# Unpacking Variability: exploring the sources of utility differences between pediatric multi-attribute utility instruments

Ashwini De Silva[a], Nancy Devlin[a], Richard Norman[b], Tianxin Pan[a], Kim Dalziel[a], Tessa Peasgood[a,c]

[a] Melbourne Health Economics, Centre for Health Policy, Melbourne School of Population and Global Health, University of Melbourne, Melbourne, Australia
[b] School of Population Health, Curtin University, Perth, Australia
[c] Division of Population Health, School of Medicine and Population Health, University of Sheffield, Sheffield, United Kingdom

## Acknowledgments

The data reported in this paper were part of the QUOKKA pediatric multi-instrument comparison study (P-MIC). The authors acknowledge and thank Renee Jones for her assistance in accessing the data.

## Funding

Ashwini De Silva's PhD is funded by EuroQol Research Foundation grant number 348-PHD.

*Supplementary files are available on request.

**Objectives:** Several pediatric multi-attribute utility instruments (MAUIs) are available, which differ in their measurement and valuation properties. Measurement differences include variations in dimensions, severity levels and descriptions, and recall periods. Valuation differences are known to arise from differences in preference elicitation and anchoring methods, statistical modelling approaches, and whose preferences and what perspective is adopted in the preference elicitation tasks. Identifying and quantifying the impact these differences have on utility values is useful for researchers aiming to improve measurement and valuation. It can also assist policymakers to select MAUIs for a specific context, improve the comparability of findings across studies and enable the mapping of MAUIs onto a common scale for cost-utility analysis comparisons. This study aims to understand (1) what drives the differences in utilities between pediatric MAUIs, (2) whether the relative importance of these sources of difference are consistent across 3 countries, and (3) if relative importance of the sources of differences are consistent between self-reported vs proxy-reported descriptive system responses.

**Methods:** Survey responses for the EQ-5D-Y-3L and CHU9D instruments, collected from children aged 7-18 years (self-report) or their parents/ caregivers (proxy-report), were obtained from QUOKKA's Australian Pediatric Multi-Instrument Comparison study. Utilities were derived using the Australian, Chinese and Dutch value sets for both instruments. To measure the sources contributing to utility differences between the instruments, the three-step attributional regression approach developed by Richardson et al. (2015) was applied. Additional robustness tests were included. This approach disaggregates the differences between utilities from different MAUIs into three key components: (1) scale effect, caused by the valuation scales of the two instruments, (2) descriptive system effect, arising from variations in the questions and response categories in each instrument and (3) discrepancy effect, representing the residual variation which may reflect the differences in relative weights allocated to dimension and levels in the value set algorithm.

**Results:** A total of 4,099 children aged 7-18 years (self-report) and 1,182 parents or caregivers (proxy-report) completed the survey. The average (absolute) differences in utilities between EQ-5D-Y-3L and CHU9D were 0.196 (Australia), 0.138 (China) and 0.114 (Netherlands) across the 3 countries. The differences between utilities were primarily attributed to the scale and descriptive system, varying by value set and reporting perspective. Using the Australian value sets, 62% (self-reports) and 60% (proxy-reports) of the difference between the two MAUIs were attributable to the scale. For the Chinese value sets, the scale accounted for 53% of self-reports, while 52% of proxy-reports were attributed to the descriptive system. In contrast, the Dutch values set showed higher contributions from the descriptive system:74% (self-reports) and 80% (proxy-reports).

**Conclusion:** Differences in utilities between the two MAUIs were driven by both scale and descriptive system effects. The scale effect dominated when the range of the utilities captured by the instrument differed significantly when the utility ranges were similar, the descriptive system effect dominated. Results highlight the importance of choices around the descriptive system for pediatric MAUIs and the methods applied in valuation; on the contrary, differences in the relative weights of domains/levels are of less importance.

# 1. Introduction

## 1.1 Introduction to Multi-attribute Utility Instruments (MAUIs)

Utilities represent societal preferences for different health states which are quantified on a scale anchored at 0-1, dead-full health. A common approach to estimating utility values involves using multi-attribute utility instruments (MAUI) which consist of two parts. First, a descriptive system which consists of several dimensions represented by one or more items that have several response categories. Second, a utility-based scoring algorithm which converts the response in the descriptive system into a utility value, based on a value set derived from the preferences of the general population.

There are many different MAUIs, which have been developed in different countries, such as EQ-5D, HUI3, SF-6D, 15-D, and AQoL-8D (Richardson et al., 2014). In the pediatric context, a corresponding growth in MAUIs has occurred, including the EQ-5D-Y, Child Health Utility 9D (CHU9D), Assessment of Quality of Life-6 Dimensions (AQoL-6D) -Adolescent and the Health Utility Index Mark 2 (HUI2) (Kwon et al., 2022).

## 1.2 Key Differences Between MAUIs

There are several key differences between MAUIs and these differences can be classified into measurement properties, respondent characteristics and valuation properties.

### 1.2.1 Measurement Properties

Differences related to measurement concern how MAUIs capture the patient's health. The first difference is due to differences in dimensions. Each MAUI consists of several dimensions, which might differ between instruments. For example, the EQ-5D-Y-3L and Y-5L instruments consists of five dimensions (Wille et al., 2010) and CHU9D consists of nine dimensions (Ratcliffe et al., 2016). CHU9D includes dimensions around sleep and schoolwork/ homework which are not explicitly included in the EQ-5D-Y. Second, even if the dimensions are similar, the number of response options and the severity labels attached to both mild and severe states might differ. For example, the worst level for mobility in the EQ-5D-Y-3L is a 'lot of problems walking about' whereas in the AQoL 6D- Adolescent version (Moodie et al., 2010) the worst level is being 'bedridden'. Third, the length of the recall period differs between instruments. The recall period is the time frame over which each individual is asked to consider their health state when completing the instrument. For example, the recall period for EQ-5D-Y is 'today' whereas it is the 'past week (7 days) for the AQoL-6D-Adolescent.

### 1.2.2 Respondent Characteristics

Differences in MAUIs may also vary based on who completes the instrument. The instrument may be completed by the individual themself (self) or by a proxy respondent. A systematic review by Khadka et al. (2019) identified several studies with self and proxy reports to derive child utility values and found that the values derived from self and proxy reports do differ but the pattern in direction was not consistent.

### 1.2.3 Valuation Properties

There are eight key differences in how utility values for MAUIs are estimated. The valuation method may differ based on (1) the elicitation method(s) used, (2) modes of valuation data collection, (3) perspective adopted, (4) the duration(s) adopted in the elicitation task, (5) the methods of reaching point of equivalence (where relevant), (6) methods for selecting which health states to value and (7) modelling and analysis approaches. Additionally, (8) the target population whose preferences are sought may differ.

The elicitation method refers to the method used to measure the preferences or health states. There are several elicitation methods such as time trade-off (TTO), discrete choice experiment (DCE), visual analogue scale (VAS) and best-worst scaling (BWS). Different elicitation methods may lead to different values (Brazier et al., 2009). Additionally, there are different modes of data collection such as face-to face, online interviews and self-complete (Mulhern et al., 2019). Differences in valuation may also occur due to the perspectives adopted in the preference elicitation tasks. For example, the CHU9D value set derived for Australia (Ratcliffe et al., 2016) involved adolescents aged 11-17 years valuing from their own perspective, whereas the value set for the Netherlands (Rowen et al., 2018) involved adults valuing from an adults' perspective. A recent systematic review by De Silva et al. (2024) found that utility values vary by perspective. The elicited values may also differ based on the durations used in the choice task. A review by Wang et al. (2023) identified 29 DCE studies with duration and the range of the duration varied from 2 months to 15 years. In relation to TTO, Attema et al. (2013) found that most studies use 10- or 20-year time frames and that values varied based on the time frame. They also identified different iteration procedures such as bisection procedure, top-down titration and a 'ping-pong' approach to reach the point of equivalence. Studies have concluded that the values could differ by iteration procedure (Jakubczyk et al., 2023).

Instruments describe many possible health states. For example, EQ-5D-Y-3L has 243 health states. Therefore, it is not feasible to ask respondents to value all the possible health states, and studies select a sub-set of states to value. The methods used to select those states can vary, as discussed by Attema et al. (2013) and Wang et al. (2023).

Various statistical models are used to interpolate values from a specific sub-set of selected states (Mulhern et al., 2019) for which preferences have been obtained. For example, a Garbage class mixed logit model[1] was used to derive the EQ-5D-Y-3L Australian value set (Pan et al., 2024) whereas a Hybrid model with an A3 constant to capture the gap between state 33333 and other states was used to derive the Chinese value set (Yang et al., 2022).

Finally, the differences between value sets may occur due to differences in the target populations completing the valuation tasks. Values for pediatric instruments may come either from adults imagining states in a child (Pan et al., 2024), or from children themselves (Ratcliffe et al., 2016). Respondents are from different countries, and different population groups within countries. Wang et al. (2023) identified that most of the DCE sampled the general population however, there were studies sampling patients, children/ adolescents and home care and aged people.

