.

[2] J. Brazier and M. Deverill, "A checklist for judging preference-based measures of health related quality of life: Learning from psychometrics," *Health Economics,* https://doi.org/10.1002/(SICI)1099-1050(199902)8:1<41::AID-HEC395>3.0.CO;2-# vol. 8, no. 1, pp. 41-51, 1999/02/01 1999.

[3] M. Davidson, "Known-Groups Validity," in *Encyclopedia of Quality of Life and Well-Being Research*, A. C. Michalos, Ed. Dordrecht: Springer Netherlands, 2014, pp. 3481-3482.

[4] D. Scott, G. D. Ferguson, and J. Jelsma, "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 and Quality of Life Outcomes,* vol. 15, no. 1, p. 12, 2017/01/19 2017.

[5] D. Rowen, O. Rivero-Arias, N. Devlin, and J. Ratcliffe, "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?," (in eng), *Pharmacoeconomics,* Jan 6 2020.

[6] J. Kwon, S. W. Kim, W. J. Ungar, K. Tsiplova, J. Madan, and S. Petrou, "Patterns, trends and methodological associations in the measurement and valuation of childhood health utilities," (in eng), *Quality of Life Research,* vol. 28, no. 7, pp. 1705-1724, Jul 2019.

[7] *Measuring & Valuing Health. A brief overview of the Child Health Utility 9D (CHU9D)*. Available: https://licensing.sheffield.ac.uk/product/CHU-9D

[8] R. Jones *et al.*, "Psychometric Performance of HRQoL Measures: An Australian Paediatric Multi-Instrument Comparison Study Protocol (P-MIC)," (in eng), *Children (Basel, Switzerland),* vol. 8, no. 8, p. 714, 2021.

[9] J. W. Varni, M. Seid, and C. A. Rode, "The PedsQL: measurement model for the pediatric quality of life inventory," (in eng), *Med Care,* vol. 37, no. 2, pp. 126-39, Feb 1999.

[10] S. Jalali-Farahani, F. A. Shojaei, P. Parvin, and P. Amiri, "Comparison of health-related quality of life (HRQoL) among healthy, obese and chronically ill Iranian children," (in eng), *BMC Public Health,* vol. 18, no. 1, p. 1337, Dec 4 2018.

[11] J. W. Varni, D. R. Globe, S. R. Gandra, D. J. Harrison, M. Hooper, and S. Baumgartner, "Health-related quality of life of pediatric patients with moderate to severe plaque psoriasis: comparisons to four common chronic diseases," (in eng), *European Journal of Pediatrics,* vol. 171, no. 3, pp. 485-92, Mar 2012.

[12] J. W. Varni, C. A. Limbers, and T. M. Burwinkle, "Impaired health-related quality of life in children and adolescents with chronic conditions: a comparative analysis of 10 disease clusters and 33 disease categories/severities utilizing the PedsQL 4.0 Generic Core Scales," (in eng), *Health Qual Life Outcomes,* vol. 5, p. 43, Jul 16 2007.

[13] K. D. Petersen, J. Ratcliffe, G. Chen, D. Serles, C. S. Frosig, and A. V. Olesen, "The construct validity of the Child Health Utility 9D-DK instrument," (in eng), *Health and Quality of Life Outcomes,* vol. 17, no. 1, p. 187, Dec 23 2019.

[14] P. Yang *et al.*, "Psychometric evaluation of the Chinese version of the Child Health Utility 9D (CHU9D-CHN): a school-based study in China," (in eng), *Quality of Life Research,* vol. 27, no. 7, pp. 1921-1931, Jul 2018.

[15] U. Ravens-Sieberer *et al.*, "Feasibility, reliability, and validity of the EQ-5D-Y: results from a multinational study," *Quality of Life Research,* vol. 19, no. 6, pp. 887-897, 2010/08/01 2010.

[16] K. D. Petersen, G. Chen, C. Mpundu-Kaambwa, K. Stevens, J. Brazier, and J. Ratcliffe, "Measuring Health-Related Quality of Life in Adolescent Populations: An Empirical Comparison of the CHU9D and the PedsQL(TM) 4.0 Short Form 15," (in eng), *Patient,* vol. 11, no. 1, pp. 29-37, Feb 2018.

[17] A. G. Canaway and E. J. Frew, "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," (in eng), *Quality of Life Research,* vol. 22, no. 1, pp. 173-83, Feb 2013.

[18] A. A. Shafie, I. K. Chhabra, J. H. Y. Wong, and N. S. Mohammed, "Mapping PedsQL™ Generic Core Scales to EQ-5D-3L utility scores in transfusion-dependent thalassemia patients," *The European Journal of Health Economics,* vol. 22, no. 5, pp. 735-747, 2021/07/01 2021.

[19] K. A. Khan, S. Petrou, O. Rivero-Arias, S. J. Walters, and S. E. Boyle, "Mapping EQ-5D Utility Scores from the PedsQL™ Generic Core Scales," *PharmacoEconomics,* vol. 32, no. 7, pp. 693-706, 2014/07/01 2014.

[20] R. Sweeney, G. Chen, L. Gold, F. Mensah, and M. Wake, "Mapping PedsQL(TM) scores onto CHU9D utility scores: estimation, validation and a comparison of alternative instrument versions," (in eng), *Qual Life Res,* vol. 29, no. 3, pp. 639-652, Mar 2020.

[21] M. Åström, C. Persson, M. Lindén-Boström, O. Rolfson, and K. Burström, "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,* vol. 27, no. 11, pp. 2859-2871, 2018/11/01 2018.

[22] G. Chen *et al.*, "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,* vol. 18, no. 4, pp. 432-438, 2015/06/01/ 2015.

[23] J. Ratcliffe, K. Stevens, T. Flynn, J. Brazier, and M. Sawyer, "An assessment of the construct validity of the CHU9D in the Australian adolescent general population," (in eng), *Quality of Life Research,* vol. 21, no. 4, pp. 717-25, May 2012.

[24] D. Rowen, A. D. Keetharuth, E. Poku, R. Wong, B. Pennington, and A. Wailoo, "A Review of the Psychometric Performance of Selected Child and Adolescent Preference-Based Measures Used to Produce Utilities for Child and Adolescent Health," (in eng), *Value Health,* vol. 24, no. 3, pp. 443-460, Mar 2021.

[25] D. Rowen, O. Rivero-Arias, N. Devlin, and J. Ratcliffe, "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?," (in eng), *Pharmacoeconomics,* vol. 38, no. 4, pp. 325-340, Apr 2020.

[26] C. Soloff, Lawrence, D., & Johnstone, R., "LSAC sample design (Technical Paper No. 1)," *Melbourne: Australian Institute of Family Studies. ,* 2005.

[27] J. W. Varni, T. M. Burwinkle, M. Seid, and D. Skarr, "The PedsQL™* 4.0 as a Pediatric Population Health Measure: Feasibility, Reliability, and Validity," *Ambulatory Pediatrics,* vol. 3, no. 6, pp. 329-341, 2003/11/01/ 2003.

[28] K. Stevens, "Developing a descriptive system for a new preference-based measure of health-related quality of life for children," (in eng), *Quality of Life Research,* vol. 18, no. 8, pp. 1105-13, Oct 2009.

[29] L. Scalone *et al.*, "Assessing quality of life in children and adolescents: Development and validation of the Italian version of the EQ-5D-Y," *Italian Journal of Public Health,* vol. 8, pp. 331-341, 12/01 2011.

[30] T. W. Anderson and J. D. Finn, *The new statistical analysis of data*. Springer Science & Business Media, 2012.

[31] C. D. Bethell, D. Read, R. E. Stein, S. J. Blumberg, N. Wells, and P. W. Newacheck, "Identifying children with special health care needs: development and evaluation of a short screening instrument," *Ambulatory Pediatrics,* vol. 2, no. 1, pp. 38-48, 2002.

[32] P. Diggle, P. J. Diggle, P. Heagerty, K.-Y. Liang, P. J. Heagerty, and S. Zeger, *Analysis of longitudinal data*. Oxford University Press, 2002.

[33] B. Mulhern *et al.*, "Using generic preference-based measures in mental health: psychometric validity of the EQ-5D and SF-6D," *British Journal of Psychiatry,* vol. 205, no. 3, pp. 236-243, 2018.

[34] B. Mulhern and K. Meadows, "The construct validity and responsiveness of the EQ-5D, SF-6D and Diabetes Health Profile-18 in type 2 diabetes," (in eng), *Health Qual Life Outcomes,* vol. 12, p. 42, Mar 24 2014.

[35] R. T. Wolf, J. Ratcliffe, G. Chen, and P. Jeppesen, "The longitudinal validity of proxy-reported CHU9D," *Quality of Life Research,* vol. 30, no. 6, pp. 1747-1756, 2021/06/01 2021.

[36] J. Verstraete, L. Ramma, and J. Jelsma, "Validity and reliability testing of the Toddler and Infant (TANDI) Health Related Quality of Life instrument for very young children," *Journal of Patient-Reported Outcomes,* vol. 4, no. 1, p. 94, 2020/11/09 2020.

[37] J. Verstraete, L. Ramma, and J. Jelsma, "Item generation for a proxy health related quality of life measure in very young children," *Health and Quality of Life Outcomes,* vol. 18, no. 1, p. 11, 2020/01/14 2020.

[38] J. Verstraete, A. Lloyd, D. Scott, and J. Jelsma, "How does the EQ-5D-Y Proxy version 1 perform in 3, 4 and 5-year-old children?," *Health and Quality of Life Outcomes,* vol. 18, no. 1, p. 149, 2020/05/24 2020.

[39] S. A. Lipman, W. B. F. Brouwer, and A. E. Attema, "A QALY loss is a QALY loss is a QALY loss: a note on independence of loss aversion from health states," *The European Journal of Health Economics,* vol. 20, no. 3, pp. 419-426, 2019/04/01 2019.

[40] D. Kahneman and A. Tversky, "Choices, values, and frames," in *Handbook of the fundamentals of financial decision making: Part I*: World Scientific, 2013, pp. 269-278.

4.8. Tables and Figures

Table 4-1 Patient characteristics of the study sample

Sample characteristics	Baseline sample		All the observations used
	B cohort wave 2 (N=4606)	K cohort wave 1 (N=4983)	(N=9774, observations=52993)
Male, yes, %	51.0	51.0	51.1
Indigenous, yes, %	3.9	3.8	3.1
Special health care needs, yes, %	11.3	13.2	15.8
Primary carer's education with bachelor, yes, %	33.9	28.1	34.4
SEIFA, mean (SD)	1008.5(74.1)	1005.7(78.3)	1009.8(75.5)

Notes: SEIFA: socio-economic index for areas. SD: standard deviation. Descriptive information is given as arithmetic means and standard deviations (SD) or frequencies and percentages (%). The baseline sample of B Cohort for this study is wave 2 since B cohort has no HRQoL data in wave 1 (0-1 year). The sample sizes of B cohorts from Wave 2 to 7 are 4606, 4386, 4242, 4085, 3764, 3381 respectively, totaling 24464. The sample sizes of K cohorts from Wave 1 to 7 are 4983, 4464, 4331, 4169, 3956, 3537, 3089 respectively, totaling 28529.

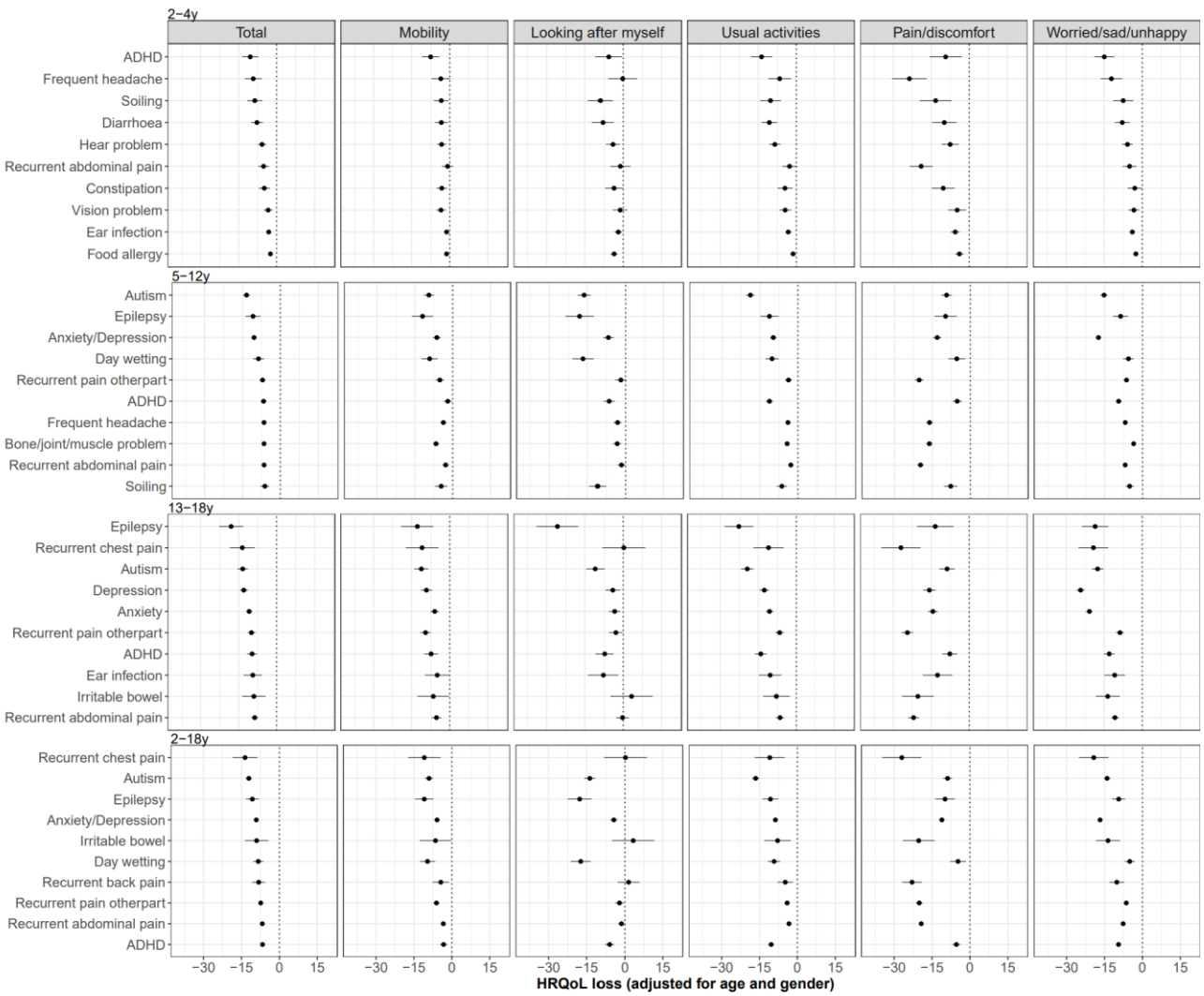


Figure 4-1 Associations between different health conditions and HRQoL across age groups based on inferred EQ-5D-Y

Notes: 1. The x axis represents the HRQoL difference between those with the condition compared with those without the condition. The points are the coefficients, with the line indicating the 95% confidence interval. The conditions are ranked according to the size of the HRQoL difference. **2.** Anxiety and depression were only measured separately from the 6th wave of LSAC (10-13 years old for B cohort and 14-17 years old for K cohort), and were measured together between 4th -7th waves (6-17 years). Thus, anxiety or depression were presented as one category in 5-12 years old age group, and as two separate categories in 13-18 years old.

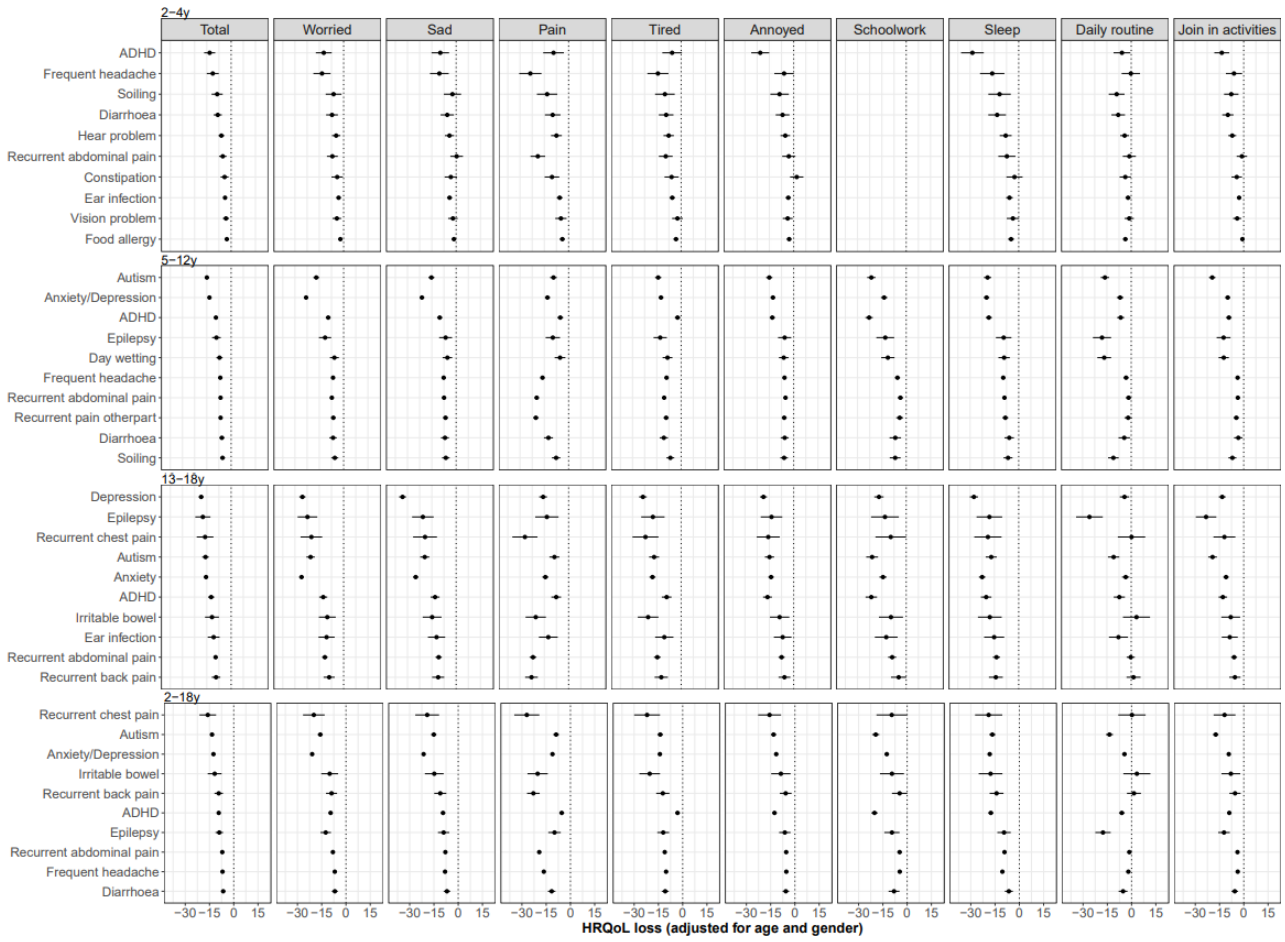


Figure 4-2 Associations between different health conditions and HRQoL across age groups based on inferred CHU9D

Notes: 1. 2-4 years old usually don't attend school so the 'schoolwork' dimension is missing. **2.** Anxiety and depression were only measured separately from the 6th wave of LSAC (10-13 years old for B cohort and 14-17 years old for K cohort), and were measured together between 4th -7th waves (6-17 years). Thus, anxiety or depression were presented as one category in 5-12 years old age group, and as two separate categories in 13-18 years old.

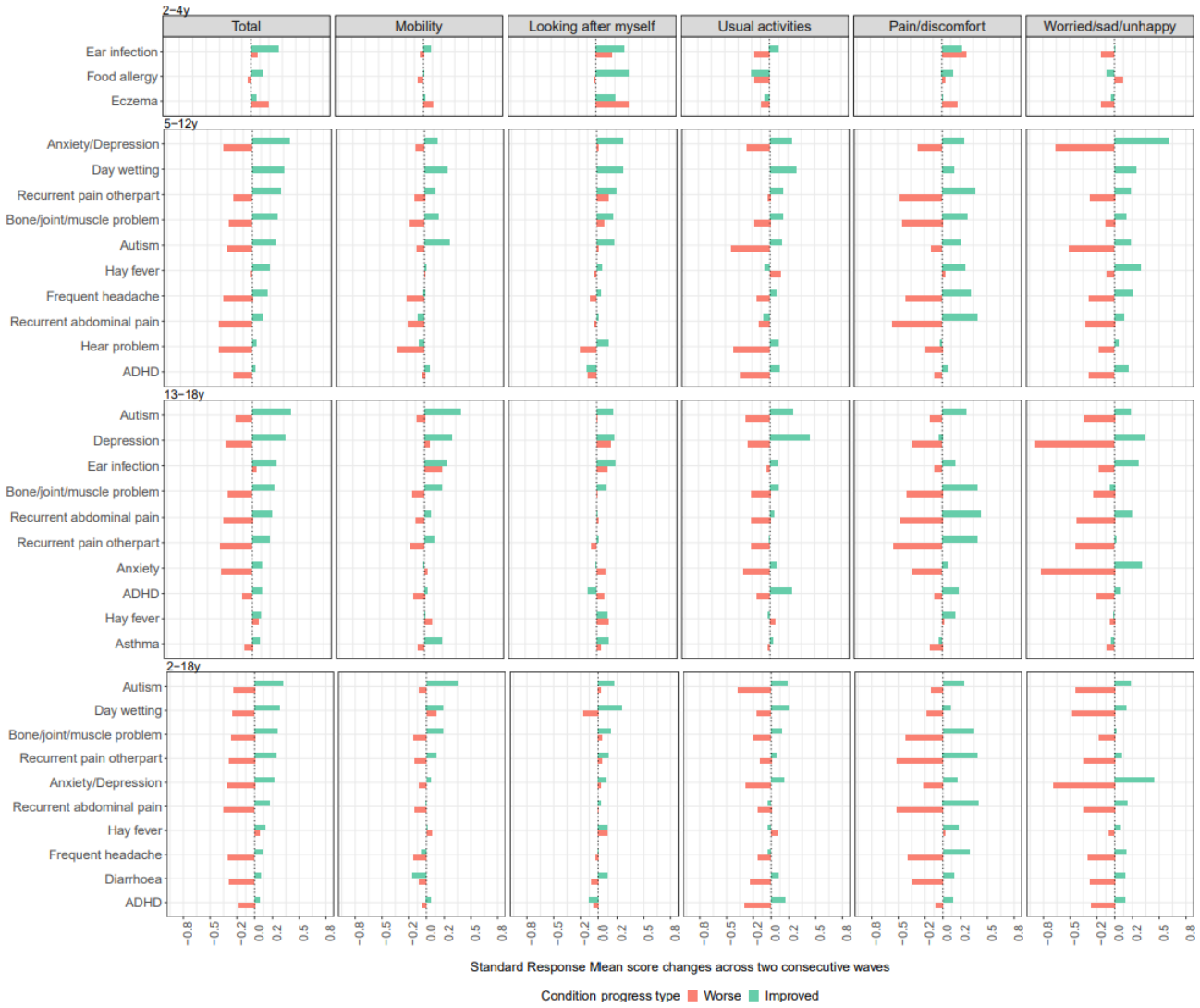


Figure 4-3 The HRQoL changes of different health conditions over a two-year period based on inferred EQ-5D-Y

Notes: 1. The x axis represents the HRQoL changes measured in standard response mean (SRM). The "Improved" represents those who recovered from the condition while the "Worse" represents those who newly developed the condition. Health conditions are ranked according to the HRQoL changes of the "Improved" group. **2.** Anxiety and depression were only measured separately from the 6th wave of LSAC (10-13 years old for B cohort and 14-17 years old for K cohort), and they were measured together between 4th -7th waves (6-17 years). Thus, anxiety or depression were presented as one category in the 5-12 years old age group, and as two separate categories in 13-18 years old. **3.** Children aged 2-4 years old had only three conditions available with status changes given no HRQoL data was available for 0–1-year-olds.

Figure 4-4 The HRQoL changes of different health conditions over a two-year period based on inferred CHU9D

Notes: 1. The x axis represents the HRQoL changes measured in standard response mean (SRM). The "Improved" represents those who recovered from the condition while the "Worse" represents those who newly developed the condition. Health conditions are ranked according to the HRQoL changes of the "Improved" group. **2.** Anxiety and depression were only measured separately from the 6th wave of LSAC (10-13 years old for B cohort and 14-17 years old for K cohort), and they were measured together between 4th -7th waves (6-17 years). Thus, anxiety or depression were presented as one category in 5-12 years old age group, and as two separate categories in 13-18 years old. **3**. Children aged 2-4 years old had only three conditions available with status changes given no HRQoL data was available for 0-1 year olds.

4.9. Supplementary materials

Appendix 1 – Relevant PedsQL items identified for EQ-5D-Y and CHU9D

Appendix 1 Table

Dimension	Relevant PedsQL items identified

EQ-5D-Y dimensions	
Mobility (Walking about)	1.1 Walking more than 100 meters
	1.2 Running
Looking after myself (washing or dressing)	1.5 Taking a bath or shower by him or herself
Doing usual activities (for example, going to school, hobbies, sports, playing, doing things with family or friends)	1.3 Participating in sports activity or exercise
	1.6 Doing chores around the house
	3.4 Not being able to do things that other children his or her age can do
	3.5 Keeping up when playing with other children
	4.3 Keeping up with schoolwork
Having pain and discomfort	1.7 Getting aches and pains
Feeling worried, sad and or unhappy	1.8 Having a low energy
	2.1 Feeling afraid or scared
	2.2 Feeling sad
	2.3 Feeling angry
	2.4 Having trouble sleeping
	2.5 Worrying about what will happen to him or her
CHU9D dimensions	
Worried	2.1 Feeling afraid or scared
	2.5 Worrying about what will happen to him or her
Sad	2.2 Feeling sad
Pain	1.7 Getting aches and pains
Tired	1.8 Having a low energy level
Annoyed	2.3 Feeling angry
School work/ homework	4.3 Keeping up with schoolwork
Sleep	2.4 Having trouble sleeping
Daily routine (eating, having a bath/ Shower, getting dressed)	1.5 Taking a bath or shower by him or herself
Able to join in activities (playing out with friends, doing sports, joining things)	1.3 Participating in sports activity or exercise 3.1 Getting along with other children 3.5 Keeping up when playing with other children

Note: 1. The description of the PedsQL items used the 8-12 years parent reported version. The different versions by age are mostly the same, with a few word differences in description. **2.** The relevant PedsQL items for EQ-5D-Y dimensions are based on the description of the dimension and referred to in the literature: Scalone, L., et al.

(2011). "Assessing Quality of Life in Children and Adolescents: Development and Validation of the Italian Version of EQ-5D-Y." **3.** We chose the PedsQL items that most closely matched the description of CHU9D dimensions. When it was unclear whether an item was relevant, we took a conservative approach and excluded the item.

Appendix 2 – Correlation of identified PedsQL items within one dimension

Appendix 2 Table

Internal correlation of PedsQL items representing CHU9D dimension 'worried'

	Feeling afraid or scared
Worrying about what will happen to him or her	0.4962*

Internal correlation of PedsQL items representing CHU9D dimension 'Join in activities'

	Participating in sports activity	Getting along with other children
Getting along with other children	0.3851*	
Keeping up when playing with other children	0.4809*	0.5218*

Internal correlation of PedsQL items representing EQ-5D-Y dimension 'mobility'

	Walking more than 100 meters
Running	0.7448*

Internal correlation of PedsQL items representing EQ-5D-Y dimension 'usual activities'

	Participating in sports activity	Doing chores	Not being able to do things that other children can do	Keeping up when playing with other children
Doing chores	0.3356*			
Not being able to do things that other children can do	0.3407*	0.2408*		
Keeping up when playing with other children	0.4809*	0.2672*	0.5272*	
Keeping up with schoolwork	0.3499*	0.3180*	0.3941*	0.5205*

Internal correlation of PedsQL items representing EQ-5D-Y dimension 'worried, sad or unhappy'

	Having a low energy	Feeling afraid or scared	Feeling sad	Feeling angry	Having trouble sleeping

Feeling afraid or scared	0.3607*				
Feeling sad	0.3998*	0.5143*			
Feeling angry	0.3013*	0.3905*	0.5024*		
Having trouble sleeping	0.3384*	0.3671*	0.3618*	0.3336*	
Worrying about what will happen to him or her	0.3304*	0.4962*	0.5110*	0.3663*	0.4042*

* *p* value <0.01.

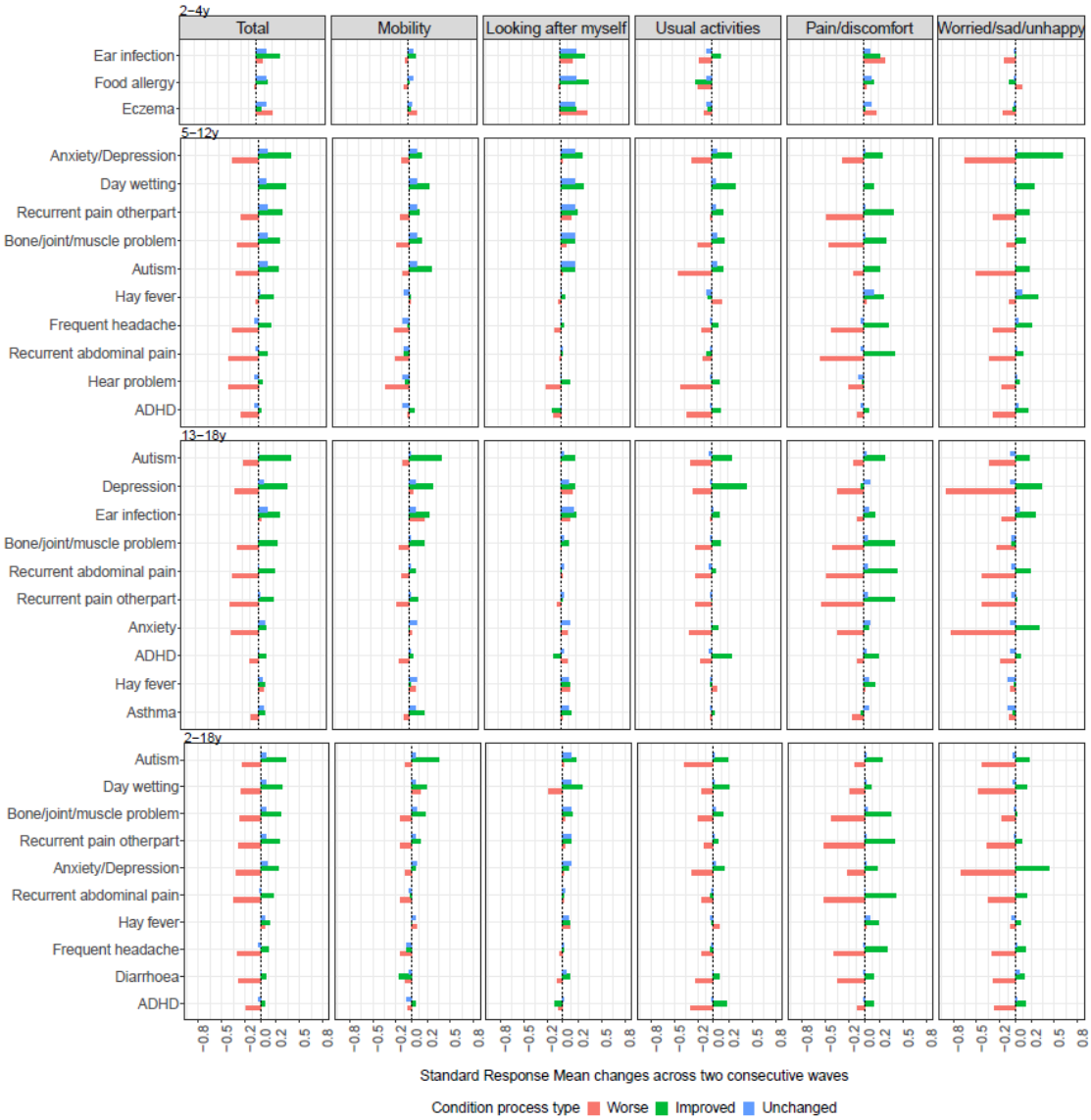
Note: Most researchers agree <0.1 indicates negligible correlation, >0.9 very strong correlation. Values in between are disputable. Many studies have used 0.30 as the cutoff between weak and moderate correlation. [1-3]

Appendix 3 – Health conditions included in the analysis

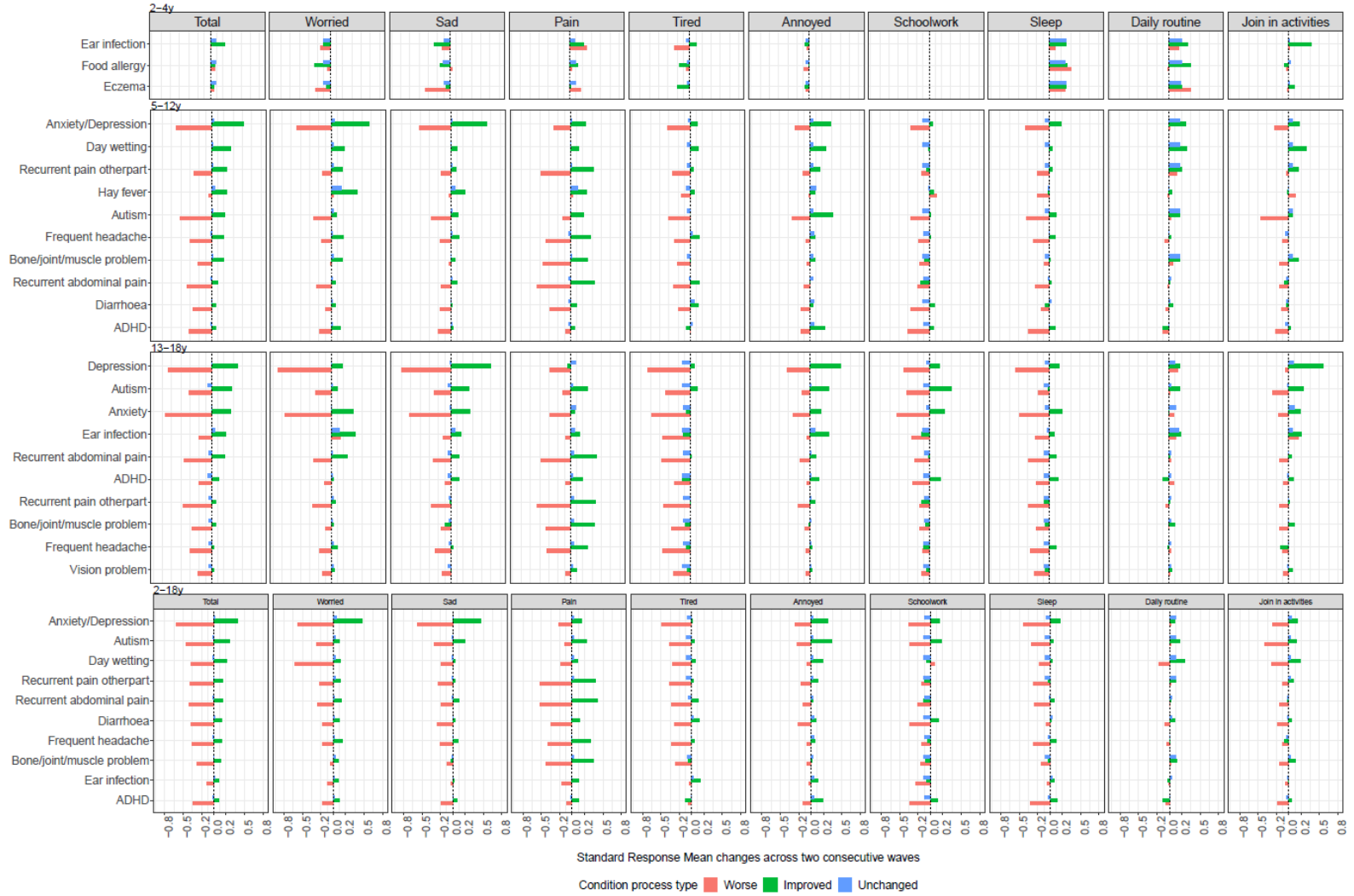
The health conditions finally included were eczema, vision problem, hay fever, asthma, acne, food allergy, ear infection, anxiety, depression, bone/joint/muscle problem, frequent headache, recurrent adnominal pain, hearing problem, tonsillitis, constipation, recurrent back pain, autism, recurrent pain in other part, attention deficit hyperactivity disorder (ADHD), soiling, day-wetting, irritable bowel, diarrhea, recurrent chest pain, epilepsy, chronic fatigue and diabetes (27 conditions in total).

Appendix 4 – The HRQoL change over time including the unchanged group (‘00’ and ‘11’)

For every two waves, there were four possible states that a child could be in: 1) not having one health condition in the previous wave or the current wave, ‘00’; 2) not having the condition in the previous wave, but having it in the current wave, ‘01’; 3) having the condition in the previous wave, but not having it in the current wave; ‘10’; 4) having the condition in both waves, ‘11’. Three groups were defined: “Worse” (‘01’), “Improved” (‘10’), and “Unchanged” (‘00’, ‘11’).



Appendix 4 Figure 1 The total and dimension score changes (SRM) of inferred EQ-5D-Y over 2 years



Appendix 4 Figure 2 The total and dimension score changes (SRM) of inferred CHU9D over 2 years

Reference:

[1] K. D. Petersen, J. Ratcliffe, G. Chen, D. Serles, C. S. Frosig, and A. V. Olesen, "The construct validity of the Child Health Utility 9D-DK instrument," (in eng), *Health and Quality of Life Outcomes,* vol. 17, no. 1, p. 187, Dec 23 2019.

[2] B. Mulhern *et al.*, "Using generic preference-based measures in mental health: psychometric validity of the EQ-5D and SF-6D," *British Journal of Psychiatry,* vol. 205, no. 3, pp. 236-243, 2018.

[3] K. Burström, Å. Bartonek, E. Broström, S. Sun, and A.-C. Egmar, "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,* vol. 103, no. 4, pp. 426-435, 2014.

Appendix 5 – Sensitivity analysis: including a narrower set of items in dimensions

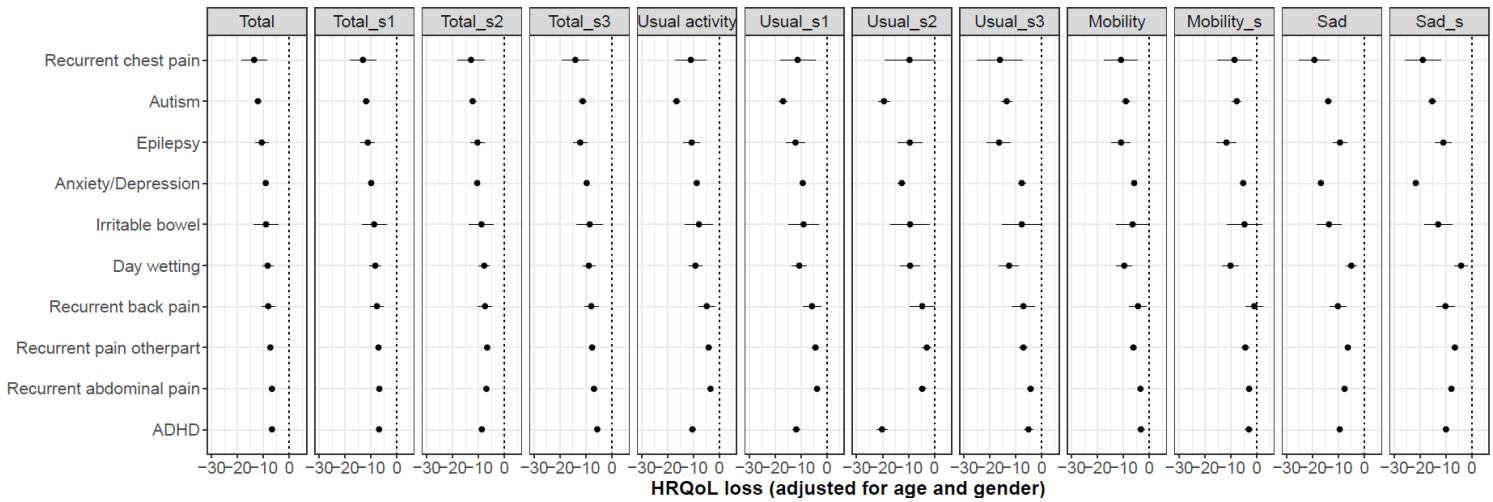
Appendix 5 Table 1 Items included in dimensions in sensitivity analysis of EQ-5D-Y and CHU9D

EQ-5D-Y dimension s1	PedsQL items
Mobility_s	1.1 Walking more than 100 meters 1.2 Running
Sad and Worried_s	1.8 Having a low energy 2.1. Feeling afraid or scared 2.2 Feeling sad 2.3 Feeling angry 2.4 Having trouble sleeping 2.5 Worrying about what will happen to him or her
Usual Activities_s1	1.3. Participating in sports activity or exercise 1.6. Doing chores around the house 3.4. Not being able to do things that other children his or her age can do 3.5 Keeping up when playing with other children 4.3 Keep up with school activities/ schoolwork
Usual Activities_s2	1.3. Participating in sports activity or exercise 1.6. Doing chores around the house 3.4. Not being able to do things that other children his or her age can do 3.5. Keeping up when playing with other children 4.3 Keep up with school activities/ schoolwork
Usual Activities_s3	1.3 Participating in sports activity or exercise 1.6. Doing chores around the house 3.4. Not being able to do things that other children his or her age can do 3.5. Keeping up when playing with other children 4.3 Keep up with school activities/ schoolwork
CHU9D dimension s1	PedsQL items
Worried_s	2.1 Feeling afraid or scared 2.5 Worrying about what will happen to him or her

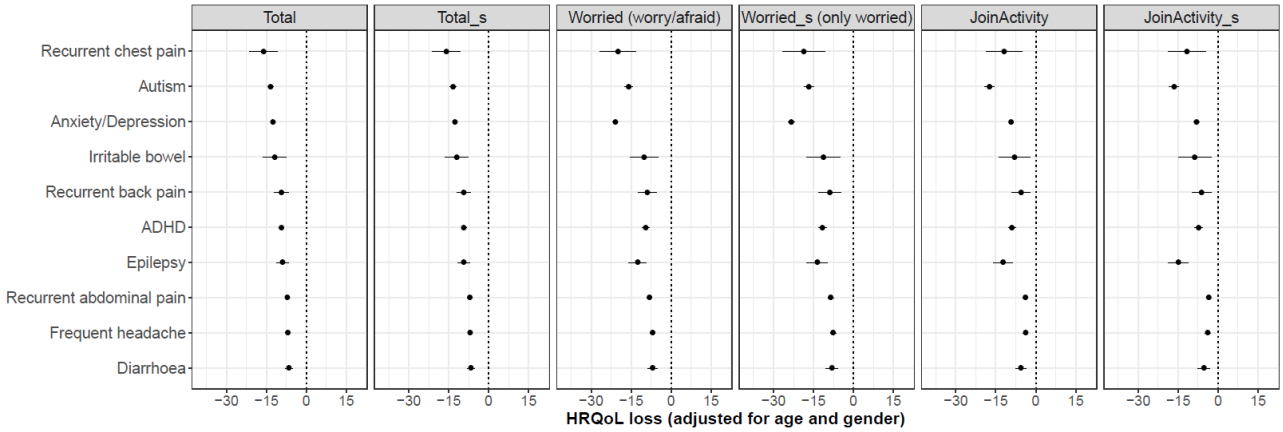
Join in activities_s (playing out with friends, doing sports, joining things)	1.3 Participating in sports activity or exercise 3.1 Getting along with other children 3.5 Keeping up when playing with other children

Note: The grey items were included in the main paper but were excluded in the sensitivity analyses.

Including a narrower set of items to calculate dimension scores generally does not change much of the total scores, although there is some difference in dimension scores.

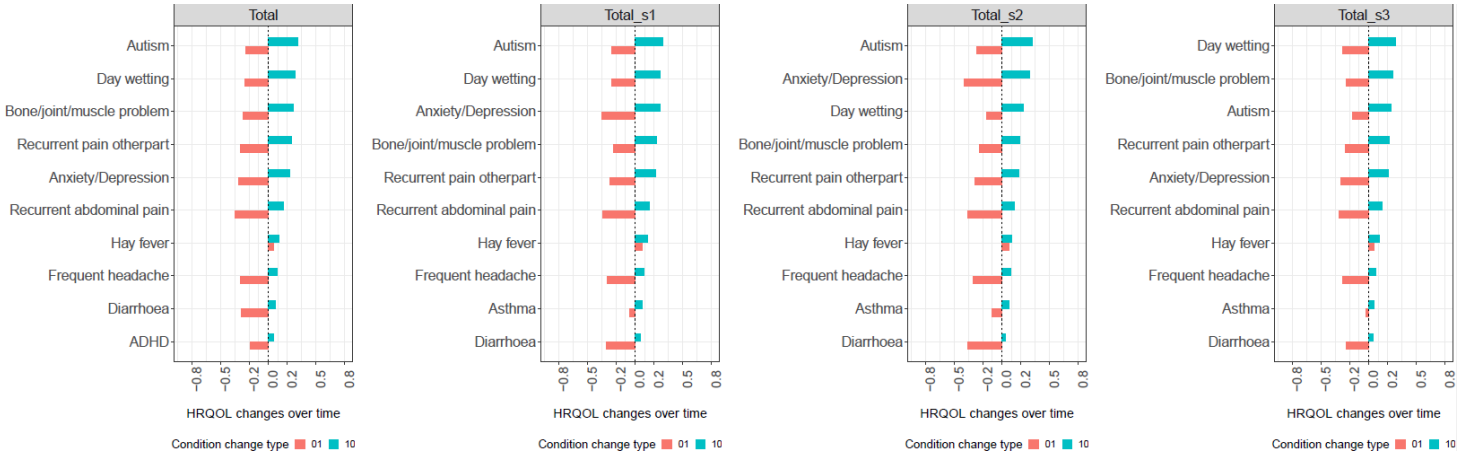


Appendix 5 Figure 1 Compare HRQoL loss measured by inferred EQ-5D-Y when including different items in 2-18 years old (Total_s1 uses Mobility_s, Sad/worry_s and Usual_s1. Total_s2 and Total_s3 only differed by using Usual_s2 and Usual_s3)



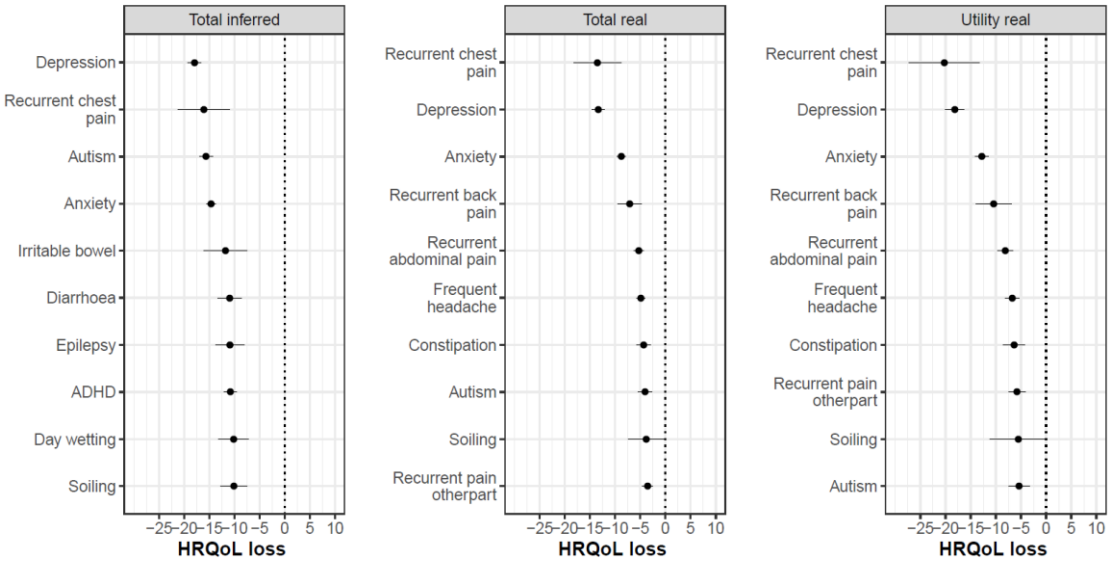
Appendix 5 Figure 2 Compare HRQoL loss measured by inferred CHU9D when including different items in 2-18 years old ('Total_s' incldued 'Worried_s' and 'JoinActivity_s".)

The top 10 conditions with largest HRQoL improvement are generally the same in different total scores, only the last condition differed. The rank of conditions differed a bit (as expected).

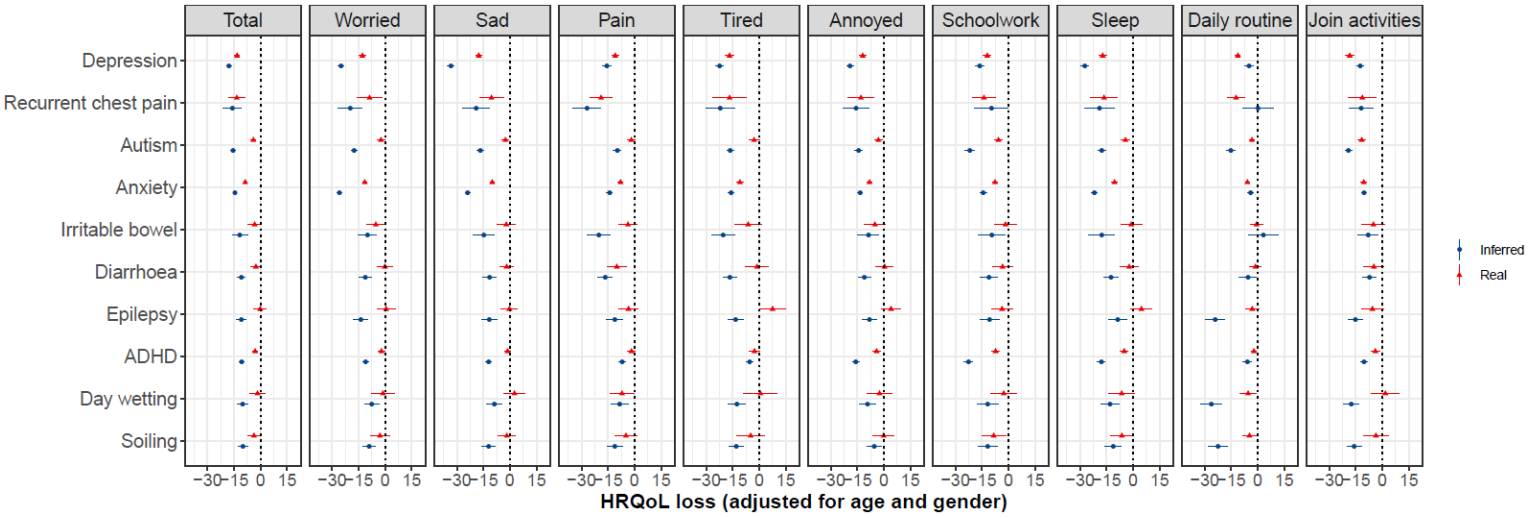


Appendix 5 Figure 3 Compare top 10 conditions with the largest inferred EQ-5D-Y total score changes over time when including different items in dimensions in 2-18 years old (Total_s1 includes Usual_s1. Total_s2 includes Usual_s2. Total_s3 includes Usual_s3)

Appendix 6 – Using real CHU9D data to estimate HRQoL impairment



Appendix 6 Figure 1 Comparing the top 10 conditions with largest HRQoL impairment associated with health conditions between inferred and real CHU9D



Appendix 6 Figure 2 Compare HRQoL impairment between inferred and real CHU9D average total scores

Notes: CHU9D utility were calculated by Australian adolescent value set[1]. CHU9D utilities was transformed to 0-100 by multiply the original utility with 100 to have comparable regression coefficients with CHU9D like score. For *daily routine*, the inferred CHU9D had a simpler description than the real CHU9D. Real CHU9D had *Daily routine (eating, having a bath/ Shower, getting dressed),* while inferred CHU9D had *4.3.Taking a bath or shower by him or herself.*

Reference

[1] J. Ratcliffe, T. Flynn, F. Terlich, K. Stevens, J. Brazier, and M. Sawyer, "Developing Adolescent-Specific Health State Values for Economic Evaluation," *PharmacoEconomics,* vol. 30, no. 8, pp. 713-727, 2012/08/01 2012.

Chapter 5: Psychometric properties of Child Health Utility 9D (CHU9D) proxy version administered to parents and caregivers of children aged 2-4 years compared with Pediatric Quality of Life Inventory™ (PedsQL)

Published in Pharmacoeconomics (2024) with Carvalho N, Huang L, Chen G, Jones R, Devlin N, Mulhern B, Dalziel K.

Citation: Xiong X, Carvalho N, Huang L, Chen G, Jones R, Devlin N, Mulhern B, Dalziel K. Psychometric Properties of Child Health Utility 9D (CHU9D) Proxy Version Administered to Parents and Caregivers of Children Aged 2-4 Years Compared with Pediatric Quality of Life Inventory™ (PedsQL). Pharmacoeconomics. 2024 Jan 27. https://doi.org/10.1007/s40273-024-01355-1

5.1. Abstract

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

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

Results: 842 caregivers completed the survey at baseline, with 513 completing the follow-up survey. The CHU9D did not demonstrate ceiling effects in the sample with special health care needs, with only 6% of respondents reporting best levels for all 9 dimensions. CHU9D correlated with PedsQL moderately to strongly between comparable items (correlation coefficients 0.34-0.70). CHU9D was able to differentiate between groups with known health differences with moderate to large effect sizes (Cohen D: 0.58-2.03). Moderate test-retest reliability was found for CHU9D in those reporting no health change at a 2-day follow-up (ICC: 0.52). A standard response mean (SRM) of 0.25–0.44 was found for children with changes in general health and SRM of 0.72-0.82 for children who reported worsened health when developing new illnesses, indicating small to large responsiveness according to different

definitions of health changes. Compared with PedsQL, CHU9D had similar known-group validity and responsiveness, and slightly poorer test-retest reliability.

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

Keywords: CHU9D, HRQoL, 2-4 years, psychometric properties, health utility instrument

5.2. Key Points for Decision Makers

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

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

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

5.3. Introduction

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

Many economic evaluations for younger ages used utilities obtained from generic pediatric preference-accompanied measures developed for older age groups or adults [11, 12]. This is problematic [13] as there is evidence that children under 5 years old have different developmental stages and may have different quality of life dimensions or constructs compared to older populations [14]. It is questionable whether instruments having common health dimensions with versions for older children or adults are suitable for use in younger children directly; they usually have adapted wording (or added guidance notes) and different report types (e.g, proxy-report or self-report), which often requires further validation evidence [15]. Health technology assessment authorities in Australia and the UK have also noted the lack of utilities used in pediatric economic evaluations and promote the use of concise, generic measures of pediatric HRQoL accompanied by relevant value sets [16, 17]. There is potential to unfairly penalize young children in the health technology assessment process due to poor quality, missing or uncertain utility evidence [18]. It is therefore important to explore appropriate HRQoL measurement in young children.

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

Recently, five HRQoL measures have become available with the potential for cost-utility analysis for children aged 2-4 years old, all with limited validation evidence. They are the EuroQol Toddler and Infant Populations (for children aged 0-3 years) instrument [21], the Health Status Classification System for Pre-School Children (for children aged 2.5-5 years)

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

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

The primary objective of this study was to assess the psychometric properties of CHU9D proxy version administered to parents or caregivers of Australian children aged 2-4 years compared with the PedsQL. Specifically, we aimed to assess the CHU9D's feasibility, ceiling/floor effects, test-retest reliability, convergent and divergent validity, known-group

validity, and responsiveness, compared with the PedsQL. We hypothesized that the CHU9D would show good convergent validity with PedsQL due to their similar constructs. Other tests were exploratory due to little previous evidence for measurement of HRQoL for children aged 2-4 years.

5.4. Method

5.4.1. Sample

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

5.4.2. Survey

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

At the beginning of the initial survey, screening questions were presented to establish the eligibility of participants [42]. Respondents who consented would proceed with the survey. The survey then asked participants about their socio-demographic characteristics including age, gender, language, income, education and general health status of their child. The survey asked if the child had any chronic conditions that have lasted or are likely to last for six months or more. If yes, then the caregivers would be prompted to select listed conditions. Only conditions with sample sizes equal to or larger than 30 were included in the analysis.

The next survey section presented multiple HRQoL instruments including CHU9D and PedsQL, with the order of these instruments randomized to minimize order and survey fatigue effects [44]. The order of the instruments was the same for the initial and follow-up survey for each participant. Questions about changes in the child's health status since the first survey were included in the follow-up survey. Time to complete sections of the survey was also recorded on the online REDcap system.

5.4.3. HRQoL instruments

The CHU9D has a proxy version with guidance notes for reporting HRQoL of children under 5 years. The CHU9D asks parents/caregivers to report their child's HRQoL today [33]. The CHU9D consists of 9 dimensions (worried, sad, pain, tired, annoyed, schoolwork/homework, sleep, daily routine, and able to join in activities), with five levels of responses for each dimension. The developer of CHU9D developed the guidance notes, with input from other health outcome researchers. The guidance notes provide additional instructions and adaptations on how to interpret schoolwork/homework, daily routine and able to join in activities questions for children aged under 5 years (**Appendix Table S1)**. In this study, the CHU9D scoring algorithms, developed based on preferences obtained from Australian adolescents, were applied to calculate and report CHU9D utilities, with the UK adult weights used for sensitivity analysis [45, 46]; no specific value set was available for CHU9D proxy version with guidance notes for children under 5 years.

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

5.4.4. Psychometric analyses

Several subgroups were defined to facilitate analysis: sub-groups defined by variables including general health status (excellent, very good, good, fair, poor), having special health care needs (yes, no), having a chronic health condition (yes, no), or general health status

change (much better, somewhat better, about the same, somewhat worse, much worse). More details of classifications are available in the relevant sections below.

Acceptability and feasibility

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

Ceiling/floor effects

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

Test-retest reliability

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

Convergent and divergent validity

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

Known group validity

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

Responsiveness

Responsiveness is used to demonstrate the extent to which an instrument's response reflects changes in underlying health status [19]. Caregivers were asked to report their child's general health status change, general health status change related specifically to the initially reported main condition, and health change related to new events occurring during follow up (e.g., new illness or treatment) at the follow-up survey. We identified two subgroups for the analysis of responsiveness: "Improved" (answer of "much better"), "Worsened" (answers of "somewhat worse" or "much worse" combined because of small sample size). Mean changes in scores between baseline and follow-up were tested by paired *t* test in each group [59]. One sided P values were used as we had specific hypothesis for the direction of the changes [60]. Standard

response means (SRM) or Cohen's D within groups is another type of effect sizes and is widely used to assess responsiveness [61, 62]. The SRM was computed by dividing the mean score change by the standard deviation of the change. The magnitude of responsiveness was evaluated using conventional threshold according to Cohen, with <0.2 deemed as trivial, 0.2 to <0.5 small, 0.5 to <0.8 medium, and >=0.8 large [55]. Both SRM and Cohen's D are methods to calculate effect sizes, but they are typically used in different contexts. SRM is most used for within-group comparisons over time to assess instrument responsiveness, while Cohen's D is more versatile and used for between-group as well as within-group comparisons. Cohen's D within groups (or paired samples Cohen's D) shares the same formula as SRM, and the two terms are sometimes used interchangeably.

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

5.5. Results

5.5.1. Basic characteristics

The total sample had a generally even distribution of gender and age, with slightly more males (54%) and children aged 4 years (39%). The characteristics of the generally healthy sample were comparable with the estimates from population representative Australian data (Longitudinal Study of Australian Children) except that the study sample had higher parental education and income (Table 5-1).

5.5.2. Acceptability and feasibility

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

5.5.3. Ceiling/floor effects

Ceiling effects were not present for CHU9D in the total sample or the sample with special health care needs, with only 12.4% and 6.1% of respondents reporting best levels for all 9 dimensions; 15.5% of respondents reported best levels for all 9 dimensions in the sample with no special health care needs, just exceeded the ceiling effects threshold. PedsQL did not demonstrate ceiling effects in any sample, with only 3%, 4% and 1% of respondents reporting best levels for all 21 items in the total sample, the sample with no health care needs and the sample with special health care needs. No floor effects were found for any sample. In terms of

CHU9D dimensions, pain dimension had over 70% of respondents reporting best level in the total sample (82.30%) and the sample with special health care needs (70.25%). In general, CHU9D had a distribution of different levels of response in the sample with special health care needs and the unwell sample (Figure 5-1), which was similarly observed for the PedsQL (Appendix Figure S2).

5.5.4. Test-retest reliability

The median days between initial and the follow-up survey completion for participants for the 2-day and 4-week follow-up were 3 days and 35 days respectively. The CHU9D had moderate test-retest reliability overall, with estimated ICCs of 0.52 (95% confidence interval (CI): 0.21, 0.72) and 0.60 (95%CI: 0.52, 0.67) for CHU9D Australian utilities in the 2-day and 4-week follow-ups respectively. PedsQL also had moderate test-retest reliability, with ICCs of 0.63 (95% CI:0.34, 0.80) and 0.80 (0.75, 0.84) for PedsQL total score in the 2-day and 4-week follow-ups respectively. The 95% confidence intervals for ICCs at 2-day follow-up were wide due to a small sample size of 53 (Appendix Table S3).

The test-retest reliability for individual dimensions were diverse for CHU9D, with 4 dimensions (worried, pain, annoyed, and schoolwork) having moderate agreement (weighted kappa ranging 0.44-0.48) and the remaining 5 dimensions (sad, tired, sleep, daily routine, and joining activities) having fair agreement (weighted kappa ranging 0.19-0.29) (Table 5-2). Results using the 4-week follow-up without health change sample had generally similar or larger agreement except “worried”, “sad” and “pain” dimensions. PedsQL generally had better test-retest reliability for individual items than CHU9D, with 13 (out of total 21) items demonstrating moderate agreement (kappa above 0.4). PedsQL generally showed similar results using the two follow-ups.

5.5.5. Convergent and divergent validity

As hypothesized, CHU9D utilities strongly correlated with PedsQL total scores (r=0.63). In addition, CHU9D and PedsQL displayed moderate correlations (r: 0.3-0.5) across all hypothesized correlated items, except for ‘sleep’ and ‘trouble sleeping’, which had a high correlation (r: 0.7). Weak correlations were found in items hypothesized not to be correlated (r< 0.3) (Table 5-3).

5.5.6. Known group validity

The CHU9D and PedsQL were both able to discriminate between groups with health difference defined as presence versus not of any chronic health conditions, or with special health care needs versus without, or having versus not having good/fair/poor general health

status (Table 5-4). The group mean differences of CHU9D utilities and PedsQL total scores were all significant, with medium to large Cohen D effect sizes. Known-group validity was also tested in 15 specific health conditions identified in this study compared with those with no health conditions. CHU9D performed well in discriminating individual chronic conditions compared with those with no health conditions, with significant utility differences (0.16-0.36) and large effect sizes (0.86-2.03). The top 5 conditions with largest effect size were behavioral/ cognitive/emotional problems, autism, genetic condition, soiling, and developmental delay. CHU9D had similar or better known-group validity compared with the PedsQL using all different definitions of health differences.

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

5.5.7. Responsiveness

In the sample with health condition(s), CHU9D had small effect sizes of responsiveness to general health change and health change to initially reported condition, with SRMs of 0.25-0.30 in the "Improved" group and SRMs of 0.21-0.44 in the "Worsened" group (Table 5-5). The results in the "Worsened" group need to be treated with caution considering the small sample sizes (n=14 and 16). PedsQL had small effect sizes (SRM: 0.26-0.41) in the "Improved" group and trivial effect size (SRM: 0.15-0.18) in the "Worsened" group. The supplementary results demonstrated that the CHU9D was able to reflect health changes in those who reported worsened health when developing new illness, with medium to large effect sizes (SRMs: 0.72-0.82). PedsQL was able to reflect this health change with medium effect size (SRM=0.50) (Appendix Table S5.2).

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

5.6. Discussion

5.6.1. Overview

Our study showed that the CHU9D with guidance notes proxy-reported for 2-4 year old Australian children was easy to complete, had no ceiling effects in a sample with special

health care needs, had moderate to high correlation with PedsQL pre-specified similar items, medium to large effect sizes of known-group validity, overall moderate test-retest reliability (with diverse results for individual dimensions), and showed some responsiveness to meaningful health changes over time (with small to large effect sizes using different definitions of health change). Compared with the PedsQL, CHU9D had similar feasibility, known-group validity, responsiveness, and slightly poorer test-retest reliability.

5.6.2. Distribution of responses

CHU9D did not exhibit ceiling effects except in the sample with no special health care needs. However, the ceiling effects issue was minor as the percentage of those reporting best levels in all dimensions (15.5%) just exceeded the criteria (15%). In addition, it may be less of a concern as good health was expected in the generally healthy sample with no special health care needs. Most CHU9D dimensions had a good distribution across different levels in the sample of children with impaired health. However, 70.3% of respondents reported the best level for dimension 'pain' even in the sample with special health care needs, which indicates that this item may not distinguish children well. This is consistent with the results from EQ-TIPs, with 73% and 88% respondents reporting the best level for 'pain' in acute and chronic condition samples [21]. This may be because pain is infrequent for young children or is difficult for parent/caregivers to observe. Recommended observable pain related behavior in young children includes grimacing, restless movement, and inconsolable crying [21]. These, or other behaviors, could be added as guidance notes for the 'pain' dimension in CHU9D to help improve the sensitivity of this item.

5.6.3. Test-retest reliability

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

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

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

5.6.4. Convergent and divergent validity

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

5.6.5. Known-group validity

The CHU9D was able to discriminate between groups with known health differences, with medium to large effect sizes, regardless of which scoring algorithm was applied. The utility difference between those 'with and without chronic conditions or disabilities' was 0.13 using an Australian adolescent algorithm and 0.07 using the UK adult scoring algorithm, with differences similar to previous studies [65, 66]. Peterson et al. found that the utility differences between 'with and without chronic conditions or disabilities' in a Danish high school student sample were 0.11 and 0.06 for Australian adolescent and UK adult scoring algorithms respectively [65]. In a similar study conducted with an Australian adolescent sample using Australian adolescent weights, the utility difference between 'with and without chronic conditions or disabilities' is 0.15 [66]. Neither of the prior studies reported Cohen D

effect sizes, however, the utility differences are similar to our findings. This suggests that CHU9D may have comparable known-group validity in children 2-4 years old compared with older age groups.

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

5.6.6. Responsiveness

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

5.6.7. Implications and limitations

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

Several limitations have been identified. First, missing data was not permitted for CHU9D based on a structural decision to not allow skipping items. This had its advantage such as reducing the percentage of missing data but might have forced people to randomly select an

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

There is an ongoing debate on the validity of proxy-reported HRQoL, particularly due to poor agreement between self-report by older children and proxy-report by adults [68]. While proxy reports are discouraged when children can self-report, they remain the only option for very young or cognitively challenged individuals. Parents of young children under 5, who usually

spend more time caring for their children, may serve as better proxies due to their close observations and connections. There is evidence that agreement is stronger in the youngest age group (5.5-6.5 years) than older age groups (6.5-8.5 years) [68]. Concerns regarding proxy-report, especially for more subjective dimensions such as “worried”, “sad” and “pain”, may be addressed by including externally observable indicators. Evaluating the validity of proxy HRQoL measures is controversial. Nevertheless, the pressure to include young children and their QALYs for cost-utility analysis and the existence of valid preference-based measures for older children continues to underscore the practical value of these investigations for younger children.

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

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

5.7. Conclusion

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

5.8. Reference

[1] G. L. Freed, S. Gafforini, and N. Carson, "Age distribution of emergency department presentations in Victoria," (in eng), *Emerg Med Australas,* vol. 27, no. 2, pp. 102-7, Apr 2015.

[2] M. Shaker, E. S. Chan, J. L. P. Protudjer, L. Soller, E. M. Abrams, and M. Greenhawt, "The Cost-Effectiveness of Preschool Peanut Oral Immunotherapy in the Real-World Setting," *The Journal of Allergy and Clinical Immunology: In Practice,* vol. 9, no. 7, pp. 2876-2884.e4, 2021/07/01/ 2021.

[3] L. Wang *et al.*, "The cost-effectiveness of alternative vision screening models among preschool children in rural China," *Acta Ophthalmologica,* https://doi.org/10.1111/aos.13954 vol. 97, no. 3, pp. e419-e425, 2019/05/01 2019.

[4] M. Tanaka, R. Okubo, S.-L. Hoshi, N. Ishikawa, and M. Kondo, "Cost-effectiveness of pertussis booster vaccination for preschool children in Japan," *Vaccine,* vol. 40, no. 7, pp. 1010-1018, 2022/02/11/ 2022.

[5] S. M. Sullivan, K. Tsiplova, and W. J. Ungar, "A scoping review of pediatric economic evaluation 1980-2014: do trends over time reflect changing priorities in evaluation methods and childhood disease?," *Expert Review of Pharmacoeconomics & Outcomes Research,* vol. 16, no. 5, pp. 599-607, 2016/09/02 2016.

[6] Pediatric Economic Database Evaluation (PEDE). *Trends in Economic Evaluation*. Available: http://pede.ccb.sickkids.ca/pede/trends.jsp

[7] S. K. Kromm *et al.*, "Characteristics and quality of pediatric cost-utility analyses," *Quality of Life Research,* vol. 21, no. 8, pp. 1315-1325, 2012/10/01 2012.

[8] D. Rowen, O. Rivero-Arias, N. Devlin, and J. Ratcliffe, "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?," (in eng), *Pharmacoeconomics,* vol. 38, no. 4, pp. 325-340, Apr 2020.

[9] J. Kwon *et al.*, "Systematic Review of Conceptual, Age, Measurement and Valuation Considerations for Generic Multidimensional Childhood Patient-Reported Outcome Measures," *PharmacoEconomics,* vol. 40, no. 4, pp. 379-431, 2022/04/01 2022.

[10] J. Kwon *et al.*, "Systematic Review of the Psychometric Performance of Generic Childhood Multi-attribute Utility Instruments," *Applied Health Economics and Health Policy,* 2023/05/03 2023.

[11] J. Kwon, S. W. Kim, W. J. Ungar, K. Tsiplova, J. Madan, and S. Petrou, "Patterns, trends and methodological associations in the measurement and valuation of childhood health utilities," (in eng), *Quality of Life Research,* vol. 28, no. 7, pp. 1705-1724, Jul 2019.

[12] J. L. Wolstenholme, D. Bargo, K. Wang, A. Harnden, U. Räisänen, and L. Abel, "Preference-based measures to obtain health state utility values for use in economic evaluations with child-based populations: a review and UK-based focus group assessment of patient and parent choices," (in eng), *Qual Life Res,* vol. 27, no. 7, pp. 1769-1780, Jul 2018.

[13] P. Kind, K. Klose, N. Gusi, P. R. Olivares, and W. Greiner, "Can adult weights be used to value child health states? Testing the influence of perspective in valuing EQ-5D-Y," *Quality of Life Research,* vol. 24, no. 10, pp. 2519-2539, 2015/10/01 2015.

[14] J. Verstraete, L. Ramma, and J. Jelsma, "Item generation for a proxy health related quality of life measure in very young children," *Health and Quality of Life Outcomes,* vol. 18, no. 1, p. 11, 2020/01/14 2020.

[15] S. Kreimeier *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,* vol. 21, no. 11, pp. 1291-1298, 2018/11/01/ 2018.

[16] R. N. Nancy Devlin, Julie Ratcliffe, Brendan Mulhern, Kim Dalziel, Gang Chen, Rosalie Viney,. (2020). *Do child QALYs = adult QALYs? Five reasons why they might not*. Available: https://www.ohe.org/news/do-child-qalys-adult-qalys-five-reasons-why-they-might-not

[17] A. G. Department of Health and Aged Care. *Preventive and Public Health Research initiative*. Available: https://www.health.gov.au/initiatives-and-programs/preventive-and-public-health-research-initiative

[18] W. J. Ungar, L. A. Prosser, and H. F. Burnett, "Values and evidence colliding: health technology assessment in child health," *Expert Review of Pharmacoeconomics & Outcomes Research,* vol. 13, no. 4, pp. 417-419, 2013/08/01 2013.

[19] L. B. Mokkink *et al.*, "COSMIN Study Design checklist for Patient-reported outcome measurement instruments," ed, 2019.

[20] D. A. Cook and T. J. Beckman, "Current Concepts in Validity and Reliability for Psychometric Instruments: Theory and Application," *The American Journal of Medicine,* vol. 119, no. 2, pp. 166.e7-166.e16, 2006/02/01/ 2006.

[21] J. Verstraete, L. Ramma, and J. Jelsma, "Validity and reliability testing of the Toddler and Infant (TANDI) Health Related Quality of Life instrument for very young children," *Journal of Patient-Reported Outcomes,* vol. 4, no. 1, p. 94, 2020/11/09 2020.

[22] S. Saigal *et al.*, "Development, reliability and validity of a new measure of overall health for pre-school children," (in eng), *Qual Life Res,* vol. 14, no. 1, pp. 243-57, Feb 2005.

[23] W. Furlong *et al.*, "Generic Health-Related Quality of Life Utility Measure for Preschool Children (Health Utilities Preschool): Design, Development, and Properties," *Value in Health,* 2022/08/26/ 2022.

[24] J. Verstraete and R. Amien, "Cross-Cultural Adaptation and Validation of the EuroQoL Toddler and Infant Populations Instrument Into Afrikaans for South Africa," *Value in Health Regional Issues,* vol. 35, pp. 78-86, 2023/05/01/ 2023.

[25] X. Fang *et al.*, "Feasibility and validity of the Health Status Classification System-Preschool (HSCS-PS) in a large community sample: the Generation R study," (in eng), *BMJ Open,* vol. 8, no. 12, p. e022449, Dec 18 2018.

[26] U. Ravens-Sieberer *et al.*, "Feasibility, reliability, and validity of the EQ-5D-Y: results from a multinational study," *Quality of Life Research,* vol. 19, no. 6, pp. 887-897, 2010/08/01 2010.

[27] J. Ratcliffe, K. Stevens, T. Flynn, J. Brazier, and M. Sawyer, "An assessment of the construct validity of the CHU9D in the Australian adolescent general population," (in eng), *Quality of Life Research,* vol. 21, no. 4, pp. 717-25, May 2012.

[28] K. J. Stevens, "Working with children to develop dimensions for a preference-based, generic, pediatric, health-related quality-of-life measure," (in eng), *Qual Health Res,* vol. 20, no. 3, pp. 340-51, Mar 2010.

[29] E. J. Frew, M. Pallan, E. Lancashire, K. Hemming, and P. Adab, "Is utility-based quality of life associated with overweight in children? Evidence from the UK WAVES randomised controlled study," *BMC Pediatrics,* Article vol. 15, no. 1, 2015, Art. no. 211.

[30] A. G. Canaway and E. J. Frew, "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," (in eng), *Quality of Life Research,* vol. 22, no. 1, pp. 173-83, Feb 2013.

[31] K. Stevens and J. Ratcliffe, "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," (in eng), *Value Health,* vol. 15, no. 8, pp. 1092-9, Dec 2012.

[32] K. Stevens, "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,* vol. 9, no. 3, pp. 157-169, 2011/05/01 2011.

[33] *Measuring & Valuing Health. A brief overview of the Child Health Utility 9D (CHU9D)*. Available: https://licensing.sheffield.ac.uk/product/CHU-9D

[34] A. F. Klassen *et al.*, "Health related quality of life in 3 and 4 year old children and their parents: preliminary findings about a new questionnaire," (in eng), *Health Qual Life Outcomes,* vol. 1, p. 81, Dec 22 2003.

[35] J. W. Varni, M. Seid, and P. S. Kurtin, "PedsQL 4.0: reliability and validity of the Pediatric Quality of Life Inventory version 4.0 generic core scales in healthy and patient populations," (in eng), *Med Care,* vol. 39, no. 8, pp. 800-12, Aug 2001.

[36] M. Solans *et al.*, "Health-Related Quality of Life Measurement in Children and Adolescents: A Systematic Review of Generic and Disease-Specific Instruments," vol. 11, no. 4, pp. 742-764, 2008.

[37] J. Paltzer, E. Barker, and W. P. Witt, "Measuring the health-related quality of life (HRQoL) of young children in resource-limited settings: a review of existing measures," *Quality of Life Research,* vol. 22, no. 6, pp. 1177-1187, 2013/08/01 2013.

[38] A. Gheissari *et al.*, "Validation of Persian Version of PedsQL™ 4.0™ Generic Core Scales in Toddlers and Children," (in eng), *Int J Prev Med,* vol. 3, no. 5, pp. 341-50, May 2012.

[39] J. W. Varni, T. M. Burwinkle, M. Seid, and D. Skarr, "The PedsQL™* 4.0 as a Pediatric Population Health Measure: Feasibility, Reliability, and Validity," *Ambulatory Pediatrics,* vol. 3, no. 6, pp. 329-341, 2003/11/01/ 2003.

[40] D. Buck, "The PedsQL™ as a measure of parent-rated quality of life in healthy UK toddlers:Psychometric properties and cross-cultural comparisons," vol. 16, no. 4, pp. 331-338, 2012.

[41] R. Jones *et al.*, "Psychometric Performance of HRQoL Measures: An Australian Paediatric Multi-Instrument Comparison Study Protocol (P-MIC)," (in eng), *Children (Basel, Switzerland),* vol. 8, no. 8, p. 714, 2021.

[42] B. M. Renee Jones, Nancy Devlin, Harriet Hiscock, Gang Chen, Rachel O'Louglin, Xiuqin Xiong, Mina Bahrampour, Kristy McGregor, Shilana Yip, and Kim Dalziel on behalf of the Quality Of Life in Kids: Key evidence to strengthen decisions in Australia (QUOKKA) project team. (2023, 26/04/2023). *Australian Paediatric Multi-Instrument Comparison (P-MIC) Study: Technical Methods Paper [Online]*. Available: https://www.quokkaresearchprogram.org/project-1-1

[43] P. A. Harris *et al.*, "The REDCap consortium: Building an international community of software platform partners," (in eng), *J Biomed Inform,* vol. 95, p. 103208, Jul 2019.

[44] G. M. Breakwell, S. E. Hammond, C. E. Fife-Schaw, and J. A. Smith, *Research methods in psychology*. Sage Publications, Inc, 2006.

[45] J. Ratcliffe, T. Flynn, F. Terlich, K. Stevens, J. Brazier, and M. Sawyer, "Developing Adolescent-Specific Health State Values for Economic Evaluation," *PharmacoEconomics,* vol. 30, no. 8, pp. 713-727, 2012/08/01 2012.

[46] K. Stevens, "Valuation of the Child Health Utility 9D Index," (in eng), *Pharmacoeconomics,* vol. 30, no. 8, pp. 729-47, Aug 1 2012.

[47] K. Dalziel, M. Catchpool, B. Garcia-Lorenzo, I. Gorostiza, R. Norman, and O. Rivero-Arias, "Feasibility, Validity and Differences in Adolescent and Adult EQ-5D-Y Health State Valuation in Australia and Spain: An Application of Best-Worst Scaling," (in eng), *Pharmacoeconomics,* Jan 24 2020.

[48] C. B. Terwee *et al.*, "Quality criteria were proposed for measurement properties of health status questionnaires," (in eng), *J Clin Epidemiol,* vol. 60, no. 1, pp. 34-42, Jan 2007.

[49] T. Peasgood *et al.*, "Developing a New Generic Health and Wellbeing Measure: Psychometric Survey Results for the EQ Health and Wellbeing," *Value in Health,* 2022/01/13/ 2022.

[50] T. K. Koo and M. Y. Li, "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research," (in eng), *Journal of chiropractic medicine,* vol. 15, no. 2, pp. 155-163, 2016.

[51] H. Brenner and U. Kliebsch, "Dependence of weighted kappa coefficients on the number of categories," (in eng), *Epidemiology,* vol. 7, no. 2, pp. 199-202, Mar 1996.

[52] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," (in eng), *Biometrics,* vol. 33, no. 1, pp. 159-74, Mar 1977.

[53] P. Yang *et al.*, "Psychometric evaluation of the Chinese version of the Child Health Utility 9D (CHU9D-CHN): a school-based study in China," (in eng), *Quality of Life Research,* vol. 27, no. 7, pp. 1921-1931, Jul 2018.

[54] W. W. Daniel and C. L. Cross, *Biostatistics: a foundation for analysis in the health sciences*. Wiley, 2018.

[55] J. Cohen, *Statistical power analysis for the behavioral sciences*. Routledge, 2013.

[56] L. Huang, G. L. Freed, and K. Dalziel, "Children with special health care needs: how special are their health care needs?," *Academic Pediatrics,* 2020.

[57] M. Hollander, D. A. Wolfe, and E. Chicken, *Nonparametric statistical methods*. John Wiley & Sons, 2013.