### 1.3 Empirical Evidence of utility differences

Existing literature has compared utilities between different instruments. For example, Xie et al. (2007) compared the EQ-5D-3L and SF-6D among patients with knee osteoarthritis and reported the EQ-5D utilities were bimodal and SF-6D utilities were normal. Whitehurst et al. (2014) compared the EQ-5D-3L and SF-6D utilities using an identical DCE approach. They concluded that since the valuation method was similar, the differences in utilities between EQ-5D and SF-6D (which was 0.253) was attributable to the difference in their descriptive systems. Richardson et al. (2015) investigated the reasons for the differences between the utilities across six different MAUIs and posited three main reasons why these differences may occur i.e., (1) due to the descriptive system, (2) due to the measurement scale (caused by the valuation scales of the two instruments) and (3) due to what they describe as a micro-utility effect which is the effect after adjusting for the descriptive system and scale effect which indicates the non-linearities between the utilities. They conducted Ordinary Least Square (OLS) regression to analyse the differences between utilities and estimate the proportion of differences attributable to each of these three causes. For example, when comparing EQ-5D vs the SF-6D the mean absolute difference was 0.111, with the descriptive system effect contributing 77%, micro-utility effect accounting for 20% and the scale effect accounting for 3% of the total difference.

### 1.4 Aims

This study aims to replicate and update the methods proposed by Richardson et al. (2015) to unpack the variability in the differences in utilities between two pediatric MAUIs. To achieve this, this study aims to

[1] Garbage Class Mixed Logit model is designed using Bayesian methods to identify low data quality data.

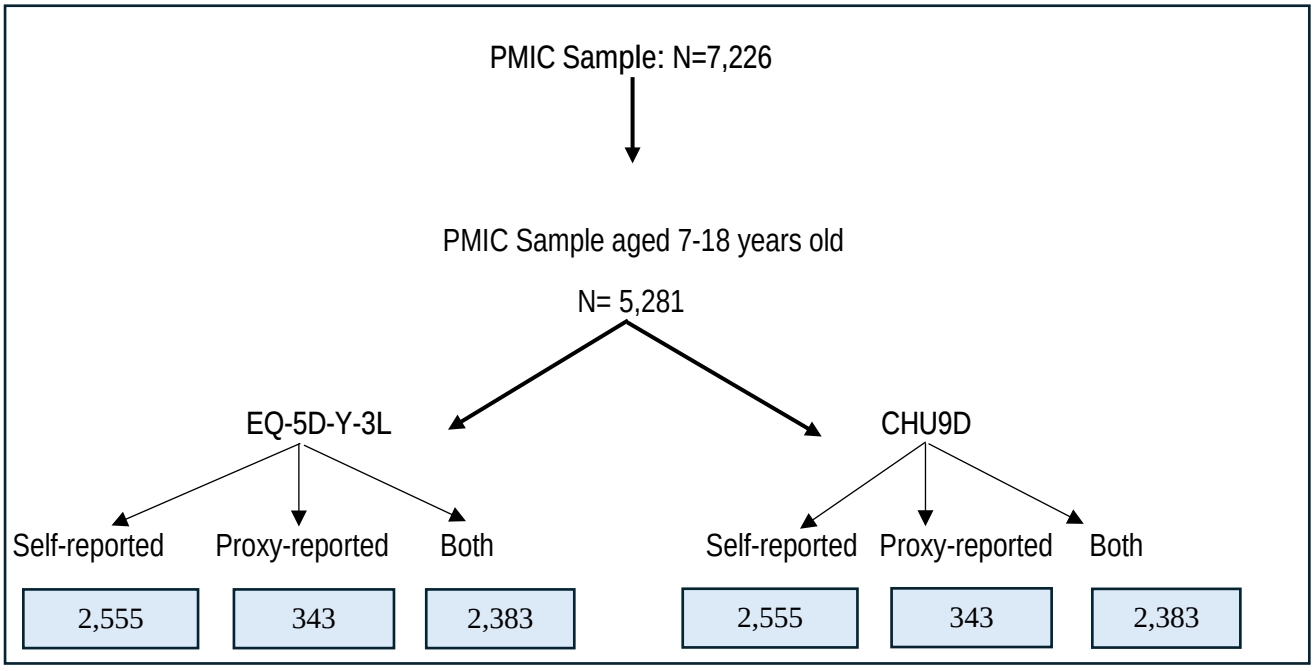
identify, (1) what drives the differences in utilities between two pediatric MAUIs, (2) whether the relative importance of the sources of these differences are common across countries, and (3) whether the relative importance of the sources of differences are consistent between self-reported vs proxy-reported descriptive system responses.

## 2. Methods

### 2.1 Study Sample

A pediatric multi-instrument comparison study (P-MIC) was carried out in Australia as part of the QUOKKA research programme (Jones et al., 2021). This study recruited parents, caregivers or guardians of a child aged 2-18 years to assess the validity and reliability of a range of health-related quality of life (HRQoL) instruments. Participants were recruited through the Royal Children's Hospital (RCH) in Melbourne, Australia or via an online survey company. For this study, survey data from Data Cut 3 (June 2023) of parents/ caregivers of children aged 7-18 years (proxy-reported), children aged 7-18 years (self-reported), including initial survey responses for the EQ-5D-Y-3L and CHU9D instruments were used. The sample flow chart is shown in Figure 1.

Figure 1. Sample Flow Chart




### 2.2 Instruments

The EQ-5D-Y-3L has, three response levels: no problems, some problems and extreme problems for each of the five dimensions (Table 1). The Health states using the EQ-5D-Y-3L could be represented by

five digits representing the responses for each of the questions e.g. 23212. The best health state is described as 11111 and the worst as 33333.

The CHU9D is based on nine dimensions (see Table 1) with five response categories represented by five digits from 1-5. CHU9D health states can be described as 231254231. The best health state is described as 111111111 and the worst as 555555555. The descriptive systems and the content of these instruments are described in Table 1.

Table 1. Comparison of the dimensions and content of pediatric MAU instruments

| | CHU9D | EQ-5D-Y-3L |
|---|---|---|
| Target Age | 7-17 years | 8-15 years |
| Recall Period | Today | Today |
| Number of items | 9 | 5 |
| Number of response levels | 5 | 3 |
| Number of different health states | 1,953,125 | 243 |
| Physical health items | | |
| • Mobility | | x |
| • Self-care | x | x |
| • Usual activities/ schoolwork/ work | x x | x |
| • Pain/ discomfort | x | x |
| Emotional and Social items | | |
| • Worried, sad, unhappy | x x | x |
| • Tired, annoyed | x x | |
| • Sleeping | x | |

## 2.3 Value Set Characteristics

A single utility can be assigned to each health state described by the EQ-5D-Y-3L and CHU9D using a value set, as described in X. A value set is developed using general population preferences in a valuation study. Value sets could be different based on countries, populations, perspectives and valuation methods. This study will use the Australian, Chinese and Dutch value sets applied to the Australian data to derive utilities for each participant in the sample. These three countries were selected because value sets are available for both pediatric MAUIs in each country and they represent a wide range of utility values. A comparison and characteristics of the three value sets are presented in Table 2.

Table 2. Comparison and characteristics of the EQ-5D-Y-3L and CHU9D value sets for Australia, China and Netherlands

| | EQ-5D-Y-3L | | |
|---|---|---|---|
| | Australia (Pan et al., 2024) | China (Yang et al., 2022) | The Netherlands (Roudijk et al., 2022) |
| Perspective | Adults from a child's perspective | Adults from a child's perspective | Adults from a child's perspective |
| Sample size | DCE: 1,002<br>TTO: 268 | DCE: 1,058<br>TTO: 418 | DCE: 959<br>TTO: 197 |
| Sampling method | Quota sampling | Quota sampling | Quota sampling |
| Mode of administration | Online via video conferencing | Face-to-face online interview | TTO: online via video conferencing<br>DCE: Online survey |
| Number of health states valued | 52 | 28 | 18 |
| Valuation method | cTTO and DCE | cTTO and DCE | cTTO and DCE |
| Modelling approach | Garbage class mixed logit model | Hybrid model with A3 constant | Mixed logit model |
| Anchoring (Anchoring methods includes anchoring on to the worst health state, mapping different valuation methods) | Mapping approach without a constant: modelling the relationship between the predicted DCE values from the garbage class MIXL model and the mean observed cTTO values | Fit the DCE and TTO data jointly in a hybrid model | Mapping approach without a constant: mapping the mean observed cTTO values onto the predicted DCE values without specifying the intercept |
| Range of values (Scale) | [0.142,1] | [-0.089,1] | [-0.218,1] |
| | CHU9D | | |
| | Australia (Ratcliffe et al., 2016) | China (Chen et al., 2019) | The Netherlands (Rowen et al., 2018) |
| Perspective | Adolescents-self | Adolescents-self | Adults-self |
| Sample size | 2,076 | BWS: 902<br>TTO:38 | 1,276 |
| Sampling method | Random sampling | Convenience sampling | Not mentioned |
| Mode of administration | Online | Self-completed | Online |
| Valuation method | BWS+TTO<br>(TTO study: (Ratcliffe et al., 2015)) | BWS+TTO | DCE |
| Anchoring method | Anchored to the QALY scale using TTO health states (Mapping) | Anchored to the QALY scale using TTO health states (Mapping) | Anchored using a DCE consistent model: DCE with duration |
| Number of health states valued | 10 | 5 | 408 different health states |
| Range of values (Scale) | [-0.1059,1] | [0.0563,1] | [-0.568,1] |

TTO: Time Trade-Off, DCE: Discrete Choice Experiment, cTTO: Composite Time Trade-Off, BWS: Best Worst Scaling

All three of the EQ-5D-Y-3L value sets used TTO and DCE. The Australian (Pan et al., 2024) and Dutch (Roudijk et al., 2022) value sets developed for the EQ-5D-Y-3L used a mixed logit model whereas the value set for China (Yang et al., 2022) used a hybrid model with a constant term and an A3 term to capture the gap between the value for health state 33333 and other states. Unlike in Australia, in the Chinese and Dutch value sets the value for the state 33333 falls below zero. Both the Australian (Ratcliffe et al., 2016) and Chinese (Chen et al., 2019) CHU9D value sets used best worst scaling to elicit preferences. The value for the pits state falls below 0 for the Australian and Dutch (Rowen et al., 2018) CHU9D value sets.