[58] J. Cohen, "A power primer," (in eng), *Psychol Bull,* vol. 112, no. 1, pp. 155-9, Jul 1992.

[59] M. Xu, D. Fralick, J. Z. Zheng, B. Wang, X. M. Tu, and C. Feng, "The Differences and Similarities Between Two-Sample T-Test and Paired T-Test," (in eng), *Shanghai Arch Psychiatry,* vol. 29, no. 3, pp. 184-188, Jun 25 2017.

[60] J. Ludbrook, "Should we use one-sided or two-sided P values in tests of significance?," *Clinical and Experimental Pharmacology and Physiology,* vol. 40, no. 6, pp. 357-361, 2013/06/01 2013.

[61] J. A. Husted, R. J. Cook, V. T. Farewell, and D. D. Gladman, "Methods for assessing responsiveness: a critical review and recommendations," *Journal of Clinical Epidemiology,* vol. 53, no. 5, pp. 459-468, 2000/05/01/ 2000.

[62] D. Lakens, "Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs," (in English), Review vol. 4, 2013-November-26 2013.

[63] J. Sim and C. C. Wright, "The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements," *Physical Therapy,* vol. 85, no. 3, pp. 257-268, 2005.

[64] D. G. Froberg and R. L. Kane, "Methodology for measuring health-state preferences—II: Scaling methods," *Journal of Clinical Epidemiology,* vol. 42, no. 5, pp. 459-471, 1989/01/01/ 1989.

[65] K. D. Petersen, J. Ratcliffe, G. Chen, D. Serles, C. S. Frosig, and A. V. Olesen, "The construct validity of the Child Health Utility 9D-DK instrument," (in eng), *Health and Quality of Life Outcomes,* vol. 17, no. 1, p. 187, Dec 23 2019.

[66] K. D. Petersen, G. Chen, C. Mpundu-Kaambwa, K. Stevens, J. Brazier, and J. Ratcliffe, "Measuring Health-Related Quality of Life in Adolescent Populations: An Empirical Comparison of the CHU9D and the PedsQL(TM) 4.0 Short Form 15," (in eng), *Patient,* vol. 11, no. 1, pp. 29-37, Feb 2018.

[67] R. T. Wolf, J. Ratcliffe, G. Chen, and P. Jeppesen, "The longitudinal validity of proxy-reported CHU9D," *Quality of Life Research,* vol. 30, no. 6, pp. 1747-1756, 2021/06/01 2021.

[68] J. Cremeens, C. Eiser, and M. Blades, "Factors influencing agreement between child self-report and parent proxy-reports on the Pediatric Quality of Life Inventory™ 4.0 (PedsQL™) generic core scales," *Health and Quality of Life Outcomes,* vol. 4, no. 1, p. 58, 2006/08/30 2006.

5.9. Tables and figures

Table 5-1 Baseline characteristics

Baseline characteristics	Total sample (N=842) [n(%)]	Generally healthy[2] (N=465) [n(%)]	With health condition(s)[3] (N=377) [n(%)]	LSAC[1] (%)
Child sex				
Male	453(53.80)	243(52.26)	210(55.70)	51.66
Female	386(45.84)	222(47.74)	164(43.50)	48.34
Other	3(0.36)		3(0.80)	
Child age (years)				
2	263(31.24)	180(38.71)	83(22.02)	
3	247(29.33)	144(30.97)	103(27.32)	
4	332(39.43)	141(30.32)	191(50.66)	
Aboriginal or Torress Strait Islander				
No	791(93.94)	442(95.05)	349(92.57)	97.35
Yes	49(5.82)	23(4.95)	26(6.90)	2.65
Prefer not to say	2(0.24)		2(0.53)	
Child having a health condition or disability that lasted or are likely to last for 6 months or more				
No	529(62.83)	368(79.14)	161(42.71)	
Yes	313(37.17)	97(20.86)	216(57.29)	
Child having special health care needs				
No	563(66.86)	391(84.09)	172(45.62)	87.26
Yes	279(33.14)	74(15.91)	205(54.38)	12.74
Caregiver education-bachelor's degree or above				
Yes	407(48.34)	227(48.82)	180(47.75)	34.95
No	435(51.66)	238(51.18)	197(52.25)	65.05
Household weekly income before tax				

Less than $500 per week ($25,999 or less per year)	40(4.75)	25(5.38)	15(3.98)	5.46
$500-$999 per week ($26,000-$51,999 per year)	151(17.93)	81(17.42)	70(18.57)	16.81
$1,000-$1,999 per week ($52,000-$103,9799 per year)	314(37.29)	172(36.99)	142(37.67)	48.17
$2,000 or more per week ($104,000 or more per year)	320(38.00)	183(39.35)	137(36.34)	29.57
Missing	17(2.02)	4(0.86)	13(3.45)	
In general, how would you say the study child's current health is?				
Excellent	287(34.09)	212(45.59)	75(19.89)	52.72
Very good	355(42.16)	198(42.58)	157(41.64)	34.32
Good	150(17.81)	49(10.54)	101(26.79)	10.99
Fair	47(5.58)	5(1.08)	42(11.14)	1.86
Poor	3(0.36)	1(0.22)	2(0.53)	0.12

Note: 1. Longitudinal Study of Australian Children (LSAC) is a nationally representative survey of Australian children aged 0 to 18 years old. LSAC estimates here are based on LSAC 2-4 years old and used population weights. 2.The generally healthy sample is composed of the online general population sample and those not receiving health care from the hospital sample. 3. The sample with health condition(s) is composed of online disease groups and those receiving healthcare at royal children's hospital of the hospital sample.

Table 5-2 Weighted-kappa of CHU9D dimensions compared with PedsQL for children reporting no health changes at different follow-ups

Dimensions	Dimensions/Items	Weighted kappa (95%CI)	
		2-day follow-up (N=53)	4-week follow up (N=265)
	CHU9D		
	1. Worried	0.45 (0.27,0.64)	0.27 (0.18,0.36)
	2. Sad	0.26 (0.07,0.46)	0.22 (0.13,0.31)
	3. Pain	0.47 (0.24,0.70)	0.35 (0.25,0.45)
	4. Tired	0.21 (0.04,0.39)	0.32 (0.24,0.40)
	5. Annoyed	0.44 (0.24,0.63)	0.38 (0.29,0.47)
	6. School Work	0.48 (0.29,0.67)	0.44 (0.35,0.54)
	7. Sleep	0.28 (0.09,0.48)	0.36 (0.27,0.45)
	8. Daily routine	0.19 (-0.03,0.41)	0.50 (0.41,0.58)
	9. Able to join in activities	0.29 (0.13,0.45)	0.47 (0.38,0.56)
	PedsQL		
Physical function	1. Walking	0.61 (0.42,0.80)	0.61 (0.52,0.71)
	2. Running	0.59 (0.38,0.80)	0.60 (0.50,0.69)
	3. Participating in active play or exercise	0.41 (0.20,0.61)	0.53 (0.44,0.62)
	4. Lifting something heavy	0.44 (0.25,0.64)	0.50 (0.41,0.59)

	5. Bathing	0.39 (0.19,0.58)	0.51 (0.42,0.60)
	6. Helping to pick up his or her toys	0.41 (0.23,0.59)	0.43 (0.34,0.51)
	7. Getting aches and pains	0.31 (0.11,0.50)	0.43 (0.34,0.52)
	8. Having a low energy level	0.28 (0.07,0.48)	0.45 (0.36,0.54)
Emotional function	1. Feeling afraid or scared	0.43 (0.23,0.63)	0.38 (0.29,0.46)
	2. Feeling sad	0.36 (0.15,0.56)	0.37 (0.28,0.45)
	3. Feeling angry	0.40 (0.23,0.58)	0.46 (0.38,0.54)
	4. Having trouble sleeping	0.52 (0.31,0.72)	0.52 (0.44,0.59)
	5. Worrying	0.54 (0.35,0.73)	0.53 (0.45,0.62)
Social function	1. Playing with other children	0.48 (0.31,0.65)	0.52 (0.44,0.61)
	2. Other children not wanting to play with him or her	0.28 (0.10,0.46)	0.43 (0.34,0.51)
	3. Getting teased by other children	0.26 (0.05,0.46)	0.50 (0.41,0.60)
	4. Not being able to do things that other children his or her age can do	0.43 (0.23,0.63)	0.66 (0.58,0.75)
	5. Keeping up when playing with other children	0.35 (0.18,0.52)	0.57 (0.48,0.66)
School function*	1. Doing the same school activities as other children his or her	0.36 (0.19,0.53)	0.50 (0.40,0.59)
	2. Missing school because of not feeling well	0.59 (0.37,0.81)	0.39 (0.30,0.48)
	3. Missing school to go to the doctor or hospital	0.49 (0.26,0.71)	0.52 (0.42,0.62)

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

Table 5-3 Convergence between CHU9D and PedsQL in total sample

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

Missing school because not well	0.18	0.17	0.28	0.19	0.16	**0.27**	0.21	0.28	0.27
Missing school to go to doctor or hospital	0.18	0.19	0.33	0.20	0.20	**0.35**	0.23	0.32	0.32

Note: High correlations: >=0.5 (green); moderate correlations: 0.3 to 0.49 (yellow), low correlation: <0.3 (white). All correlation significant at 0.05 level. **Bold** indicates expected moderate or high correlations (r>=0.3) based on highly similar items in line with published technical guide. *Italic and underscore* indicate items hypothesized not to be correlated or weak correlations (r<0.3). Correlation coefficients were calculated by Spearman rank correlation.

Table 5-4 Known group validity (Cohen D effect size) of CHU9D and PedsQL for different health difference groups

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

Note: The diseases included in this table have sample sizes>=30. Standard thresholds 0.2 to <0.5, 0.5 to <0.8, and 0.8 or more denote small, medium, and large effect sizes, respectively. Cohen D ES (effect sizes): the numerator is the difference between means of the two groups, and the denominator is the pooled standard deviation. Healthy sample=with no chronic condition. P values are obtained from Man-Whitney U test as scores were not normally distributed. The health conditions are ordered from large to small by their effect sizes calculated using Australia adolescent weights.

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

Index	Health status change[3]	Sample size	Baseline (Mean, SD)	At 4-week Follow-up (Mean, SD)	Paired Difference (Mean, SD)	*P*[4]	SRM[1]
General health change[2]							
CHU9D utility Australia adolescents (higher score reflects better health)	Improved	33	0.76(0.23)	0.83(0.23)	0.07 (0.26)	0.080	0.25
	Worsened	14	0.42(0.22)	0.32(0.21)	-0.09 (0.22)	0.063	-0.44
PedsQL Total score (higher score reflects better health)	Improved	33	74.83(21.51)	80.88(17.59)	6.05 (14.58)	0.012	0.41
	Worsened	14	53.97(18.89)	50.48(27.28)	-3.49 (22.67)	0.287	-0.15
Health change related to initially reported condition[2]							
CHU9D utility Australia adolescents (higher score reflects better health)	Improved	32	0.76(0.23)	0.84(0.20)	0.08 (0.26)	0.052	0.30
	Worsened	16	0.44(0.23)	0.38(0.24)	-0.06 (0.27)	0.207	-0.21
PedsQL Total score (higher score reflects better health)	Improved	32	78.22(20.75)	82.11(17.08)	3.88 (14.66)	0.072	0.26
	Worsened	16	54.53(22.57)	50.98(26.64)	-3.56 (20.13)	0.245	-0.18

Note: 1. SRM=Standard response mean, dividing the mean score change (i.e., follow-up minus baseline) by the standard deviation of the change. The interpretation for SRM were defined as trivial for <0.2, small for >=0.2 and <0.5, medium for >0.5 and <0.8, and large for >=0.8. 2. General health change: how would you rate the study child's health in general now? Health change related to initially reported condition: thinking about the Study Child's main health condition, how would you say this is going now compared to when you completed the first survey for this study? (with answers: Much better, Somewhat better, About the same, Somewhat worse, Much worse). 3. "Improved" includes *Much better*, "Worsened" includes *Somewhat worse* and *Much worse*. 4. P values were one sided P from paired t test.

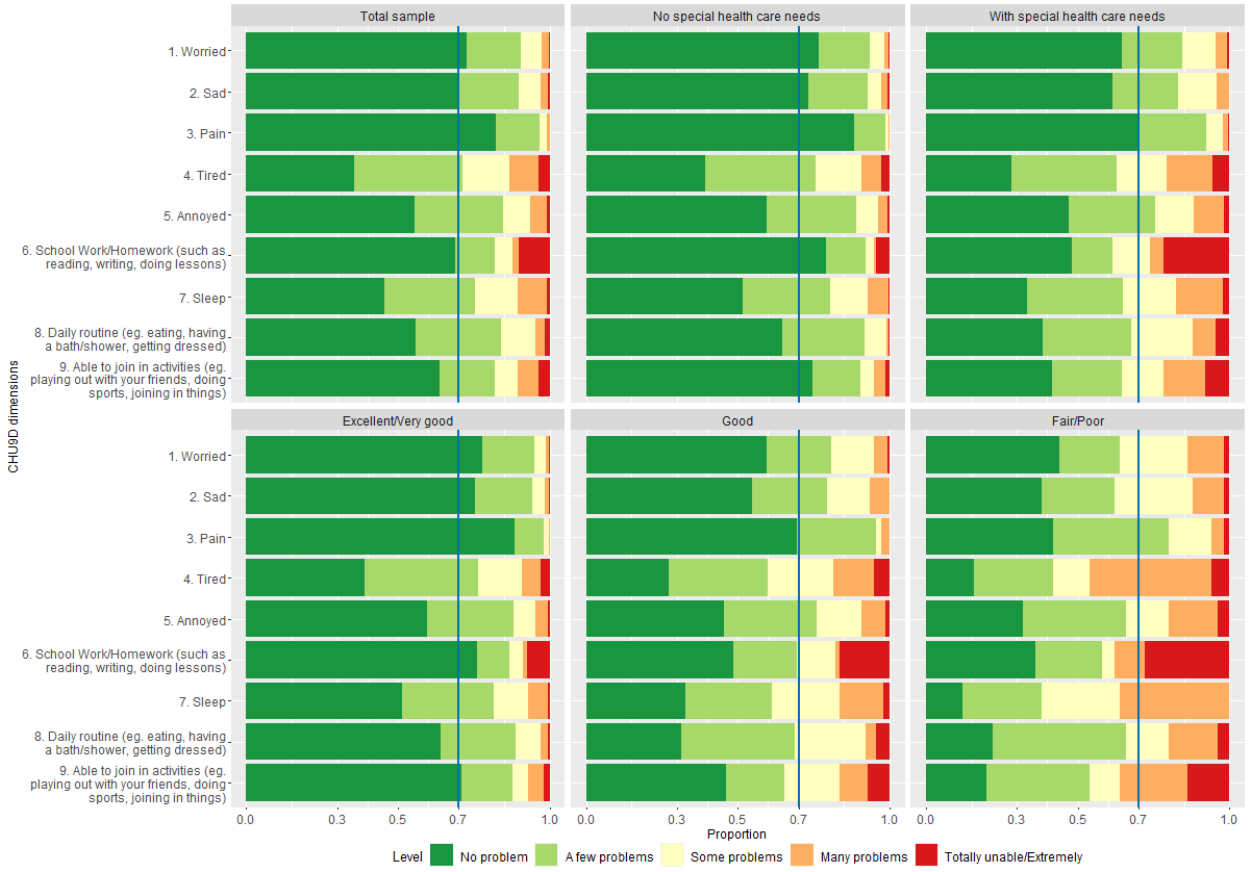


Figure 5-1 Distribution of CHU9D response in different samples

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

5.10. Supplementary Materials

Detailed guidance notes for CHU9D proxy version for children under 5 years old

Table S1: guidance notes

CHU9D dimension/question	**Guidance notes for children under 5**
Worried	NA
Sad	NA
Pain	NA
Tired	NA
Annoyed	NA
School Work/Homework	If your child is at preschool/nursery/kindergarten then please think about that. If your child didn't go today because of their health and they usually would have, please tick the last option "My child can't do their schoolwork/homework today". If today is not a day they usually would have gone, then please think about how you think they would have been had they gone. If your child does not go to preschool/nursery/kindergarten, then please think about whether they have had any problems with activities such as colouring, looking at books/reading, and concentrating, as appropriate for their age.
Sleep	NA
Daily Routine	Please think about this question in terms of eating, drinking, toileting, washing and teeth cleaning, as appropriate for their age.
Able to join in activities	Please think about this question in terms of the activities your child would usually be doing today.

Note: It is necessary to obtain a license to use CHU9D proxy version with guidance notes. https://licensing.sheffield.ac.uk/product/CHU-9D

Acceptability and feasibility

Completion time

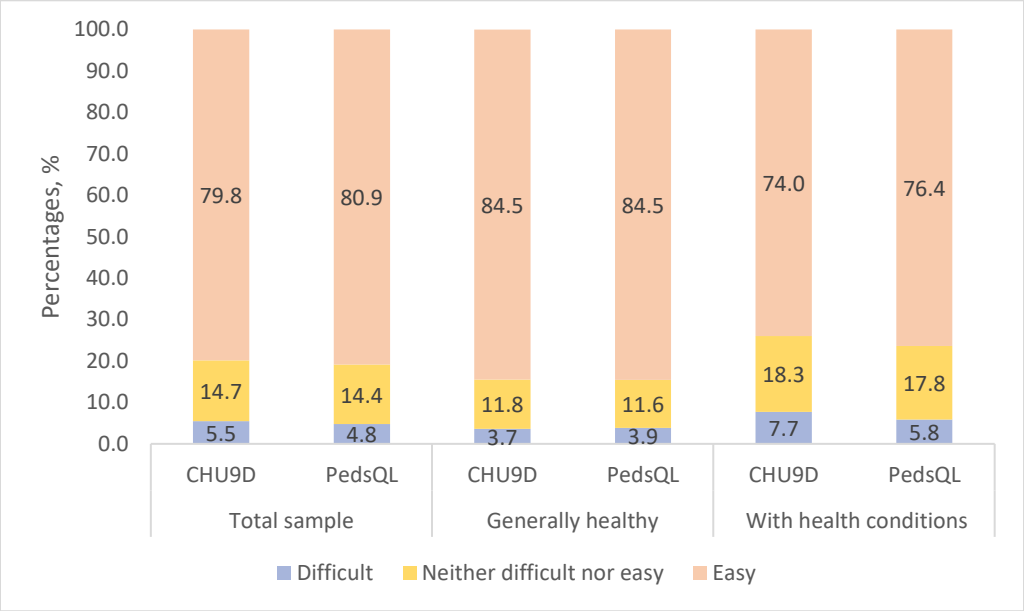
Table S2: Time to complete

		Time to complete (minutes) [c], Median (IQR)		
Sample	N	Total survey [b]	CHU9D (9 items)	PedsQL (21 items)
Generally healthy sample	465	9.49(7.21,12.64)	1.03(0.74,1.45)	1.32(1.00,1.74)
Sample with health condition(s)	377	11.92(9.15,17.34)	1.28(0.96,1.86)	1.60(1.26,2.24)
Total sample [a]	842	10.59(7.96,14.40)	1.12(0.81,1.65)	1.44(1.10,1.98)

Note: IQR=interquartile range, the first quartile to third quartile. All data are given as median time in minutes (25th, 75th). [a] PedsQL 2-4 sample size is 807, as children not attending childcare or school may not fill out PedsQL instrument. [b] The components of the total survey can be found in technical paper. [c] For CHU9D, the 95% and 99% percentile of completion time are 3.85 minutes and 21.23 minutes; for PedsQL, the 95% and 99% percentile are 3.88 minutes and 20.25 minutes; for total survey, the 95% and 99% percentile are 43.09 minutes and 174.01 minutes. The completion time is recorded automatically by the online system. The outliers happened when participants left the instrument open on their electronic device and decided to come back to the survey at a later point this time would be included. We only reported median and IQR which were not impacted by these outliers.

Reported difficulty

Figure S1. Reported difficulty completing CHU9D and PedsQL instruments



Note: Pearson chi2 test showed no significant difference in reported difficulty between CHU9D and PedsQL in all samples.

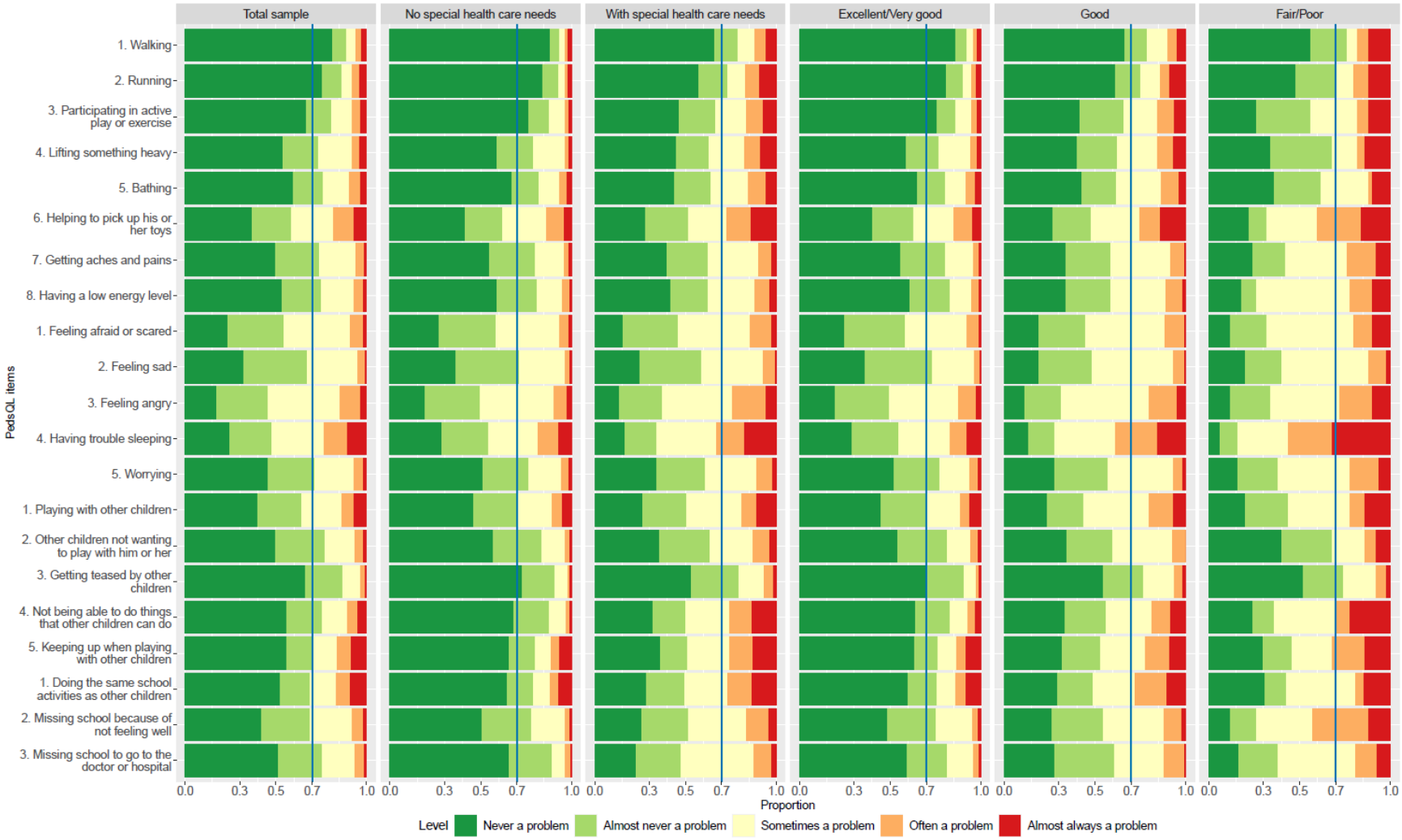

Total sample
No special health care needs
With special health care needs
Excellent/Very good
Good
Fair/Poor
PedsQL items
1. Walking
2. Running
3. Participating in active play or exercise
4. Lifting something heavy
5. Bathing
6. Helping to pick up his or her toys
7. Getting aches and pains
8. Having a low energy level
1. Feeling afraid or scared
2. Feeling sad
3. Feeling angry
4. Having trouble sleeping
5. Worrying
1. Playing with other children
2. Other children not wanting to play with him or her
3. Getting teased by other children
4. Not being able to do things that other children can do
5. Keeping up when playing with other children
1. Doing the same school activities as other children
2. Missing school because of not feeling well
3. Missing school to go to the doctor or hospital
0.0
0.3
0.5
0.7
1.0
Proportion
Level
Never a problem
Almost never a problem
Sometimes a problem
Often a problem
Almost always a problem

Figure S2: Distribution of PedsQL response (ceiling effects and floor effects)

Note: 1. The subgroups are defined by variables asking about special health care needs (yes, no) and the general health status of the child (with responses of excellent, very good, good, fair and poor). 2. School function only applies to children attending daycare, preschool/kindergarten, or school.

Test-retest reliability

Table S3: ICC for summary scores (continuous score) for different samples and intervals

Index	ICC (95%CI)	
	2-day follow-up with unchanged health (N=53)	4-week follow-up with unchanged health (N=265)
CHU9D utility (Australia adolescents)	0.52(0.21,0.72)	0.60(0.52,0.67)
CHU9D utility (UK adults)	0.53(0.19,0.73)	0.63(0.55,0.70)
PedsdQL Total score	0.63(0.34,0.80)	0.80(0.75,0.84)

Note: It is suggested that ICC values less than 0.5 are indicative of poor reliability, between 0.50 and 0.75 indicate moderate reliability, and values between 0.75 and 0.90 indicate good reliability. 2-day follow up survey were only sent to the online general population sample. 4-week follow up survey were sent to the whole sample (hospital recruited, online general population and online disease groups samples).

Known group validity

Table S4.1 Known group validity (Cohen D effect size) of CHU9D using different utilities

Groups	Sample size	CHU9D utilities Australia adolescents (range 0-1, lower utility reflects more health problems)				CHU9D utilities UK adults (range 0-1, lower utility reflects more health problems)			
		Mean	Diff	P value	Cohen D ES	Mean	Diff	P value	Cohen D ES
Any medical condition or disabilities lasting for 6 months or more	Yes=313	0.65	-0.13	<0.001	0.58	0.81	-0.07	<0.001	0.66
	No=529	0.78				0.88			
Special health care needs	Yes=279	0.62	-0.16	<0.001	0.75	0.80	-0.09	<0.001	0.82
	No=563	0.78				0.89			
General health status (good/fair/poor)	Yes=200	0.57	-0.21	<0.001	1.02	0.78	-0.10	<0.001	1.03
	No=642	0.78				0.88			
Healthy (children with no chronic conditions, as comparison)	267	0.84				0.92			
Behavioral, cognitive, emotional problems	98	0.49	-0.36	<0.001	2.03	0.73	-0.18	<0.001	2.11
Autism	54	0.48	-0.36	<0.001	2.00	0.73	-0.19	<0.001	2.13
Genetic condition	30	0.50	-0.34	<0.001	1.92	0.74	-0.18	<0.001	2.05
Soiling	33	0.51	-0.33	<0.001	1.91	0.73	-0.18	<0.001	2.11
Developmental delay	96	0.53	-0.31	<0.001	1.69	0.75	-0.17	<0.001	1.87
Bone, joint or muscle problem	40	0.56	-0.28	<0.001	1.59	0.77	-0.15	<0.001	1.73
ADHD	64	0.56	-0.28	<0.001	1.56	0.78	-0.14	<0.001	1.61
Ear infection	44	0.59	-0.25	<0.001	1.45	0.79	-0.13	<0.001	1.52
Sleep problems	190	0.58	-0.26	<0.001	1.37	0.79	-0.13	<0.001	1.37
Anxiety	42	0.60	-0.24	<0.001	1.34	0.80	-0.12	<0.001	1.38
Constipation	62	0.60	-0.24	<0.001	1.27	0.79	-0.13	<0.001	1.40
Hay fever	56	0.64	-0.20	<0.001	1.18	0.81	-0.10	<0.001	1.27
Asthma	108	0.68	-0.16	<0.001	0.87	0.83	-0.08	<0.001	0.95
Eczema	149	0.68	-0.16	<0.001	0.87	0.84	-0.08	<0.001	0.88
Food or digestive allergies	84	0.68	-0.16	<0.001	0.86	0.83	-0.09	<0.001	0.96

Table S4.2: Known group validity for dimensions (Cohen D effect size) for different health conditions compared with no health conditions

Health conditions	Sample size	CHU9D dimensions									PedsQL dimensions			
		1. Worried	2. Sad	3. Pain	4. Tired	5. Annoyed	6. School Work	7. Sleep	8. Daily routine	9. Able to join in activities	Physical Function	Emotional function	Social function	School function
Sleep problems	190	0.55	0.66	0.64	0.98	0.78	0.60	1.57	0.98	0.83	0.75	1.27	0.86	0.67
Anxiety	42	1.00	0.78	0.58	0.59	0.86	1.01	1.03	1.11	1.35	0.95	1.53	1.41	1.13
Behavioral, cognitive, emotional problems	98	0.75	0.94	0.86	0.99	1.28	1.33	1.34	1.62	1.79	1.19	1.41	1.44	1.16
ADHD	64	0.64	0.77	0.45	0.97	1.18	0.95	1.01	1.34	1.40	0.90	1.27	1.22	0.81
Autism	54	0.93	1.01	0.99	0.96	1.19	1.64	1.41	1.89	2.53	1.34	1.60	2.11	1.62
Bone, joint or muscle problem	40	0.49	0.54	1.26	1.23	0.93	1.02	1.32	1.47	1.70	1.44	0.81	1.30	1.13
Constipation	62	0.43	0.61	1.10	0.88	0.87	0.85	0.87	1.28	1.23	1.07	0.85	0.98	0.88
Soiling	33	0.61	1.13	1.16	1.00	1.42	1.19	1.29	2.26	1.74	1.16	1.32	1.62	1.37
Ear infection	44	0.52	0.60	1.32	0.94	0.99	0.76	1.20	1.34	0.86	0.57	0.89	0.70	0.68
Eczema	149	0.31	0.46	0.55	0.50	0.55	0.43	0.66	0.63	0.61	0.42	0.62	0.55	0.56
Food or digestive allergies	84	0.37	0.43	0.75	0.62	0.51	0.53	0.66	0.78	0.73	0.43	0.59	0.56	0.67
Hay fever	56	*0.24*	0.39	0.90	0.76	0.71	0.57	1.20	0.82	1.00	0.74	0.93	0.84	0.87
Asthma	108	0.22	0.45	0.71	0.59	0.40	0.43	0.89	0.62	0.70	0.47	0.55	0.55	0.89
Developmental delay	96	0.55	0.66	0.89	0.92	0.93	1.44	1.06	1.68	1.68	1.36	0.99	1.66	1.48
Genetic condition	30	0.97	0.82	1.48	1.30	1.39	1.67	1.55	1.92	2.11	1.54	0.88	1.57	1.56

Note: All disease groups were compared with sample with no chronic condition. P values are obtained from Mann-Whitney U test. In tests for significant difference between disease groups compared with healthy sample without any condition, all P values were <0.05 except 'worried' dimension in Hayfever. The P values >0.05 were italic. Standard thresholds 0.2 to <0.5, 0.5 to <0.8, and 0.8 or more denote small, medium, and large effect sizes, respectively. Similar health conditions were roughly grouped together to help interpretation of effect sizes for different dimensions.

Table S4.3 Known-group validity (Cohen D effect size) of CHU9D and PedsQL for different health difference groups by age

Child age (years)	Groups	CHU9D utilities Australian adolescent weights					PedsQL Total score				
		M- Yes	M- No	Diff	P	ES	M- Yes	M- No	Diff	P	ES
2	Chronic conditions (yes=105, no=158)	0.66	0.81	-0.15	<0.001	-0.71	70.27	83.86	-13.59	<0.001	-0.80
3	Chronic conditions (yes=80, no=167)	0.67	0.78	-0.11	0.002	-0.53	70.98	82.08	-11.10	<0.001	-0.72
4	Chronic conditions (yes=128, no=204)	0.63	0.75	-0.11	<0.001	-0.52	66.79	78.30	-11.52	<0.001	-0.71
2	Special health care needs (yes=93, no=170)	0.64	0.81	-0.17	<0.001	-0.79	69.20	83.49	-14.29	<0.001	-0.85
3	Special health care needs (yes=66, no=161)	0.64	0.79	-0.15	<0.001	-0.73	68.25	82.21	-13.96	<0.001	-0.93
4	Special health care needs (yes=120, no=212)	0.60	0.76	-0.16	<0.001	-0.73	65.06	78.85	-13.79	<0.001	-0.87
2	General health status good/fair/poor (yes=49, no=214)	0.54	0.80	-0.26	<0.001	-1.27	62.27	82.14	-19.86	<0.001	-1.20
3	General health status good/fair/poor (yes=67, no=180)	0.61	0.80	-0.19	<0.001	-0.95	68.73	82.11	-13.38	<0.001	-0.88
4	General health status good/fair/poor (yes=84, no=248)	0.56	0.75	-0.19	<0.001	-0.93	64.19	77.14	-12.95	<0.001	-0.80

Note: Standard thresholds 0.2 to <0.5, 0.5 to <0.8, and 0.8 or more denote small, medium, and large effect sizes, respectively. Cohen D ES (effect sizes): the numerator is the difference between means of the two groups, and the denominator is the pooled standard deviation. Healthy sample=with no chronic condition. P values are obtained from Man-Whitney U test as scores were not normally distributed.

Responsiveness

In group of having new illness and reporting worsened health status, there is significant difference of HRQoL between baseline and follow-up, with large effect size for CHU9D level sum score and CHU9D utilities (both Australian adolescent and UK adults). PedsQL also had almost significant changes (p=0.058), with smaller effect sizes than CHU9D (Table S5.2). This added to the evidence that CHU9D was able to response to the health changes due to having a new illness in 2-4 years old.

Table S5.1 Responsiveness of CHU9D using UK utilities

Type of health changes	Health status change	Sample size	Baseline (Mean, SD)	At 4-week Follow-up (Mean, SD)	Paired Difference (Mean, SD)	*P*	SRM
Self-reported general health change	Improved	33	0.87(0.12)	0.91(0.11)	0.04(0.14)	0.064	0.27
	Worsened	14	0.70(0.12)	0.63(0.14)	-0.07(0.13)	0.030	-0.55
	All changed	47			0.046(0.13)	0.010	0.35
Self-reported health change related to initially reported condition	Improved	32	0.88(0.12)	0.92(0.10)	0.04(0.13)	0.047	0.31
	Worsened	16	0.71(0.12)	0.65(0.15)	-0.06(0.15)	0.078	-0.37
	All changed	48			0.05(0.14)	0.013	0.33

Note: 1. SRM=Standard response mean, dividing the mean score change (i.e., follow-up minus baseline) by the standard deviation of the change. 2. Groups defined by Self-reported general health change: how would you rate the study child's health in general now? (with answers: Much better, Somewhat better, About the same, Somewhat worse, Much worse). Improved includes *Much better*, Unchanged includes *About the same*, Worsened includes *Somewhat worse* and *Much worse*. 3. The interpretation for SRM were defined as trivial for <0.2, small for >=0.2 and <0.5, medium for >0.5 and <0.8, and large for >=0.8. 4. P value was obtained by paired difference t test. 4. "All changed" combined "Improved" and "Worsened" to increase the sample size of people with changes in order to increase statistical power. The "Improved" generally has higher utility in the follow up while the "Worsened" generally has lower utility in the follow up. We used the follow up utility to minus the baseline utility to calculate the paired differences in both the "Improved" and the "Worsened" groups. In the "All changed" group, to keep the changes in the same direction to calculate effect sizes, we

reversed the calculation for the participants with worsened health (using the baseline utility to minus the follow up utility). The baseline utility and the follow-up utility are not meaningful for the combined “All changed” group and are thus not displayed. P values were one sided P from paired t test.

Table S5.2: Responsiveness of CHU9D and PedsQL related with health changes due to new events happened during follow up

New event happened	Index	Self-reported health change related to the new event	Sample size	Baseline (Mean, SD)	At 4-week Follow-up (Mean, SD)	Paired Difference (Mean)
Start a new treatment	CHU9D utility Australia adolescents	Improved	27	0.6(0.2)	0.6(0.2)	-0.02
		Worsened	2	0.4(0.2)	0.2(0.2)	-0.21
	CHU9D utility UK adults	Improved	27	0.8(0.1)	0.8(0.1)	0.00
		Worsened	2	0.7(0.1)	0.6(0.1)	-0.10
	PedsQL Total score	Improved	27	65.0(18.0)	63.4(19.9)	-1.61
		Worsened	2	40.5(25.3)	45.7(27.6)	5.26
Had an accident or injury	CHU9D utility Australia adolescents	Improved	2	1.0(0.1)	0.9(0.1)	-0.04
		Worsened	2	0.7(0.5)	0.5(0.4)	-0.22
	CHU9D utility UK adults	Improved	2	1.0(0.0)	0.9(0.0)	-0.03
		Worsened	2	0.8(0.3)	0.7(0.2)	-0.09
	PedsQL Total score	Improved	2	85.7(3.4)	89.9(2.5)	4.17
		Worsened	2	42.3(16.0)	55.4(41.2)	13.10
New condition diagnosis	CHU9D utility Australia adolescents	Improved	4	0.7(0.1)	0.7(0.3)	-0.02
		Worsened	1	0.3(.)	0.7(.)	0.38
	CHU9D utility UK adults	Improved	4	0.9(0.1)	0.9(0.2)	-0.01
		Worsened	1	0.6(.)	0.8(.)	0.23
	PedsQL Total score	Improved	4	89.0(3.9)	89.6(15.5)	0.60
		Worsened	1	31.0(.)	84.5(.)	53.57
New illness	CHU9D utility Australia adolescents	Improved	2	0.9(0.2)	0.2(0.0)	-0.65
		Worsened	17	0.6(0.3)	0.4(0.2)	-0.19
	CHU9D utility UK adults	Improved	2	0.9(0.1)	0.6(0.0)	-0.30
		Worsened	17	0.8(0.2)	0.7(0.1)	-0.10
	PedsQL Total score	Improved	2	71.9(26.0)	45.7(27.6)	-26.19
		Worsened	17	62.6(22.9)	54.6(25.9)	-7.96
Unplanned doctor visit	CHU9D utility Australia adolescents	Improved	24	0.7(0.2)	0.7(0.2)	-0.02
		Worsened	1	0.6(.)	0.6(.)	0.09

	CHU9D utility UK adults	Improved	24	0.8(0.1)	0.8(0.1)	-0.01
		Worsened	1	0.7(.)	0.8(.)	0.09
	PedsQL Total score	Improved	24	73.5(18.9)	69.5(21.1)	-4.00
		Worsened	1	73.6(.)	69.4(.)	-4.17
Unplanned hospital visit	CHU9D utility Australia adolescents	Improved	9	0.7(0.1)	0.7(0.2)	-0.03
		Worsened	3	0.5(0.2)	0.3(0.1)	-0.22
	CHU9D utility UK adults	Improved	9	0.9(0.1)	0.8(0.1)	-0.02
		Worsened	3	0.8(0.1)	0.7(0.1)	-0.10
	PedsQL Total score	Improved	9	78.7(11.3)	77.2(9.5)	-1.50
		Worsened	3	54.0(29.4)	58.7(29.7)	4.70
Start a new medication	CHU9D utility Australia adolescents	Improved	27	0.7(0.2)	0.7(0.2)	0.03
		Worsened	2	0.3(0.2)	0.4(0.2)	0.05
	CHU9D utility UK adults	Improved	27	0.8(0.1)	0.8(0.1)	0.01
		Worsened	2	0.6(0.0)	0.7(0.1)	0.04
	PedsQL Total score	Improved	27	72.3(18.8)	75.3(17.3)	2.96
		Worsened	2	34.3(17.1)	32.0(23.7)	-2.28

Note: Parent/caregivers report of major health event between initial and follow-up survey (new treatment/therapy, new medication, accident or injury, new diagnosis, new illness, unplanned doctor/hospital visit). If any major health event selected, parent/caregiver were asked to report if this event made the child's health better or worse or no change.

Section III: Valuation of health-related quality of life

Chapter 6: Test-Retest Reliability of EQ-5D-Y-3L Best-Worst Scaling Choices of Adolescents and Adults.

Published in Value in Health (2023) with Dalziel, K., Huang, L., & Rivero-Arias, O.

Citation: Xiong, X., Dalziel, K., Huang, L., & Rivero-Arias, O. (2023). Test-Retest Reliability of EQ-5D-Y-3L Best-Worst Scaling Choices of Adolescents and Adults. Value in health: the journal of the International Society for Pharmacoeconomics and Outcomes Research, 26(1), 50–54. https://doi.org/10.1016/j.jval.2022.07.007

6.1. Abstract

Background: There is an increasing interest to obtain adolescents' own health state valuation preferences and to understand how these differ from adult preferences for the same health state. An important question in health state valuation is whether adolescents can report preferences reliably, yet research remains limited.