## 2.4 Statistical Analysis

In this study we replicate and update the methodology used by Richardson et al. (2015). Richardson et al. (2015) investigated pairwise differences in utilities derived from six MAUIs. Specifically, we follow their analysis approach however, we make several modifications both to the regression model and the terminology.

### 2.4.1 Amendments to Richardson's Regression Approach

The first modification is to conduct diagnostic tests to check several assumptions made in the Ordinary Least Squares regression (OLS). More detail on the diagnostic tests are provided in the section below.

The second modification is to conduct the analysis using the same instrument but different value sets (for example, analyze the differences in utilities derived from the EQ-5D-Y-3L_Australia vs EQ-5D-Y-3L_China) to analyze the three effects. This modification will allow us to confirm that the OLS regression approach does work to identify the different sources. The results of these analysis will be available in Supplementary materials.

### 2.4.2 Diagnostic Tests

The Ordinary Least Squares (OLS) regression has several limitations: First, it only produces the best linear unbiased estimates when a number of stringent assumptions are met. The second limitation is that the regression may predict values outside the range of data. This will be tested in step 5 where the utility from CHU9D and the fitted value of CHU9D are transformed onto the same scale. In this step, we will test if the predicted values $U_j^{\wedge}(u_i)$ and $V_j^{\wedge}(u_i)$ are on the same scale as the EQ-5D-Y-3L. Table 1 in Supplementary materials mentions the results of the diagnostic tests and the problems caused by the data not satisfying the underlying assumptions of OLS.

(1) Normality

The residuals should follow a normal distribution with a zero mean in OLS regression. Normality can be checked by plotting a histogram of the residuals or by performing a Shapiro-Wilk test. If the p-

value is greater than the significance level of 5% ($p > 0.05$) the null hypothesis is not rejected and conclude that the residuals are judged to be normally distributed.

(2) No heteroskedasticity

In OLS regression, the variance of errors should be constant for all observations. If there is no constant variance, this refers to heteroskedasticity. This will be tested using the Breusch-Pagan test. If the probability of the test statistic is greater than 5% ($p > 0.05$), the null hypothesis is not rejected suggesting that the residual variance is constant.

$$Var\ (\varepsilon|X) = \sigma^2$$

(3) No endogeneity

The independent variables should not be correlated with the error terms. A correlation test will be conducted between the residuals and the independent variable and if the correlation is above 0.5 it will be considered a strong correlation.

$$Cov\ (X, \varepsilon) = 0$$

The third limitation of OLS is the sensitivity to outliers. We drew box plots to identify if there are any outliers in the original data (Table 6).

2.4.3 Regression Approach

Results are produced using a 10-step approach, as detailed below.

Step #1. Calculate utilities for both instruments using the value sets

Utilities will be calculated for both self and proxy reports using the Australian, Chinese and Dutch value sets.

Utility from EQ-5D-Y-3L: $U_{EQ}$

Utility from CHU9D: $U_{CHU}$

The letters 'EQ' indicates the instrument EQ-5D-Y-3L and 'CHU' indicates the CHU9D.

Step #2. Calculate the rank order score (R)

In the case of the EQ-5D-Y-3L; for the mobility dimension the best response (1, no problems) will be given a score of 3 and the worst response (3, a lot of problems) will be given a score of 1. If the participant responded to a health state as 11213, the rank order score would be, 3+3+2+3+1= 12. $R_{min}$, $R_{max}$ are the minimum and maximum 'rank order' scores which are obtained from the instrument.

$R_{max}$ would be 3x5= 15 and $R_{min}$ would be 1x5 =5.

If we consider CHU9D: for worried dimension the best response (1, no problems) would be given a score of 5 and the worst response (5, extreme problems) would be given a score of 1. If the participant responded to a health state as 543112345, the rank order score would be, 1+2+3+5+5+4+3+2+1= 26.

$R_{max}$ would be 5x9= 45, $R_{min}$ is 1x9= 9.

Higher rank order scores represent better quality of life. Within these scores all domains are given equal weight and a move between any adjacent response levels is given the same weight.

Step #3. Transform the rank order score for each instrument to a scale from 0-1 to obtain a re-scaled rank order score-S

$$S_i = (R_i - R_{min})/ (R_{max} - R_{min}) \quad (1)$$

$S_{EQ}$ for EQ-5D-Y-3L health state 11213 would be= (12-5)/(15-5)= 0.7

$S_{CHU}$ for CHU9D health state 543112345 would be= (26-9)/ (45-9)= 0.47

Both instruments are now scored using the same scale (0-1) scale. Note that although this is a 0 to 1 scale it does not represent a utility, it is a re-scaled rank order score based on treating all domains and all movements between response levels as equally weighted.

Step #4. Subject $\mathbf{S_{EQ}}$ and $\mathbf{S_{CHU}}$ scores to a linear transformation and calculate values $\mathbf{\hat{V}^2_{EQ}}$ and $\mathbf{\hat{V}_{CHU}}$

The values ($\hat{V}_{EQ}$ and $\hat{V}_{CHU}$) are calculated by getting the predicted values from Eq. 2 and 3. OLS regression will be conducted to estimate, values $\hat{V}_{EQ}$ and $\hat{V}_{CHU}$ through the two regressions. The value $\hat{V}_{EQ}$ will be predicted through the regression in Eq 2. based on the re-scaled rank score $S_{EQ}$. The value $\hat{V}_{CHU}$ will be predicted through the regression in Eq. 3 based on the re-scaled rank score $S_{CHU}$. The value ($\hat{V}$) is an estimate of the utility score that can be predicted by the re-scaled rank score alone.

$$U_{EQ} = a + bS_{EQ} + res_{EQ} \quad (2)$$

$$U_{CHU} = a + bS_{CHU} + res_{CHU} \quad (3)$$

[2] A modification to the predicted value in the Richardson's approach was made which will be indicated by '^' symbol.

Step #5. Transform the utilities and fitted values onto a common scale

For example, let's say we are transforming the utilities ($U_{CHU}$) and values ($\hat{V}_{CHU}$) from CHU9D onto the scale of EQ-5D-Y-3L ($U_{EQ}$). Regress $U_{EQ}$ firstly on $U_{CHU}$ and secondly on $\hat{V}_{CHU}$. The dependent variable would be the utilities from the EQ-5D-Y-3L.

$$U_{EQ} = a_1 + b_1 U_{CHU} + res_1 \quad (4)$$

$$U_{EQ} = a_2 + b_2 \hat{V}_{CHU} + res_2 \quad (5)$$

Then predict $\hat{U}_{CHU}$ ($u_{EQ}$) and $\hat{V}_{CHU}$ ($u_{EQ}$) through the below two regression models. In this step a test is conducted to check if the predicted values: $\hat{U}_{CHU}$ ($u_{EQ}$) and $\hat{V}_{CHU}$ ($u_{EQ}$) fall within the appropriate range ($U_{EQ}$).

Rotated utilities and values for CHU9D is obtained from the regression.

$$\hat{U}_{CHU}\ (u_{EQ}) = a_1 + b_1 U_{CHU} \quad (6)$$

$$\hat{V}_{CHU}\ (u_{EQ}) = a_2 + b_2 \hat{V}_{CHU} \quad (7)$$

$\hat{U}_{CHU}$ ($u_{EQ}$): Predicted utility from CHU9D transformed to the same scale as the EQ-5D-Y-3L

$\hat{V}_{CHU}$ ($u_{EQ}$): Predicted value from CHU9D transformed to the same scale as the EQ-5D-Y-3L

The three diagnostic tests (Normality, heteroskedasticity and endogeneity) will be run in steps 4 and 5.

Step 6, 7, 8,9 and 10 includes measuring the three components affecting the utility differences.

Step #6. Calculate Pairwise difference in utilities: A

Absolute utility difference between the utilities calculated from the EQ-5D-Y-3L and CHU9D instruments.

$$A = U_{EQ} - U_{CHU}$$

Step #7. Calculate Scale-free differences in utility: B

Absolute differences in utility measured on a common scale (scale of EQ-5D-Y-3L). This difference eliminates the effect of the measurement scale.

$$B = U_{EQ} - \hat{U}_{CHU}\ (u_{EQ})$$

Step #8. Calculate Scale effect: C

From the absolute difference between utilities (A), the amount of difference explained by the scale. This is calculated by first calculating the absolute differences between the utilities from EQ-5D-Y-3L and

CHU9D and then subtracting the absolute utility differences between the utility from EQ-5D-Y-3L and the utility of CHU9D transformed to the same scale of the EQ-5D-Y-3L.

$$C = [U_{EQ} - U_{CHU}] - [U_{EQ} - \hat{U}_{CHU}(u_{EQ})] \text{ or } A\text{-}B$$

Step #9. Calculate Descriptive system effect: D

The absolute scale-free difference in fitted values attributable to the differences in the descriptive system.

$$D = \hat{V}_{EQ} - \hat{V}_{CHU}(u_{EQ})$$

Step #10. Calculate Discrepancy effect: E

The effect of the utility formula after taking account of the scale effect and descriptive system effect (non-linearity).

$$E = [U_{EQ} - \hat{U}_{CHU}(u_{EQ})] - [\hat{V}_{EQ} - \hat{V}_{CHU}(u_{EQ})] \text{ or } B\text{-}D$$

Hence the absolute utility difference between the instruments (A) is comprised of the Scale Effect (C) plus the Descriptive Systems Effect (D) plus the Discrepancy Effect (E).

Table 3. Explanation of the terms of the three components

| Terminology | Definition |
|---|---|
| (1) The Scale Effect (C) | The effect caused by differences in the measurement scale of the two instruments. *Assumptions: The scale effect might be affected depending on the scale of the value set* |
| (2) Descriptive system Effect (D) | The effect caused by the design (domains included and the number of response options) and the content of each of the instruments. *Assumptions: The descriptive system effect might be affected depending on the dimension distribution and response patterns* |
| (3) Discrepancy Effect (E) | The residual effect after accounting for the scale and descriptive systems effect. *Assumptions: Non-linearities in the relationship between utilities may be the cause of the discrepancy effect* |

We provide a descriptive summary of sample characteristics, the distribution of responses and range of utilities for each instrument.

# 3. Results

## 3.1 Sample Characteristics and Response Patterns

A total of 4,099 children aged 7-18 years completed the survey from their own perspective (self-report) and 1,182 parents or caregivers completed the survey from a child's perspective (proxy-report). Table 4 presents the sociodemographic characteristics of the children and parents who participated in the survey.

Table 4. Sociodemographic characteristics of children and parents in the P-MIC sample

| | N(%) |
|---|---|
| Child Characteristics | |
| Age: 7-10 yrs | 1,715 (42%) |
| 11-14 yrs | 1,300 (32%) |
| 15-18 yrs | 1,084 (26%) |
| Gender: Male | 2,112 (52%) |
| Female | 1,921 (47%) |
| Other | 66 (1%) |
| Special health care needs | 1,792 (47%) |
| Presence of health conditions | 1,786 (44%) |
| Parent/ caregiver characteristics | |
| Age: 18-25 yrs | 16 (1%) |
| 26-35 yrs | 200 (17%) |
| 36-45 yrs | 511 (43%) |
| 46-60 yrs | 424 (36%) |
| >60 yrs | 31 (3%) |
| Gender: Male | 235 (20%) |
| Female | 940 (79%) |
| Other | 7 (1%) |

Figures 2 and 3 present the distribution of child-self and parent-proxy responses for the EQ-5D-Y-3L and CHU9D dimensions respectively. Using the EQ-5D-Y-3L, self-reports indicated higher prevalence in the worried, sad or unhappy (WSU) dimension, while proxy-reports indicated more issues in usual activities (UA). Using the CHU9D, self-reports highlighted the schoolwork/ homework and tired dimension with the most problems whereas for proxy-reports it was schoolwork/ homework and activities with the most problems.

Figure 2. Distribution of response frequencies using EQ-5D-Y-3L: A comparison between Self and Proxy

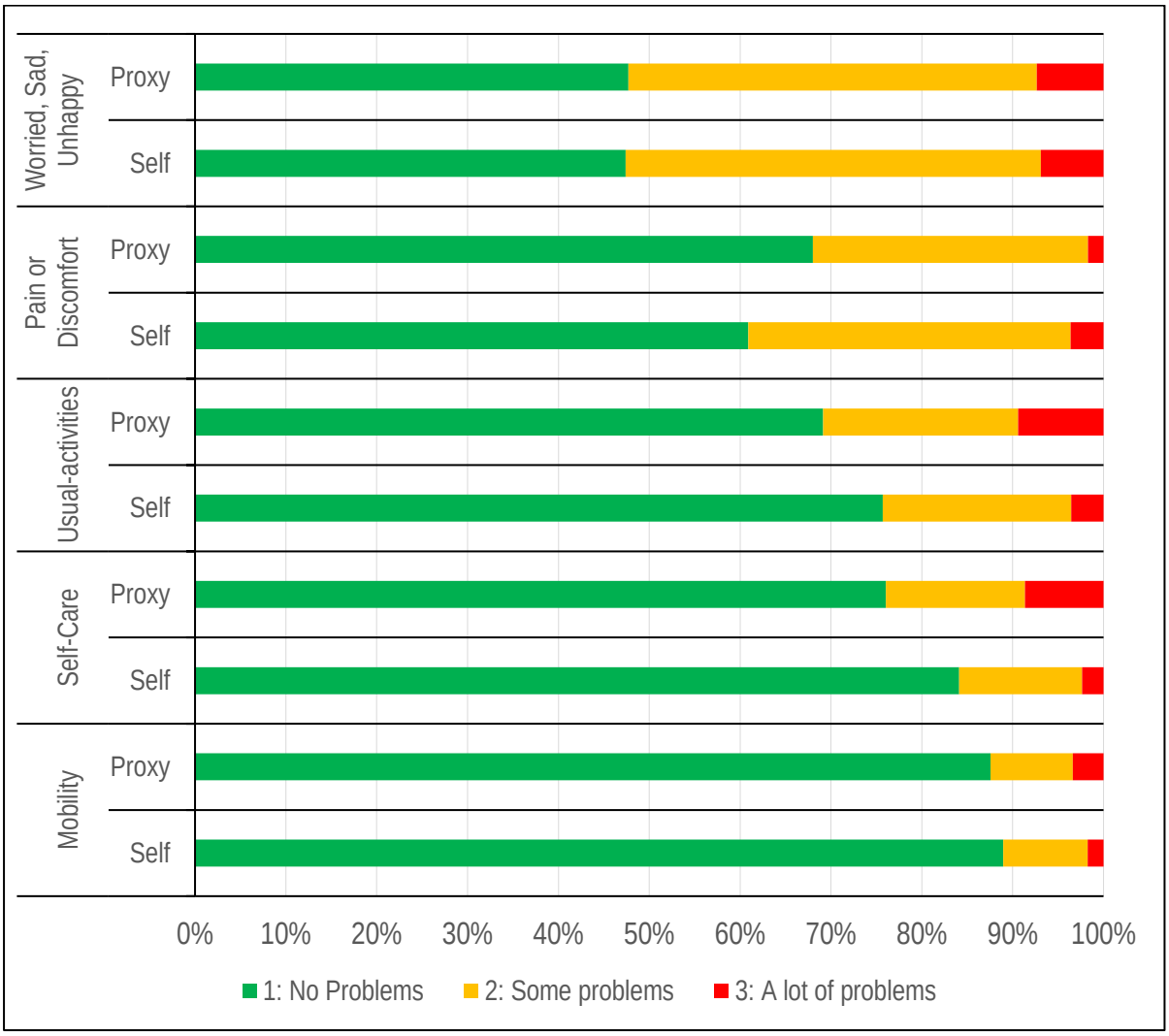
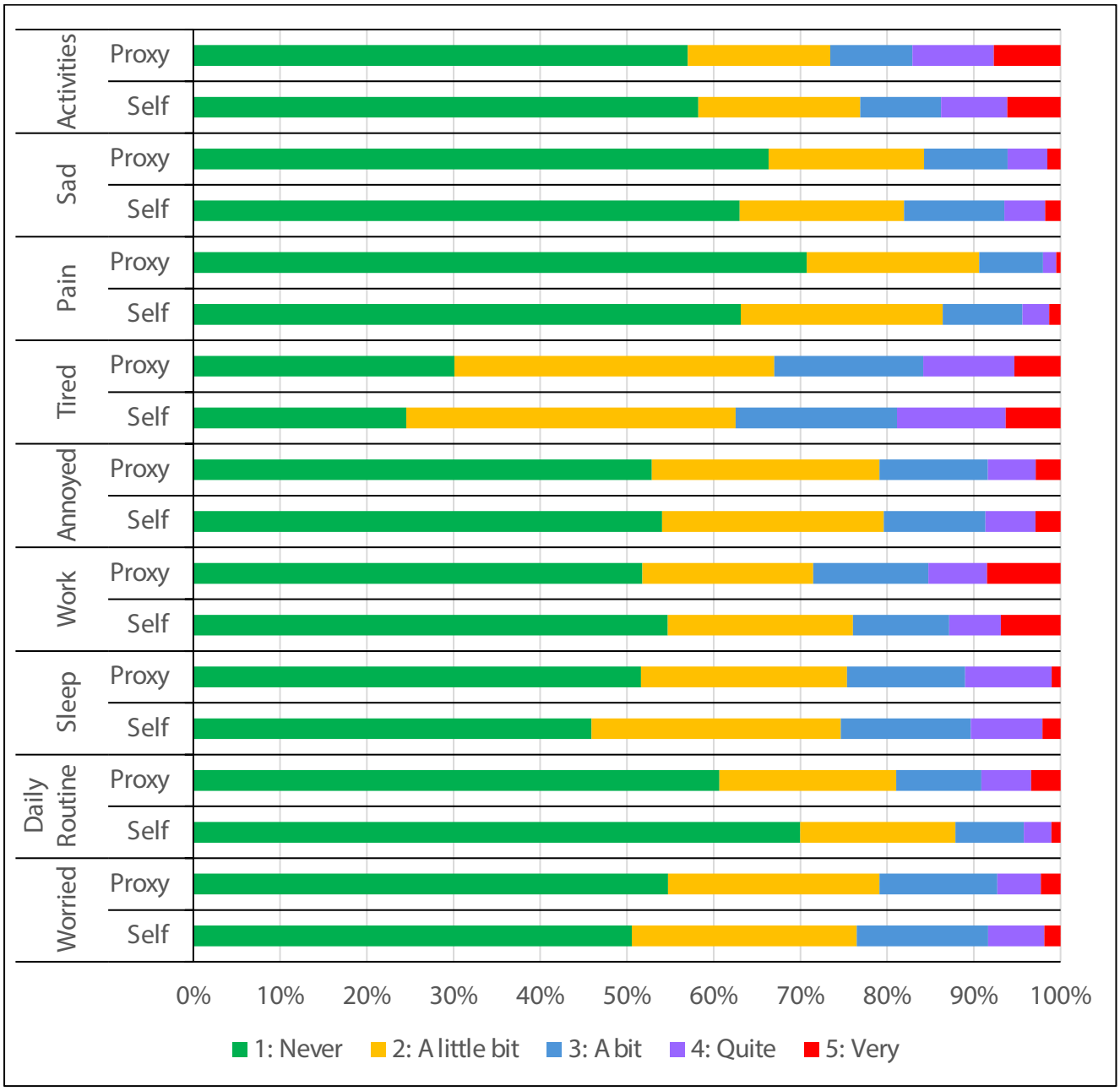