Objective: This study aims to investigate the test-retest reliability of best-worst scaling (BWS) to elicit adolescent preferences compared with adults.

Methods: Identical BWS tasks designed to value EQ-5D-Y-3L health states were administered online in samples of 1000 adolescents (aged 11-17 years) and 1006 adults in Spain. The valuation survey was repeated approximately three days later. We calculated (1) simple percentage agreement, (2) kappa statistic as measures of test-retest reliability. We also compared BWS marginal frequencies, and relative attribute importance (RAI) between baseline and follow-up to explore similarities in the obtained preferences.

Results: We found that both adolescents and adults were able to report their preferences with moderate reliability (kappa: 0.46 for adolescents, 0.46 for adults) for best choices and fair to moderate reliability (kappa: 0.39 for adolescents, 0.41 for adults) for worst choices. No notable difference was observed across years of child age. Higher consistency was observed for best choices compared to worst in some dimensions for both populations. No significant differences were found in the RAI between baseline and follow-up in both populations.

Conclusion: Our results suggest that BWS is a reliable elicitation technique to value EQ-5D-Y-3L health states in both adolescents and adults.

Keywords: test-retest reliability, best-worst scaling, EQ-5D-Y-3L, adolescents, adults, preference

Highlights: There is a lack of evidence whether adolescents can report health state preferences reliably, although there is an increasing interest in including the young population's views in health decision making. This is the first study reporting the test-retest reliability of best-worst scaling (BWS) for health state preference elicitation in adolescents compared to adults. Our study adds to the evidence that adolescents as young as 11-12 years old can complete BWS tasks reliably.

6.2. Introduction

Obtaining preferences for health states is essential to generate utility values for economic evaluation and inform resource allocation decisions.[1] It is commonly accepted that preferences for adult health states should be elicited from the general adult population.[2] Preferences for child health states, however, have been obtained from both adult and child samples for a variety of reasons including normative considerations and concerns about children's cognitive ability.[2] Mounting evidence suggests that child preferences differ from adults'.[3, 4] Where feasible, directly obtaining child and adolescent preferences is increasingly preferred, due to an awareness of the importance of children's own views about intervention and program outcomes.[4]

Ordinal techniques such as discrete choice experiments (DCE) and best-worst scaling (BWS) tasks are relatively easy in terms of comprehension and administration.[5] Previous studies have demonstrated that adolescents can provide internally valid responses in DCE and BWS, for example, their responses to dominant choices are rational.[3, 6] BWS tasks have been increasingly used in health care.[7] Profile case BWS is considered to have a lower cognitive burden than standard DCE[8] and has been used to elicit preferences from adolescents.[9]

There is a research gap related to whether children can report preferences using BWS reliably. The test-retest reliability of a valuation method, also termed repeatability, refers to its the ability to provide consistent utility elicitation over time.[10] Good test-retest reliability is important in reducing measurement error and boosting statistical power.[11] To the best of our knowledge, test-retest reliability of BWS in eliciting preferences for health states has not been examined in samples of adults or children. Beyond health state valuation, and in the field of psychology, only two studies have explored the test-

retest reliability of BWS in the measurement of facial impression and found that BWS is more reliable than Likert ratings.[11, 12] This study aims to investigate the test-retest reliability of using BWS to elicit preferences for EQ-5D-Y-3L in adolescents compared with adults.

6.3. Methods

Two community-based samples, one of adults and the other of adolescents aged 11-17 years, were recruited in Spain via an online panel company in February and March 2016. Full details of the study design can be found elsewhere (ref blinded for review). The 11-17 years age range was chosen as this is when a transitional stage of physical and mental development generally occurs for children [13] and has been used in other preferences elicitation studies. [6, 14, 15] Briefly, the process began with screening questions about age, sex and region to facilitate selection of a representative general Spanish population. The first survey section included the self-completed EQ-5D-Y-3L, consisting of five dimensions: mobility (MO), looking after myself (SC), usual activities (UA), pain or discomfort (PD) and worried, sad or unhappy (SW), with three levels in each dimension. In the second section, participants completed a profile case BWS experiment where participants were presented with single profiles EQ-5D-Y-3L health states and were asked to indicate the dimension level they considered best and worst (see **Appendix S1** for example BWS task). Adolescents and adults completed the survey from their own perspective. A full factorial design was adopted dividing the 243 health states into 20 blocks to generate the profiles for the BWS experiment. Each block included 13 BWS tasks except for one block that included 14. The complete experimental design has been published elsewhere.[3] Each participant was randomly allocated to complete one block. The final section of the survey asked participants about their socio-demographic characteristics.

All participants were invited to repeat the survey within a week and were allocated to their original block of the BWS tasks. To analyze the test-retest reliability, we used the sample completing both the baseline and the follow-up surveys. This study received ethics approval from the [blinded for review].

We calculated (1) simple percentage agreement, (2) kappa coefficient[16] as measures of test-retest reliability at the input data level. Each participant was given 13 or 14 BWS choice tasks represented by a single EQ-5D-Y-3L profile.[3] Participants were asked to choose a level in one of the dimensions (mobility, looking after myself, usual activities, pain/discomfort, worried/sad/unhappy) that they considered best and a level of a different dimension that they considered worst. For example, for best choices each participant had 13/14 observations at each survey, with one observation for each choice task (one health state). The simple percentage agreement between the two repeated surveys was calculated dividing the number of matched answers by the number of all answers available. For kappa estimation,

we compared the dimension chosen at baseline and follow up as the level for each dimension was the same between the two surveys. Unweighted Kappa was calculated as the choices were categorical variables with five categories (mobility, looking after myself, usual activities, pain/discomfort, worried/sad/unhappy).

We estimated simple agreement and unweighted kappa separately for best choice and worst choice as there is literature indicating that worst choices may be less reliable than best choices.[4, 9, 17] Strength of agreement for kappa statistic was judged according to the recommended classification from Landis and Koch[16]: <0.00 indicates poor agreement, 0.00-0.20 indicates slight agreement, 0.21-0.40 indicates fair agreement, 0.41-0.60 indicates moderate agreement, 0.61-0.80 indicates substantial agreement and 0.81-1.00 indicates almost perfect agreement.

We also compared BWS marginal frequencies and relative attribute importance (RAI) between baseline and follow-up to explore similarities in the obtained preferences as supplementary measures of test-retest reliability.[18, 19] These measures investigated test-retest reliability evaluating the stability of choices and model coefficients over time.[18, 19]

The marginal frequency was computed by dividing the number of times a dimension level was chosen as best (or worst) by the number of times that dimension level was available for selection. The Pearson correlation coefficients of the marginal frequencies at the two time points were calculated.

The RAI was calculated based on recommended methods.[20] First, we used a conditional logit model and the pooled best-worst data to estimate latent scale values associated with each dimension level, where the choice responses were treated as a binary dependent variable (1 and 0 for being chosen or not respectively).[21] We used a linear additive utility function (see Equation 1) and assumed that the value of the worst choices was the negative of the value for a best choice. We therefore used variables dummy coded for each dimension assigning 1 to best and -1 to worst. Level 1 for each EQ-5D-Y-3L dimension was used as reference level. All standard errors are cluster-robust, which allows for arbitrary correlation between the error terms at the individual level.

$$V= \beta_1 MO2+\beta_2 MO3+\beta_3 SC2+\beta_4 SC3+\beta_5 UA2+\beta_6 UA3+\beta_7 PD2+\beta_8 PD3+\beta_9 \mathrm{SW2}+\beta_{10} \mathrm{SW3} \quad \text{(Eq. 1)}$$

The beta coefficients in Equation 1 are not directly interpretable and comparable because they represent within-attribute importance referring to the reference levels and must be interpreted in the context of all the other attributes presented to respondents.[20] To aid in their interpretation and comparison across

different groups, we used attribute-based normalization to obtain the RAIs, with attribute importance calculated as a proportion of the reference attribute importance.

$$RAI_y = \frac{\beta y}{\beta x}$$

RAI_y is the RAI score for attribute Y. Attribute X is the reference attribute, and in this study, this is '*worried, sad or unhappy'* as it was the least important dimension from the pooled best-worst model in both samples. β_y and β_x are the coefficients for the level 3 of attribute Y and attribute X, respectively. For example, RAI_MO=β_MO3/β_SW3, RAI_SC=β_SC3/β_SW3, with other attributes following the same process.

All analyses were conducted in Stata SE 16.

6.4. Results

The baseline survey included 1006 adults and 1000 adolescents, with 470 adults and 323 adolescents completing the repeated survey (average 3.36 days for adults and 2.93 days later for adolescents, detailed frequency distribution in Appendix Table S2.2). The sample completing both baseline and follow-up were broadly representative of the general Spanish adult population in terms of gender and age (ref 3, blinded for review), with slightly higher male and older population (Appendix Table S2.1).

The simple percentage agreements were similar between adolescents and adults and were slightly higher for best choices compared to worst choices (adolescent best: 0.571, worst: 0.513; adult best: 0.570, worst: 0.531). There were no notable differences in agreement between different age groups of adolescents (Appendix Table S3.1).

Table 6-1 presents the estimated kappa for adults and adolescents. For best choice, the kappa was 0.46 for adults and adolescents indicating moderate test-retest reliability. For worst choice, the kappa was 0.41 for adults indicating moderate reliability, and 0.39 for adolescents indicating fair reliability. Adolescents had almost the same test-retest reliability in best choice and slightly worse reliability in worst choice than adults. The test-retest reliability of worst choices was worse than best choices in both adults and adolescents. Adolescents as young as 11-12 years old had moderate test-retest reliability in best choices (kappa=0.44) and fair reliability in worst choices (kappa=0.39). The kappa estimates were generally similar in different age groups. Similarly, the test-retest reliability of worst choices was worse than best choices in all sub age groups.

The sample with longer baseline completion time (minimum total completion time, adult: 2.24 minutes, adolescents: 1.66 minutes; median BWS tasks completion time, adults: 4.7 minutes, adolescents: 4 minutes) had higher absolute agreement and kappa estimates, which indicates better test-retest reliability (details in Appendix Table S3.2 and Table S3.3).

The marginal frequencies between baseline and follow-up were similar for both adolescents and adults, with *pain or discomfort* being the most frequently chosen dimension as both best and worst. Baseline and follow-up marginal frequencies were highly correlated (correlation coefficients greater than 0.9). Correlation coefficients were slightly higher for best choices than worst choices, for both adolescents (correlation for best: 0.996, worst: 0.993) and adults (correlation for best: 0.999, worst: 0.986). Please see Appendix Table S3.4 for the detailed marginal frequency results.

The RAI score results of baseline and follow-up were presented in Table 6-2. Take the RAI score of 1.72 for dimension *pain or discomfort* from adolescents at baseline for example, it can be interpreted as respondents consider *pain or discomfort* to be 1.72 times more important than *worried, sad or unhappy* on average. There were no significant differences in RAIs between baseline and follow-up for both adolescents and adults.

6.5. Discussion

This is the first study to our knowledge reporting test-retest reliability of BWS for health state preference elicitation by both adolescents and adults. We found that adolescents aged 11-17 years were able to self-report preferences for EQ-5D-Y-3L health states with a level of reliability similar to adults. The results suggest that it is reliable to directly elicit preference from adolescents as young as 11-12 years old (Appendix Table S3.3) using profile case BWS valuation tasks.

Previous studies have explored test-retest stability of BWS for adult responses.[22, 23] However, test-retest stability only measures consistency for a few tasks within a survey in comparison to test-retest reliability which provides a complete capture of preference reliability measured through a follow-up survey.[24]

Moderate test-retest reliability of BWS was found for adults for both best and worst choices, with kappa ranging from 0.41 to 0.46. Moderate test-retest reliability was found for adolescents in best choices and fair reliability was found for adolescents in worst choices, with kappa ranging from 0.39 to 0.46. Compared with evidence reported previously by discrete choice experiments (see Appendix Table S4.1 for detailed kappa results for DCE reported in previous studies in health care area), the kappa we reported

suggest that BWS has comparable or slightly less reliability in adults.[19, 25-31] For example, Xie et al. (2022) reported kappa of 0.528 for discrete choice experiments with duration in valuing SF-6Dv2 health states.[25] Gamper et al. (2018) reported kappa of 0.411(in France) and 0.605 (in Germany) for valuing QLU-C10D health states.[26] Bryan et al. (2000) reported kappa of 0.65 for preference measurement in treatment of knee injuries.[19] To the best of our knowledge, no previous study reported kappa for DCE for adolescents. Nevertheless, caution is required when comparing kappa across different valuation techniques and/or studies. BWS tasks focus on choices among dimension levels while DCE focuses on choices among health states (combined dimension levels). This may lead to BWS being less cognitive demanding for certain population such as adolescents. Also, the interpretation and comparison of kappa should be exercised with caution as there are other factors that can influence kappa coefficients including prevalence, bias and nonindependence of ratings.[32]

Besides kappa, the high correlation between baseline and follow-up marginal frequencies of BWS choices and the non-significant differences between baseline and follow-up RAI added to the evidence of stability of preferences obtained by BWS in our current study. This is similar with previous DCE studies too.[19, 30] For example, Bryan et al. (2000) reported that the coefficients from models between test and retest were similar and had overlapping 95% confidence intervals.[19]

The interval between the initial survey and follow-up in our study is around 3 days, which is short enough to avoid any significant changes in preferences or health status and long enough to minimize memory effects. Previous studies investigating test-retest reliability adopted intervals ranging from several days to several months.[12, 19, 28] One study compared the test-retest reliability at 2 days and 2 weeks in order to inform interval selection and found no statistically significant differences in the test-retest reliability for the two time intervals, although the study was with adults.[33] The memory effect can be partially tested by comparing the time taken to complete each experiment. If memory effects exist, the time taken for the follow-up experiment is hypothesized to be shorter. Unfortunately, we only collected time to complete the survey at baseline which precluded this analysis which will be a valuable consideration for future experiments. However, we found that people with longer baseline completion time had better reliability. The reason may be that longer completion time signifies careful thinking and thus increased reliability. Another interesting finding is that adolescents took shorter time to complete the BWS tasks than adults. Similar results were seen in previous studies[6, 34] however further investigation may be needed as to why this is the case. Given that in our study longer completion time appears to be associated with a higher kappa, we speculate that adults may tend to think more carefully about their choices.

Another factor that may impact the reliability of preferences is preference construction that occurs during an elicitation task.[35] In the retest survey, individual's preferences may be affected by the thoughts provoked by the initial evaluation tasks. This may partly explain why the responses between the initial and follow-up surveys are never 100% the same. Considering this issue, the true reliability may be higher than our estimates.

We found that best choices were slightly more reliable than worst choices. This echoes with previous research findings that worst choices tend to be less consistent.[4, 9, 17] Therefore, caution should be taken when combining best and worst choice responses. Further research is warranted to investigate the implications and options for managing differences in best and worst values when eliciting health state preferences.

Our study has several strengths. The sample size of our study is relatively large among similar studies evaluating test-retest reliabilities. Second, we included both adolescents and adults, enabling the comparison between them. In addition, the test-retest reliability was assessed at different levels, including choice-set level (e.g. simple agreement, kappa) and level of parametric models (e.g. relative attribute importance estimates), making the conclusion more robust. However, our study is not without limitations. A higher percentage of the adult participants compared to adolescent participants completed the follow-up survey. This may be related to internet accessibility on a day-to-day basis, which would be unlikely to correlate with reliability and preferences. Although the unequal samples may imply more precise estimates for adults, variability around estimates in terms of 95% CI of the kappa suggest that the impact was minimal and unlikely to impact our conclusion. Additionally, the test-retest reliability of BWS may be different in valuing health states of other multiple-attribute utility instruments (MAUIs) and future similar studies using other MAUIs would be valuable.

6.6. Conclusion

Our study adds to the evidence that adolescents as young as 11-12 years old can complete BWS tasks reliably.

6.7. Reference

[1] J. Brazier, D. Rowen, M. Karimi, T. Peasgood, A. Tsuchiya, and J. Ratcliffe, "Experience-based utility and own health state valuation for a health state classification system: why and how to do it," *The European Journal of Health Economics,* vol. 19, no. 6, pp. 881-891, 2018/07/01 2018.

[2] D. Rowen, O. Rivero-Arias, N. Devlin, and J. Ratcliffe, "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?," (in eng), *Pharmacoeconomics,* Jan 6 2020.

[3] K. Dalziel, M. Catchpool, B. Garcia-Lorenzo, I. Gorostiza, R. Norman, and O. Rivero-Arias, "Feasibility, Validity and Differences in Adolescent and Adult EQ-5D-Y Health State Valuation in Australia and Spain: An Application of Best-Worst Scaling," (in eng), *Pharmacoeconomics,* Jan 24 2020.

[4] J. Ratcliffe, E. Huynh, K. Stevens, J. Brazier, M. Sawyer, and T. Flynn, "Nothing About Us Without Us? A Comparison of Adolescent and Adult Health-State Values for the Child Health Utility-9D Using Profile Case Best-Worst Scaling," (in eng), *Health Econ,* vol. 25, no. 4, pp. 486-96, Apr 2016.

[5] S. Ali and S. Ronaldson, "Ordinal preference elicitation methods in health economics and health services research: using discrete choice experiments and ranking methods," *British Medical Bulletin,* vol. 103, no. 1, pp. 21-44, 2012.

[6] D. J. Mott, K. K. Shah, J. M. Ramos-Goñi, N. J. Devlin, and O. Rivero-Arias, "Valuing EQ-5D-Y-3L Health States Using a Discrete Choice Experiment: Do Adult and Adolescent Preferences Differ?," *Medical Decision Making,* p. 0272989X21999607, 2021.

[7] K. L. Cheung *et al.*, "Using Best–Worst Scaling to Investigate Preferences in Health Care," *PharmacoEconomics,* vol. 34, no. 12, pp. 1195-1209, 2016/12/01 2016.

[8] H. J. Rogers, Z. Marshman, H. Rodd, and D. Rowen, "Discrete choice experiments or best-worst scaling? A qualitative study to determine the suitability of preference elicitation tasks in research with children and young people," *Journal of Patient-Reported Outcomes,* vol. 5, no. 1, p. 26, 2021/03/10 2021.

[9] J. Ratcliffe *et al.*, "Valuing the Child Health Utility 9D: Using profile case best worst scaling methods to develop a new adolescent specific scoring algorithm," (in eng), *Soc Sci Med,* vol. 157, pp. 48-59, May 2016.

[10] H. M. E. van Agt, M.-L. Essink-Bot, P. F. M. Krabbe, and G. J. Bonsel, "Test-retest reliability of health state valuations collected with the EuroQol questionnaire," *Social Science & Medicine,* vol. 39, no. 11, pp. 1537-1544, 1994/12/01/ 1994.

[11] N. Burton, M. Burton, C. Fisher, P. G. Peña, G. Rhodes, and L. Ewing, "Beyond Likert ratings: Improving the robustness of developmental research measurement using best–worst scaling," *Behavior Research Methods,* 2021/04/05 2021.

[12] N. Burton, M. Burton, D. Rigby, C. A. M. Sutherland, and G. Rhodes, "Best-worst scaling improves measurement of first impressions," *Cognitive Research: Principles and Implications,* vol. 4, no. 1, p. 36, 2019/09/23 2019.

[13] S. M. Sawyer, P. S. Azzopardi, D. Wickremarathne, and G. C. Patton, "The age of adolescence," (in eng), *Lancet Child Adolesc Health,* vol. 2, no. 3, pp. 223-228, Mar 2018.

[14] J. Ratcliffe, T. Flynn, F. Terlich, K. Stevens, J. Brazier, and M. Sawyer, "Developing Adolescent-Specific Health State Values for Economic Evaluation," *PharmacoEconomics,* vol. 30, no. 8, pp. 713-727, 2012/08/01 2012.

[15] J. Ratcliffe *et al.*, "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," (in eng), *Appl Health Econ Health Policy,* vol. 9, no. 1, pp. 15-27, 2011.

[16] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," (in eng), *Biometrics,* vol. 33, no. 1, pp. 159-74, Mar 1977.

[17] G. Chen, F. Xu, E. Huynh, Z. Wang, K. Stevens, and J. Ratcliffe, "Scoring the Child Health Utility 9D instrument: estimation of a Chinese child and adolescent-specific tariff," (in eng), *Qual Life Res,* vol. 28, no. 1, pp. 163-176, Jan 2019.

[18] U. Liebe, J. Meyerhoff, and V. Hartje, "Test–Retest Reliability of Choice Experiments in Environmental Valuation," *Environmental and Resource Economics,* vol. 53, no. 3, pp. 389-407, 2012/11/01 2012.

[19] S. Bryan, L. Gold, R. Sheldon, and M. Buxton, "Preference measurement using conjoint methods: an empirical investigation of reliability," *Health Economics,* https://doi.org/10.1002/1099-1050(200007)9:5<385::AID-HEC533>3.0.CO;2-W vol. 9, no. 5, pp. 385-395, 2000/07/01 2000.

[20] J. M. Gonzalez, "A Guide to Measuring and Interpreting Attribute Importance," *The Patient - Patient-Centered Outcomes Research,* vol. 12, no. 3, pp. 287-295, 2019/06/01 2019.

[21] A. C. Mühlbacher, A. Kaczynski, P. Zweifel, and F. R. Johnson, "Experimental measurement of preferences in health and healthcare using best-worst scaling: an overview," *Health economics review,* vol. 6, no. 1, p. 2, 2016.

[22] N. Krucien, V. Watson, and M. Ryan, "Is Best–Worst Scaling Suitable for Health State Valuation? A Comparison with Discrete Choice Experiments," *Health Economics,* https://doi.org/10.1002/hec.3459 vol. 26, no. 12, pp. e1-e16, 2017/12/01 2017.

[23] F. Xie, E. Pullenayegum, K. Gaebel, M. Oppe, and P. F. M. Krabbe, "Eliciting preferences to the EQ-5D-5L health states: discrete choice experiment or multiprofile case of best–worst scaling?," *The European Journal of Health Economics,* vol. 15, no. 3, pp. 281-288, 2014.

[24] E. M. Janssen, D. A. Marshall, A. B. Hauber, and J. F. P. Bridges, "Improving the quality of discrete-choice experiments in health: how can we assess validity and reliability?," (in eng), *Expert Rev Pharmacoecon Outcomes Res,* vol. 17, no. 6, pp. 531-542, Dec 2017.

[25] S. Xie, J. Wu, and G. Chen, "Discrete choice experiment with duration versus time trade-off: a comparison of test–retest reliability of health utility elicitation approaches in SF-6Dv2 valuation," *Quality of Life Research,* 2022/05/25 2022.

[26] E.-M. Gamper *et al.*, "Test-Retest Reliability of Discrete Choice Experiment for Valuations of QLU-C10D Health States," *Value in Health,* vol. 21, no. 8, pp. 958-966, 2018/08/01/ 2018.

[27] D. Bijlenga, G. J. Bonsel, and E. Birnie, "Eliciting willingness to pay in obstetrics: comparing a direct and an indirect valuation method for complex health outcomes," vol. 20, no. 11, pp. 1392-1406, 2011.

[28] U. S. Skjoldborg, J. Lauridsen, and P. Junker, "Reliability of the Discrete Choice Experiment at the Input and Output Level in Patients with Rheumatoid Arthritis," *Value in Health,* vol. 12, no. 1, pp. 153-158, 2009/01/01/ 2009.

[29] D. Bijlenga, E. Birnie, and G. J. Bonsel, "Feasibility, Reliability, and Validity of Three Health-State Valuation Methods Using Multiple-Outcome Vignettes on Moderate-Risk Pregnancy at Term," *Value in Health,* vol. 12, no. 5, pp. 821-827, 2009/07/01/ 2009.

[30] M. Ryan, A. Netten, D. Skåtun, and P. Smith, "Using discrete choice experiments to estimate a preference-based measure of outcome—An application to social care for older people," *Journal of Health Economics,* vol. 25, no. 5, pp. 927-944, 2006/09/01/ 2006.

[31] F. San Miguel, M. Ryan, and A. Scott, "Are preferences stable? The case of health care," *Journal of Economic Behavior & Organization,* vol. 48, no. 1, pp. 1-14, 2002/05/01/ 2002.

[32] J. Sim and C. C. Wright, "The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements," *Physical Therapy,* vol. 85, no. 3, pp. 257-268, 2005.

[33] R. G. Marx, A. Menezes, L. Horovitz, E. C. Jones, and R. F. Warren, "A comparison of two time intervals for test-retest reliability of health status instruments," *Journal of Clinical Epidemiology,* vol. 56, no. 8, pp. 730-735, 2003/08/01/ 2003.

[34] V. Prevolnik Rupel, J. M. Ramos-Goñi, M. Ogorevc, S. Kreimeier, K. Ludwig, and W. Greiner, "Comparison of Adult and Adolescent Preferences Toward EQ-5D-Y-3L Health States," (in eng), *Value in*

health : the journal of the International Society for Pharmacoeconomics and Outcomes Research, vol. 24, no. 9, pp. 1350-1359, 2021.

[35] A. J. Lloyd, "Threats to the estimation of benefit: are preference elicitation methods accurate?," *Health Economics,* https://doi.org/10.1002/hec.772 vol. 12, no. 5, pp. 393-402, 2003/05/01 2003.

6.8. Tables and figures

Table 6-1 Kappa for best choice and worst choice between baseline and follow up

	Kappa for best choice (95% confidence interval)	**Kappa for worst choice (95% confidence interval)**
Adults	0.46 (0.45,0.47)	0.41 (0.40,0.42)
Adolescents	0.46 (0.44,0.47)	0.39 (0.37,0.40)
11-12y	0.44 (0.41,0.47)	0.39 (0.36,0.42)
13-14y	0.49 (0.46,0.52)	0.41 (0.38,0.44)
15-17y	0.45 (0.43,0.47)	0.37 (0.35,0.39)

Note: Landis and Koch proposed the following standards for strength of agreement for the kappa coefficient: <=0=poor, 0.01-0.2=slight, 0.21-0.40=fair, 0.41-0.60=moderate, 0.61-0.80=substantial, 0.80-1=almost perfect.

Table 6-2 RAI scores and differences between baseline and follow-up

	Baseline		Follow up		RAI difference (95% CI)	*P*-value
	RAI	SE	RAI	SE		
Adolescent (n=323)						
Mobility	1.49	0.09	1.34	0.07	0.14 (-0.09, 0.37)	0.229
Looking after myself	1.26	0.08	1.18	0.06	0.08 (-0.12, 0.28)	0.448
Usual activities	1.49	0.09	1.42	0.07	0.06 (-0.17, 0.29)	0.606
Pain or discomfort	1.72	0.10	1.56	0.07	0.17 (-0.07, 0.41)	0.171
Worried, sad or unhappy†	1.00		1.00			
Adult (n=470)						
Mobility	1.17	0.05	1.17	0.05	0.00 (-0.14, 0.13)	0.958
Looking after myself	1.09	0.05	1.12	0.04	-0.03 (-0.16, 0.09)	0.613
Usual activities	1.29	0.06	1.31	0.05	-0.02 (-0.16, 0.13)	0.825
Pain or discomfort	1.42	0.06	1.40	0.05	0.02 (-0.13, 0.17)	0.801
Worried, sad or unhappy†	1.00		1.00			

Note: Coefficients obtained from conditional logistic regression model; standard errors (SE) calculated using the Delta method.
†Worried, sad or unhappy was the reference attribute.

6.9. Supplementary materials

Appendix S1: Example BWS task

Select the part that you consider is the best of living in that health state, and then the part that you consider is the worst for each question.

Best	**Example 1**	**Worst**
	I have a lot of problems walking about	**X**
	I have some problems washing or dressing myself	
	I have a lot of problems doing my usual activities	
	I have some pain or discomfort	
X	I am not worried, sad, or unhappy	

Appendix S2: Participant characteristics and time interval between baseline and follow-up surveys

Table S2.1 Socio-demographic characteristics of test-retest respondents

Characteristics	Adolescents		Adults		Spanish adult general population %
	Baseline (n=1000)	Follow up (n=323)*	Baseline (n=1006)	Follow up (n=470)*	
Sex, n (%)					
Male	448 (44.80)	180 (55.73)	491 (48.81)	268 (57.02)	49.35
Female	552 (55.20)	143 (44.27)	515 (51.19)	202 (42.98)	50.65
Age, n (%)					
11 to 12	252 (25.20)	78 (24.15)			
13 to 14	275 (27.50)	91 (28.17)			
15 to 17	473 (47.30)	154 (47.68)			
18-24			83 (8.25)	49 (10.43)	9.36
25-34			157 (15.61)	40 (8.51)	18.07
35-44			217 (21.57)	69 (14.68)	20.52
45-54			189 (18.79)	103 (21.91)	17.67
55 and older			360 (35.79)	209 (44.47)	34.38
Participant with chronic condition, n (%)					
No	864 (86.40)	279 (86.38)	634 (63.02)	293 (62.34)	
Yes	134 (13.40)	43 (13.31)	372 (36.98)	177 (37.66)	
Missing	2 (0.20)	1 (0.31)			
Family Affluence Scale (FAS), n (%)					
Medium	408 (40.80)	124 (38.39)			
High	584 (58.40)	195 (60.37)			
Missing	8 (0.80)	4 (1.24)			
Income group, n (%)					
Less than 1000 euros			191 (18.99)	82 (17.45)	
1000 to less than 2000 euros			390 (38.77)	179 (38.09)	
2000 to less than 3000 euros			268 (26.64)	134 (28.51)	
3000 euros or more			155 (15.41)	73 (15.53)	
Missing			2 (0.20)	2 (0.43)	
Education level, n (%)					
Primary			129 (12.82)	61 (12.98)	
Secondary			207 (20.58)	104 (22.13)	
Vocational training			218 (21.67)	96 (20.43)	
Higher education			450 (44.73)	208 (44.26)	
Missing			2 (0.20)	1 (0.21)	
Employment status, n (%)					
Employed			517 (51.39)	248 (52.77)	
Unemployed			166 (16.50)	63 (13.40)	

Student	70 (6.96)	31 (6.60)
Retired/Early retired	105 (10.44)	64 (13.62)
Other	145 (14.41)	63 (13.40)
Missing	3 (0.30)	1 (0.21)
Marital status, n (%)		
Single	328 (32.60)	127 (27.02)
Married	556 (55.27)	279 (59.36)
Widow/er	36 (3.58)	23 (4.89)
Separated	29 (2.88)	18 (3.83)
Divorced	57 (5.67)	23 (4.89)
Number of children in household under 18, n (%)		
No children	688 (68.39)	320 (68.09)
One child	179 (17.79)	81 (17.23)
Two children	112 (11.13)	59 (12.55)
Three children or more	22 (2.19)	9 (1.91)
Missing	5 (0.50)	1 (0.21)
Living with someone with disabilities, n (%)		
No	888 (88.27)	415 (88.30)
Yes	114 (11.33)	53 (11.28)
Missing	4 (0.40)	2 (0.43)

*Our retest samples constitute 32%-47% of the baseline, which is neither low nor high compared to similar studies, e.g. 356/390 [1], 50/294 [2], 105/500 [3].

Note: A composite FAS score is calculated for each adolescent based on his or her responses to four items (car, holidays, computer, bedroom). For most analysis, we use a 3-point ordinal scale, where FAS low (score = 0, 1, 2) indicates low affluence, FAS medium (score = 3, 4, 5) indicates middle affluence and FAS high (score = 6, 7, 8, 9) indicates high affluence.

Table S2.2 Time interval between baseline and follow-up surveys

	Adult		Adolescent	
Interval, days	Frequency	Percent	Frequency	Percent
1	2	0.43	-	-
2	74	15.74	69	21.36
3	217	46.17	213	65.94
4	118	25.11	36	11.15
5	47	10	4	1.24
6	12	2.55	1	0.31
Total	470	100	323	100

Appendix S3: Other reliability measures

Table S3.1 Simple agreement by proportion with exact match in choices

	Adolescents				Adults
	All ages	11-12y (n=78)	13-14y (n=91)	15-17y (n=154)	All ages
Best	0.571	0.556	0.594	0.565	0.570
Worst	0.513	0.514	0.533	0.501	0.531

Note: Agreement is equal to 1 when the choice of best or worst of each set in the two surveys are the same, and 0 when not the same. For example, if participant A chose 'No problems in pain/discomfort' as the best choice in both the first and second surveys, then the answers are the same.

Table S3.2 Simple agreement by baseline completion time

		Simple agreement by baseline completion time		
Population	Best/Worst	<p25	p25-p75	>p75
Adolescent	Best	0.463	0.595	0.654
	Worst	0.399	0.551	0.578
Adult	Best	0.502	0.592	0.620
	Worst	0.448	0.564	0.575

Note: Agreement is equal to 1 when the choice of best or worst of each set in the two surveys are the same, and 0 when not the same. Baseline completion time was categorized into three groups with 25% percentile (p25) and 75% percentile (p75) as cutoffs.

Table S3.3 Kappa for best and worst choices by baseline completion time

	Kappa (95% confidence interval)		
Sample	Baseline completion time <p25	Baseline completion time p25-p75	Baseline completion time >p75
Adults			
Best choice	0.37 (0.35,0.40)	0.49 (0.47,0.50)	0.52 (0.49,0.55)
Worst choice	0.31 (0.28,0.33)	0.45(0.43,0.47)	0.46 (0.43,0.49)
Adolescents			
Best choice	0.32 (0.29,0.35)	0.49 (0.47,0.51)	0.56 (0.53,0.59)
Worst choice	0.25 (0.22,0.27)	0.43 (0.41,0.45)	0.46 (0.43,0.50)

* The sample with longer baseline completion time generally have higher kappa for both best and worst choices. The sample with the lowest 25 percentile baseline completion time has lower kappa than the ones with longer completion time. The sample in the highest 25 percentile completion time group has higher or similar kappa compared with the sample in the middle 50 percentile group.

Table S3.4 Marginal frequencies of BWS comparing baseline and follow-up

Attribute level	Adolescent (n=323)						Adult (n=470)					
	Best			Worst			Best			Worst		
	Baseline	Follow up	Corr	Baseline	Follow up	Corr	Baseline	Follow up	Corr	Follow up	Baseline	Corr
MO1	0.377	0.408		0.063	0.046		0.383	0.387		0.044	0.057	
MO2	0.085	0.061		0.179	0.184		0.088	0.091		0.162	0.168	
MO3	0.061	0.053		0.321	0.336		0.068	0.056		0.295	0.280	
SC1	0.295	0.311		0.064	0.050		0.327	0.345		0.030	0.040	
SC2	0.111	0.094		0.138	0.131		0.080	0.083		0.145	0.153	
SC3	0.095	0.085		0.240	0.259		0.073	0.074		0.277	0.234	
UA1	0.447	0.488		0.087	0.062		0.475	0.502		0.047	0.055	
UA2	0.084	0.082		0.193	0.194		0.084	0.079		0.206	0.206	
UA3	0.087	0.060		0.353	0.400		0.068	0.050		0.394	0.363	
PD1	**0.491**	**0.536**		0.078	0.066		**0.505**	**0.528**		0.071	0.051	
PD2	0.097	0.096		0.215	0.188		0.136	0.128		0.178	0.200	
PD3	0.042	0.051		***0.426***	***0.475***		0.062	0.053		***0.464***	***0.443***	
SW1	0.409	0.432		0.098	0.063		0.401	0.421		0.069	0.101	
SW2	0.230	0.204		0.177	0.136		0.199	0.185		0.153	0.200	
SW3	0.217	0.187	0.996	0.232	0.241	0.993	0.194	0.175	0.999	0.289	0.297	0.986

Note: MO: Mobility, SC: looking after myself, UA: doing usual activities, PD: having pain or discomfort, SW: feeling worried, sad or unhappy. MO1, MO2, and MO3 indicate three levels of dimension MO (1: no problems, 2: some problems, 3: a lot of problems) with the same pattern applied for the other dimensions. **Bold =** most often being selected as best; ***bold italic*** = most often being selected as worst. 'Corr' = Pearson's correlation coefficient of baseline and follow-up marginal frequencies.

Reference

[1] W. J. Brown, S. G. Trost, A. Bauman, K. Mummery, and N. Owen, "Test-retest reliability of four physical activity measures used in population surveys," *Journal of Science and Medicine in Sport,* vol. 7, no. 2, pp. 205-215, 2004/06/01/ 2004.

[2] X. Badia, S. Monserrat, M. Roset, and M. Herdman, "Feasibility, validity and test-retest reliability of scaling methods for health states: the visual analogue scale and the time trade-off," (in eng), *Qual Life Res,* vol. 8, no. 4, pp. 303-10, Jun 1999.

[3] S.-H. Kim, S.-i. Lee, and M.-W. Jo, "Feasibility, comparability, and reliability of the standard gamble compared with the rating scale and time trade-off techniques in Korean population," *Quality of Life Research,* vol. 26, no. 12, pp. 3387-3397, 2017/12/01 2017.

Appendix S4: Supplementary material about previous literature on DCE reliability

This is not a systematic review, and the below studies are based on an unstructured literature search to identify relevant DCE studies evaluating test-retest reliability and reported kappa results in health care to enhance our discussion. We have also checked 130 studies citing a relevant DCE reliability study "Preference measurement using conjoint methods: an empirical investigation of reliability" which was published in 2000 (accessed in google scholar on June 23rd, 2022). A summary in a recent book chapter also confirmed that main references were identified.[1]

Table S4.1 Kappa of DCE previously reported in health care

Author	Year	Population	Sample size	Age	Administration mode	Interval between two tests	Description of choice tasks	Method	Number of choice tasks	Absolute agreement	Kappa
Xie et al. [2]	2022	Representative sample of the Chinese general population	162	Range 18–80 years	Face-to-face interviews	2 weeks	SF-6Dv2 health states	*DCE_{TTO}	10	76.40%	0.528
Gamper et al.[3]	2018	General population samples	300 German respondents, and 305 French respondents	Mean age 48 and 47 years for German and French	Web-based self-completed survey using online panels	4 to 6 weeks	QLU-C10D Health States (a cancer-specific multi-attribute utility instrument)	DCE	16	Germany: 80.2%; France: 70.6%	Germany: 0.605; France: 0.411
Bijlenga et al.[4]	2011	Laypersons	88	Between 21 and 79 years of age	Questionnaires by postal mail	within one week	Willingness to pay (WTP) elicitation in obstetrics	DCE	6	NA	0.49 (95% CI: 0.38–0.60)
Bijlenga et al.[5]	2009	People from community	97	Mean 51.5, range 21-79.	Panel session and individual home assignment	ranged from 3 to 21 days, with a	Multiple-Outcome Vignettes on Moderate-Risk Pregnancy at Term	DCE	11	NA	0.78

						median of 5 days.					
Skjoldborg et al.[6]	2009	Patients with rheumatoid arthritis	145, 130	Ranging between 18 and 70 years	Face-to-face interviews	4 months	Scenarios that describe the effect of treating patients with rheumatoid arthritis	DCE	10	75.8% or 87.2%	0.496 or 0.725
Ryan et al. [7]	2006	People aged 60 and older	47	Mean 76.3 (SD 8.2)	Face-to-face interviews	11-60-days	Social care outcome measure	DCE	7	82%	0.64
San Miguel et al. [8]	2002	Parents of children under 13y	731	Mean age 35 years, ranging from 18 to 53 years.	The experiment was mailed to the parents or guardians of these children	2 months	Parents' preferences for out-of-hours health care for their children	DCE	7 or 8	NA	0.75, 0.68, 0.65, 0.68
Bryan et al. [9]	2000	University students	134	Mean: 21.8 years, range 18-40.	Questionnaires completed at the end or at the beginning of a lecture, postal questionnaire of the retest survey.	2 weeks	Preferences for treatment options for patients with knee injuries	DCE	12	86%	0.65

*DCE_{TTO}: discrete choice experiments with duration, a variant of DCE

Reference

[1] P. Mariel *et al.*, "Validity and Reliability," in *Environmental Valuation with Discrete Choice Experiments: Guidance on Design, Implementation and Data Analysis*, P. Mariel *et al.*, Eds. Cham: Springer International Publishing, 2021, pp. 111-123.

[2] S. Xie, J. Wu, and G. Chen, "Discrete choice experiment with duration versus time trade-off: a comparison of test–retest reliability of health utility elicitation approaches in SF-6Dv2 valuation," *Quality of Life Research,* 2022/05/25 2022.

[3] E.-M. Gamper *et al.*, "Test-Retest Reliability of Discrete Choice Experiment for Valuations of QLU-C10D Health States," *Value in Health,* vol. 21, no. 8, pp. 958-966, 2018/08/01/ 2018.

[4] D. Bijlenga, G. J. Bonsel, and E. Birnie, "Eliciting willingness to pay in obstetrics: comparing a direct and an indirect valuation method for complex health outcomes," vol. 20, no. 11, pp. 1392-1406, 2011.