Figure 3. Distribution of response frequencies using CHU9D: A comparison between Self and Proxy



### 3.2 Utility Decrements and Distributions

The utility decrement for each attribute level of EQ-5D-Y-3L and CHU9D derived using Australian, Chinese, and Dutch value sets are shown in Figures 4 and 5. Figures 4 and 5 are author-created graphs from the existing value sets. Of the EQ-5D-Y-3L dimensions, pain or discomfort has the highest utility decrement in all three value sets. The Dutch value set showed the largest decrement for this dimension.

Figure 4. Utility decrements of the EQ-5D-Y-3L for Australian (Pan et al., 2024) , Chinese (Yang et al., 2022) and Dutch (Roudijk et al., 2022) value sets

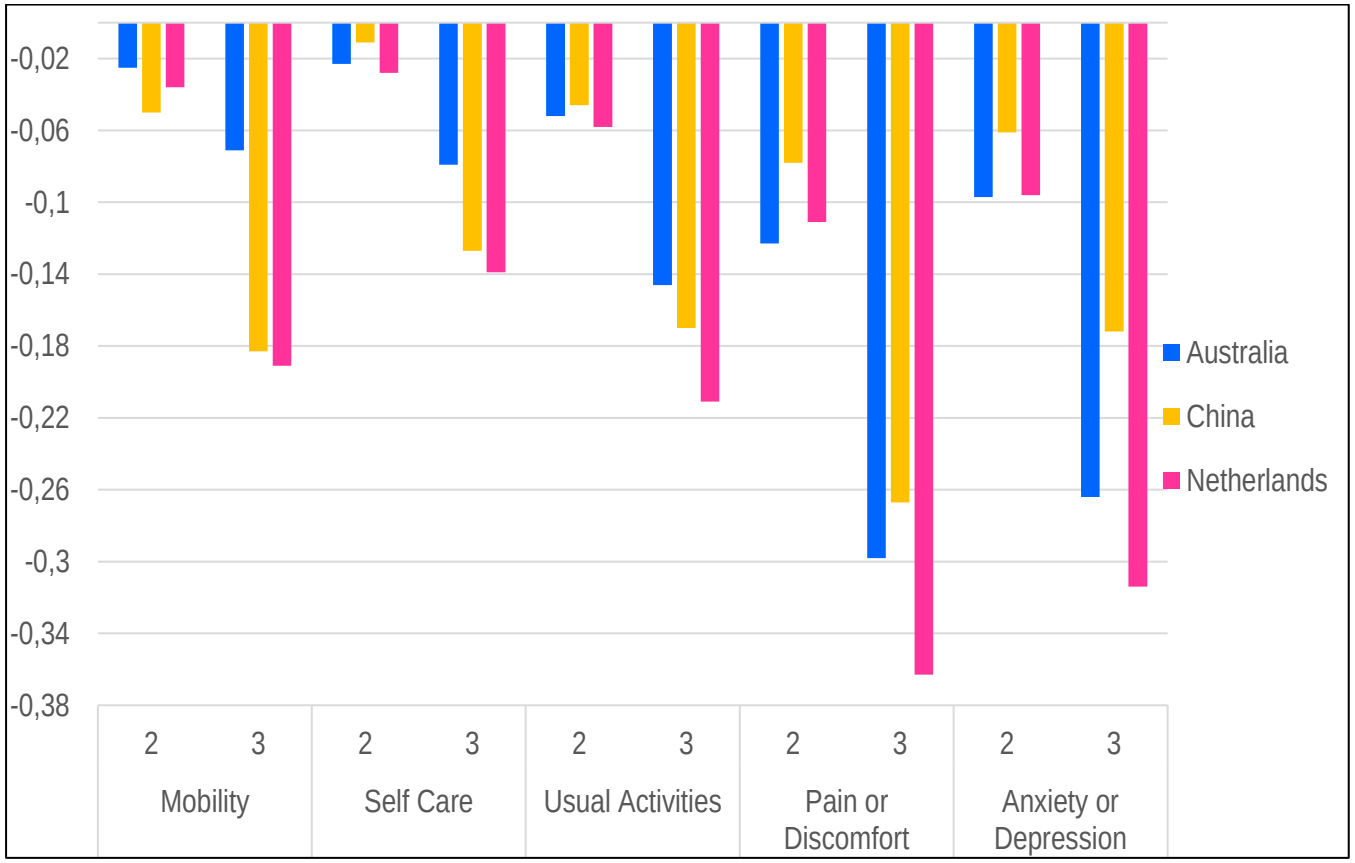
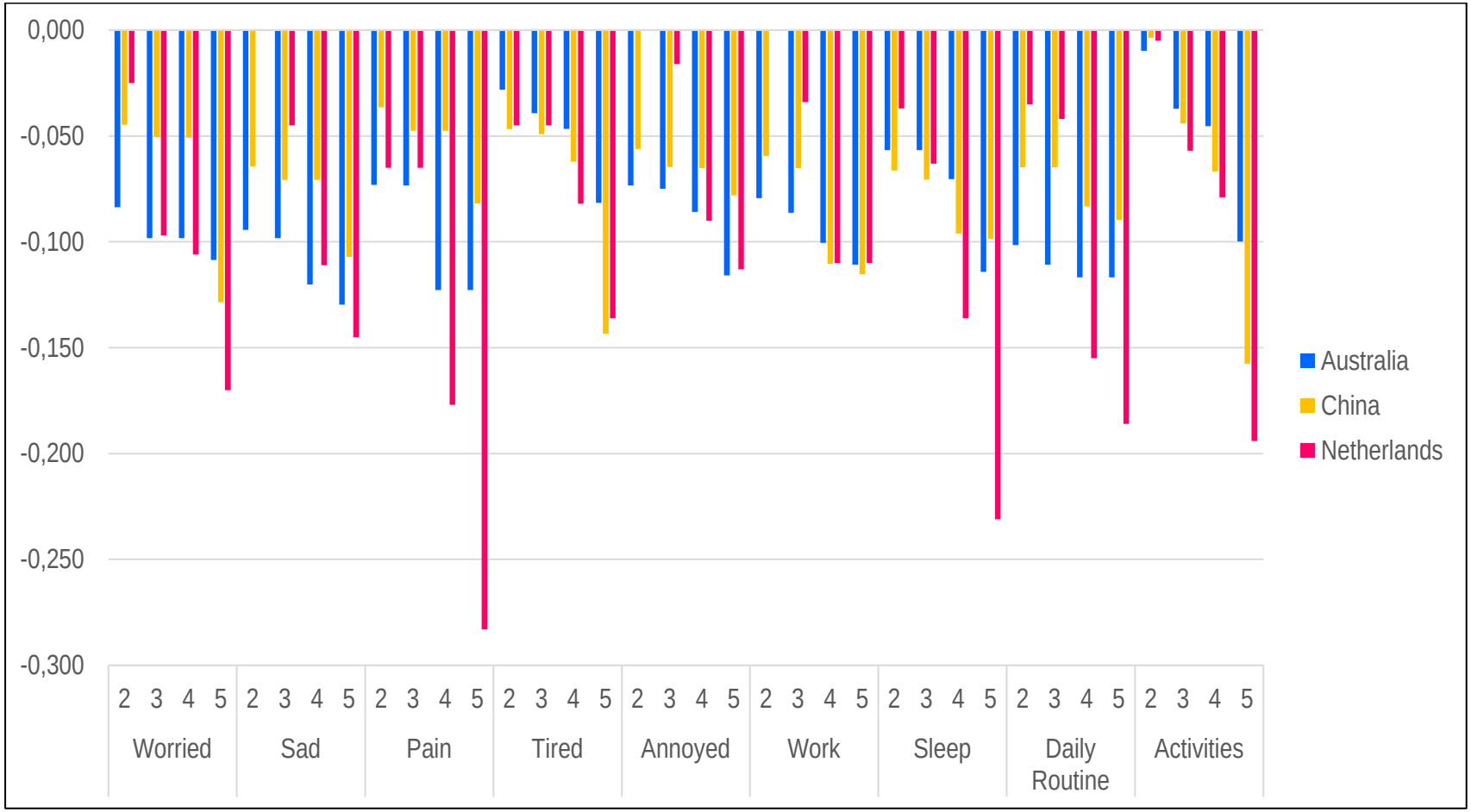


Figure 5. Utility decrements of the CHU9D for Australian (Ratcliffe et al., 2016), Chinese (Chen et al., 2019) and Dutch value (Rowen et al., 2018) sets




Of the CHU9D dimensions, pain has the highest utility decrement in the Dutch value set, the activities dimension in the Chinese value set, and the sad dimension in the Australian value set.

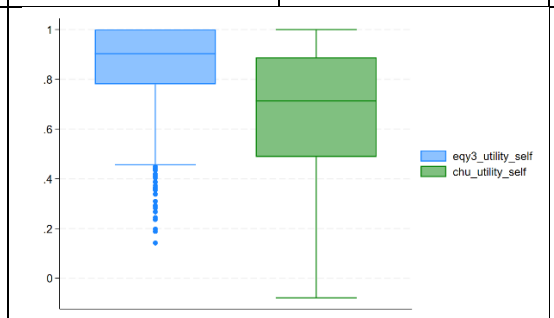
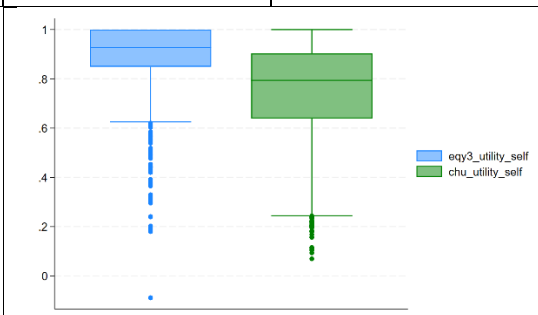
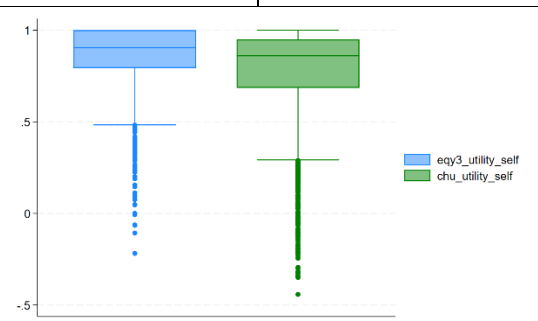
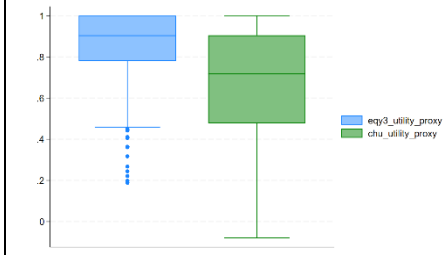
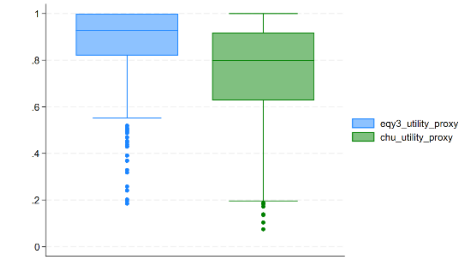
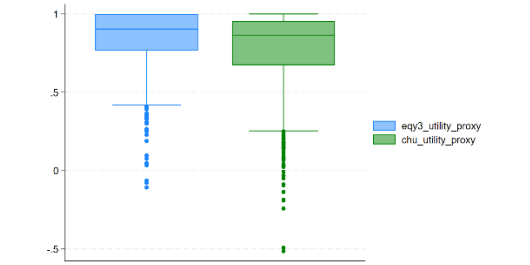
Table 5 reports the summary statistics for the MAU instruments. For both self-reports and proxy-reports, the wider range of utilities are reported using the Dutch value sets for both EQ-5D-Y-3L and the CHU9D.

Table 5. Summary statistics for the MAUIs using Australian, Chinese and Dutch value sets

| Country specific utilities | Instrument | Mean | SD | Median | IQR | Range |
|---|---|---|---|---|---|---|
| Australian | $EQ_{self}$ | 0.86 | 0.15 | 0.90 | 0.78-1.00 | 0.86 |
| | $EQ_{proxy}$ | 0.85 | 0.16 | 0.90 | 0.78-1.00 | 0.81 |
| | $CHU_{self}$ | 0.67 | 0.25 | 0.71 | 0.49-0.89 | 1.08 |
| | $CHU_{proxy}$ | 0.68 | 0.26 | 0.72 | 0.47-1.00 | 1.08 |
| Chinese | $EQ_{self}$ | 0.89 | 0.14 | 0.93 | 0.85-1.00 | 1.09 |
| | $EQ_{proxy}$ | 0.87 | 0.15 | 0.93 | 0.82-1.00 | 0.82 |
| | $CHU_{self}$ | 0.76 | 0.18 | 0.79 | 0.64-0.90 | 0.93 |
| | $CHU_{proxy}$ | 0.76 | 0.19 | 0.79 | 0.63-0.92 | 0.93 |
| Dutch | $EQ_{self}$ | 0.85 | 0.18 | 0.90 | 0.79-1.00 | 1.22 |
| | $EQ_{proxy}$ | 0.84 | 0.19 | 0.90 | 0.77-1.00 | 1.11 |
| | $CHU_{self}$ | 0.78 | 0.23 | 0.86 | 0.69-0.95 | 1.44 |
| | $CHU_{proxy}$ | 0.78 | 0.24 | 0.87 | 0.67-0.96 | 1.51 |

The utility values for each health state reported by the respondents for both self-reports and proxy-reports were calculated using the utility decrements of each of the value sets. Table 6 presents the range of the utilities for each instrument, value set and report types (self vs proxy) accompanied by box plots for visual representation. The range of utility values varies across value sets for both the EQ-5D-Y-3L and CHU9D instruments. When comparing the two instruments, the Dutch value set produces lower utilities to the worst health states than the Chinese and Australian value sets for both self and proxy reports. When comparing the utilities for the two instruments within each country, using the Australian value set, the EQ-5D-Y-3L shows positive utilities for the worst health state while the CHU9D shows negative utilities for both self and proxy reports. With the Dutch value set, both instruments show negative utilities for worst health states across self and proxy reports. In contrast, the Chinese value set results in negative utilities for the EQ-5D-Y-3L in self-reports but positive utilities in proxy reports, while the CHU9D shows positive utilities for both self and proxy reports.

Table 6. Utility distribution for self-reports and proxy-reports across value sets and instruments

| | | Utility (Australian) | | Utility (Chinese) | | Utility (Dutch) | |
|---|---|---|---|---|---|---|---|
| Self-report | Range | EQ-5D-Y-3L [0.142, 1] | CHU9D [-0.078, 1] | EQ-5D-Y-3L [-0.089, 1] | CHU9D [0.069, 1] | EQ-5D-Y-3L [-0.218, 1] | CHU9D [-0.442, 1] |
| | Box Plot |  | |  | |  | |
| Proxy-report | Range | EQ-5D-Y-3L [0.188, 1] | CHU9D [-0.079, 1] | EQ-5D-Y-3L [0.185, 1] | CHU9D [0.075, 1] | EQ-5D-Y-3L [-0.107, 1] | CHU9D [-0.514, 1] |
| | Box Plot |  | |  | |  | |

## 3.3 Regression Analysis: Richardson's attributional regression approach

### 3.3.1 Rescaling utilities and predicted values

Some of the results of the diagnostics tests indicate that the OLS assumptions were not satisfied. The presence of outliers was detected in the data as shown in the graphs in Table 6. To address this issue, a robust regression analysis was conducted, and the results of the robust regression are included in the supplementary materials for reference. Although the OLS assumptions were not met, Richardson's approach relies on the fitted values from the regression analysis, and the violation of assumptions related to the error term does not impact the validity of the fitted values. Furthermore, the robust regression results did not deviate significantly from the standard OLS regression, indicating that the presence of outliers did not substantially alter the key findings. Consequently, the results of the standard OLS regression are considered appropriate and retained for presentation in the main analysis.

The linear regressions used to transform the utilities and fitted values of the CHU9D instrument onto the same scale as the EQ-5D-Y-3L (Step 5) are reported in Table 7. The coefficient '$b_1$' represents the adjustment factor (increase/ decrease) required to align the utility of the CHU9D to the same scale as the utility of the EQ-5D-Y-3L. The '$b_2$' coefficient represents the adjustment factor required to align the fitted value of the CHU9D to the same scale as the EQ-5D-Y-3L. For example, from the regression between EQ-5D-Y-3L and CHU9D utilities, the utility of CHU9D must be scaled down by a factor of 0.43 to be equivalent to the EQ-5D-Y-3L scale.