[5] D. Bijlenga, E. Birnie, and G. J. Bonsel, "Feasibility, Reliability, and Validity of Three Health-State Valuation Methods Using Multiple-Outcome Vignettes on Moderate-Risk Pregnancy at Term," *Value in Health,* vol. 12, no. 5, pp. 821-827, 2009/07/01/ 2009.

[6] U. S. Skjoldborg, J. Lauridsen, and P. Junker, "Reliability of the Discrete Choice Experiment at the Input and Output Level in Patients with Rheumatoid Arthritis," *Value in Health,* vol. 12, no. 1, pp. 153-158, 2009/01/01/ 2009.

[7] M. Ryan, A. Netten, D. Skåtun, and P. Smith, "Using discrete choice experiments to estimate a preference-based measure of outcome—An application to social care for older people," *Journal of Health Economics,* vol. 25, no. 5, pp. 927-944, 2006/09/01/ 2006.

[8] F. San Miguel, M. Ryan, and A. Scott, "Are preferences stable? The case of health care," *Journal of Economic Behavior & Organization,* vol. 48, no. 1, pp. 1-14, 2002/05/01/ 2002.

[9] S. Bryan, L. Gold, R. Sheldon, and M. Buxton, "Preference measurement using conjoint methods: an empirical investigation of reliability," *Health Economics,* https://doi.org/10.1002/1099-1050(200007)9:5<385::AID-HEC533>3.0.CO;2-W vol. 9, no. 5, pp. 385-395, 2000/07/01 2000.

Chapter 7: Valuing the Child Health Utility 9D (CHU9D) for children under 5 years in Australia.

Authors: Xiuqin Xiong, Kim Dalziel, Li Huang, Natalie Carvalho, Nancy Devlin

Suggested reference (manuscript finished): Xiong X, Dalziel K, Huang L, Carvalho N, Devlin N. Valuing the Child Health Utility 9D (CHU9D) for children under 5 years in Australia.

7.1. Abstract

Objective: Children under 5 years are relatively high health service users, so evidence to inform decisions about value of care and outcomes is important. There is a lack of research on the valuation of child health-related quality of life (HRQoL) for children under 5 years. This study aims to explore whether the adult general population's preferences for health states as described by the Child Health Utility 9D (CHU9D) differ when they consider a 2–4-year-old child compared to a 10-year-old child, and to develop a value set appropriate for scoring the CHU9D for children under 5 years.

Methods: An online survey using a discrete choice experiment (DCE) comprising 12 choice tasks was administered between September and November 2023 to a representative sample of Australian adults from the general population, where each task asked respondents to choose between two health states. Participants were randomly allocated to two arms: one considering the health of a 2–4-year-old child, and the other a 10-year-old child. A conditional logit model was used to analyze the responses and produce preference weights. Comparisons between the 2-4- and 10-year-old child perspectives were conducted using the relative attribute importance (RAI), a poolability test of preference, and a pooled model with interaction terms. Visual Analogue Scale (VAS) responses were used to anchor the latent values onto a 0-1 scale.

Results: In total, 2428 participants completed the survey:1176 in the 2–4-year-old arm and 1252 in the 10-year-old arm. The RAIs, the poolability test and the pooled model with interactions all suggested that there is no appreciable difference between the two age perspectives. The pooled data were then used to estimate health state values. Most (27 out of 36) coefficients showed desired monotonicity where worse health states are associated with larger value impairments. The final value set was based on the consistent model and the value range from 0.23 for the worst state to 1 for the best state.

Conclusion: The adult general population's preferences for health states described by CHU9D are similar when asked to consider a 2–4-year-old compared to a 10-year-old child. A value set was developed based on the preferences of the adult general population in Australia. This makes it possible for cost-utility analysis to include the quality of life of children under 5 years old using the CHU9D. The value set based on adults' preferences complements the existing Australian CHU9D value set based on adolescents' preferences and may be suitable for population-level decision-making requiring an adult perspective.

7.2. Introduction

Children under 5 years of age are essential users of healthcare services, requiring dedicated attention and consideration in healthcare planning.[1] Many emerging healthcare technologies focus on addressing diseases prevalent in early childhood including congenital and genetic disorders.[2-4] Consequently, evidence informed decisions regarding the allocation of health resources for this age group is crucial. The utilization of economic evaluation to assess childhood interventions to aid resource allocation decisions, particularly through cost-utility analysis, has increased in recent years.[5, 6] However, there is a limited availability of health-related quality of life (HRQoL) instruments appropriate for utility measurement of young children; a required input to cost-utility analyses.[7-9]

Many economic evaluations including younger children used utilities obtained from generic pediatric HRQoL measures developed for older age groups or adults. It is acknowledged that HRQoL instruments developed for older children may not be suitable for use in young children due to possible differences in relevant health dimensions and framing. Using adapted wording or adding guidance notes is assumed to enhance the suitability of these instruments for young children.[10, 11] Additionally, applying adapted versions from instruments used with older children has the advantage of enabling consistent HRQoL measurement throughout childhood. The child age people consider and/or the adapted wording (or added guidance notes) in valuation tasks may impact people's preferences for health states.[12]

Health technology assessment (HTA) authorities in Australia and the UK have noted the lack of available health state values for use in pediatric economic evaluations and promote the use of concise, generic measures of pediatric HRQoL accompanied by relevant value sets. There is potential to unfairly penalize young children in the HTA process due to poor quality, missing or uncertain utility evidence. A recent review found that the current evidence on child health-related quality of life being submitted to PBAC is lacking,[13] and rarely uses appropriate HRQoL instruments. Importantly, instruments appropriate for

health state value measurement in pre-school children are lacking, compared to their older peers and adults.

Child Health Utility 9D (CHU9D) is a concise, generic measure of HRQoL, which was developed specifically for children to facilitate the estimation of quality-adjusted life-years (QALYs) for the economic evaluation.[14] It has been well validated for use for children between 7-17 years of age.[15-18] CHU9D developers offer a proxy version with guidance notes for measuring HRQoL in children under 5 years old. Its psychometric performance has been assessed and the findings published(see chapter 5).[19] The CHU9D proxy version with guidance notes demonstrated good psychometric performance in 2–4-year-old Australian children and shows potential as a valid and reliable instrument for assessing their HRQoL. However, the preference-weighted scoring for this instrument in this age group is still lacking. Preference-weighted scoring will allow health state values to be accurately and consistently produced by the CHU9D in children as young as 2 years old to facilitate economic evaluation of pediatric interventions.

Several preference-weighted scoring systems or value sets exist for CHU9D. The earliest value set was developed by Steven etc. 2012 based on preferences obtained from UK adults for the CHU9D original version suitable for 7-17 years.[20] Ratcliffe etc. 2016 compared the values between Australian adults and adolescents using best-worst scaling and provided estimates that can be potentially used as value sets for QALY calculation.[21] However, the limitations are that they used the worst health state value from the existing UK adult scoring algorithm to anchor the estimates on a 0-1 scale, and the sample of the adult general population is not representative, and thus no formal value sets was developed based on this study.[21] The formal Australian adolescent value set was developed based on another study using a large sample of adolescents recruited from an Australia wide online panel company.[22] Recently, there are also value sets available for China and the Netherlands.[23, 24] Different preferences across countries are often found due to different culture contexts and thus it is important to develop a specific value set for each country.

Clearly, there is a lack of value sets for younger children under 5 years old for CHU9D, although CHU9D for 2-4 years old has already demonstrated good psychometric performance. To solve this problem, the first thing is to understand whether the existing value sets for older children can be used for 2-4 years as well. In other words, there is a need to understand whether there are differences in values for health states for 2–4-year-olds compared with older children. There is currently no evidence about this to our knowledge.

In addition, whose preferences to use when valuing child HRQOL is an ongoing debate. Adolescents' preferences are considered valuable to best reflect children's own experience and views. It also empowers adolescents by giving them a voice in decisions impacting their own health and leads to more relevant and effective policies. However, when developing a value set for use in country-level decision making it is common practice for preferences to be obtained from the general adult population as they are taxpayers [25]. Discrete choice experiments (DCEs) have become widely used in recent years in the valuation of health-related quality of life (HRQoL) instruments, and are the method recommended in the international valuation protocol for a widely used instrument, theEQ-5D-Y-3L.[26] Its advantages include relatively low cognitive burden compared to traditional valuation techniques such as time-trade off (TTO) and standard gamble (SG), and a lower cost of data collection with self-completion online. The disadvantage of the DCE is that it produced values on a latent scale that are unanchored to a 0-1 utility scale, so this is usually conducted via a separate anchoring task, for example using a TTO. However, when valuing child health, people are often reluctant to sacrifice life years for children leading to inflated utility values derived using TTO methods. Another alternative for producing anchored utility values is to include duration in the DCE valuation task, however for health states described for children under 5 years old this is problematic. For example, a 2–4-year-old child would be an adolescent in scenarios where participants are asked to consider a 10-year duration, which may make the health state descriptor no longer appropriate. The visual analogue scale (VAS) is a simple and pragmatic approach previously used to anchor DCE latent scales to a 0-1 utility scale.[27]

The rational of this study is to first explore whether the general population adults' preferences for health states of children aged 2-4 years old differ from preferences for older children. To be more specific, the impact of the added guidance notes of the CHU9D and the age of child participants are asked to consider was investigated. If the general population adults' preferences are not different for a 2-4-year-old child compared with older children, a value set would be developed based on pooled preferences. This value set will be appropriate for use for 2-4 years old and supplement the existing Australian adolescent value set. If the general population adults' preferences are different between considering a 2–4-year-old versus older children, separate value sets for these two age groups will be developed, with the value set for 2–4-year-old for use in 2-4 years old and the value set for older children to supplement for the existing adolescent value set.

In summary, the study aimed to 1) explore if general population adults have different preferences for health states described for a child 2-4 years old, compared to older children, and 2) develop a CHU9D

value set appropriate for use in children aged 2-4-year-old, 3) develop a CHU9D value set based on general adult population's preferences for older children to supplement the existing value set based on adolescents' preferences in Australia.

7.3. Method

7.3.1. Overview

This study compared preferences from two randomly allocated arms of study participants valuing health states described by the CHU9D instrument for different aged children. For each study arm, a DCE and VAS were used to obtain the latent scale and anchor the values on a 0-1 utility scale. Ethics approval was obtained from the University of Melbourne (reference: 2023-27323-42636-3). This study is reported following the RETRIEVE checklist for Studies Reporting the Elicitation of Stated Preferences for Child Health-Related Quality of Life;[28] see attached RETRIEVE checklist in Appendix 3.

7.3.2. Sample

Data were collected from a sample of the adult general population (including both parents and non-parents) in Australia between September to March, with a target sample size of 2400, recruited by an online survey company Cint. The analysis for this study was conducted using data collected up to November 30th, 2023. Quotas were implemented to guarantee that the recruited sample represents the general population, with respect to gender, age, and region. The sample was randomly allocated to two arms: 1) valuing CHU9D for 2–4-year-olds; 2) valuing CHU9D for older children.

7.3.3. Survey

The survey (Appendix 2) comprised the following elements in order: information sheet and informed consent, screening questions (age, gender and region), reported child health using CHU9D, instructions for the DCE tasks, 13 DCE tasks (including 1 dominant task for logic checking), VAS (for use in anchoring), debrief questions, and background questions (including education, income, employment status, experience with young children, serious illness experience, rating own general health status, Aboriginal or Torres Strait Islander origin, and marital status). Logic check questions (for data quality control) were included (refer to quality control section below for details). Depending on the study arm, CHU9D with guidance notes for children under 5 years old or CHU9D original version for children 7-17 years old (with permission from the University of Sheffield) were used to frame health states.[29] The health description of the two versions of CHU9D were identical except the optional pop-up guidance notes for the younger age group which appear only when participants put the cursor to that dimension for

further information. The survey was set up using the Qualtrics platform and data were collected and stored through Qualtrics.

Two pilot surveys were conducted. First, 10 colleagues or friends were asked to check for comprehension and errors, display of DCE tasks, difficulty understanding or completing the tasks, and to monitor time taken to finish the survey. Slight modifications were made according to the comments from the first pilot survey. Second, the survey was administered to 10% of the target sample size to estimate priors for the DCE design, and the estimated priors were used to generate the DCE design for the full launched survey. After the second pilot survey, the survey was fully launched in the remaining sample.

7.3.4. DCE

Experimental Design

The CHU9D consists of nine dimensions, and each dimension contains five levels, resulting in 5^9=1,953,125 possible health states. Please see Appendix 1 Table S1.1 and Table S1.2 for the CHU9D descriptive system. Including all health states in a valuation study was not feasible from a practical standpoint. Choice sets were chosen to guarantee that the gathered data would facilitate the estimation of a predefined regression model.[30]

The experimental design was a Bayesian efficient design using the following approach. No prior values were available given that the CHU9D with guidance notes for preschool children had not been valued using DCE method.[24] When there are no priors, it is recommended 1) to create an initial design using zero priors; 2) then apply the initial design in a pilot survey with 10% of respondents; 3) estimate parameters; 4) create the Bayesian efficient design using pilot parameter estimates as priors; 5) apply the Bayesian efficient design in the main survey with the remaining 90% participants.[30]

In Bayesian efficient designs, random prior distributions instead of fixed priors were assumed, reducing the loss of efficiency due to incorrect priors. In other words, Bayesian efficient designs are more robust against misspecification of priors.[30]

A modified Fedorov algorithm was employed to create a Bayesian D-efficient design, based on an initial design from a pool of candidates and iteratively refining it to minimize the d-error. The final design was chosen when there was no improvement in the d-error after 2 minutes of additional iterations.[24]

There were nine dimensions in each profile, which represented a high cognitive burden for respondents being asked to simultaneously consider and choose between alternatives. To reduce the cognitive burden

for respondents and to reduce attribute non-attendance, a partial design of DCE was adopted,[31, 32] with four of the nine dimensions having the same severity levels and five dimensions varying in each choice task. The text of the varied levels was presented in bold font, and the dimensions with the same severity levels (overlapped dimensions) were presented with background shading to visually assist respondents in assessing the information presented.

Sample size for DCE

No single optimal method exists for determining the sample size required for a DCE in valuation studies. The sample size for a DCE is determined by three factors: 1) the number of choice pairs needed, 2) the number of observations needed for each choice pair, 3) and the number of choice tasks each respondent can manage without extensive cognitive burden.

There is no fixed rule for the number of choice pairs needed to estimate the models. The practical minimum for DCEs with two alternatives is at least as large as the number of parameters to be estimated.[33, 34] In this study, the linear utility function with only main effects needed to estimate 9*(5-1) = 36 parameters, which resulted in a minimum of 36 choice sets. If interaction effects were to be included, more choice sets would be needed. Finally, 204 choice sets were used according to a previous similar study valuing CHU9D using DCE with duration.[24]

Regarding the number of choice pairs assigned to each respondent, previous similar DCE studies ranged from 10 to 15.[24, 26, 35] In this study, each respondent was asked to complete 12 choice tasks considering the cognitive burden. The DCE design for 204 choice tasks resulted in 17 blocks, with 12 choice pairs in each block. Each respondent was randomly assigned to one block.

There are two recommendations for setting a minimum sample size for the number of observations per pair: Lancsar and Louviere (2008) suggest a minimum of 20 observations per pair [36], and Hensher and colleagues (2005) recommended a minimum of 30 observations per pair.[30] The international valuation protocol for the EQ-5D-Y doubled the average of the two rules and used 50 observations per pair as the minimum,[26] which was followed in this study.

A design comprising 17 blocks would require a minimum of 17*50=850 individual respondents. Previously published literature usually recruited a larger sample than the minimum required given the low marginal cost of including extra online respondents compared with the fixed cost of implementing the survey and to ensure adequate quality sample for analysis.[26] A similar study valuing CHU9D using

DCE with duration used 1200 as the final sample size, ending up with 70 observations per choice pair.[24]

In this study, the target sample size of 1200 per study arm was prespecified to be consistent with the previous similar study and to be above the minimum sample size.

Perspective

As the adult general population doesn't necessarily have children, they were asked to consider a hypothetical child. One study arm asked participants to consider a 2–4-year-old child and the other study arm were asked to consider a 10-year-old child. The reasons of choosing 10-year-old are 1) 10 year old is approximately in the middle of 5-17 age range; 2) to be consistent with previous similar studies (e.g., EQ-5D-Y international protocol[26]); 3) a specific age is easier for people to think than an age range; 4) to maximize respondent consistency in the child age they are thinking of. Respondents were asked to give their own views regarding the health of a hypothetical child (not imagine what the child's views/preference would be). For example, in the 2–4-year-old study arm, the DCE asked "Considering your views about a 2–4-year-old child: which do you prefer?". Sample DCE tasks can be seen in **Appendix 1 Figure S1.**

7.3.5. Anchoring

Anchoring latent preferences obtained from a DCE on a 0-1 scale can be performed using several methods. Duration can be added to a DCE to provide anchoring,[24] or add on time-trade off (TTO) tasks can be conducted with a smaller set of participants for this reason.[26] These were deemed difficult options due to the framing of a young child in the current experiment. Considering the valuation of health for children aged 2-4 years, specifying a duration would cause problems (a child aged 2-4 years old will be an adolescent with a 10-year duration which may cause the health state description to no longer be appropriate). A published paper [27] in European Journal of Health Economics indicated that it is feasible to use VAS to anchor values without specifying the duration as an alternative method.

A VAS was therefore included for the purpose of anchoring. In the VAS task, respondents were asked to rate three health states on a scale of 0-100, where 0 indicated the worst health that they could imagine while 100 indicated the best health that they could imagine. Please see Appendix 1 Figure S2 for the sample VAS task.

7.3.6. Data quality control

Several criteria were created to guarantee the quality of data used for the main analysis. Responses that failed any criteria in categories 1 and 2 were not included in the main analysis.

Category 1: Responses that failed these criteria were deleted (due to being classified as obviously illegitimate responses).

- Total completion time: completion time was recorded by Qualtrics. Responses with less than 1/3 of median completion time were deleted.
- Straight-liner: Responses that always chose A or B were deleted.
- Traffic lights test: This question asked respondents to describe the color of a traffic light picture. This is a test to ensure that the respondent is a real person and attentive respondent. Responses that failed the test were deleted.
- Dominant choice task: those who failed the dominant choice task were deleted.
- Age consistency: Age band multiple choice task was presented at the front of the survey and a free text age question was presented later in the survey. Those whose answers didn't match were deleted.

Category 2: Responses that failed this type of criteria were not included in the main analysis but could be used as sensitivity analysis.

- Completion time for individual choice tasks: From the pilot study, the median time was 12 seconds for each CHU9D choice task. If a participant had more than half of (>=7) the choice tasks with completion time less than 1/3 median (4 seconds), the quality of data for this person was deemed problematic and was not included in the main analysis.
- Respondent engagement: At the end of the survey, respondents were asked their engagement to the survey, with choices of "fully engaged", "partially engaged", and "not engaged". Those reporting "not engaged" were not included in the main analysis.

Category 3: These categories were monitored but responses were included in the main analysis regardless.

- Honesty oath: A previous study indicated that use of an honesty oath improved results.[37] The question in this survey was: “Before we begin, do you promise to answer the following questions truthfully? (you will be allowed to continue with this survey regardless of your answer to this question)” with answers yes or no. All data was included in the main analysis.

7.3.7. Statistical analysis

The demographic characteristics of the sample were summarized for both arms separately and pooled.

Model specification

The choice data were analyzed using the conditional logit model which is consistent with the random utility model of choice.[38] The conditional logit model was chosen as the aim was to estimate the average preference of the general population, consistent with previous similar studies.[24, 39] The mixed logit model is another common type of model for analyzing choice data that can help understanding individual-level heterogeneity, and is used in this study for sensitivity analyses. A linear additive utility function was used. The choice responses were treated as a binary dependent variable (1 and 0 for being chosen or not. Independent variables (dimension levels) were dummy coded, with level 1 for each CHU9D dimension used as a reference level. All standard errors were cluster-robust, allowing for arbitrary correlation between the error terms at the individual level. Equation 1 describes the model specification.

$$V = \sum_{i=1}^{9}(\beta_{i2}X_{i2} + \beta_{i3}X_{i3} + \beta_{i4}X_{i4} + \beta_{i5}X_{i5}) \quad \text{Eq. 1}$$

In equation 1, V represents the choice variable (0 or 1), i represents the 9 attributes of CHU9D, and X_{i2} represents the attribute i and level 2. The coefficient β_{i2} can be interpreted as marginal probabilities compared with the reference level. In other words, it indicates that attribute i at level 2 exhibits a preference difference of β_{i2} relative to attribute i at level 1.

Preference comparison

The preference weights from the DCE data across different samples cannot be compared directly considering that different samples may have different preferences and scales.[40-42] Two approaches were adopted to compare the preferences of the two samples while controlling for scale and preference heterogeneity.

The first approach used relative attribute importance (RAI).[40] Attribute-based normalization was used to obtain the RAIs, with attribute importance calculated as a proportion of the reference attribute importance (Equation 2).

$$RAI_y = \frac{\beta_y}{\beta_x} \quad \text{Eq. 2}$$

RAI_y is the RAI score for attribute Y. Attribute X is the reference attribute, and in this study, it was "annoyed" as it was the least important dimension in both samples. β_y and β_x were the coefficients for level 5 of attribute Y and attribute X, respectively. For example, RAI_worried=β_worried5/β_annoyed5, RAI_sad5=β_sad5/β_annoyed5, with other attributes following the same process.

The second approach was to estimate a pooled model with each variable (dimension level) interacting with the study arm variable (1 = 2–4-year-old arm, 2 =10-year-old arm). A pooled model with interactions can indicate differences in preferences between the samples by dimension levels, rather than dimensions alone in RAI. In this pooled model with interactions, the coefficients on the main parameters reflect the health state preferences for a 2–4-year-old child. The coefficients on the interaction terms represent the change of average health state preference from a 2–4-year-old child to a 10-year-old child.

In addition, Swait-Louviere tests [43] (i.e., poolability test) were also used to test for differences in preference and scale between samples or subgroups (e.g., 2–4-year-old arm vs 10-year-old arm).

Consistent model

To estimate a value set for utility generation, a consistent model is needed. The model should ensure that utility either remains the same or decreases when health or quality of life deteriorates. Coefficients from the latent scale DCE were examined for any logically mis-ordered coefficients. To produce a consistent model, adjacent inconsistent levels was merged or constrained to be equal, which has also been widely used in prior studies.[22, 24, 44]

VAS anchoring

The raw VAS valuation of the worst health state defined by the CHU9D descriptive system was anchored to be on a full health=1, dead=0 scale using the Equation below (Equation 3**)**:

$$\frac{Anchored\_VAS_{555555555} - 0}{1-0} = \frac{Raw\_VAS_{555555555} - Raw\_VAS_{dead}}{Raw\_VAS_{111111111} - Raw\_VAS_{dead}} \quad \text{Eq. 3}$$

The anchored value of the worst health state (555555555) could be below zero if $Raw\_VAS_{555555555}$ was smaller than Raw_VAS_{dead}.

The latent preference values estimated from the conditional logit model based on DCE responses were linearly transformed to a 0-1 scale where 0=dead and 1=full health according to equation below (Equation 4):

$$\tilde{\beta}_k = \left(\frac{1-\widetilde{VAS}_{555555555}}{\sum_{j=1}^{9}\beta_{j5}}\right)\beta_k \quad \text{Eq. 4}$$

where the left sided $\tilde{\beta}_k$ is the rescaled coefficient and the right sided β_k is the original estimates of the conditional logit model. The β_{j5} are the latent scale level 5 coefficients. $\widetilde{VAS}_{555555555}$ is the anchored value for the worst health state obtained from Eq.3.

Only the logical VAS scores were used. The VAS scores were used if $VAS_{111111111} > VAS_{555555555}$ and $VAS_{111111111} > VAS_{dead}$. For more details of the anchoring process, please see reference.[27]

7.3.8. Sensitivity analysis

Sensitivity analyses were conducted by relaxing quality control criteria (quality criteria category 2) to test the robustness of the model. Another type of model for choice data, i.e., mixed logit model, was used as sensitivity analysis to test the robustness of the results.

7.3.9. Comparing Value Sets

The key characteristics and Kernel density distribution of the developed value set were compared with the existing Australian CHU9D adolescent value set.[22]

7.4. Results

7.4.1. The sample characteristics

In total, 2428 responses meeting the pre-specified quality criteria were included in the analysis, with 1176 and 1252 from the 2–4-year-old arm and 10-year-old arm respectively. The sample was generally representative of the general population adults in Australia in terms of age, gender, and region, with more people in 25-44 years age group and less people 55+ years. The sample characteristics between the two arms were generally comparable. The percentage of Aboriginal or Torress Strait Island origin respondents was higher than the Australian average.

Table 7-1 Sample characteristics

Sample characteristics	Study Arm 1 (2-4y) N=1176	Study Arm 2 (10y) N=1252	Pooled data (N=2428)	Australian general population [a,b]
Gender				
Female	628(53.40)	635(50.72)	1263(52.02)	50.7%
Male	546(46.43)	613(48.96)	1159(47.73)	49.3%
Other	2(0.17)	4(0.32)	6(0.25)	
Age groups				
18-24 years	162(13.78)	132(10.54)	294(12.11)	13.0%
25-34 years	337(28.66)	333(26.60)	670(27.59)	18.0%
35-44 years	252(21.43)	308(24.60)	560(23.06)	17.1%
45-54 years	164(13.95)	181(14.46)	345(14.21)	16.0%
55-64 years	187(15.90)	219(17.49)	406(16.72)	15.0%
65+	74(6.29)	79(6.31)	153(6.30)	21.5%
Aboriginal or Torress Strait Island origin				
No	1050(89.29)	1133(90.50)	2183(89.91)	
Yes	126(10.71)	119(9.50)	245(10.09)	3.2%
Region combined				
NSW/ACT	388(32.99)	433(34.58)	821(33.81)	33.6%
VIC/TAS	363(30.87)	351(28.04)	714(29.41)	27.8%
QLD	210(17.86)	263(21.01)	473(19.48)	20.1%
SA/NT	94(7.99)	112(8.95)	206(8.48)	7.9%
WA	121(10.29)	93(7.43)	214(8.81)	10.5%
Employment				
Full time	605(51.45)	637(50.88)	1242(51.15)	57.7%
Part-time	262(22.28)	274(21.88)	536(22.08)	30.4%
Away from work	72(6.12)	88(7.03)	160(6.59)	5.0%
Unemployed	237(20.15)	253(20.21)	490(20.18)	6.9%
Marital status				
Single	364(30.95)	376(30.03)	740(30.48)	35%
Married/Partner	692(58.84)	759(60.62)	1451(59.76)	48.1%
Separated	29(2.47)	28(2.24)	57(2.35)	3.2%
Divorced	61(5.19)	57(4.55)	118(4.86)	8.5%
Widowed	23(1.96)	20(1.60)	43(1.77)	5.2%
Prefer not to say	7(0.60)	12(0.96)	19(0.78)	

Qualification			
None	214(18.20)	228(18.21)	442(18.20)
Trade/Apprenticeship	74(6.29)	74(5.91)	148(6.10)
Certificate/Diploma	360(30.61)	364(29.07)	724(29.82)
Undergraduate degree	369(31.38)	424(33.87)	793(32.66)
Postgraduate degree	145(12.33)	156(12.46)	301(12.40)
Household weekly income before tax			
Less than $300 per week ($15,999 or less per year)	27(2.30)	38(3.04)	65(2.68)
$300-$499 per week ($15,600-$25,999 per year)	80(6.80)	85(6.79)	165(6.80)
$500-$999 per week ($26,000-$51,999 per year)	196(16.67)	216(17.25)	412(16.97)
$1,000-$1,999 per week ($52,000-$103,999 per year)	378(32.14)	379(30.27)	757(31.18)
$2,000-$2,999 per week ($104,000-$155,999 per year)	251(21.34)	267(21.33)	518(21.33)
$3,000-$3,999 per week ($156,000-$207,999 per year)	146(12.41)	150(11.98)	296(12.19)
$4,000 or more per week ($208,000 or more per year)	98(8.33)	117(9.35)	215(8.86)
General health status			
Excellent	152(12.93)	156(12.46)	308(12.69)
Very good	374(31.80)	387(30.91)	761(31.34)
Good	411(34.95)	429(34.27)	840(34.60)
Fair	188(15.99)	224(17.89)	412(16.97)
Poor	51(4.34)	56(4.47)	107(4.41)
Have children 0-18y			
No	797(67.77)	887(70.85)	1684(69.36)
Yes	379(32.23)	365(29.15)	744(30.64)
Have experienced illness			
No	344(29.25)	367(29.31)	711(29.28)
Yes	832(70.75)	885(70.69)	1717(70.72)

Note: a. Age, region, gender Australia data were from 2021 Census.[45] b. Employment and marital data were from the paper Dalziel 2020.[46] The chi2 tests were non-significant between the two arms except the region.

The median time to complete the whole survey was 11.6 minutes, with no difference between the arms. The median time to complete an individual choice task was 13.2 seconds (Appendix 1 Table S2).

Only 5.5% of respondents strongly agreed that "I found the DCE tasks difficult". Only 3.4% of respondents strongly agreed that "I found it difficult to tell the difference between the descriptions". Only 4.7% of respondents strongly agreed that "I found it difficult to imagine the health problems described". See Appendix 1 Figure S3.

7.4.2. Regression analysis results

Preference comparison

In the conditional logit models for both study arms, the coefficients for every dimension level were negative and statistically significant except "Tired" level 2, "Annoyed" level 3 and "Join in activities" level 2. "Annoyed" was the least important dimension in both arms (smallest level 5 coefficient). Most coefficients were logically consistent. Inconsistent estimates were observed for "Sad" level 2, "Annoyed" level 3, "Daily routine" level 3 in the 2–4-year-old arm, and "Sad" level 2, "Tired" level 3, "Annoyed" level 3, and "Daily routine" level 3 in the 10-year-old arm.

Table 7-2 Discrete Choice Modelling Estimation Results by Study Arm

Mean of Coefficients	2-4y arm	10y arm
	Conditional logit	Conditional logit
Worried level 2	-0.157**	-0.310***
Worried level 3	-0.243***	-0.396***
Worried level 4	-0.579***	-0.646***
Worried level 5	-0.881***	-0.919***
Sad level 2	-0.327***	-0.270***
Sad level 3	*-0.244***	*-0.184***
Sad level 4	-0.615***	-0.622***
Sad level 5	-1.052***	-0.953***
Pain level 2	-0.629***	-0.506***
Pain level 3	-0.691***	-0.688***
Pain level 4	-1.785***	-1.682***
Pain level 5	-1.695***	-1.453***
Tired level 2	**-0.096**	-0.115*

Tired level 3	-0.139**	*-0.103**
Tired level 4	-0.311***	-0.302***
Tired level 5	-0.516***	-0.519***
Annoyed level 2	-0.134**	-0.144**
Annoyed level 3	**-0.110**	**-0.074**
Annoyed level 4	-0.159**	-0.242***
Annoyed level 5	-0.512***	-0.518***
Schoolwork level 2	-0.153**	-0.150**
Schoolwork level 3	-0.198***	-0.264***
Schoolwork level 4	-0.684***	-0.740***
Schoolwork level 5	-0.719***	-0.780***
Sleep level 2	-0.288***	-0.259***
Sleep level 3	-0.295***	-0.343***
Sleep level 4	-0.788***	-0.785***
Sleep level 5	-1.062***	-1.022***
Daily routine level 2	-0.354***	-0.394***
Daily routine level 3	*-0.316***	*-0.343***
Daily routine level 4	-0.835***	-0.780***
Daily routine level 5	-0.952***	-1.088***
Join in activities level 2	**0.054**	**0.104**
Join in activities level 3	-0.233***	-0.148**
Join in activities level 4	-0.240***	-0.162**
Join in activities level 5	-0.654***	-0.638***

Note: Bold indicates insignificant or inconsistent estimates.

The relative attribute importance scores were normalized using the least important dimension, which was "Annoyed". The RAI scores can be interpreted as the relative importance compared to "Annoyed". For example, the RAI score of 3.31 for "Pain" in the 2–4-year-old arm can be interpreted as "Pain" being more than 3 times as important as "Annoyed" on average. The difference in RAIs between the two arms were not statistically significant for any dimensions, indicating no difference in preferences for relative importance of dimensions between the two arms.

Table 7-3 Relative Attribute Importance Scores by study arm and RAI differences with 95% Confidence Intervals

	2–4-year-old		10-year-old		RAI difference (95% CI)	*P*-value
	RAI	SE	RAI	SE		
Worried	1.72	0.21	1.77	0.21	-0.05(-0.63,0.53)	0.873
Sad	2.06	0.23	1.84	0.22	0.22(-0.41,0.85)	0.506
Pain	3.31	0.36	2.80	0.30	0.51(-0.41,1.43)	0.282
Tired	1.01	0.14	1.00	0.13	0.01(-0.37,0.38)	0.971
Annoyed	1.00		1.00			
Schoolwork	1.41	0.17	1.50	0.18	-0.10(-0.59,0.39)	0.707
Sleep	2.08	0.24	1.97	0.22	0.10(-0.53,0.74)	0.761
Daily routine	1.86	0.21	2.10	0.23	-0.24(-0.84,0.37)	0.453
Join in activities	1.28	0.18	1.23	0.17	0.05(-0.44,0.53)	0.860

Note: RAI scores were calculated based on the conditional logit models from Table 2. An attribute-based normalization was applied using the least important attribute (here is annoyed) and a scaling factor of 1. Standard errors were calculated using the Delta method.

The pooled model with interaction terms by study arm (2–4-year-old vs 10-year-old) was used to explore differences in preferences by dimension levels instead of dimensions alone. Figure 7-1 presents the coefficients of the interaction terms and the 95% confidence intervals. More detailed results are presented in the supplementary material (Appendix 1 Table S3). Out of 36 interaction terms, only one coefficient was statistically significant (*P=0.028*) which was “Pain” level 5, indicating respondents rated that the 10-year-old framing had a smaller magnitude utility decrease than the 2–4-year-old framing for “pain” level 5. The other 35 interaction terms had non-significant differences, indicating that the two arms were generally the same in preference weights in dimension levels.

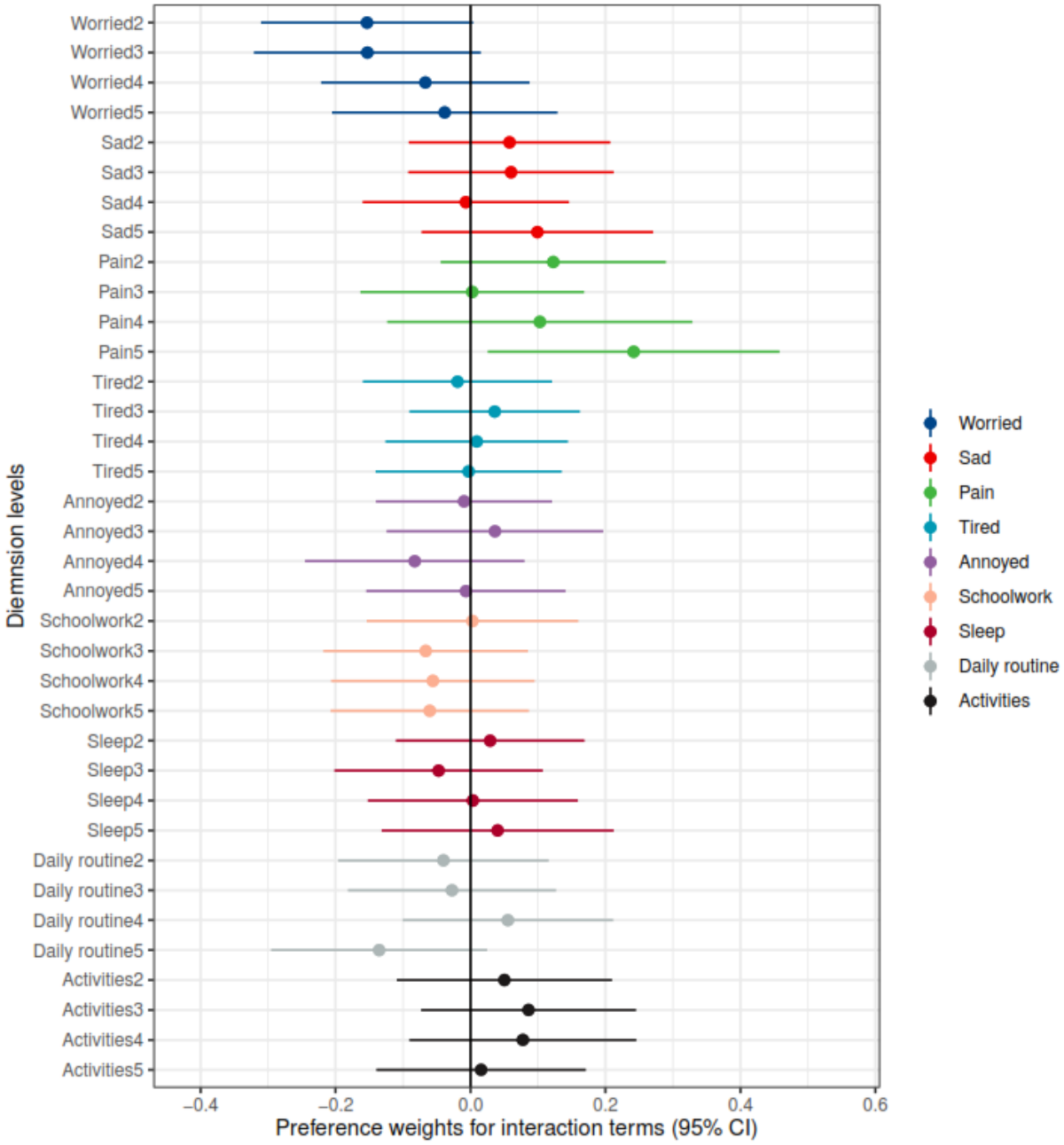


Figure 7-1 Mean preference weights for interaction terms by study arm from the pooled model and associated 95% confidence intervals.

The Swait-Louviere test (poolability tests) results indicated no difference in preference or scale by study arm (Appendix 1 Table S4). Adults' stated preferences for HRQoL in 2–4-year-old children can therefore be pooled with preferences for 10-year-old children.

In summary, the relative attribute importance scores, pooled model with interaction terms, and poolability tests all indicated that there were no differences in adults' stated preferences for HRQoL in a 2–4-year-old child and preferences for a 10-year-old child, except for pain level 5 in the pooled model. However, given this was only 1 out of 36 interaction terms and the non-significance of the other test results, the data of the two arms were pooled together for the estimates of the value sets for CHU9D based on preferences of general population adults' preferences.

Estimate the value set

Model 1 was the original conditional logit model using pooled data, with no constraints or weights. In model 1, all coefficients were significant, and the majority of coefficients (27 out of 36) had the expected sign and were logically consistent. The coefficient for "Join in activities" level 2 had the wrong sign (with level 1 as the reference), but the magnitude of the coefficient was very small. The monotonicity also failed for "Sad" level 2 and 3, "Pain" level 4 and 5, "Annoyed" level 2 and 3, "Daily routine" level 2 and 3. These suggested that for these dimension levels, respondents had little distinction between them in terms of utility decreasing.

The consistent model (Model 2) merges "Join in activities" level 1 and 2 as the reference level. It then applies constraints to make the inconsistent adjacent coefficients in each dimension equal. For example, the coefficients of "Sad" level 2 and 3 were equal in the consistent model.

Model 3 further applied sample weights on the base of the consistent model to account for the relatively high percentages of Aboriginal or Torress Strait Island origin respondents, and 25-44 years old to guarantee representative Australian population estimates. Model 3 was still consistent and with all coefficients negative and significant.

Table 7-4 Pooled model and consistent model

	Model 1	Model 2	Model 3
	clogit model using pooled data	Consistent model	Consistent model with sample weights
Worried level 2	-0.237***	-0.221***	-0.223***
Worried level 3	-0.325***	-0.315***	-0.331***
Worried level 4	-0.612***	-0.616***	-0.657***

Worried level 5	-0.902***	-0.897***	-0.951***
Sad level 2	**-0.297*** **	*-0.235****	-0.271***
Sad level 3	**-0.211*** **	*-0.235****	-0.271***
Sad level 4	-0.617***	-0.601***	-0.666***
Sad level 5	-0.999***	-0.957***	-1.077***
Pain level 2	-0.565***	-0.542***	-0.600***
Pain level 3	-0.688***	-0.666***	-0.723***
Pain level 4	**-1.729*** **	*-1.637****	-1.798***
Pain level 5	**-1.565*** **	*-1.637****	-1.798***
Tired level 2	-0.105**	-0.106**	-0.121**
Tired level 3	-0.121***	-0.107***	-0.126***
Tired level 4	-0.306***	-0.301***	-0.338***
Tired level 5	-0.516***	-0.520***	-0.566***
Annoyed level 2	**-0.139*** **	*-0.131****	-0.109***
Annoyed level 3	**-0.092***	*-0.131****	-0.109***
Annoyed level 4	-0.201***	-0.220***	-0.197***
Annoyed level 5	-0.515***	-0.512***	-0.546***
Schoolwork level 2	-0.149***	-0.125**	-0.139***
Schoolwork level 3	-0.231***	-0.213***	-0.229***
Schoolwork level 4	-0.713***	-0.700***	-0.739***
Schoolwork level 5	-0.751***	-0.749***	-0.802***
Sleep level 2	-0.271***	-0.306***	-0.296***
Sleep level 3	-0.320***	-0.332***	-0.341***
Sleep level 4	-0.785***	-0.800***	-0.845***
Sleep level 5	-1.038***	-1.046***	-1.091***
Daily routine level 2	**-0.374*** **	*-0.355****	-0.365***
Daily routine level 3	**-0.328*** **	*-0.355****	-0.365***
Daily routine level 4	-0.806***	-0.820***	-0.871***
Daily routine level 5	-1.022***	-1.017***	-1.057***
Join in activities level 2	**0.080***		
Join in activities level 3	-0.189***	-0.219***	-0.249***
Join in activities level 4	-0.198***	-0.236***	-0.277***
Join in activities level 5	-0.645***	-0.681***	-0.753***

Note: Bold = inconsistency in coefficient values. *Italic* = consistent estimates after applying constraints and merging levels. Significance levels: * p<0.05; ** p<0.01; *** p<0.001.

The final value sets were generated using estimates from the consistent model with sample weights (i.e., Model 3 in Table 7-4). The raw estimates for dimension levels from Model 3 were anchored onto the

scale of 0-1 through VAS scores from the same sample. The mean PITS health state value (reflecting the lowest or worst level of health state for all nine CHU9D dimensions, here the anchored_$VAS_{555555555}$) was 0.28. In Table 7-5, the column "Coefficients" are the model coefficients from the consistent model with sample weights, and the column "Utility decrements" are the anchored values of those coefficients using VAS. The utility for a health state is 1 plus the sum of the utility decrements for each corresponding severity level for each dimension. For example, health state 214511111 is generated using 1+(-0.020+0-0.160-0.050+0+0+0+0+0)=0.77, with utility decrements shown in Table 7-5.

Table 7-5 Value sets from Australia general population adults for children for CHU9D (consistent model applying population weights and after anchoring using VAS)

Dimension	Level	Coefficients	Utility Decrements
Worried	1		0
	2	-0.223	-0.020
	3	-0.331	-0.030
	4	-0.657	-0.059
	5	-0.951	-0.085
Sad	1		0
	2	-0.271	-0.024
	3	-0.271	-0.024
	4	-0.666	-0.059
	5	-1.077	-0.096
Pain	1		0
	2	-0.600	-0.054
	3	-0.723	-0.065
	4	-1.798	-0.160
	5	-1.798	-0.160
Tired	1		0
	2	-0.121	-0.011
	3	-0.126	-0.011
	4	-0.338	-0.030
	5	-0.566	-0.050
Annoyed	1		0
	2	-0.109	-0.010
	3	-0.109	-0.010
	4	-0.197	-0.018
	5	-0.546	-0.049

School work/homework	1		0
	2	-0.139	-0.012
	3	-0.229	-0.020
	4	-0.739	-0.066
	5	-0.802	-0.072
Sleep	1		0
	2	-0.296	-0.026
	3	-0.341	-0.030
	4	-0.845	-0.075
	5	-1.091	-0.097
Daily routine	1		0
	2	-0.365	-0.033
	3	-0.365	-0.033
	4	-0.871	-0.078
	5	-1.057	-0.094
Able to join in activities	1		0
	2		0
	3	-0.249	-0.022
	4	-0.277	-0.025
	5	-0.753	-0.067

Note: The coefficients used for final anchoring were from Model 3 in Table 4. The severest health state has a utility of 0.23.

7.4.3. Sensitivity analysis

The results were robust when relaxing the quality control criteria with similar coefficients. The main model has the lowest AIC and BIC among all models, indicating a better fit (Appendix 1 Table S7). The coefficients were also similar to those estimated using mixed logit model; the AIC and BIC are also similar (Appendix 1 Table S8).

7.4.4. Comparison with the existing Australian value set

The two different value sets have different utility ranges. The value set produced in the current study using adult preferences has a narrower utility range than the previously published value set using adolescent preferences.[22] See kernel density plots comparing two value sets in Appendix 1 Figure S5.

The adult preference values generated from this study are generally higher than the existing published Australian adolescent preference values for the same health states except for the pain dimension when holding other dimensions at best level. The difference in values between the two value sets were smallest at mild health states and were larger on moderate and worst health states (Appendix 1 Figure S6).

7.5. Discussion

7.5.1. Summary of findings

This study compared the preferences for health states described by the CHU9D when asked to consider 2–4-year-olds and 10-year-olds in the adult general population in Australia. It found that almost all analyses and tests showed no difference aside from one of the 36 interaction effects in the pooled analysis (pain level 5). Consequently, it was concluded that the health states only exhibited negligible difference and therefore considered poolable. On this basis a value set (scoring algorithm) for the CHU9D was developed using a representative sample of the adult general population in Australia, which is appropriate for use for both 2–4-year-old and older children.

7.5.2. Comparison with previous studies investigating different child age framing

This study found that the general population adults have no appreciable difference in preferences for health states for 2–4-year-old compared with a 10-year-old. There are a few previous studies exploring the impact of different age framing in valuation tasks and evidence is mixed. Most prior research found no impact of different age framing although most studies only included ages above 5 years. Ramos etc. 2022 found that the impact is minimal on DCE modeled latent scale values when using different age framing (5-7 years old, 8-10 years old, 11-13 years old, and 14-15 years old) of a hypothetical child in valuing EQ-5D-Y-3L in the United Kingdom and United States using a representative sample with 1000 adults in each country.[47] Retra etc. 2020 explored the influence of different age descriptions (child age 4, 10, 16 years) on scores of EQ-5D-Y health states by using the EQ-5D VAS scale in a convenience sample of university students in the Netherlands, and found that except for one moderate and one severe health state, other EQ-5D-Y health states were not valued significantly differently when description of age

differed.[48] Craig etc. 2016 explored the difference by age (7 and 10 years old) in valuing EQ-5D-Y using paired comparisons and found that the overall differences by age were not statistically significant and the differences between age 7 and 10 were judged to be minor.[49, 50] The above quantitative studies found no difference of age framing for valuing child health. One qualitative study found that age matters in valuation tasks. Reckers-Droog etc. 2022 found that a 10 year old may not represent a 15 year old adolescent as evidenced in a think aloud study.[51] However, it might be that people think there is a difference qualitatively but that this is not reflected in their quantitative valuations. In addition, there may indeed be differences in values between age groups, but instruments such as the CHU9D or EQ-5D-Y might be too general and brief to capture these differences.

7.5.3. Comparison with previous studies valuing the CHU9D

Valuation methodology comparison

Table S5 in Appendix 1 compared the key characteristics of the current valuation study with previous studies valuing CHU9D, describing the various preference elicitation techniques, anchoring methods if used, source of preferences, perspective, preference ranking of dimensions, and values for some example health states. Only the Australia adolescent value sets and the Chinese adolescent values employed the same methods.

The relative importance of dimensions

This study found that pain was the most important dimension, followed by sleep, daily routine, sad, worried, and schoolwork/homework, while tired and annoyed were the least important dimensions. Previous studies valuing CHU9D in UK and Netherlands also found that pain was the most important dimension, with some differences in the order of importance of the other dimensions.[20, 24] This may be because the current study, the UK and the Netherland value sets were all based on adult general population's preferences. On the contrary, the value sets developed based on Australia adolescents' preferences and the value set developed using Chinese adolescent preferences had "sad" and "activities" as the most important dimensions respectively.[23] Dalziel 2020 also found that Australian adults place less weight on being worried, sad or unhappy and more weight on having pain or discomfort than Australian adolescents when valuing EQ-5D-Y.[46] These differences suggest the potential impact of who's preferences are elicited (adults vs adolescents).

Compare utilities from different value sets

The utilities generated using Australian adult preferences in this current study were higher than the utilities generated using the Australian adolescent preferences, suggesting that adults were less concerned than adolescents for the same worse health states. This is consistent with previous studies comparing child/adolescent own values with adult/parent values for children, with the majority (four out of five) of studies reporting that children/adolescents provided lower values than those provided by adults/parents valuing the same child health states.[25] The utilities generated in this study were closest to the utilities generated using the UK general population. The utilities in this study are higher than the utilities generated using the Dutch value sets. One potential reason might be due to the different perspective as there is evidence that adults taking the perspective of a child (this current study) usually have higher values than adults taking the perspective of themselves (Dutch value set).[52] A second potential reason might be due to the anchoring method as the Dutch value set used DCE with duration and this current study used VAS.

Compare the values for worst health state- anchoring

The rescaled value of the worst health state (555555555) in this study (0.23) falls between estimates from previous CHU9D valuation studies, where the UK value sets is 0.34 and the Australian adolescent value sets is -0.1059.[20, 22] The UK value sets used the standard gamble (SG) approach in the general adult population with all ages (age range: 16-87 years; n=300) in the UK to generate the values for health states including the worst health state [20] and required no further anchoring. The Australian adolescent value sets used values generated from a separate time-trade off (TTO) study in young adults (a convenience sample of Flinders University undergraduate students aged 18–29 years in Australia; n=38) to rescale or anchor the latent values.[53] The rescaled value for the worst health state in this study is closest to that from the UK value sets. This may be because this study used a population of similar age ranges with the UK study for anchoring, which differed from the population age in the Australian anchoring study. The Dutch value set developed by Rowen et al. 2018 using DCE with duration had a rescaled value of -0.568 for the worst health state,[24] which is substantially lower than this study. The differences are likely due to different anchoring methods (DCE with duration vs VAS). Future studies exploring the impact of different anchoring methods in valuing child HRQoL are valuable.

Inconsistent estimates

The first estimated model in this current study generated mostly ordered logical results, with 5 dimensions sad, pain, annoyed, daily routine, join in activities having some logical inconsistencies between adjacent levels. For example, sad level 2 (a little bit sad) had a utility decrease of 0.297, while

sad level 3 (a bit sad) had a utility decrease of 0.211 compared with level 1 (not sad). Pain level 4 (quite a lot of) had a utility decrement of 1.729 while pain level 5 (a lot of) had a utility decrement of 1.565, compared with level 1 (not any pain). Join in activities level 2 (can join in with most activities) had a significantly positive estimate compared with level 1 (can join in with any activities), indicating a utility increase. These inconsistent responses suggest on average the sample didn't distinguish or have a clear preference trend between these adjacent levels or variance in responses. Some comments from the pilot survey echoed these findings. For example, one colleague commented that she had difficulty telling the difference of some levels and needed to refer to the level definitions. Although CHU9D has been validated, the inconsistency problem may relate to the format of how these adjacent levels were presented to respondents. The presentation of individual response levels in a DCE choice task lacked the context of the original level descriptors from the CHU9D questionnaire, which typically presents a clear trend from level 1 to level 5. Consequently, participants may encounter difficulty distinguishing between similar adjacent levels when these levels are presented independently. This seems to be more of a problem for instruments with subtle differences between levels or with more levels described, such as CHU9D and EQ-5D-Y-5L compared with EQ-5D-Y-3L. Previous studies valuing CHU9D also find some similar inconsistencies.[24] For example, Rowen et al. 2018 found inconsistency in sad level 2 and 3, and annoyed level 2 and 3 [24]. Also, Ratcliffe et al. 2016 found inconsistency in join in activities level 1 and 2.[22]

The current study has fewer dimension levels with inconsistency (9 versus 23) and thus fewer merged adjacent levels than the UK valuation study which used a SG approach where many levels have the same utility decrement. This current study has similar number of inconsistencies with the Dutch valuation study (9 versus 10) which used a DCE with duration approach and the Australian adolescent valuation study (9 versus 8) using a best-worst scaling (BWS) approach. These findings suggest a potential advantage of using ordinal approach (DCE, BWS) compared with traditional SG, however further research is necessary.

7.5.4. Strength and limitations

This study has several strengths. It used a large generally representative sample (around 2400) of Australia general adult population to generate preferences, larger than previous similar studies (see Appendix 1 Table S5). It used extensive pre-specified data quality control criteria to ensure only legitimate data are included in the study and therefore the main analysis. Another strength is that this study applied a partial design with 4 overlapped dimensions which has greatly improved its feasibility compared to a previous similar study by Rowen 2018 which used DCE with duration and a partial design

with 3 overlapped dimensions. In the current study, only 5.5% strongly agree that they found the DCE tasks difficult and only 3.4% strongly agree that they found it difficult to tell the difference between the descriptions (Appendix 1 Figure S3). This is quite low compared with Rowen 2018 using DCE with duration which had nearly 13% of respondents reporting it being very difficult to choose between the different health descriptions, although the Rowen study also had the added complexity of duration.

This study has several limitations. First, despite best efforts to recruit a nationally representative sample, the final sample still had higher percentages of young adults (age: 25-34 years) and people of Aboriginal or Torress Strait Island origin compared to the Australian population. Sample weights were therefore applied to generate nationally representative estimates. Sample weights adjust for the fact that some groups might be overrepresented or underrepresented in the sample relative to the general population. By applying these weights, the analysis can more accurately reflect the characteristics and proportions of the broader population, thus improving the generalizability of the findings. Secondly, it might be difficult to ask adults to value health from the perspective of a child if they have limited experience of caring for children. Some studies chose to ask adults to value health from the perspective of themselves. However, for the current study it wasn't deemed appropriate to ask adults to think of themselves experiencing some very child-specific health states described by the CHU9D with guidance notes for under 5 years, such as "Daily Routine (eating, drinking, toileting, washing and teeth cleaning, as appropriate for their age)". This study has collected preferences from parents and people with illness experience, and future analyzes exploring the impact of having children or having experience caring for children on valuation is guaranteed. Thirdly, there exist some logical inconsistencies in the estimates, as has also been observed in previous similar studies. The logical inconsistency suggests that people have difficulty distinguishing between these adjacent levels. Another limitation is the choice of 10-year-old children to represent 5-17 years old children in the framing of the valuation task. It is acknowledged that the choice of 10-year-old is arbitrary, although it followed published international protocol.[26] Finally, this study only used VAS as the anchoring method due to time and budget constraint.

7.5.5. Implications for policy and research

This is the first study developing a value set for a common generic HRQoL instrument appropriate for use for 2-4 year olds. The developed value set makes it possible to measure utilities for 2–4-year-old children, which contributes to the accuracy of calculating QALYs in economic evaluations including young children. In addition, this study highlights that consistent values across childhood from 2-18 years can be generated using the same scoring. This value set, developed based on the general adult population in Australia also complements the existing CHU9D value set developed based on adolescent preferences,

allowing policy makers to include different sources of preferences for different uses. This study also provides important insights regarding the feasibility of using DCEs to value HRQoL states described by instruments for children under 5 years of age and for use of the VAS to anchor preferences. There are also implications for future research. To better understand the impact of anchoring methods on values, future research should explore various techniques or incorporate multiple anchoring methods to aid comparisons. Additionally, considering the observed logical consistencies in the estimates, qualitative studies investigating the reasons and identifying situations where fewer levels may be more appropriate would be valuable. Moreover, studies employing alternative valuation methods could provide further insights on the impact of different methods on utilities generated.

7.6. Conclusion

This research demonstrated there is no appreciable difference in the general population adults' preferences for health states described by CHU9D when considering a 2–4-year-old child versus a 10-year-old child. A value set was developed that is considered appropriate for use to calculate QALYs for 2-4 years old children as well as older children. This enables economic evaluation to include this young age group and provides reassurance about consistency in scoring from 2 to 18 years. The developed value set supplements the existing Australian value set derived from adolescent preference, providing decision makers greater choices on which perspective can be used for different decision contexts.

Funding: This study was supported by Kim Dalziel's NHMRC Investigator Award (GNT1198047), which supported data collection and analysis.

Acknowledgement: Thanks go to the QUOKKA study researchers. Valuable collegial support in child health valuation was gained from the affiliation with QUOKKA. Special thanks to Yan Meng who has shared the instruction for importing DCE tasks to Qualtrics. Many thanks to colleagues for piloting the survey, providing feedback, and providing instructions and advice for experimental design, and data analysis.

7.7. Reference

[1] G. L. Freed, S. Gafforini, and N. Carson, "Age distribution of emergency department presentations in Victoria," (in eng), *Emerg Med Australas,* vol. 27, no. 2, pp. 102-7, Apr 2015.

[2] M. Shaker, E. S. Chan, J. L. P. Protudjer, L. Soller, E. M. Abrams, and M. Greenhawt, "The Cost-Effectiveness of Preschool Peanut Oral Immunotherapy in the Real-World Setting," *The Journal of Allergy and Clinical Immunology: In Practice,* vol. 9, no. 7, pp. 2876-2884.e4, 2021/07/01/ 2021.

[3] L. Wang *et al.*, "The cost-effectiveness of alternative vision screening models among preschool children in rural China," *Acta Ophthalmologica,* https://doi.org/10.1111/aos.13954 vol. 97, no. 3, pp. e419-e425, 2019/05/01 2019.

[4] M. Tanaka, R. Okubo, S.-L. Hoshi, N. Ishikawa, and M. Kondo, "Cost-effectiveness of pertussis booster vaccination for preschool children in Japan," *Vaccine,* vol. 40, no. 7, pp. 1010-1018, 2022/02/11/ 2022.

[5] S. M. Sullivan, K. Tsiplova, and W. J. Ungar, "A scoping review of pediatric economic evaluation 1980-2014: do trends over time reflect changing priorities in evaluation methods and childhood disease?," *Expert Review of Pharmacoeconomics & Outcomes Research,* vol. 16, no. 5, pp. 599-607, 2016/09/02 2016.

[6] Pediatric Economic Database Evaluation (PEDE). *Trends in Economic Evaluation*. Available: http://pede.ccb.sickkids.ca/pede/trends.jsp

[7] S. K. Kromm *et al.*, "Characteristics and quality of pediatric cost-utility analyses," *Quality of Life Research,* vol. 21, no. 8, pp. 1315-1325, 2012/10/01 2012.

[8] D. Rowen, O. Rivero-Arias, N. Devlin, and J. Ratcliffe, "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?," (in eng), *Pharmacoeconomics,* vol. 38, no. 4, pp. 325-340, Apr 2020.

[9] J. Kwon *et al.*, "Systematic Review of Conceptual, Age, Measurement and Valuation Considerations for Generic Multidimensional Childhood Patient-Reported Outcome Measures," *PharmacoEconomics,* vol. 40, no. 4, pp. 379-431, 2022/04/01 2022.

[10] X. Xiong *et al.*, "Psychometric Properties of Child Health Utility 9D (CHU9D) Proxy Version Administered to Parents and Caregivers of Children Aged 2-4 Years Compared with Pediatric Quality of Life Inventory™ (PedsQL)," (in eng), *Pharmacoeconomics,* Jan 27 2024.

[11] A. van Heusden *et al.*, "Psychometric Performance Comparison of the Adapted versus Original Versions of the EQ-5D-Y-3L and -Y-5L in Proxy Respondents for 2- to 4-Year-Olds," *PharmacoEconomics,* 2024/01/18 2024.

[12] S. Kreimeier *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,* vol. 21, no. 11, pp. 1291-1298, 2018/11/01/ 2018.

[13] C. Bailey, K. Dalziel, P. Cronin, N. Devlin, and R. Viney, "How are Child-Specific Utility Instruments Used in Decision Making in Australia? A Review of Pharmaceutical Benefits Advisory Committee Public Summary Documents," (in eng), *Pharmacoeconomics,* vol. 40, no. 2, pp. 157-182, Feb 2022.

[14] K. J. Stevens, "Working with children to develop dimensions for a preference-based, generic, pediatric, health-related quality-of-life measure," (in eng), *Qual Health Res,* vol. 20, no. 3, pp. 340-51, Mar 2010.

[15] E. J. Frew, M. Pallan, E. Lancashire, K. Hemming, and P. Adab, "Is utility-based quality of life associated with overweight in children? Evidence from the UK WAVES randomised controlled study," *BMC Pediatrics,* Article vol. 15, no. 1, 2015, Art. no. 211.

[16] A. G. Canaway and E. J. Frew, "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," (in eng), *Quality of Life Research,* vol. 22, no. 1, pp. 173-83, Feb 2013.

[17] K. Stevens and J. Ratcliffe, "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," (in eng), *Value Health,* vol. 15, no. 8, pp. 1092-9, Dec 2012.

[18] K. Stevens, "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,* vol. 9, no. 3, pp. 157-169, 2011/05/01 2011.

[19] X. Xiong *et al.*, "Psychometric Properties of Child Health Utility 9D (CHU9D) Proxy Version Administered to Parents and Caregivers of Children Aged 2–4 Years Compared with Pediatric Quality of Life Inventory™ (PedsQL)," *PharmacoEconomics,* 2024/01/27 2024.

[20] K. Stevens, "Valuation of the Child Health Utility 9D Index," (in eng), *Pharmacoeconomics,* vol. 30, no. 8, pp. 729-47, Aug 1 2012.

[21] J. Ratcliffe, E. Huynh, K. Stevens, J. Brazier, M. Sawyer, and T. Flynn, "Nothing About Us Without Us? A Comparison of Adolescent and Adult Health-State Values for the Child Health Utility-9D Using Profile Case Best-Worst Scaling," (in eng), *Health Econ,* vol. 25, no. 4, pp. 486-96, Apr 2016.

[22] J. Ratcliffe *et al.*, "Valuing the Child Health Utility 9D: Using profile case best worst scaling methods to develop a new adolescent specific scoring algorithm," (in eng), *Soc Sci Med,* vol. 157, pp. 48-59, May 2016.

[23] G. Chen, F. Xu, E. Huynh, Z. Wang, K. Stevens, and J. Ratcliffe, "Scoring the Child Health Utility 9D instrument: estimation of a Chinese child and adolescent-specific tariff," (in eng), *Qual Life Res,* vol. 28, no. 1, pp. 163-176, Jan 2019.

[24] D. Rowen, B. Mulhern, K. Stevens, and J. H. Vermaire, "Estimating a Dutch Value Set for the Pediatric Preference-Based CHU9D Using a Discrete Choice Experiment with Duration," (in eng), *Value Health,* vol. 21, no. 10, pp. 1234-1242, Oct 2018.

[25] C. Bailey *et al.*, "Preference Elicitation Techniques Used in Valuing Children's Health-Related Quality-of-Life: A Systematic Review," *PharmacoEconomics,* vol. 40, no. 7, pp. 663-698, 2022/07/01 2022.

[26] J. M. Ramos-Goñi *et al.*, "International Valuation Protocol for the EQ-5D-Y-3L," *PharmacoEconomics,* pp. 1-11, 2020.

[27] E. J. D. Webb, J. O'Dwyer, D. Meads, P. Kind, and P. Wright, "Transforming discrete choice experiment latent scale values for EQ-5D-3L using the visual analogue scale," (in eng), *Eur J Health Econ,* vol. 21, no. 5, pp. 787-800, Jul 2020.

[28] C. Bailey *et al.*, "The RETRIEVE Checklist for Studies Reporting the Elicitation of Stated Preferences for Child Health-Related Quality of Life," *PharmacoEconomics,* 2024/01/13 2024.

[29] *Measuring & Valuing Health. A brief overview of the Child Health Utility 9D (CHU9D).* Available: https://licensing.sheffield.ac.uk/product/CHU-9D

[30] D. A. Hensher, J. M. Rose, J. M. Rose, and W. H. Greene, *Applied choice analysis: a primer*. Cambridge university press, 2005.

[31] R. Kessels, B. Jones, and P. Goos, "Bayesian optimal designs for discrete choice experiments with partial profiles," *Journal of Choice Modelling,* vol. 4, no. 3, pp. 52-74, 2011/01/01/ 2011.

[32] M. F. Jonker, B. Donkers, E. de Bekker-Grob, and E. A. Stolk, "Attribute level overlap (and color coding) can reduce task complexity, improve choice consistency, and decrease the dropout rate in discrete choice experiments," *Health Economics,* https://doi.org/10.1002/hec.3846 vol. 28, no. 3, pp. 350-363, 2019/03/01 2019.

[33] E. W. de Bekker-Grob, B. Donkers, M. F. Jonker, and E. A. Stolk, "Sample Size Requirements for Discrete-Choice Experiments in Healthcare: a Practical Guide," (in eng), *Patient,* vol. 8, no. 5, pp. 373-84, Oct 2015.

[34] B. K. J. Orme, "Getting started with conjoint analysis: strategies for product design and pricing research," 2006.

[35] N. Bansback, A. R. Hole, B. Mulhern, and A. Tsuchiya, "Testing a discrete choice experiment including duration to value health states for large descriptive systems: Addressing design and sampling issues," *Social Science & Medicine,* vol. 114, pp. 38-48, 2014/08/01/ 2014.

[36] E. Lancsar and J. J. P. Louviere, "Conducting discrete choice experiments to inform healthcare decision making: a user's guide," vol. 26, pp. 661-677, 2008.

[37] N. Jacquemet, A. James, S. Luchini, J. F. J. E. Shogren, and r. economics, "Referenda under oath," vol. 67, pp. 479-504, 2017.

[38] A. B. Hauber *et al.*, "Statistical Methods for the Analysis of Discrete Choice Experiments: A Report of the ISPOR Conjoint Analysis Good Research Practices Task Force," *Value in Health,* vol. 19, no. 4, pp. 300-315, 2016/06/01/ 2016.

[39] J. Ratcliffe, T. Flynn, F. Terlich, K. Stevens, J. Brazier, and M. Sawyer, "Developing Adolescent-Specific Health State Values for Economic Evaluation," *PharmacoEconomics,* vol. 30, no. 8, pp. 713-727, 2012/08/01 2012.

[40] J. M. Gonzalez, "A Guide to Measuring and Interpreting Attribute Importance," *The Patient - Patient-Centered Outcomes Research,* vol. 12, no. 3, pp. 287-295, 2019/06/01 2019.

[41] C. M. Vass, S. Wright, M. Burton, and K. Payne, "Scale Heterogeneity in Healthcare Discrete Choice Experiments: A Primer," (in eng), *Patient,* vol. 11, no. 2, pp. 167-173, Apr 2018.

[42] S. J. Wright, C. M. Vass, G. Sim, M. Burton, D. G. Fiebig, and K. Payne, "Accounting for Scale Heterogeneity in Healthcare-Related Discrete Choice Experiments when Comparing Stated Preferences: A Systematic Review," (in eng), *Patient,* vol. 11, no. 5, pp. 475-488, Oct 2018.

[43] J. Swait and J. Louviere, "The Role of the Scale Parameter in the Estimation and Comparison of Multinomial Logit Models," *Journal of Marketing Research,* vol. 30, no. 3, pp. 305-314, 1993/08/01 1993.

[44] J. E. Brazier and J. Roberts, "The estimation of a preference-based measure of health from the SF-12," (in eng), *Med Care,* vol. 42, no. 9, pp. 851-9, Sep 2004.

[45] A. B. o. Statistics. (Jan 2nd.). *Australian 2021 Census data*.

[46] K. Dalziel, M. Catchpool, B. Garcia-Lorenzo, I. Gorostiza, R. Norman, and O. Rivero-Arias, "Feasibility, Validity and Differences in Adolescent and Adult EQ-5D-Y Health State Valuation in Australia and Spain: An Application of Best-Worst Scaling," (in eng), *Pharmacoeconomics,* Jan 24 2020.

[47] J. M. Ramos-Goñi *et al.*, "Does Changing the Age of a Child to be Considered in 3-Level Version of EQ-5D-Y Discrete Choice Experiment–Based Valuation Studies Affect Health Preferences?," *Value in Health,* vol. 25, no. 7, pp. 1196-1204, 2022/07/01/ 2022.

[48] J. G. A. Retra, B. A. B. Essers, M. A. Joore, S. M. A. A. Evers, and C. D. Dirksen, "Age dependency of EQ-5D-Youth health states valuations on a visual analogue scale," *Health and Quality of Life Outcomes,* vol. 18, no. 1, p. 386, 2020/12/12 2020.

[49] B. M. Craig, W. Greiner, D. S. Brown, and B. B. Reeve, "Valuation of Child Health-Related Quality of Life in the United States," vol. 25, no. 6, pp. 768-777, 2016.

[50] B. M. Craig, D. S. Brown, and B. B. Reeve, "Valuation of Child Behavioral Problems from the Perspective of US Adults," (in eng), *Med Decis Making,* vol. 36, no. 2, pp. 199-209, Feb 2016.

[51] V. Reckers-Droog, M. Karimi, S. Lipman, and J. Verstraete, "Why Do Adults Value EQ-5D-Y-3L Health States Differently for Themselves Than for Children and Adolescents: A Think-Aloud Study," *Value in Health,* vol. 25, no. 7, pp. 1174-1184, 2022/07/01/ 2022.

[52] S. A. Lipman, V. T. Reckers-Droog, and S. Kreimeier, "Think of the Children: A Discussion of the Rationale for and Implications of the Perspective Used for EQ-5D-Y Health State Valuation," (in eng), *Value Health,* vol. 24, no. 7, pp. 976-982, Jul 2021.

[53] J. Ratcliffe *et al.*, "Valuing Child Health Utility 9D Health States with Young Adults: Insights from a Time Trade Off Study," (in eng), *Appl Health Econ Health Policy,* vol. 13, no. 5, pp. 485-92, Oct 2015.

7.8. Appendix 1: Additional results

Table S1.1: CHU9D Instrument (Dimension Levels)

Dimensions	Levels	Wording
Worried	1	Don't feel worried today
	2	Feels a little bit worried today.
	3	Feels a bit worried today.
	4	Feels quite worried today.
	5	Feels very worried today.
Sad	1	Don't feel sad today
	2	Feels a little bit sad today.
	3	Feels a bit sad today.
	4	Feels quite sad today.
	5	Feels very sad today.
Pain	1	Don't have any pain today.
	2	Have a little bit of pain today.
	3	Have a bit of pain today.
	4	Have quite a lot of pain.
	5	Have a lot of pain.
Tired	1	Don't feel tired today
	2	Feels a little bit tired today.
	3	Feels a bit tired today.
	4	Feels quite tired today.
	5	Feels very tired today.
Annoyed	1	Don't feel annoyed today
	2	Feels a little bit annoyed today.
	3	Feels a bit annoyed today.
	4	Feels quite annoyed today.
	5	Feels very annoyed today.
School work/homework	1	Have no problems with schoolwork/homework today.
	2	Have a few problems with schoolwork/homework today.
	3	Have some problems of schoolwork/homework today.
	4	Have many problems with their schoolwork/homework today.
	5	Can't do schoolwork/homework today.
Sleep	1	Have no problems sleeping last night.
	2	Have a few problems sleeping last night.
	3	Have some problems sleeping last night.
	4	Have many problems sleeping last night.
	5	Couldn't sleep at all last night.
Daily routine	1	Have no problems with daily routine today.
	2	Have a few problems with daily routine today.

	3	Have some problems with daily routine today.
	4	Have many problems with daily routine today.
	5	Can't do their daily routine today.
Able to join in activities	1	Can join in with any activities today.
	2	Can join in with most activities today.
	3	Can join in with some activities today.
	4	Can join in with a few activities today.
	5	Can join win with no activities today.

Note: It is necessary to obtain a license to use CHU9D. https://licensing.sheffield.ac.uk/product/CHU-9D

Table S1.2: CHU9D Instrument notes for under 5 years old

Dimension	**Guidance notes for children under 5**
Worried	NA
Sad	NA
Pain	NA
Tired	NA
Annoyed	NA
School Work/Homework	If your child is at preschool/nursery/kindergarten then please think about that. If your child didn't go today because of their health and they usually would have, please tick the last option "My child can't do their schoolwork/homework today". If today is not a day they usually would have gone, then please think about how you think they would have been had they gone. If your child does not go to preschool/nursery/kindergarten, then please think about whether they have had any problems with activities such as colouring, looking at books/reading, and concentrating, as appropriate for their age.
Sleep	NA
Daily Routine	Please think about this question in terms of eating, drinking, toileting, washing and teeth cleaning, as appropriate for their age.
Able to join in activities	Please think about this question in terms of the activities your child would usually be doing today.

Note: It is necessary to obtain a license to use CHU9D proxy version with guidance notes.
https://licensing.sheffield.ac.uk/product/CHU-9D

Table S2: completion time

		Completion time, Median (25th, 75th interquartile range)	
Sample	N	Total survey (Minutes)	Average time for individual choice task (Seconds)
Total sample	2,428	11.6(8.8, 16.2)	13.2(8.3, 20.7)
2-4y arm	1,176	11.6(8.9, 16.2)	13.2(8.3, 20.6)
10y arm	1,252	11.6(8.7, 16.1)	13.1(8.2, 20.7)

Note: For time to complete individual choice tasks, N=number of participants*12 choice tasks.

Considering your views about a 2-4 year old child: which health state do you prefer?

Health Dimensions	Health State A	Health State B
Worried	**Very** worried today	**Quite** worried today
Sad	**A little bit** sad today	**A bit** sad today
Pain	**A lot of** pain today	**A lot of** pain today
Tired	**Not** tired today	**Not** tired today
Annoyed	**Very** annoyed today	**A bit** annoyed today
Schoolwork/Homework (such as reading, writing, doing lessons)	**A few problems with** their schoolwork/homework today	**A few problems with** their schoolwork/homework today
Sleep	**Some problems** sleeping last night	**A few problems** sleeping last night
Daily routine (things like eating, having a bath/shower, getting dressed)	**Can't do** their daily routine today	**A few problems with** their daily routine today
Able to join in activities (things like playing out with their friends, doing sports, joining in things)	Join in with **no** activities today	Join in with **no** activities today
	Health State A	Health State B
Your choice:	○	○

Figure S1: example DCE tasks

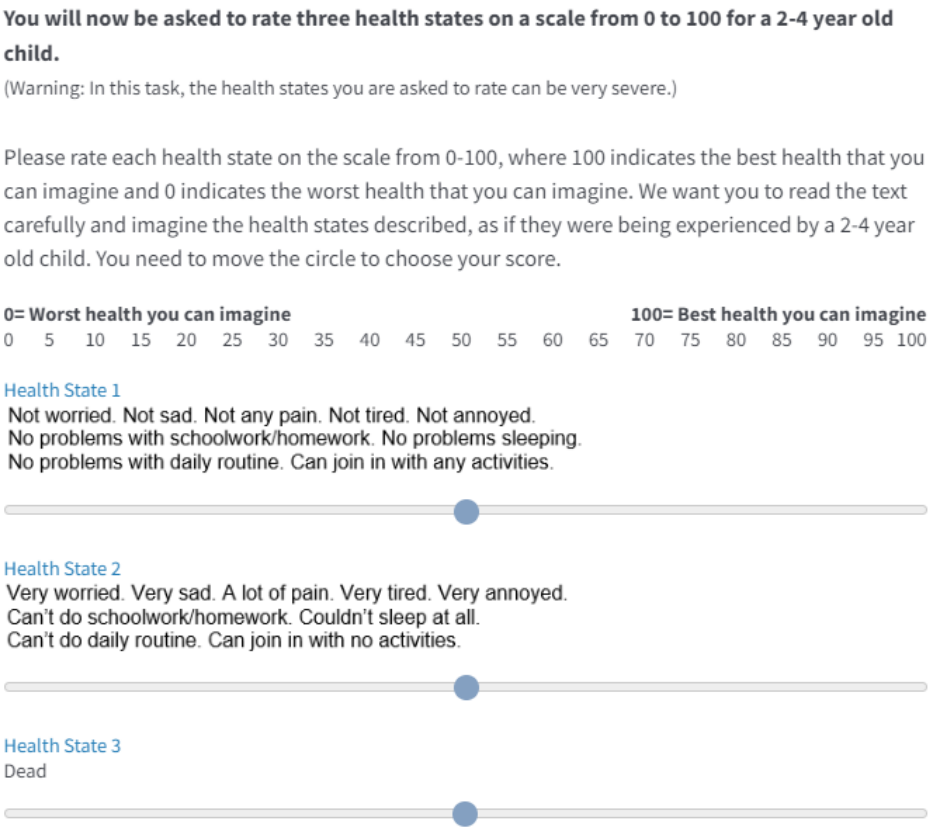


Figure S2: VAS valuation anchoring task

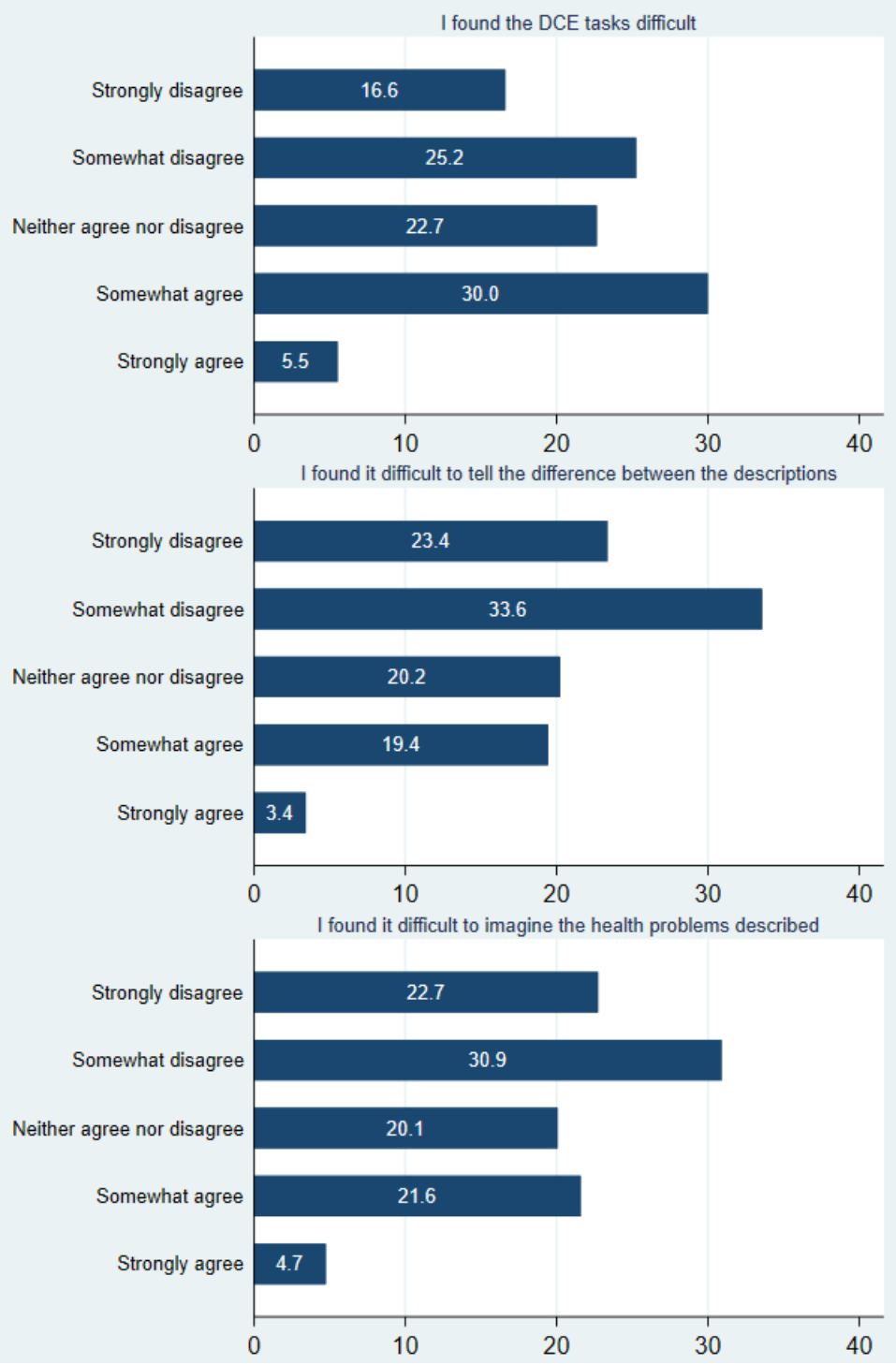


Figure S3: DCE task difficulty

Table S3: 2-4y vs 10y: P values of interaction terms

Interaction terms	Coef.	Robust Std. Err.	P values	95% Confidence Interval Low	High
Worried#sample1					
level 2#10y	-0.153	0.080	0.056	-0.310	0.004
level 3#10y	-0.153	0.086	0.075	-0.321	0.015
level 4#10y	-0.067	0.079	0.395	-0.221	0.087
level 5#10y	-0.038	0.085	0.653	-0.205	0.129
Sad#sample1					
level 2#10y	0.058	0.076	0.448	-0.091	0.207
level 3#10y	0.060	0.078	0.439	-0.092	0.212
level 4#10y	-0.007	0.078	0.929	-0.160	0.146
level 5#10y	0.099	0.088	0.257	-0.072	0.271
Pain#sample1					
level 2#10y	0.123	0.085	0.149	-0.044	0.289
level 3#10y	0.003	0.084	0.976	-0.163	0.168
level 4#10y	0.103	0.115	0.373	-0.123	0.329
level 5#10y	0.242	0.110	**0.028**	0.026	0.458
Tired#sample1					
level 2#10y	-0.019	0.071	0.788	-0.159	0.121
level 3#10y	0.036	0.064	0.578	-0.090	0.162
level 4#10y	0.009	0.069	0.894	-0.126	0.144
level 5#10y	-0.003	0.070	0.968	-0.141	0.135
Annoyed#sample1					
level 2#10y	-0.010	0.067	0.886	-0.140	0.121
level 3#10y	0.036	0.082	0.658	-0.124	0.197
level 4#10y	-0.083	0.083	0.319	-0.245	0.080
level 5#10y	-0.007	0.075	0.928	-0.154	0.141
Schoolwork#sample1					
level 2#10y	0.003	0.080	0.972	-0.154	0.160
level 3#10y	-0.066	0.077	0.391	-0.218	0.085
level 4#10y	-0.056	0.077	0.467	-0.206	0.095
level 5#10y	-0.060	0.075	0.419	-0.207	0.086
Sleep#sample1					

level 2#10y	0.029	0.071	0.683	-0.110	0.169
level 3#10y	-0.047	0.079	0.548	-0.202	0.107
level 4#10y	0.003	0.079	0.965	-0.152	0.159
level 5#10y	0.040	0.088	0.647	-0.132	0.212
Routine#sample1					
level 2#10y	-0.040	0.079	0.614	-0.196	0.116
level 3#10y	-0.027	0.079	0.727	-0.181	0.127
level 4#10y	0.055	0.080	0.486	-0.100	0.211
level 5#10y	-0.135	0.082	0.097	-0.295	0.025
Activities#sample1					
level 2#10y	0.050	0.081	0.537	-0.109	0.210
level 3#10y	0.086	0.081	0.290	-0.073	0.245
level 4#10y	0.078	0.086	0.365	-0.090	0.246
level 5#10y	0.016	0.079	0.843	-0.139	0.171

Table S4: Poolability testresults for 2-4y vs 10y.