Table 7. Regression analysis results of the utility of EQ-5D-Y-3L ($U_{EQ}$) on the utility of CHU9D ($U_{CHU}$) and the utility of EQ-5D-Y-3L ($U_{EQ}$) on the fitted value of CHU9D ($\hat{V}_{CHU}$)

| Country specific Utilities | Report-type | $U_{EQ} = a_1 + b_1 U_{CHU}$ (Eq 4) | $R^2$ | $U_{EQ} = a_2 + b_2 \hat{V}_{CHU}$ (Eq 5) | $R^2$ |
|---|---|---|---|---|---|
| Australian | Self | EQ-5D = 0.57 + 0.43 CHU9D | 0.51 | EQ-5D = 0.54 + 0.47 $\hat{V}$CHU9D | 0.54 |
| | Proxy | EQ-5D = 0.55 + 0.45 CHU9D | 0.54 | EQ-5D = 0.52 + 0.49 $\hat{V}$CHU9D | 0.59 |
| Chinese | Self | EQ-5D = 0.49 + 0.51 CHU9D | 0.46 | EQ-5D = 0.47 + 0.55 $\hat{V}$CHU9D | 0.50 |
| | Proxy | EQ-5D = 0.44 + 0.57 CHU9D | 0.50 | EQ-5D = 0.41 + 0.60 $\hat{V}$CHU9D | 0.54 |
| Dutch | Self | EQ-5D = 0.40 + 0.57 CHU9D | 0.54 | EQ-5D = 0.39 + 0.58 $\hat{V}$CHU9D | 0.52 |
| | Proxy | EQ-5D = 0.34 + 0.63 CHU9D | 0.59 | EQ-5D = 0.33 + 0.64 $\hat{V}$CHU9D | 0.58 |

$U_{EQ}$= utility of EQ-5D-Y-3L, $U_{CHU}$= utility of CHU9D, $\hat{V}_{CHU}$= fitted value

Table 8 reports the results of the analysis conducted to verify whether the rescaling of the utilities and fitted values of the CHU9D are aligned with the scale of the EQ-5D-Y-3L utilities. The intercept term a=0 indicates the means of the variables in the regression is equal to the mean of $U_{EQ}$ (utility of EQ-5D-Y-3L). The slope 'b' represents values that are very close to 1.00 indicating that non-linearities exist in the relationship. These non-linearities will result in discrepancy effects.

Table 8. Regression results of scale-free difference between utilities and difference between fitted values

| Country specific Utilities | Report-type | Regression Y= *a* + *bX* | | |
|---|---|---|---|---|
| | | *a* | *b* | $R^2$ |
| Australian | Self | 0.01 | 0.81 | 0.57 |
| | Proxy | 0.02 | 0.77 | 0.54 |
| Chinese | Self | 0.00 | 1.00 | 0.68 |
| | Proxy | 0.00 | 0.98 | 0.72 |
| Dutch | Self | 0.01 | 0.91 | 0.71 |
| | Proxy | 0.01 | 0.84 | 0.69 |

$Y= [U_{EQ} - \hat{U}_{CHU}(u_{EQ})]$; $X= [V_{EQ} - \hat{V}_{CHU}(u_{EQ})]$

$\hat{U}_{CHU}(u_{EQ})$: Predicted utility from CHU9D transformed to the same scale as the EQ-5D-Y-3L

$\hat{V}_{CHU}(u_{EQ})$: Predicted value from CHU9D transformed to the same scale as the EQ-5D-Y-3L

### 3.3.2 Unpacking the variability in utilities

The measurement results of the three components affecting the utility differences are reported in Table 9. The average absolute pairwise difference in utilities ($U_{EQ} - U_{CHU}$) is 0.150. It ranges from 0.111 (Dutch-proxy-report) to 0.199 (Australian-self-report). The largest component is the descriptive system effect which accounts for 53.8% of the difference and the smallest component is the discrepancy effect averaging 1.3% of the difference. The differences between utilities were primarily attributed to the scale and descriptive system, varying by value set and reporting perspective. Using the Australian value sets, 61.9% (self-reports) and 60.3% (proxy-reports) of differences between the EQ-5D-Y-3L and CHU9D were attributable to the scale. For the Chinese value sets, the scale accounted for 53.0% of self-reports (45.6% proxy-report), while 51.9% of proxy-reports (45.9% self-report) were attributed to the descriptive system. In contrast, the Dutch values set showed higher contributions from the descriptive system: 73.5% (self-reports) and 80.2% (proxy-reports).

Table 9. Unpacking variability in Utility Differences

| Country Specific Utilities | Report type | Absolute Differences | | | | | % of ($U_{EQ}$ -$U_{CHU}$) | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **Pairwise difference in utilities** ($U_{EQ}$– $U_{CHU}$) A | **Scale-free differences in utility** ($U_{EQ}$ – $\hat{U}_{CHU}$ ($u_{EQ}$)) B | **Scale Effect** (A-B) C | **Descriptive system** ($\hat{V}_{EQ}$– $\hat{V}_{CHU}$($u_{EQ}$)) D | **Discrepancy Effect** (B-D) E | Scale Effect (C/A)*100 | Descriptive System Effect (D/A)*100 | Discrepancy Effect (E/A)*100 |
| Australian | Self | 0.199 | 0.076 | 0.123 | 0.071 | 0.005 | 61.9 | 35.7 | 2.4 |
| | Proxy | 0.194 | 0.077 | 0.117 | 0.069 | 0.008 | 60.3 | 35.8 | 3.9 |
| Chinese | Self | 0.141 | 0.066 | 0.075 | 0.065 | 0.002 | 53.0 | 45.6 | 1.4 |
| | Proxy | 0.136 | 0.073 | 0.062 | 0.070 | 0.003 | 45.9 | 51.9 | 2.2 |
| Dutch | Self | 0.116 | 0.085 | 0.031 | 0.086 | -0.000 | 26.7 | 73.5 | -0.2 |
| | Proxy | 0.111 | 0.087 | 0.024 | 0.089 | -0.002 | 21.6 | 80.2 | -1.9 |
| Average | | 0.150 | 0.078 | 0.072 | 0.075 | 0.003 | 44.9 | 53.8 | 1.3 |

# 4. Discussion

## 4.1 Key Findings

This study provides a comparative analysis of the EQ-5D-Y-3L vs the CHU9D for a pediatric population, accounting for differences in their descriptive systems and the preference weights and showing how the relative importance of each depends on the characteristics of three local value sets and on whether the descriptive data are obtained via self or proxy-report. The results show that across the pairwise comparison, the average difference in utilities among the three value sets for the 4,099 self-reported survey respondents was 0.152 and for the 1,182 proxy-reported survey respondents the average difference was 0.147. When exploring the sources of utility differences between the two pediatric multi-attribute utility instruments, we found that the descriptive system effect accounted for the largest variation in utilities on average for both self (51.6%) and proxy (55.9%) reports.

Examining the differences in utilities that arise when applying different country value sets to the P-MIC data revealed different patterns. A larger part of the differences was accounted by the scale effect when using the Australian value sets for both self and proxy reports and when using the Chinese value sets for self-reports. The descriptive system effect contributed notably when using Dutch value sets for both self and proxy reports and when using the Chinese value sets for proxy-reports. As highlighted in Table 6, parents (proxy respondents) tend to complete the instruments on behalf of the child in a way that reflects more favourable health states compared to children's own assessments (self-reports). Using the Chinese value set as an example, the utility of the worst possible health state (PITS state) for the EQ-5D-Y-3L is -0.089 for self-reported data, compared to 0.185 for proxy-reported data.

Several expectations were made prior to the analysis. One expectation was that the scale effect might vary depending on the scale of the value set used. This aligns with the results. As shown in Table 6, a comparison of self-reports using the Australian and Dutch value sets reveals notable differences in the range of the utilities using the Australian value set compared to the Dutch value set. For example, the range of utilities for the EQ-5D-Y-3L and CHU9D using the Australian value sets are 0.142 -1.00 and -0.078 -1.00 respectively, whereas the corresponding range using the Dutch value set are -0.218 – 1.00 and -0.442 – 1.00. The range of utilities are spread across a wide spectrum for both instruments using the Dutch value set. These differences contribute to variations in the scale and descriptive system effects. Using the Australian value sets the scale effect accounts for a larger contribution of the utility differences between EQ-5D-Y-3L and CHU9D (61.9%) whereas by using the Dutch value set shows a larger contribution from descriptive system effect (73.5%) for self-reports as highlighted in Table 9.

## 4.2 Comparison with Existing Literature

Whitehurst et al. (2014) compared the utilities derived using DCE for the EQ-5D and the SF-6D. They concluded that since both instruments allowed a negative health state the range does not significantly influence the variations. Instead, differences in the descriptive system were a major contribution factor to the variation in the utilities. Similarly, our analysis found that the descriptive system effect is more pronounced when the range of the utilities between the two instruments does not vary significantly. This is evident in Table 6, which presents utilities using the Chinese value set for proxy-reports and Dutch value sets for both self and proxy reports.

This study employed an updated analysis approach from Richardson et al. (2015), who compared the utilities between the EQ-5D-5L, SF-6D, HUI3, 15D and AQoL- 8D. They reported the scale effect is larger when the instrument utilities have lower standard deviations which implies a greater compression of utilities. They concluded that when one instrument (15D) has a lower standard deviation and is compared to another with a higher standard deviation (AQoL-8D), the scale effect could be larger. Similarly, our results (Table 9) show a larger scale effect in utilities reported using the Australian value set. This is consistent with the observed standard deviations (Table 5): for $EQ_{self}$, the standard deviation is 0.15, compared to 0.25 for $CHU_{self}$; and for $EQ_{proxy}$ it is 0.16, compared to 0.26 for $CHU_{proxy}$.

## 4.3 Strengths and Limitations

A significant strength of this study is the use of different country value sets (Australian, Chinese, Dutch) for each of the two descriptive systems. This allowed us to analyse the differences in contributions from the three sources (scale, descriptive system and discrepancy effect) among the three value sets. This approach allowed us to investigate whether the range of the utilities influences the scale effect across different value sets. Additionally, the study benefits from a large and varied sample, including a sample recruited through a hospital. This diverse sample provides a broad range of utilities indicating both the most extreme health states and the best possible health states as well.

However, a limitation of the study is that the sample was recruited exclusively in Australia and different country value sets were then applied to that data. While this was informative and helped interpret our result, the findings may not generalize to data collected in other countries.

## 4.4 Future Research Implications

The main aim of this study was to identify the sources of differences in utilities between two pediatric MAUIs. The present study confirms the range of the utilities derived from the value set directly impacts the scale effect. This finding suggests that the length of the scale plays a more crucial role in resulting

utility estimates than the weights assigned to the dimensions. As we could see from Table 2, the anchoring methods for each of the value sets are quite different. Future research could explore how different anchoring methods may affect the differences in utilities among different instruments. A more thorough understanding of anchoring methods and how it affects the utility scaling may help us standardize valuation methods in pediatric health evaluations. It will also be interesting to compare the results when the CHU9D is compared against the EQ-5D-Y-5L, which introduces further changes to the descriptive system.

## 5. Conclusion

A difference in utilities might occur when using different instruments to derive the utilities. The present study investigated the potential reason for the inconsistencies between the EQ-5D-Y-3L and CHU9D using the Australian, Chinese and Dutch value sets. We find differences in utilities attributable to the descriptive system, suggesting it is important to use instruments that are comparable in terms of their descriptive systems for future health economic evaluations. Additionally, this study highlights the importance of distinguishing between self-reported and proxy-reported responses in pediatric instruments. Since the parents tend to describe child health states better, which is reflected in the higher utilities applied to those states, both self and proxy perspectives need to be considered in child health economic evaluations. Furthermore, the choice of value set also contributes to the utility differences between MAUIs. While, in practice, each country relies mainly on the value set of their own, it is important to acknowledge that different value sets may lead to variations in utilities.

## References


Attema, A. E., Edelaar-Peeters, Y., Versteegh, M. M., & Stolk, E. A. (2013). Time trade-off: one methodology, different methods. *Eur J Health Econ*, *14 Suppl 1*(Suppl 1), S53-64. https://doi.org/10.1007/s10198-013-0508-x

Brazier, J. E., Rowen, D., Yang, Y., & Tsuchiya, A. (2009). Using rank and discrete choice data to estimate health state utility values on the QALY scale.

Chen, G., Xu, F., Huynh, E., Wang, Z., Stevens, K., & Ratcliffe, J. (2019). Scoring the Child Health Utility 9D instrument: estimation of a Chinese child and adolescent-specific tariff. *Qual Life Res*, *28*(1), 163-176. https://doi.org/10.1007/s11136-018-2032-z

De Silva, A., van Heusden, A., Lang, Z., Devlin, N., Norman, R., Dalziel, K., & Peasgood, T., & Pan, T. (2024). *How do health state values differ when respondents consider adults vs children living in those states? A systematic review* [Unpublished manuscript].

Jakubczyk, M., Lipman, S. A., Roudijk, B., Norman, R., Pullenayegum, E., Yang, Y., Gu, N. Y., & Stolk, E. (2023). Modifying the Composite Time Trade-Off Method to Improve Its Discriminatory Power. *Value in Health*, *26*(2), 280-291. https://doi.org/10.1016/j.jval.2022.08.011

Jones, R., Mulhern, B., McGregor, K., Yip, S., O'Loughlin, R., Devlin, N., Hiscock, H., Dalziel, K., & On Behalf Of The Quality Of Life In Kids Key Evidence To Strengthen Decisions In Australia Quokka Project, T. (2021). Psychometric Performance of HRQoL Measures: An Australian Paediatric Multi-Instrument Comparison Study Protocol (P-MIC). *Children (Basel)*, *8*(8). https://doi.org/10.3390/children8080714

Khadka, J., Kwon, J., Petrou, S., Lancsar, E., & Ratcliffe, J. (2019). Mind the (inter-rater) gap. An investigation of self-reported versus proxy-reported assessments in the derivation of childhood utility values for economic

evaluation: A systematic review. *Soc Sci Med*, *240*, 112543. https://doi.org/10.1016/j.socscimed.2019.112543

Kwon, J., Freijser, L., Huynh, E., Howell, M., Chen, G., Khan, K., Daher, S., Roberts, N., Harrison, C., Smith, S., Devlin, N., Howard, K., Lancsar, E., Bailey, C., Craig, J., Dalziel, K., Hayes, A., Mulhern, B., Wong, G., . . . Petrou, S. (2022). Systematic Review of Conceptual, Age, Measurement and Valuation Considerations for Generic Multidimensional Childhood Patient-Reported Outcome Measures. *PharmacoEconomics*, *40*(4), 379-431. https://doi.org/10.1007/s40273-021-01128-0

Moodie, M., Richardson, J., Rankin, B., Iezzi, A., & Sinha, K. (2010). Predicting time trade-off health state valuations of adolescents in four Pacific countries using the Assessment of Quality-of-Life (AQoL-6D) instrument. *Value Health*, *13*(8), 1014-1027. https://doi.org/10.1111/j.1524-4733.2010.00780.x

Mulhern, B., Norman, R., Street, D. J., & Viney, R. (2019). One Method, Many Methodological Choices: A Structured Review of Discrete-Choice Experiments for Health State Valuation. *PharmacoEconomics*, *37*(1), 29-43. https://doi.org/10.1007/s40273-018-0714-6

Pan, T., Devlin, N., & Roudjik, B., & Norman, R. (2024). *An Australian Value Set for the EQ-5D-Y-3L* [Unpublished manuscript].

Ratcliffe, J., Chen, G., Stevens, K., Bradley, S., Couzner, L., Brazier, J., Sawyer, M., Roberts, R., Huynh, E., & Flynn, T. (2015). Valuing Child Health Utility 9D Health States with Young Adults: Insights from a Time Trade Off Study. *Appl Health Econ Health Policy*, *13*(5), 485-492. https://doi.org/10.1007/s40258-015-0184-3

Ratcliffe, J., Huynh, E., Chen, G., Stevens, K., Swait, J., Brazier, J., Sawyer, M., Roberts, R., & Flynn, T. (2016). Valuing the Child Health Utility 9D: Using profile case best worst scaling methods to develop a new adolescent specific scoring algorithm. *Social Science & Medicine*, *157*, 48-59. https://doi.org/https://doi.org/10.1016/j.socscimed.2016.03.042

Richardson, J., Iezzi, A., & Khan, M. A. (2015). Why do multi-attribute utility instruments produce different utilities: the relative importance of the descriptive systems, scale and 'micro-utility' effects. *Qual Life Res*, *24*(8), 2045-2053. https://doi.org/10.1007/s11136-015-0926-6

Richardson, J., McKie, J., & Bariola, E. (2014). Multiattribute Utility Instruments and Their Use. In A. J. Culyer (Ed.), *Encyclopedia of Health Economics* (pp. 341-357). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-12-375678-7.00505-8

Roudijk, B., Sajjad, A., Essers, B., Lipman, S., Stalmeier, P., & Finch, A. P. (2022). A Value Set for the EQ-5D-Y-3L in the Netherlands. *PharmacoEconomics*, *40*(2), 193-203. https://doi.org/10.1007/s40273-022-01192-0

Rowen, D., Mulhern, B., Stevens, K., & Vermaire, J. H. (2018). Estimating a Dutch Value Set for the Pediatric Preference-Based CHU9D Using a Discrete Choice Experiment with Duration. *Value Health*, *21*(10), 1234-1242. https://doi.org/10.1016/j.jval.2018.03.016

Wang, H., Rowen, D. L., Brazier, J. E., & Jiang, L. (2023). Discrete Choice Experiments in Health State Valuation: A Systematic Review of Progress and New Trends. *Applied Health Economics and Health Policy*, *21*(3), 405-418. https://doi.org/10.1007/s40258-023-00794-9

Whitehurst, D. G., Norman, R., Brazier, J. E., & Viney, R. (2014). Comparison of contemporaneous EQ-5D and SF-6D responses using scoring algorithms derived from similar valuation exercises. *Value Health*, *17*(5), 570-577. https://doi.org/10.1016/j.jval.2014.03.1720

Wille, N., Badia, X., Bonsel, G., Burström, K., Cavrini, G., Devlin, N., Egmar, A.-C., Greiner, W., Gusi, N., Herdman, M., Jelsma, J., Kind, P., Scalone, L., & Ravens-Sieberer, U. (2010). Development of the EQ-5D-Y: a child-friendly version of the EQ-5D. *Quality of Life Research*, *19*(6), 875-886. https://doi.org/10.1007/s11136-010-9648-y

Xie, F., Li, S. C., Luo, N., Lo, N. N., Yeo, S. J., Yang, K. Y., Fong, K. Y., & Thumboo, J. (2007). Comparison of the EuroQol and short form 6D in Singapore multiethnic Asian knee osteoarthritis patients scheduled for total knee replacement. *Arthritis Rheum*, *57*(6), 1043-1049. https://doi.org/10.1002/art.22883

Yang, Z., Jiang, J., Wang, P., Jin, X., Wu, J., Fang, Y., Feng, D., Xi, X., Li, S., Jing, M., Zheng, B., Huang, W., & Luo, N. (2022). Estimating an EQ-5D-Y-3L Value Set for China. *PharmacoEconomics*, *40*(Suppl 2), 147-155. https://doi.org/10.1007/s40273-022-01216-9