	No scaling applied
LL pooled model	-16940.536
LL group 1	-8178.46
LL group 2	-8742.86
Degrees of freedom b/w (grp 1 + grp 2) vs pooled	37
Test statistic	38.432
Critical value	52.19231973
p value	0.404499563

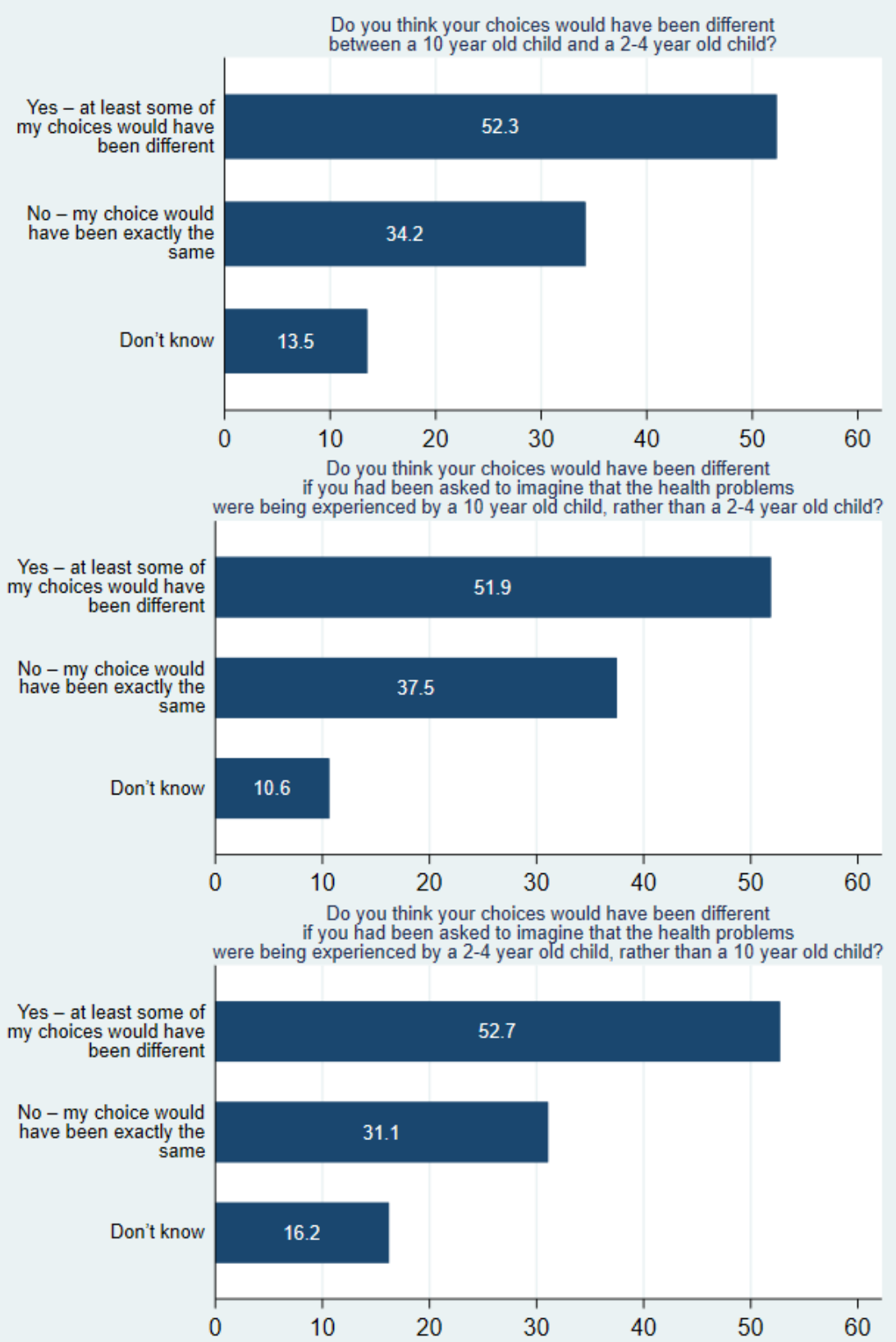


Figure S4: revealed preference: whether adults' preference differ by age

Compare with other CHU9D value sets

Table S5: utilities for some health states: comparison between different value sets

Methodologies	Utilities in this study using Australian adult value sets	Utilities using Australian adolescent value sets	Utilities using Chinese value sets	Utilities using UK adult value sets	Utilities Netherlands adult value sets
Preference elicitation method	DCE	BWS	BWS	SG	DCE with duration
Preference elicitation population	National representative general adults (age: 18y+), n=2428	A community-based sample of adolescents (age: 11-17y), n=1982	Students from primary and high schools (mean age: 13y), n=902	A random street sample in Sheffield and Huddersfield, n=300	A representative sample of Netherlands general adult population (n=1276)
Perspective	Adults think of a child	Adolescent think of themselves	Adolescent think of themselves	Adults think of themselves	Adults think of themselves
Anchoring method	VAS	TTO	TTO	/	/
Anchoring population	Same population as preference elicitation	A convenient sample of undergraduate students (age: 18-29y), n=38	A convenient sample of undergraduate students (mean age: 18y), n=38	/	/
Preference ranking of dimensions (ordered from highest to lowest utility loss at level 5)	Pain, Sleep, Daily routine, Sad, Worried, Schoolwork, Activities, Tired, Annoyed.	Sad, Pain, Daily routine, Annoyed, Sleep, Schoolwork, Worried, Activities, Tired.	Activities, Tired, Worried, Schoolwork, Sad, Pain, Sleep, Daily routine, Annoyed.	Pain, Activities, Daily routine, Sleep, Sad, Schoolwork, Tired, Annoyed, Worried.	Pain, Sleep, Activities, Daily routine, Worried, Sad, Tired, Annoyed, Schoolwork.
Health states values					
414355432	0.541	0.2505	0.5231	0.559	0.266
231345314	0.800	0.4250	0.5448	0.730	0.624
423141114	0.809	0.5606	0.7219	0.834	0.741
555555555	0.230	-0.1059	0.0563	0.326	-0.568

Thorough comparison with existing Australian adolescent value sets for CHU9D

Australian adolescent value set (Julie Ratcliff 2016) [1]

The two different value sets have different utility ranges. The adult one has a narrower utility range than the adolescent one.

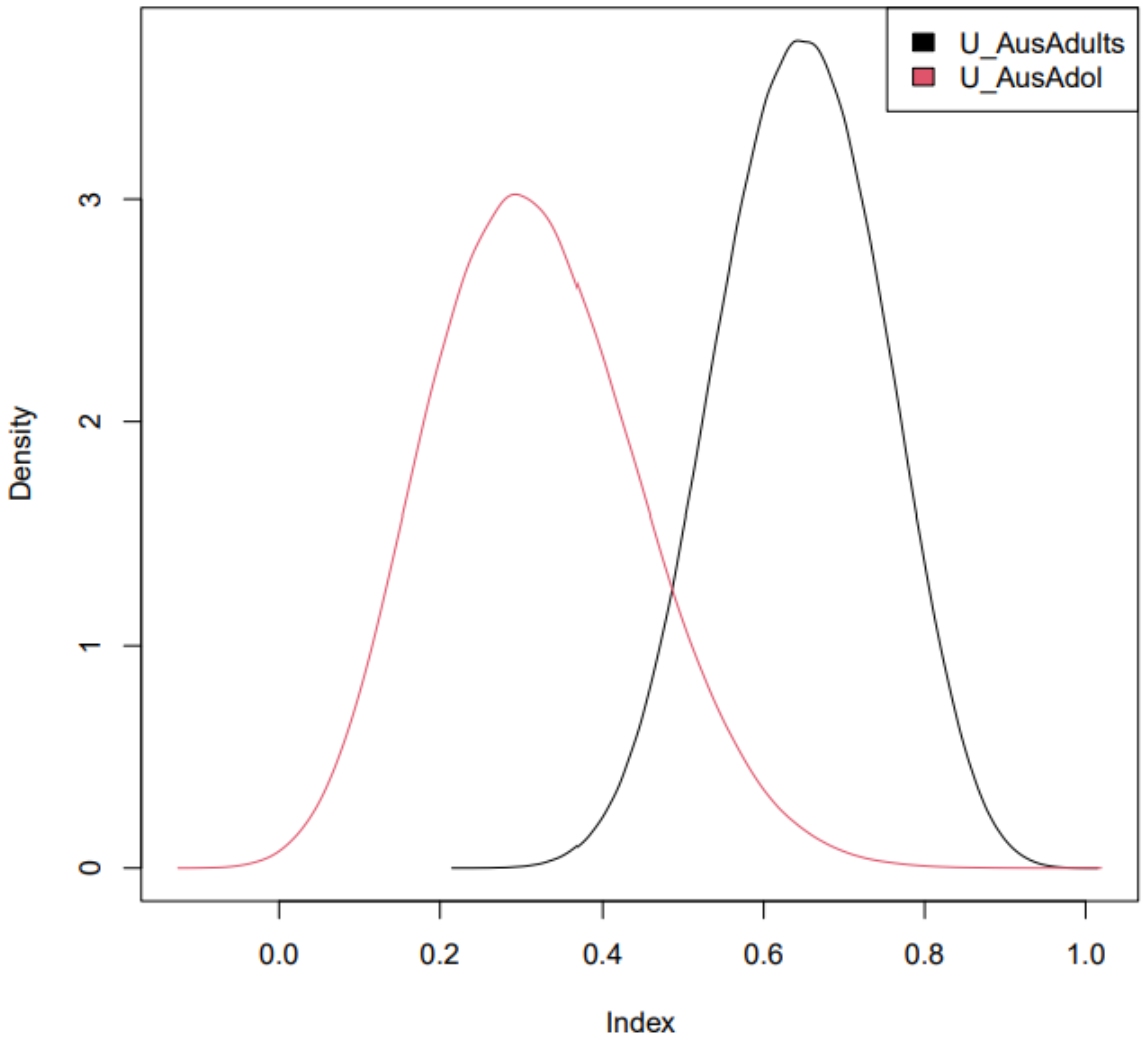


Figure S5 Kernel density plot: comparing the current value sets with existing Australian adolescent value set

Note: Kernel density plots. Included all possible CHU9D health states: 1953125 health states. Kernel density plots refer to Tianxin Pan 2022 [2] and Nicolas Boukaert 2022 [3].

Table S6 Comparison of key characteristics of the Australian adult and Australian adolescent value sets for CHU9D

Characteristics	Australian adult preferences (current study)	Australian adolescent preferences (existing value set, Julie 2016)
Value range	Min: 0.23; (state 555555555) Maximum value except full health: 0.990 (111121111) Max:1	Min: -0.1059, (state 555555555) Maximum value except full health: 0.969 (111211111) Max:1
Preference ranking of dimensions (ordered from highest to lowest utility loss at level 5)	Pain, Sleep, Sad, Daily routine, Worried, Schoolwork, Activities, Tired, Annoyed.	Sad, Pain, Daily routine, Annoyed, Sleep, Schoolwork, Worried, Activities, Tired.
Percentage of negative values (% health states valued worse than dead)	0	3721/1953125=0.2%

Note: refer to Table 3 *An EQ-5D-5L Value Set for Belgium*. [3]

The adult preference values are generally higher than the existing Australian adolescent preference values for the same health states except for the pain dimension when holding other dimensions at best level. The difference in values between the two value sets were smallest at mild health states and were larger on moderate and worst health states.

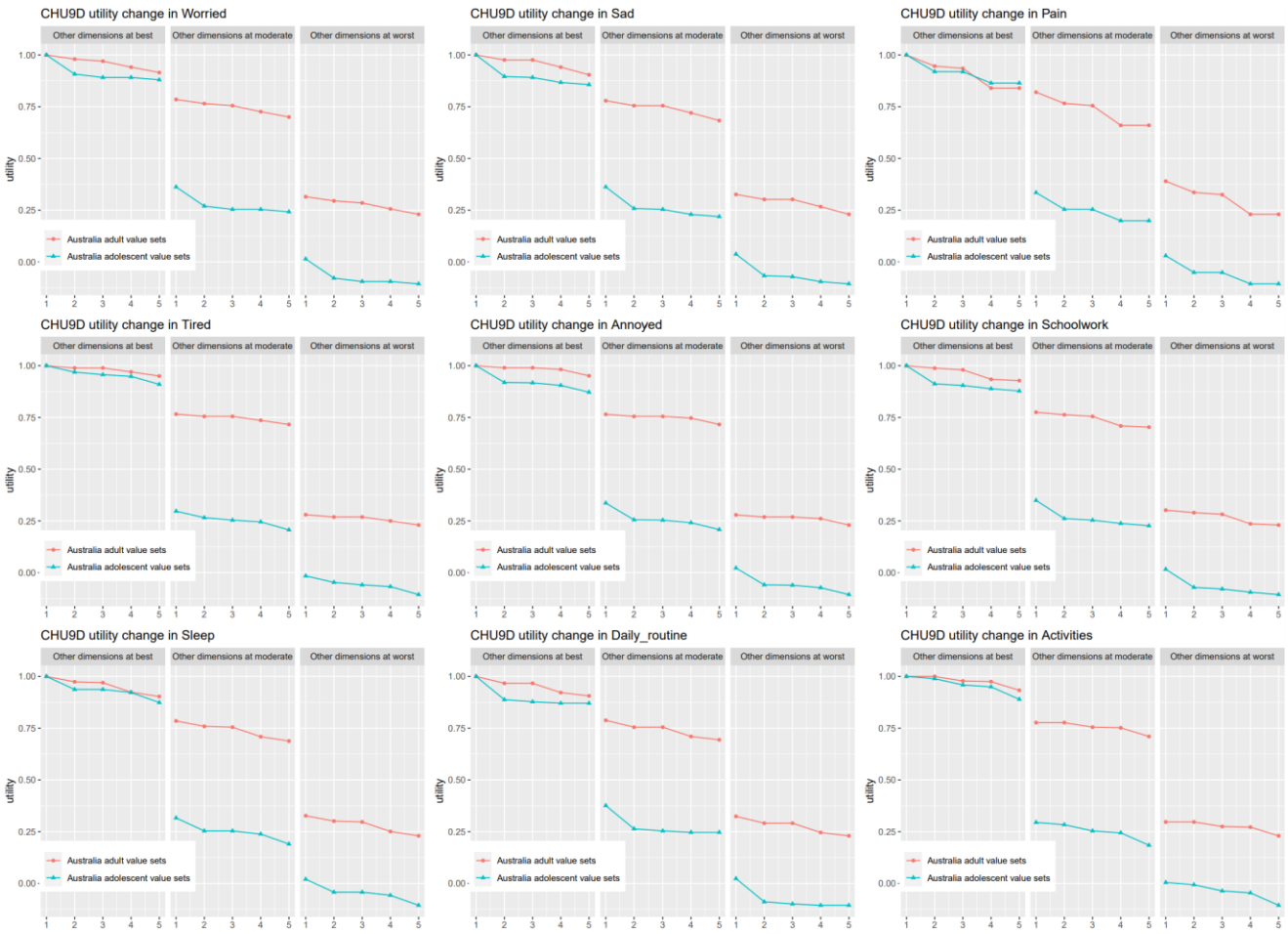


Figure S6: Changes in values between adjacent states when holding other domains at best, moderate and worst level respectively: comparing the Australian adult preference values and Adolsecnet preference values

Note: learn from Tianxin Pan 2022.[2]

Sensitivity analyses (robustness check)

Table S7: Sensitivity analyses by relaxing quality control criteria (Robustness check by including different sample)

Variable	Main model	Add not enaged responses	Add individual choice task speeders
A_1dum2	-0.237***	-0.241***	-0.198***
A_1dum3	-0.325***	-0.325***	-0.276***
A_1dum4	-0.612***	-0.609***	-0.531***

A_1dum5	-0.902***	-0.900***	-0.769***
A_2dum2	-0.297***	-0.300***	-0.261***
A_2dum3	-0.211***	-0.214***	-0.163***
A_2dum4	-0.617***	-0.614***	-0.508***
A_2dum5	-0.999***	-1.001***	-0.835***
A_3dum2	-0.565***	-0.556***	-0.477***
A_3dum3	-0.688***	-0.679***	-0.586***
A_3dum4	-1.729***	-1.712***	-1.451***
A_3dum5	-1.565***	-1.554***	-1.297***
A_4dum2	-0.105**	-0.104**	-0.093**
A_4dum3	-0.121***	-0.119***	-0.100***
A_4dum4	-0.306***	-0.304***	-0.282***
A_4dum5	-0.516***	-0.501***	-0.424***
A_5dum2	-0.139***	-0.127***	-0.142***
A_5dum3	-0.092*	-0.086*	-0.112**
A_5dum4	-0.201***	-0.184***	-0.215***
A_5dum5	-0.515***	-0.503***	-0.455***
A_6dum2	-0.149***	-0.144***	-0.131***
A_6dum3	-0.231***	-0.230***	-0.215***
A_6dum4	-0.713***	-0.710***	-0.602***
A_6dum5	-0.751***	-0.745***	-0.635***
A_7dum2	-0.271***	-0.269***	-0.238***
A_7dum3	-0.320***	-0.323***	-0.275***
A_7dum4	-0.785***	-0.783***	-0.675***
A_7dum5	-1.038***	-1.023***	-0.886***
A_8dum2	-0.374***	-0.373***	-0.325***
A_8dum3	-0.328***	-0.324***	-0.286***
A_8dum4	-0.806***	-0.807***	-0.700***
A_8dum5	-1.022***	-1.025***	-0.867***
A_9dum2	0.080*	0.083*	0.071
A_9dum3	-0.189***	-0.182***	-0.131***
A_9dum4	-0.198***	-0.194***	-0.150***
A_9dum5	-0.645***	-0.643***	-0.529***
Sample size	N=2408	N=2501	N=2784
Observations	58272	60024	66816
Log likelihood	-16940.54	-17474.80	-20324.02
AIC	33953.07	35021.60	40720.03
BIC	34276.10	35345.69	41047.98

Table S8 Sensitivity analyses comparing conditional logit model with mixed logit model

Variable	Main model	Mixed logit model
A_1dum2	-0.237***	-0.259***
A_1dum3	-0.325***	-0.358***
A_1dum4	-0.612***	-0.688***
A_1dum5	-0.902***	-1.061***
A_2dum2	-0.297***	-0.350***
A_2dum3	-0.211***	-0.239***
A_2dum4	-0.617***	-0.715***
A_2dum5	-0.999***	-1.169***
A_3dum2	-0.565***	-0.615***
A_3dum3	-0.688***	-0.783***
A_3dum4	-1.729***	-2.009***
A_3dum5	-1.565***	-1.854***
A_4dum2	-0.105**	-0.140***
A_4dum3	-0.121***	-0.155***
A_4dum4	-0.306***	-0.350***
A_4dum5	-0.516***	-0.567***
A_5dum2	-0.139***	-0.150***
A_5dum3	-0.092*	-0.111*
A_5dum4	-0.201***	-0.229***
A_5dum5	-0.515***	-0.569***
A_6dum2	-0.149***	-0.205***
A_6dum3	-0.231***	-0.254***
A_6dum4	-0.713***	-0.812***
A_6dum5	-0.751***	-0.854***
A_7dum2	-0.271***	-0.307***
A_7dum3	-0.320***	-0.370***
A_7dum4	-0.785***	-0.910***
A_7dum5	-1.038***	-1.218***
A_8dum2	-0.374***	-0.395***
A_8dum3	-0.328***	-0.370***
A_8dum4	-0.806***	-0.887***
A_8dum5	-1.022***	-1.147***
A_9dum2	0.080*	0.113*
A_9dum3	-0.189***	-0.192***
A_9dum4	-0.198***	-0.221***
A_9dum5	-0.645***	-0.758***
Sample size	N=2408	N=2408
Observations	58272	58272
Log likelihood	-16940.54	-16773.54

AIC	33953.07	33691.09
BIC	34276.10	34337.14

Reference:

[1] J. Ratcliffe *et al.*, "Valuing the Child Health Utility 9D: Using profile case best worst scaling methods to develop a new adolescent specific scoring algorithm," (in eng), *Soc Sci Med,* vol. 157, pp. 48-59, May 2016.

[2] T. Pan, B. Mulhern, R. Viney, R. Norman, J. Hanmer, and N. J. P. Devlin, "A comparison of PROPr and EQ-5D-5L value sets," pp. 1-11, 2022.

[3] N. Bouckaert, I. Cleemput, S. Devriese, and S. Gerkens, "An EQ-5D-5L Value Set for Belgium," *PharmacoEconomics - Open,* vol. 6, no. 6, pp. 823-836, 2022/11/01 2022.

7.9. Appendix 2: Survey questionnaire

(This is for one arm for 2-4 year old version; the survey for the other arm for the 10 year old version is mostly the same, only using the original CHU9D and changing 2-4y to 10y throughout.)

[Section 1: Initial screening questions – for participant quotas]

1. What is your gender:
   Male
   Female
   Other

2. How old are you?
   - ☐ Under 18 (end of survey)
   - ☐ 18-24
   - ☐ 25-34
   - ☐ 35-44
   - ☐ 45-54
   - ☐ 55-64
   - ☐ 65+

3. Where do you live?
   - ☐ NSW
   - ☐ VIC
   - ☐ QLD
   - ☐ SA
   - ☐ WA
   - ☐ ACT
   - ☐ TAS
   - NT

4. Do you have any children?
   Yes
   No

5. [If yes to above question] How old are these children now? (If you have more than one child, tick all that apply)
   - ☐ 0 - 1 years
   - 2 - 4 years
   - ☐ 5 – 7 years
   - 8- 10 years
   - 11-18 years
   - ☐ Over 18 years

Honesty oath:

Before we begin, do you promise to answer the following questions truthfully? (you will be allowed to continue with this survey regardless of your answer to this question)

Yes
No

Detect bot: traffic light

Please select the correct description of the colors in the traffic light **from top to bottom**:

Red; green; yellow
Yellow; green; red
Red; yellow; green (if not selected, end of survey)

[Section 2: Reporting health using CHU9D]

Think of a child aged 2-4 years old. The child can be your own child, or a child you know (for example, a child of your friend or relative), or any child you can imagine. You can make up the health situation if you are thinking of a hypothetical child. This section is for you to be familiar with the instrument we are valuing in this survey.

These questions ask about how the child is **today.** For each question, read all the choices and decide which one is most like the child **today**. Some questions have extra guidance with them as the child is under 5 years of age.

1. Worried

☐ My child doesn't feel worried today

☐ My child feels a little bit worried today

☐ My child feels a bit worried today

☐ My child feels quite worried today

☐ My child feels very worried today

2. Sad

☐ My child doesn't feel sad today

☐ My child feels a little bit sad today

☐ My child feels a bit sad today

☐ My child feels quite sad today

☐ My child feels very sad today

3. Pain

☐ My child doesn't have any pain today

☐ My child has a little bit of pain today

☐ My child has a bit of pain today

☐ My child has quite a lot of pain today

☐ My child has a lot of pain today

4. Tired

☐ My child doesn't feel tired today

☐ My child feels a little bit tired today

☐ My child feels a bit tired today

☐ My child feels quite tired today

☐ My child feels very tired today

5. Annoyed

☐ My child doesn't feel annoyed today

☐ My child feels a little bit annoyed today

- ☐ My child feels a bit annoyed today
- ☐ My child feels quite annoyed today
- ☐ My child feels very annoyed today

6. School Work/Homework (such as reading, writing, doing lessons)

If your child is at preschool/nursery/kindergarten then please think about that. If your child didn't go today because of their health and they usually would have, please tick the last option "My child can't do their schoolwork/homework today". If today is not a day they usually would have gone, then please think about how you think they would have been had they gone. If your child does not go to preschool/nursery/kindergarten, then please think about whether they have had any problems with activities such as colouring, looking at books/reading, and concentrating, as appropriate for their age.

- ☐ My child has no problems with their schoolwork/homework today
- ☐ My child has a few problems with their schoolwork/homework today
- ☐ My child has some problems with their schoolwork/homework today
- ☐ My child has many problems with their schoolwork/homework today
- ☐ My child can't do their schoolwork/homework today

7. Sleep

- ☐ Last night my child had no problems sleeping
- ☐ Last night my child had a few problems sleeping
- ☐ Last night my child had some problems sleeping
- ☐ Last night my child had many problems sleeping
- ☐ Last night my child couldn't sleep at all

8. Daily routine (things like eating, having a bath/shower, getting dressed)

Please think about this question in terms of eating, drinking, toileting, washing and teeth cleaning, as appropriate for their age.

- ☐ My child has no problems with their daily routine today
- ☐ My child has a few problems with their daily routine today
- ☐ My child has some problems with their daily routine today
- ☐ My child has many problems with their daily routine today
- ☐ My child can't do their daily routine today

9. Able to join in activities (things like playing out with their friends, doing sports, joining in things)
Please think about this question in terms of the activities your child would usually be doing today.

- ☐ My child can join in with any activities today
- ☐ My child can join in with most activities today
- ☐ My child can join in with some activities today
- ☐ My child can join in with a few activities today
- ☐ My child can join in with no activities today

[Section 3: DCE tasks]

Instruction for DCE task

You will now be asked to complete 13 choice tasks (choosing between two health states).

In each of the task, two different health states 'A' and 'B' will be shown (example task below). Each health state describes 9 aspects of health and wellbeing (worried, sad, pain, tired, annoyed, sleep, daily routine, schoolwork/homework, and able to join in activities), with 5 levels of severity for each aspect.
Please assume that the child will have no other health problems besides what is indicated in the health states.
Please read the text carefully and imagine the health states described, as if they were being experienced by a **2-4 year old child**, and then select the health state that **you would prefer** for a 2-4 year old child.

Only **5 of the 9 health dimensions vary in levels** between health states 'A' and 'B', and **the remaining 4 dimensions have the same levels** between 'A' and 'B' (which are grey shaded).

When you click (or touch if using tablet) on each health dimension (Worried, Sad, ...), instructions for severity levels and **further guidance notes will pop out** to help you understand these dimensions for a 2-4 year old children and make a choice.

Example choice task (not displayed here)

Here is a summary of the severity levels for different dimensions; the deeper the cell color, the bigger the health problem.

Health Dimensions	Level 1	Level 2	Level 3	Level 4	Level 5
Worried/Sad/Pain/Tired/Annoyed	Not	A little bit	A bit	Quite/Quite a lot	Very/A lot
Schoolwork/Sleep/Daily routine	No problems	A few problems	Some problems	Many problems	Can't/Couldn't
Able to join in activities	Join in any activities	Most activities	Some activities	A few activities	No activities

Before proceeding, please ensure that you are viewing the survey using the 'full screen' **of your computer or tablet**, so that the options display correctly. Note that once you have answered a question you will not be able to go back to the previous page and change your answers.

DCE choice tasks (in total 13 tasks; only showing 1 example task here)

Considering your views about a 2-4 year old child: which health state do you prefer?

Health Dimensions	Health State A	Health State B
Worried	**Very** worried today	**Quite** worried today
Sad	**A little bit** sad today	**A bit** sad today
Pain	**A lot of** pain today	**A lot of** pain today
Tired	**Not** tired today	**Not** tired today
Annoyed	**Very** annoyed today	**A bit** annoyed today
Schoolwork/Homework (such as reading, writing, doing lessons)	**A few problems with** their schoolwork/homework today	**A few problems with** their schoolwork/homework today
Sleep	**Some problems** sleeping last night	**A few problems** sleeping last night
Daily routine (things like eating, having a bath/shower, getting dressed)	**Can't do** their daily routine today	**A few problems with** their daily routine today
Able to join in activities (things like playing out with their friends, doing sports, joining in things)	Join in with **no** activities today	Join in with **no** activities today

	Health State A	Health State B
Your choice:	○	○

DCE Debrief question:

1. Please let us know what you thought of the DCE tasks (choosing between A and B) you just performed, by indicating how strongly you agree or disagree with the statement shown in bold.

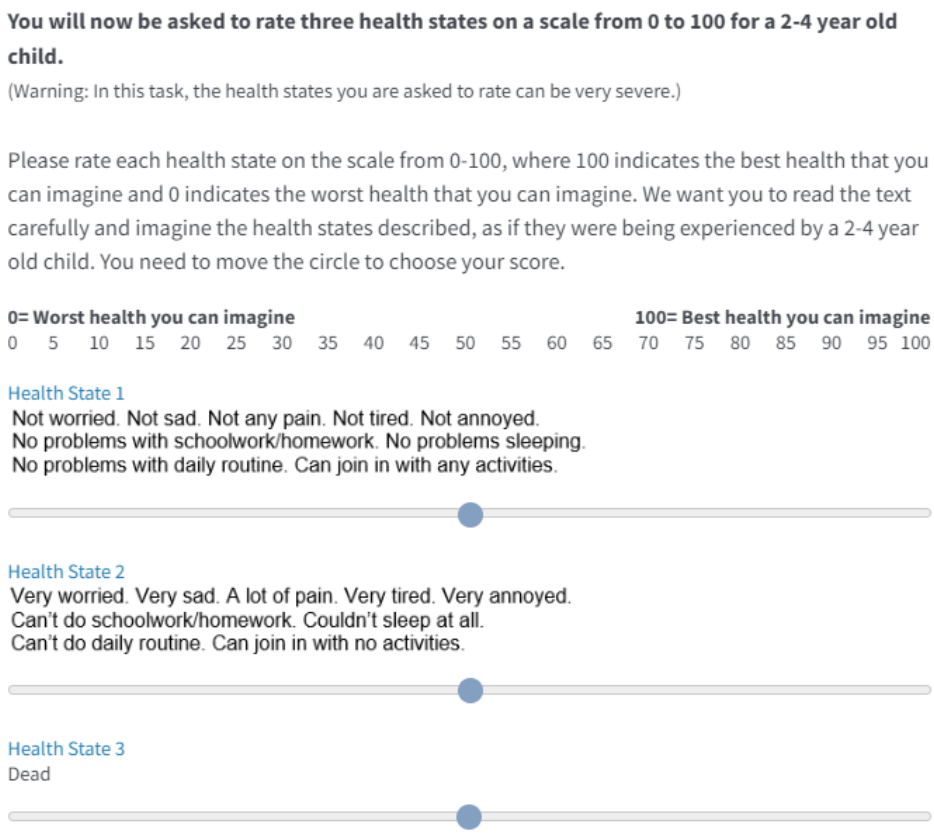
	Strongly disagree	Somewhat Disagree	Neither agree nor disagree	Somewhat Agree	Strongly agree
I found the tasks difficult					
I found it difficult to tell the difference between					

the descriptions					
I found it difficult to imagine the health problems described					

2. Do you think your choices would have been different if you had been asked to imagine that the health problems were being experienced by a 10 year old child, rather than a 2-4 year old child?

- ☐ Yes – at least some of my choices would have been different
- ☐ No – my choice would have been exactly the same
- ☐ Don't know

Session 4: VAS task



VAS Debrief question:

Do you think you understand the tasks (rating from 0-100 task) you just performed?

Fully understand,
Somewhat understand,
Not understand.

[Section 5: Demographic questions]

This is the last section of the survey, we will ask you 12 questions about simple background information or experience about illness.

1. What is your age (in years)? ____

2. What is the highest level of school you have completed?
   Primary school
   Some high school
   Finished high school (Year 12 in Victoria)

3. What is your highest post-school qualification?
   None
   Trade/Apprenticeship
   Certificate, Diploma
   Undergraduate degree (e.g., Bachelors, Honours)
   Postgraduate degree (e.g., Masters, Doctorate)
   Other_____

4. What is your current employment status? If you had more than one job or business, please think about the one in which you usually work the most hours.
   Full-time employment
   Part-time employment
   Away from work
   Unemployed

5. Which of the following categories best describes your **individual income**, from all sources, before taxes in 2022?

   Less than $300 per week ($15,599 or less per year)
   $300-$499 per week ($15,600-$25,999 per year)
   $500-$999 per week ($26,000-$51,999 per year)
   $1,000-$1,999 per week ($52,000-$103,999 per year)
   $2,000-$2,999 per week ($104,000-$155,999 per year)
   $3,000-$3,999 per week ($156,000-$207,999 per year)
   $4,000 or more per week ($208,000 or more per year)

6. Which of the following categories best describes your total **household income**, from all sources, before taxes in 2022?

   Less than $300 per week ($15,599 or less per year)
   $300-$499 per week ($15,600-$25,999 per year)

$500-$999 per week ($26,000-$51,999 per year)
$1,000-$1,999 per week ($52,000-$103,999 per year)
$2,000-$2,999 per week ($104,000-$155,999 per year)
$3,000-$3,999 per week ($156,000-$207,999 per year)
$4,000 or more per week ($208,000 or more per year)

7. Have you ever been involved in (e.g., employed/volunteering) a role that involved working directly with young children under 5 years old **in the last 5 years**?
   - ☐ Yes
   - ☐ No

8. Have you experienced serious illness?

   a1. In yourself as an adult

   - ☐ Yes
   - ☐ No

   a2. In yourself when you were a child

   - ☐ Yes
   - ☐ No

   b. In your close family members or friends

   - ☐ Yes
   - ☐ No

   c1. In caring for others- caring for an adult

   - ☐ Yes
   - ☐ No

   c1. In caring for others- caring for a child

   - ☐ Yes
   - No

9. In general, how would you rate your own health?
   Excellent
   Very good
   Good
   Fair
   Poor

10. Are you of Aboriginal or Torres Strait Islander origin?
    Yes
    No

11. Are you:
- Single
- Married/Partner
- Separated
- Divorced
- Widowed

Prefer not to say

12. Please rate your engagement to the survey you just performed.
Fully engaged
Partially engaged
Not engaged

Free text page

Thank you for completing the survey. If you have any comments about the survey, e.g., how you felt about the task or whether you had any problems viewing or understanding the information presented, please enter them in the box below:

Please click "To finish the survey" to submit your responses!

7.10. Appendix 3: RETRIEVE checklist.

Reference: The RETRIEVE checklist for studies reporting the elicitation of stated preferences for child health related quality of life. Bailey, Howell et al.[1]

Table S1 - The RETRIEVE long checklist

LONG FORM:

This checklist is modular, not all sections will apply to all papers.

Section A - Stated preferences considered relevant to valuing child HRQoL and sample characteristics	
A1 – Stated preferences	
A1a	Whose preferences were sought? ☐ Adults [x] *A1b then A2* ☐ Children and young people (CYP) <18 years *A1b then A3* ☐ Mixed adults and CYP *A1b then A2 and A3*
A1b	Did the authors provide a rationale for whose preference were sought? ☐ Yes [x] ☐ No
A2 Adults' stated preferences	
A2a	Which adults were the focus of preference elicitation? ☐ General population [x] ☐ Parent or caregiver of child ☐ Health care professionals ☐ Adult with a health condition ☐ Other adults, please specify _____
A2b	What perspective were adults asked to take in considering the child states to be valued? e.g. thinking about the health states as experienced by: ☐ Own child (parent) ☐ Another child they know ☐ A hypothetical child [x] ☐ Their own health, thinking back to when they were a child ☐ Their own health, as if they were a child now ☐ Their own health, but blinded to the states under consideration being specific to children ☐ Person with a health condition (e.g. a health professional asked to take the person with a health condition's perspective) ☐ Other, please specify: _____
A2c	Was the age of the child, for whom respondents were asked to imagine health states to be

	valued, specified? ☐ Yes [x] *Go to A2d* ☐ No *Go to A4* ☐ Not applicable *Go to A4*
A2d	If yes, what was the age of the child? 10 years, 2-4 years
A2e	Was the rationale for the choice of the age of child provided? ☐ Yes [x] Prior studies and following the EQ-5D-Y valuation protocol ☐ No
A3 Children and young people's stated preferences Section A3 is not relevant to the value set reported by this study.	
A3a	From which child/young person were preferences elicited? [N/A] ☐ General population ☐ Person with a health condition ☐ Other children, please specify: _____
A3b	What perspective was the (child/young person) respondent asked to take? e.g. thinking about the health states as experienced by: [N/A] ☐ Themselves (i.e. their own perspective) ☐ Another known child ☐ A hypothetical child ☐ Other, please specify:_____
A3c	Was the age of the child/young person, for whom respondents were asked to imagine health states to be valued, specified? [N/A] ☐ Not applicable (i.e. own perspective/themselves) *Go to A4* ☐ It was applicable but not stated *Go to A4* ☐ Yes
A3d	If the age was specified, what was the age? [N/A]
A3e	Was the rationale for the choice of the age of child/young person provided? [N/A] ☐ Yes ☐ No
A4 Sample	
A4a	Was the population or sample frame defined from which the sample was drawn? (e.g., country, age, condition) ☐ Yes [x] Australian adults general population DCE survey and VAS anchoring: Country (Australia) and representative of the general adult population (age, gender, region). ☐ No
A4b	Is information provided on how the sample was recruited (e.g., field-based recruitment, online

	panel, convenience sample)? ☐ Yes [x] Online panel. ☐ Partial ☐ No
A4c	If data were collected online, were efforts made to avoid on-line panel fraud? (eg, related to bots or automated software posing as participants and completing surveys) ☐ Yes [x] ☐ No ☐ Not applicable
A4d	Was there a target sample size (or sample sizes if by block – e.g. number of tasks per block (e.g. DCE) or health state (e.g. TTO))? ☐ Yes [x] Stated as 1200 for the DCE and VAS. ☐ No *Go to A4g*
A4e	Was the target sample justified? ☐ Yes [x] Based on established guidelines and previous studies. ☐ No
A4f	Was the target sample achieved? ☐ Yes [x] ☐ No ☐ Unclear
A4g	Were the characteristics of the final sample described? ☐ Yes [x] ☐ No *Go to A4i*

A4h	Did the sample characteristics match the intended population? ☐ Yes ☐ No [x] The sample of adults slightly under-represented adults aged>65 years and slightly over-represented adults from Aboriginal or Torress Strait Island origin. All other groups were well represented. ☐ Unclear
A4i	Was the year the data collected stated? ☐ Yes – what year(s) were the data collected? [x] Sep 2023 to Nov 2023 (may updated to March 2024 after all data colleced) ☐ No
A4j	Was information provided on missing data? (non-completion, withdrawals)? ☐ Yes [x] All included data based on data quality have no missing data. The DCE is designed to not allow skipping questions. ☐ Partial ☐ No

Section B - Child HRQoL states to be valued	
B1 Type of study	
B1	Did the values reported in this paper comprise: ☐ A value set? [x] *Go to B2* ☐ Values for a limited number of health states (e.g. vignette)? *Go to B3*
B2 Value Sets	
B2a	Which HRQoL instrument was valued? CHU9D
B2b	Were the domains and response options of the instrument clearly described? ☐ Yes [x] ☐ No
B2c	What experimental design approach was used to choose the health states (combination of dimension levels) to be valued? Bayesian D-efficient design with main effects, overlapping of health states in 4 dimensions levels. The randomly selected 204 pairs of health states were divided into 17 blocks of 12 DCE tasks.
B2d	How were the health states assigned to respondents? Each respondent was asked to complete 1 of the 17 blocks of 12 DCE tasks. Each of the 12 DCE tasks presented 2 health states and the respondent was asked to choose their preferred state (i.e. a forced choice). Respondents were randomly assigned to complete 1 out of the 17 blocks of 12 DCE tasks.
B3 Specific health states Section B3 is not relevant to this study.	
B3a	How were the health states described? [N/A] ☐ Disease specific vignettes

	☐ From a disease-specific HRQoL instrument ☐ Other, please specify _____
B3b	How many health states were preferences elicited for? [N/A]
B3c	Was the rationale for the selection of these health states specified? [N/A] ☐ Yes – What was the rationale? _____ ☐ No

Section C – Methods used to elicit stated preferences for child HRQoL	
C1	**Which method or methods were used to elicit stated preferences?** ☐ DCE [x] ☐ TTO ☐ SG ☐ BWS ☐ VAS [x] ☐ Other, please specify_____
C2	**Was a rationale for the choice of method(s) provided?** ☐ Yes [x] ☐ No
C2a	If yes, what was the rationale? Refer to The International Valuation Protocol for the EQ-5D-Y-3L, and refer to previous similar studies valuing CHU9D, with specific considerations for valuing health for 2-4 years old. 2-4 years old is a narrow age range and thus valuation method with duration is not feasible. Therefore, DCE was used for valuation and VAS was used for anchoring. Consider the 9 attributes of CHU9D and respondents' cognitive burden, partial DCE design with 4 overlapped dimensions were adopted.
C3	**Was the duration of the states to be valued reported (e.g 'x years in this state, followed by death')?** ☐ Yes ☐ No [x] No duration for this study as 2-4 year old is a narrow age range and duration is not feasible. *Go to C4*
C3a	Was the duration fixed? ☐ Yes ☐ No
C3b	What duration(s) was used?
C4	**Did the method(s) allow values to be elicited that were < 0 ('worse than dead')?** ☐ Yes [x] ☐ No *Go to C5*
C4a	How were values < 0 elicited? Using a VAS. If the VAS value for health states is lower than the VAS value for "dead", the values would be <0.

C4b	What was the minimum value possible? (may vary according to the method used so should be clearly stated) infinite. The minimum value can be a very large negative value using VAS to anchor. For example, if the sample value dead at 99, value worst state at 0, and full health at 100, then the anchored value for the worst state is (0-99)/(100-99)=-99
C4c	What determined how the task was terminated? This is not relevant with the VAS anchoring method which has no mechanism for termination.
C5	**How were the values anchored on a utility scale?** Using VAS; All variable dummy coded and DCE coefficients divided by the overall utility range and re-scaled to the value of the pits state (555555555) obtained from VAS.
C6	**What was the mode of administration for the stated preference tasks?** ☐ Online self-completion by the respondent [x] DCE and VAS ☐ Self-completion of mailed questionnaires ☐ Online computer assisted personal interview (CAPI) ☐ In person CAPI ☐ In person interview ☐ Other, please specify _____
C7	**How was the quality of stated preference data assessed?** Several criteria were created to guarantee the quality of data used for the main analysis. Responses that failed any criteria in categories 1 and 2 were not included in the main analysis. Category 1: to be deleted. Responses that failed this type of criteria were deleted (obviously illegitimate responses). • Total completion time. Responses with less than 1/3 of median completion time were deleted. • Straight-liner: Responses that always chose A or B were deleted. • Traffic lights test: This question asked respondents to describe the color of a traffic light picture. This is a test to ensure that the respondent is a real person and attentive respondent. Responses that failed the test were deleted. Category 2: to flag as problematic. Responses that failed this type of criteria were not included in the main analysis but could be used as sensitivity analysis. • Age consistency: Age band multiple choice task was presented at the front of the survey and a free text age question was presented later in the survey. Those whose answers didn't match were not included in the main analysis. • Completion time for individual choice tasks: From the pilot study, the median time was 12 seconds for each CHU9D choice task. If a participant had more than half of (>=7) choice tasks with completion time less than 1/3 median (4 seconds), the quality of data for this person was deemed problematic and was not included in the main analysis. • Dominant choice task: those who failed the dominant choice task were not included in the main analysis. • Respondent engagement: At the end of the survey, respondents were asked their engagement to the survey, with choices of "fully engaged", "partially engaged", and "not engaged". Those reporting "not engaged" were not included in the main analysis. Category 3: quality check only. Responses were included in the main analysis no matter what the answers were. • Honesty oath: A previous study indicated that use of an honesty oath improved results.[29] The question in this survey was: "Before we begin, do you promise to answer

	the following questions truthfully? (you will be allowed to continue with this survey regardless of your answer to this question)" with answers yes or no. No matter what the answer was, data was included in the main analysis.	
C8	**Were any exclusions made to the preference data (eg used to represent average preferences)?** ☐ Yes [x] excluded data with poor quality based on criteria listed above. ☐ No ☐ Unclear	 *Go to C9* *Go to C9*
C8a	**Were reasons for the exclusions provided?** ☐ Yes [x] ☐ No ☐ Unclear	
C9	**Were the health states randomly assigned?** ☐ Yes [x] ☐ No ☐ Unclear	
C10	**Was ethics approval for the study obtained from an appropriate research ethics committee?** ☐ Yes [x] ☐ No ☐ Unclear ☐ Not stated	
C11	**Were sources of funding and non-monetary support and the role of the funder(s) in the design described?** ☐ Yes [x] ☐ No	

Section D – Econometric modelling and statistical methods		
D1 – Did the values reported comprise:		
	☐ A value set? [x]	*Go to D2*
	☐ values for a limited number of health states (vignette or condition-specific)?	*Go to D3*
D2 Econometric modelling of value sets for HRQoL instruments		
D2a	What was the theoretical model? OR What models were estimated? e.g. OLS, Tobit etc. Random utility model – Linear additive utility with all variables dummy coded. Conditional logit model was used to estimate the average preferences.	
D2b	Were the main assumptions of the model stated? (e.g. assumptions about preference homogeneity/heterogeneity) ☐ Yes ☐ No [x] The aim is to obtain the population average preferences, and thus the heterogeneity of the preferences is not explored. ☐ Unclear	
D2c	How was the constant term treated (if included)? The constant term was not included as this is a two option forced choice DCE, and no "status quo" or "opt out option" were provided.	

D2d	How were missing data handled (e.g.: imputation, complete case analysis) No missing data as no skipping questions were allowed for the DCE and VAS.
D2e	Were subgroup analyses completed? ☐ Yes ☐ No ☐ Not applicable [x]
D2f	Were interaction terms included? ☐ Yes [x] Interactions were included to explore the preference difference between two perspectives, not for the final value set developed. *If no, go to D2h* ☐ No
D2g	Were details of the interactions provided? ☐ Yes [x] For the exploration of the preference by child age, the interaction of each dimension level with the subgroups of child age were included. ☐ No ☐ Not applicable [x]
D2h	Were non-linear specifications considered? ☐ Yes ☐ No [x]
D2i	Was more than one model described? ☐ Yes [x] original model and consistent model were described. ☐ No *If no, go to D2m*
D2j	Were goodness-of-fit statistics for each model reported? ☐ Yes [x] ☐ No
D2k	Was the preferred model clearly stated? ☐ Yes [x] ☐ No
D2l	Were the criteria used to select the preferred model described? ☐ Yes [x] Sensitivity analyses by relaxing data quality control criteria were conducted. Sensitivity analyses using mixed logit model were conducted. Model were chosen by research aim and AIC, BIC. ☐ No
D2m	Do the preference parameters for the health states follow a logical order (monotonic)? ☐ Yes *If yes, go to D2p* ☐ No [x]
D2n	Was any post estimation undertaken to force monotonicity (e.g. collapsing levels)? ☐ Yes [x]

	☐ No ☐ Unclear/not stated
D2o	How were insignificant differences between adjacent levels managed (e.g. collapsed/ forced to be different)? To produce a consistent model, adjacent inconsistent levels are to be merged (collapsed) or constrained to be the equal.
D2p	Were robustness checks conducted? [N/A] As only one model was used and one anchoring method was used. No place to check robustness. ☐ Yes ☐ No
D2q	Was uncertainty around values reported? ☐ Yes [x] Standard errors ☐ No
D3 Analysis of values for specific HRQoL states Not relevant to the value set in this study.	
D3a	Have the statistical methods been described? [N/A] ☐ Yes ☐ No *If no, go to D3c*
D3b	Have the statistical methods been justified? [N/A] ☐ Yes ☐ No
D3c	How were missing data handled (e.g.: imputation, complete case analysis)? [N/A]
D3d	Have subgroup analyses and interactions been undertaken? [N/A] ☐ Yes ☐ No *If no, go to D3h*
D3e	Were sub-groups and interaction variable chosen for assessment justified? [N/A] ☐ Yes ☐ No
D3f	Were sensitivity analyses undertaken? [N/A] ☐ Yes [x] *If no, go to Section E* ☐ No
D3g	Were sensitivity analyses described? [N/A] ☐ Yes [x] ☐ No

Section E - Characteristics of values	
E1	**Was qualitative or quantitative evidence reported that demonstrates the extent to which respondents**

	engaged with and understood the valuation tasks? ☐ Yes [x] ☐ No
E2	**Where a value was reported, were the values generated by the final model logically consistent?** ☐ Yes [x] ☐ No ☐ Unclear
E3	**Did authors report the distribution of values over all states defined by the HRQoL instrument (e.g. as per Figure 1 in Pan et al 2022.)** ☐ Yes [x] ☐ No
E4	**Key characteristics of the values**
E4a	How many percentage values less than zero were possible? 0
E4b	What was the maximum possible value less than one? 0.990
E4c	Where in the descriptive system does the biggest change in values occur, when shifting between adjacent states? Between the adjacent states in pain dimension, see *Figure changes in utility between adjacent states*.
E5	**Was the order of importance of dimensions (domains) suggested by the value set discussed?** ☐ Yes [x] ☐ No

Reference:

[1] C. Bailey *et al.*, "The RETRIEVE Checklist for Studies Reporting the Elicitation of Stated Preferences for Child Health-Related Quality of Life," *PharmacoEconomics,* 2024/01/13 2024.

Chapter 8: Discussion and Conclusion

8.1. Brief chapter summary

In summary, this thesis included 6 studies (chapter 2 to 7) focusing on HRQoL in children. It spans various research areas and methodologies including multilevel modeling with longitudinal HRQoL data, cost-effectiveness analysis, psychometric property assessment, and valuation of HRQoL instruments. The thesis started by exploring the current application of HRQoL, progressed to make contributions in the measurement of HRQoL for young children, and ended up by providing a value set appropriate for use for economic evaluation for children aged 2-4 years old.

Study 1 and 2 demonstrated two important applications of HRQoL, which highlighted the uniqueness of children under 5 years old in HRQoL measurement and the caveat of the lack of specific HRQoL measures for this age group. More specifically, Study 1 used data from LSAC and investigated the association between time use behaviors and HRQoL children. This study demonstrated how non-preference-weighted HRQoL is used in longitudinal studies to measure health improvement with its multidimensional attributes. The findings highlighted that the impact of time use behavior on HRQoL differs for children of different ages, especially for children aged 2-4 years old. Study 2 used data from a randomized controlled trial and evaluated the cost-effectiveness of using prednisolone to treat Bell's Palsy in children. It found that prednisolone compared to placebo has a high probability of being cost-effective in older children aged 12-17 years while not cost-effective in children younger than 12 years old. The findings provide new evidence to decision-makers regarding whether to make recommendations for using prednisolone to treat Bell's palsy in children aged 12-17 years. The different cost-effectiveness outcomes between using mapped and real utilities in the sensitivity analyses highlighted the research gap of lacking HRQoL measures with preference-weighted scoring systems in children aged under 5 years.

Study 3 and 4 focused on the measurement of HRQoL. Specifically, Study 3 investigated the impact of a variety of common conditions on child HRQoL in children aged 2-18 years old, providing pragmatic evidence for the validation of pediatric HRQoL instruments' descriptive systems. The early results from study 3 have guided the selection of patients for a large pediatric multi-instrument comparison study (P-MIC). Study 4, using data from the P-MIC study, evaluated the psychometric performance of CHU9D with guidance notes for children under 5 years. It found that the CHU9D with guidance notes was valid and reliable and supported the widespread use of it to measure HRQoL for children aged 2-4 years old.

Study 5 and 6 focused on the valuation of HRQoL. In certain scenarios, the preferences of children themselves are sought after, whereas in other instances, the preferences of the broader taxpayer population are prioritized. However, methods for eliciting preferences from children and adolescents are not as well-established, and it remains uncertain whether adolescents can reliably provide their preferences. Study 5 focused on valuation of child HRQoL by adolescents. It explored the test-retest reliability of a relatively new valuation technique, best-worst scaling (BWS). It found that adolescents (as young as 11-12 years old) can report reliable preferences for health states using BWS. This study also identified that the worst choice from BWS may be less reliable than the best choice, which contributed to the selection of a more reliable method, DCE, in the final valuation study. Study 6 focused on valuation of child HRQoL by general population adults. It aimed to value CHU9D with guidance notes for children aged 2-4 years old. It first explored whether the adult general population's preferences differ according to the child age framing of the task, finding that the general population adults' preferences didn't differ for a 10-year-old from a 2–4-year-old child. A value set based on the general adults' preferences was then developed using the pooled data (both age framing: 2-4-year-old and 10-year-olds), which is suitable for use in children 2-18 years old, given that there was no clear difference between health state preferences for the young and older age groups. With the value set developed, the measurement of HRQoL in 2-4-year-old children for economic evaluations becomes possible, which supports and encourages the use of QALYs consistently in economic evaluation throughout childhood. The newly developed value set also complements the existing value set based on adolescents' preferences in Australia.

8.2. Implications

In addition to the specific findings in individual studies, the collected body of work also provides thoughts for the measurement and valuation of child health in general. Instead of solely examining one piece of research, stepping back allows an understanding of broader patterns and connections. This marks a move towards a more comprehensive understanding, considering various viewpoints and placing research into context. The figure below demonstrates the connections between sections and studies.

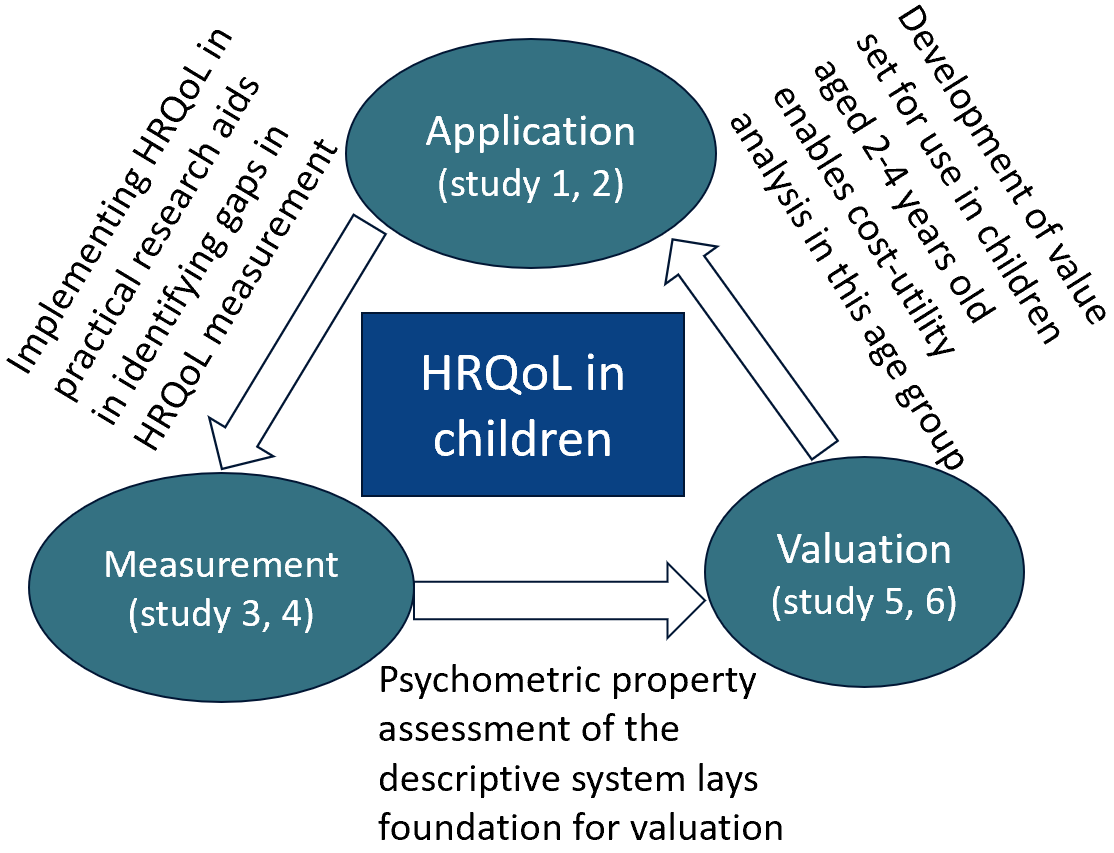


Figure 8-1 Key connections between studies

8.2.1. Child HRQoL instrument selection

For children above 5 years old, child HRQoL measures vary, and there are no consistent recommendations regarding instrument selection. Different instruments usually have distinct dimensions, response levels, recall periods, and preference elicitation techniques. Studies comparing HRQoL across multiple instruments usually find that no one instrument outperforms others consistently.[1, 2] In practical applications, it would be advantageous to include a variety of HRQoL measures and varying value sets in economic evaluations, to effectively illustrate the impact.

For children under 5 years old of age, there were few choices of HRQoL measures at the start of the PhD. This thesis made key contributions to this by providing validity evidence for CHU9D to measure HRQoL in children aged 2-4 years old (study 4) and provides social preferences from an Australian adult population to score CHU9D at this young age group (study 6). Consequently, the CHU9D can be chosen in clinical trials, population research for all children, and probably may be useful for those wishing to

understand the impact of patient care on children. It has important implications for economic evaluations wishing to include this young age group.

8.2.2. Child age in HRQoL measurement

Deciding which age groups should use separate instruments, and how different should the instrument be, is a challenge. For instance, there is evidence that developmental differences exist in the 2-4 age group compared to older (≥5 years) children.[3] However, the question remains: should this age group be provided with newly developed instruments, potentially containing entirely distinct health dimensions? Alternatively, should they utilize adapted measures from existing versions designed for older children to maintain consistency in the measurement of HRQoL across different stages of childhood? There is a trade-off: on the one hand, developing new instruments with relevant stakeholders can have domains very specific to young children; on the other hand, adapted measures from existing measures can have a health state measurement system that is consistent across ages.

Studies 1, 2, and 3 indicate the significance of age in measuring HRQoL, with variations in findings across different age groups. A consistent observation across these studies is that the HRQoL impacts, such as changes or coefficients, appear to be less pronounced in young children under 5 years old when utilizing instruments designed for older children. This suggests that employing instruments tailored for older children may not fully capture the entirety of HRQoL implications for younger children. This emphasizes the necessity of using appropriate HRQoL instruments for children under 5 years old. Study 4 contributed to the accurate and consistent measurement of HRQoL by validating the descriptive system of CHU9D with guidance notes for 2-4 years old.

8.2.3. Child age in HRQoL valuation

It is unclear whether children under 5 years old should use different value sets from older children. Using separate value sets for different stages of childhood may add further complexity to the economic evaluations, i.e., the impact of a discontinuity in HRQoL values. The problem of discontinuity in HRQoL appears when applying one value set for 2-4 year ages and applying another value set in older ages. The health state values could change significantly at the transition time point (the age cut off for the two different instruments) even if their health state description were similar.[4] This would cause HRQoL "cliff edge" which means that changes in utility could be produced that are not related to health changes. When these utility values across the life course are included in economic evaluation modeling it may result in misleading conclusions. Study 6 in this thesis found that the general population adults' preferences for child health described by CHU9D don't differ significantly between considering a 2–4-

year-old child versus older children. This means that these two age groups can use the CHU9D with the same value set, which avoids the discontinuity problem from 2–4-year-old to older children.

Age is an important consideration when generating values for HRQoL instruments, and how decision makers use this information has significant implications. There is a risk of double counting using such age-specific value sets when HTA bodies put special considerations on children under 5 years old as well. For example, the HTA bodies may consider factors like equity,[5] and may prioritize QALY gains for children under 5 years old as most societies prioritize children's well-being since they are vulnerable and are the future of society. Again, this problem was reduced as the evidence in study 6 provides support that the adult general population has no appreciable difference for a 2–4-year-old child compared with older children when valuing a child's health, and thus avoid double counting.

In summary, the findings in study 6 simplify things for those wishing to score HRQoL across childhood by avoiding "HRQoL cliff edge" and "double counting by HTA". This could have important implications for other instruments too if the findings are replicable.

8.2.4. Validity of child HRQoL measurement and valuation

It is critical that child HRQoL instruments and their scoring (value sets) are well-tested for validity. There are some important gaps that this thesis addresses.

One critical aspect of HRQoL measurement that requires improvement is responsiveness, which pertains to the instrument's ability to accurately detect meaningful changes in health status over time. Unfortunately, many studies lack conclusive evidence on responsiveness due to the omission of a reference measure, essential for monitoring significant and meaningful health changes.[1] Additionally, the small effect sizes observed in responsiveness across common general HRQoL instruments[2] raise doubts about their effectiveness in capturing HRQoL changes in both clinical practice and economic evaluations. To evaluate the responsiveness of instruments, studies need to focus on populations experiencing variations in HRQoL. Study 3 contributes valuable insights by identifying the samples undergoing significant and meaningful HRQoL changes over time. This study's findings had its immediate practical implications by informing the selection of the conditions for a pediatric multi-instrument comparison study.

Another important gap is the lack of valid instruments to measure HRQoL in young children under 5 years old. Study 4 provides new evidence about the validity of CHU9D for measuring HRQoL for

children under 5 years old. This has important implications for those wanting to measure HRQoL in young children.

The validity of adolescent values for health is also important as there is more interest in adolescent values.[6, 7] One important aspect of this, test-retest reliability, was tested in study 5. Study 5 found that adolescents could provide reliable preferences for health states. This has implications that adolescents' values could potentially be included and considered in health care decisions.

8.3. Limitations and challenges

8.3.1. Measurement of HRQoL

One limitation relates to conceptualizing HRQoL for young children. To be specific, qualitative research exploring the content validity was lacking for the CHU9D with guidance notes, with no publicly available evidence. In other words, the relevance of each item for the construct of interest and the comprehensiveness of the instrument,[8] has not been tested. Study 4 evaluated the psychometric properties of CHU9D with guidance notes. The problem is that it is still unknown if the construct of HRQoL described by the CHU9D reflects the most important things for younger children aged 2-4 years old. The developer of CHU9D recommended testing the existing option, i.e., the developed guidance notes, in quantitative psychometric property assessment studies first. However, it would be ideal to have qualitative studies investigating the relevance and appropriateness of the added guidance notes and the construct of the tool itself first before the assessment of the psychometric performance. For example, Daziel etc. 2023 published a qualitative study using the views of parents/caregivers of children aged 2 to 4 years to adapt the EQ-5D-Y instrument for appropriate use in the 2-4-year age range.[9] Future qualitative studies aimed at refining and testing CHU9D guidance notes and the construct would still be valuable.

A recall period of one day (asking about today or last night) in CHU9D may occasionally fail to capture a child's recent health status. There are cases where children were sick the day before but had recovered by the present day, making the "today" recall less indicative of very recent experiences. As evidenced by Study 3, CHU9D exhibited limited capacity to discern HRQoL impairments in health conditions with infrequent but severe outcomes, like epilepsy.[10] A recent systematic review echoed this finding. It found that symptoms were typically described as more intense, and the quality of life was perceived to be poorer when evaluated through a weekly recall compared to a one-day recall.[11] The systematic review also found diverse preferences among participants regarding recall accuracy: while some valued the precision of one-day recall, others favored seven-day recall, especially for conditions with fluctuating

symptoms or significant variability in health effects.[11] The short recall period also appears to impact CHU9D's test-retest reliability. As noted in study 5, when compared to PedsQL, which with a one-month recall period, CHU9D exhibited lower test-retest reliability. This was evident across both shorter (2-day) and longer (4-week) re-testing timeframe, especially in dimensions like 'worried' and 'sad,' which may be sensitive to acute incidents. These results highlight the need for more comprehensive exploration and future research concerning the recall period's role in HRQoL measurements.

Proxy-report for children under 5 years old might be another limitation. When measuring HRQoL, self-reporting is generally favored over proxy assessment whenever it is validly feasible,[12] as there is evidence that proxy-report and self-report can be different.[13] See detailed contents in the *Introduction, 1.2.2, self-report vs proxy report*. For children under 5 years old, proxy reports are the only option because these young children are unable to self-report. For children with developmental delays or a specific medical condition such as neurodevelopmental disorders, proxy reporting may still be relevant for older ages, and parents or carers should be consulted if their children are able to self-report. However, this approach has limitations, as there may be differences between the proxy reported HRQoL provided by the caregiver or parents and the child's actual HRQoL. It is important to validate these instruments across various pediatric populations, including those with specific medical conditions. More evidence is needed to establish clear guidelines on when and how to use self-report and proxy-report when measuring HRQoL among children.[13]

The sample (sample size and characteristics) used for testing the responsiveness of HRQoL instruments is another limitation. Although the total sample for the psychometric testing is not small (n=842) in study 4, the number for the responsiveness testing (those who finished the second test and reported change in health) is small (n: from 14 to 33 depending on the health change indicator). This can limit the statistical power of the responsiveness tests. In addition, there is no intervention employed on the sample, which resulted in small effect sizes of responsiveness and limited the ability of interpreting the responsiveness. The sensitivity analyses in this study using the variable "major health event between initial and follow-up survey" partly supplemented this limitation of lacking intervention, and those reporting worsened health when developing new illness demonstrated medium-to-large effect sizes of responsiveness even with a small sample size (n=17). These emphasize the importance of sample size and the necessity of selecting the appropriate sample with real health change to ensure valid responsiveness testing results.

8.3.2. Valuation of HRQoL

Completing a valuation task is not easy. According to a qualitative study, several factors contribute to the difficulty of the task, including unclear dimensions, perceptions of certain health states as realistic, and

challenges in interpreting and assessing health states for participants lacking relevant memories or experiences.[14] In study 6, 4.7% of participants strongly agreed, and 21.6% somewhat agreed that they found it difficult to imagine the health states experienced by a child. This difficulty may arise from the complexity of the health states and the emotional stress associated with imaging a child experiencing severe health states. This becomes more challenging when asking participants to imagine a very young child experiencing a severe health state, especially for those who had similar painful experiences themselves. Efforts have been made to avoid or reduce such problems. For example, at the very beginning of the survey, the potential risks of mental stress completing the survey were pointed out in the Plain Language Statement. Reminders were also set before the VAS valuation task (for anchoring) which involved 'dead' health state. Another problem is the difficulty in distinguishing some severity levels in the valuation tasks. In the pilot survey for study 6, participants reported that they had difficulty distinguishing between certain levels, such as "a bit" and "a little bit", "quite" and "very", "a few problems" and "some problems". This difficulty and confusion were observed among individuals whose first language is English. Every effort has been taken to mitigate this limitation, such as adding instructions for severity levels via pop-up boxes and enhancing the visibility of levels through bold fonts. The response levels perform well in the psychometric performance evaluation study (study 4). The difference may relate to the display of these response levels, which were displayed in order in the questionnaire but randomly in the valuation tasks. In the final survey, 3.4% of respondents strongly agree and 19.4% somewhat agree that they found it difficult to tell the difference between the health states. These findings underscore that careful consideration should be given to designing the valuation tasks to value health, especially for young children, and pilot tests help.

Age framing is another important factor in valuation studies. In study 6, a 10-year-old child was chosen in the valuation task for valuing CHU9D for 5–17-year-old children, compared with valuing CHU9D for 2–4-year-old child. Although 10-year-old is recommended in the international protocol for valuing EQ-5D-Y, it is unclear whether the10-year old framing can represent the whole child age group. Evidence regarding whether age matters in valuation tasks is mixed. For example, a 10 year old may not represent a 15 year old adolescent as evidenced in a think aloud study.[14] Another study examined whether the use of different child ages has an impact on the valuation of EQ-5D-Y health states, and found that except for one moderate and one severe health state, other EQ-5D-Y health states showed no significant variation in valuation when descriptions of age varied.[15] The different conclusions might be that people thought that they have difference in valuations in qualitative studies but these differences were not reflected in quantitative studies. For the 2–4-year-old, no specific age was chosen in the valuation tasks due to the narrow age range. However, it is unknown whether a different age specification for this young age group

would make a difference. Future studies employing more age definitions might be helpful to confirm the findings of study 6.

The source of preferences is a subject of ongoing debate. In the context of study 6, which focused on valuing CHU9D for children under 5 years with guidance notes, the opinions of the general adult population were gathered, with consideration for their role as taxpayers and stakeholders in the healthcare system. However, the general adult population may not all have experience with children or caring for children and may lack the ability to value child health. Thus, there are opinions that parents' viewpoints should be considered, particularly due to their direct experience with their children's health conditions. This becomes especially important when considering children under the age of 5 who are quite different from older children and adults, and individuals without children might have trouble understanding a very young child's health experiences. While current practices lean towards incorporating general adult opinions, there is a growing interest in exploring the merits of including parent perspectives or those with experience caring for ill children, considering their unique insights. Study 6 has collected data from parents and those with illness experience. Future analyses investigating the implications of utilizing different sources of preferences in valuing HRQoL measures, especially within the context of childhood health, are planned.

Another important challenge pertains to selecting an appropriate technique for valuing the health of children under 5 years old. Discrete Choice Experiment (DCE) has gained popularity in health state valuation due to its simplicity compared to traditional methods such as standard gamble (SG) and time trade off (TTO).[16] BWS could be a potential alternative,[17] although the reliability of its worst choices might be problematic according to findings in study 5. When valuing health in adolescents, BWS is less cognitive demanding than DCE and may be preferred. However, DCE, a maturer method than BWS, is preferred when valuing health for children under 5 years old who obviously cannot perform valuation tasks themselves. DCE yields latent preference values, necessitating a separate anchoring study/task to rescale utilities values to a scale of 0-1 where 1 represents full health and 0 represents dead. While DCE with duration could potentially be a solution by incorporating duration as an additional attribute (thus doesn’t require separate anchoring study), its suitability is limited in valuation studies for a population with a narrow age range. This issue arises because the attribute duration varies in years (e.g., 1y, 4y, 7y, 10y)[18], which becomes problematic as the target age range evolves over time (e.g., 2-4 years old will not be 2-4 years old after 4 or more years). In addition, DCE with duration is also more challenging to complete than DCE. Consequently, DCE with duration was not chosen for valuing CHU9D with guidance notes in study 6. Instead, DCE was employed to derive latent scales, followed by a distinct VAS

anchoring task. Emerging valuation techniques like Online Personal Utility Function (OPUF), which uses more efficient compositional elicitation methods, might offer alternatives in future studies.[19]

Using DCE to value CHU9D, which comprises 9 attributes, presents a challenge. One previous study valuing CHU9D employed a partial design with 3 overlapped attributes and 6 varying attributes plus a time duration attribute, and 13% participants reported it being very difficult making choices between health states, and about 56% stated it being difficult.[18] In study 6, a DCE was used to elicit preferences from the general adult population in Australia. Efforts were made to reduce people's cognitive burden by increasing the number of overlapped attributes to 4, bolding font for different levels between health states, and adding extra reminders for response levels in DCE choice tasks. It is reassuring that only around 5% of respondents strongly agree that the DCE tasks are difficult, indicating the feasibility of using DCE with 4 overlapped attributes and with 13 DCE tasks (one dominant task). However, there were still some logical inconsistencies in estimates. Future studies may consider further reducing the number of DCE tasks for one respondent, for example 10 DCE tasks plus 1 dominant task, to further reduce respondents' cognitive burden.

Anchoring the latent preference scales for health states for children under 5 years old poses another challenge. As mentioned above, valuation techniques involving long time duration are not suitable for valuing child health especially for children with a very narrow age range. SG might be an option, but it is difficult to conduct online as it is more difficult to understand. VAS doesn't require a time duration, which solved this problem. However, not including a time duration is not without its disadvantage. Being without a time duration might cause participants to associate severe health states with shorter duration and impact the valuation.[20] Another concern is that VAS does not consider the opportunity cost as participants don't need to trade life-years for better health.[20] However, this turns out to be an advantage for valuing health for very young children for whom people may be unwilling to sacrifice life years. Finally, VAS was chosen as the anchoring method due to its advantage of being simple and low participant burden, and suitable for valuing health for this young age group. The final anchored value for the worst health states in study 6 falls at the high end of existing value sets. Edward Webb etc. 2020 stated that "it is not uncommon for VAS values to be higher than TTO or SG based values".[20] It is also acknowledged that VAS only valued three health states (the worst health state, the best health state and the dead) to anchor the values. It would be beneficial to investigate the impact of valuing intermediate health states, although this means potentially increased respondent burden.[20]

8.4. Future work

In summary, many unresolved issues persist in the realm of health state measurement and valuation for children, particularly very young children, ensuring the necessity for future research.

8.4.1. Measurement of child health

Exploration of concepts of HRQoL for younger children, and research to understand more about trade-off between age specific and consistent HRQoL domains would be valuable. Psychometric performance comparative studies with other HRQoL instruments for young children would also help facilitate selection of instruments. Future studies testing responsiveness in larger samples with clear interventions resulting in health state changes is valuable too.

8.4.2. Valuation of child health

It is worthwhile to explore the impact on value sets using different valuation techniques. In addition, the exploration of the impact of different anchoring methods (e.g., including dead in DCE) on value sets would be valuable as well. Finally, future work exploring the difference between parents compared with general adults or non-parents would be valuable and is planned through existing data collected in Study 6.

8.4.3. Application of HRQoL

It is worthwhile to apply the instruments with consistent descriptive system across childhood (e.g., CHU9D) in future studies to accurately measure and monitor the longitudinal trend and change of preference weighted HRQoL(i.e., natural history) across the whole childhood including 2-4 years old. Research to understand the value of young child HRQoL in clinical care is of growing interest. In clinical contexts, routine HRQoL measurement during medical appointments aids in identifying health issues. More work is required to understand how routine HRQoL measurement can benefit young children, families and quality of care.

8.5. Conclusions

This thesis featured six distinct health economics studies employing diverse methodologies, collectively enriching the literature in measuring and valuing HRQoL in children, particularly the young children under 5 years of age.

Specifically, the lack of a HRQoL utility measure for children under 5 years old may compromise the ability of economic evaluations to inform resource allocation across the whole population. This thesis

showed that the CHU9D proxy version with guidance notes is both valid and reliable for assessing HRQoL in children aged 2-4 years. As preferences for health as described by the CHU9D do not significantly differ between 2-4-year-old children and 10-year-olds in the general adult population, a value set has thus been developed based on pooled preferences. This new value set is suitable to measure health state values for children aged 2-4 years, thereby addressing a critical gap.

8.6. Reference

[1] J. Kwon *et al.*, "Systematic Review of the Psychometric Performance of Generic Childhood Multi-attribute Utility Instruments," *Applied Health Economics and Health Policy,* 2023/05/03 2023.

[2] R. Jones *et al.*, "Comparative Psychometric Performance of Common Generic Paediatric Health-Related Quality of Life Instrument Descriptive Systems: Results from the Australian Paediatric Multi-Instrument Comparison Study," *PharmacoEconomics,* 2023/11/13 2023.

[3] L. Berk, *Child development*. Pearson Higher Education AU, 2015.

[4] N. J. Devlin *et al.*, "Using Age-Specific Values for Pediatric HRQoL in Cost-Effectiveness Analysis: Is There a Problem to Be Solved? If So, How?," (in eng), *Pharmacoeconomics,* vol. 41, no. 10, pp. 1165-1174, Oct 2023.

[5] R. Cookson, S. Griffin, O. F. Norheim, and A. J. Culyer, *Distributional cost-effectiveness analysis: quantifying health equity impacts and trade-offs*. Oxford University Press, 2020.

[6] J. Ratcliffe, E. Huynh, K. Stevens, J. Brazier, M. Sawyer, and T. Flynn, "Nothing About Us Without Us? A Comparison of Adolescent and Adult Health-State Values for the Child Health Utility-9D Using Profile Case Best-Worst Scaling," (in eng), *Health Econ,* vol. 25, no. 4, pp. 486-96, Apr 2016.

[7] J. Ratcliffe *et al.*, "Valuing the Child Health Utility 9D: Using profile case best worst scaling methods to develop a new adolescent specific scoring algorithm," (in eng), *Soc Sci Med,* vol. 157, pp. 48-59, May 2016.

[8] L. B. Mokkink *et al.*, "COSMIN Study Design checklist for Patient-reported outcome measurement instruments," ed, 2019.

[9] K. Dalziel *et al.*, "A Qualitative Investigation to Develop an Adapted Version of the EQ-5D-Y-3L for Use in Children Aged 2-4 Years," (in eng), *Value Health,* vol. 26, no. 10, pp. 1525-1534, Oct 2023.

[10] X. Xiong, K. Dalziel, L. Huang, B. Mulhern, and N. Carvalho, "How do common conditions impact health-related quality of life for children? Providing guidance for validating pediatric preference-based measures," *Health and Quality of Life Outcomes,* vol. 21, no. 1, p. 8, 2023/01/25 2023.

[11] T. Peasgood, J. M. Caruana, and C. Mukuria, "Systematic Review of the Effect of a One-Day Versus Seven-Day Recall Duration on Patient Reported Outcome Measures (PROMs)," *The Patient - Patient-Centered Outcomes Research,* vol. 16, no. 3, pp. 201-221, 2023/05/01 2023.

[12] J. Brazier, J. Ratcliffe, J. Saloman, and A. Tsuchiya, *Measuring and valuing health benefits for economic evaluation*. OXFORD university press, 2016.

[13] C. Mpundu-Kaambwa *et al.*, "A Systematic Review of International Guidance for Self-Report and Proxy Completion of Child-Specific Utility Instruments," *Value in Health,* vol. 25, no. 10, pp. 1791-1804, 2022/10/01/ 2022.

[14] V. Reckers-Droog, M. Karimi, S. Lipman, and J. Verstraete, "Why Do Adults Value EQ-5D-Y-3L Health States Differently for Themselves Than for Children and Adolescents: A Think-Aloud Study," *Value in Health,* vol. 25, no. 7, pp. 1174-1184, 2022/07/01/ 2022.

[15] J. G. A. Retra, B. A. B. Essers, M. A. Joore, S. M. A. A. Evers, and C. D. Dirksen, "Age dependency of EQ-5D-Youth health states valuations on a visual analogue scale," *Health and Quality of Life Outcomes,* vol. 18, no. 1, p. 386, 2020/12/12 2020.

[16] H. Wang, D. L. Rowen, J. E. Brazier, and L. Jiang, "Discrete Choice Experiments in Health State Valuation: A Systematic Review of Progress and New Trends," (in eng), *Appl Health Econ Health Policy,* vol. 21, no. 3, pp. 405-418, May 2023.

[17] E. Lancsar, J. Louviere, C. Donaldson, G. Currie, and L. Burgess, "Best worst discrete choice experiments in health: Methods and an application," *Social Science & Medicine,* vol. 76, pp. 74-82, 2013/01/01/ 2013.

[18] D. Rowen, B. Mulhern, K. Stevens, and J. H. Vermaire, "Estimating a Dutch Value Set for the Pediatric Preference-Based CHU9D Using a Discrete Choice Experiment with Duration," (in eng), *Value Health,* vol. 21, no. 10, pp. 1234-1242, Oct 2018.

[19] P. P. Schneider, B. van Hout, M. Heisen, J. Brazier, and N. J. W. O. R. Devlin, "The Online Elicitation of Personal Utility Functions (OPUF) tool: a new method for valuing health states," vol. 7, p. 14, 2022.

[20] E. J. D. Webb, J. O'Dwyer, D. Meads, P. Kind, and P. Wright, "Transforming discrete choice experiment latent scale values for EQ-5D-3L using the visual analogue scale," (in eng), *Eur J Health Econ,* vol. 21, no. 5, pp. 787-800, Jul 2020.