UNIVERSITY OF CAPE TOWN

DEPARTMENT OF HEALTH AND REHABILITATION SCIENCES

DIVISION OF PHYSIOTHERAPY

Performance of the EQ-5D-Y Interviewer Administered Version in young children

by

RAZIA AMIEN

AMNRAZ001

In fulfilment of the requirements for the degree

MSc in Physiotherapy

Date of submission: 08/07/2022

Supervisors: Dr Janine Verstraete and Mrs Desiree Scott

Word count: 29812 (excluding abstract, figures, graphs, references and appendices)

Declaration

I, Razia Amien, hereby declare that the work on which this dissertation/thesis is based on my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university.

I empower the university to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signature: Signed by candidate

Date: 07/07/2022

Acknowledgements

Firstly, I would like to acknowledge the EuroQoL group for identifying the need for this research and their continuous hard work in developing new paediatric Health-Related Quality of Life instruments. The variations in modes of administration these instruments offer, have the ability to allow all children, despite their age or level of education, to feel heard and empowered when their wellbeing is concerned. It also reminds us as healthcare professionals to always address patient care in a holistic manner. I would also like to acknowledge the EuroQoL group for graciously providing the funds to carry out this project.

I would like to express my gratitude to my two supervisors, Dr Janine Verstraete and Mrs Desiree Scott, for granting me the opportunity to pursue my postgraduate studies at UCT and to conduct this research project. The world of research was a very new experience for me, in saying that, it was an absolute honour to work with and learn from them both. Without their continuous support and guidance throughout this project, none of it would have been possible. They have both assisted me in developing a very keen interest in research, particularly paediatric-related research.

A huge thank you is needed for both my supervisors and the UCT Health Sciences Writing Centre for their input during the write-up of this project. I have always found writing to be challenging but with their help, I believe my writing has improved tremendously.

To the clinicians and nursing staff at the healthcare facilities, thank you for all assistance in the recruitment process of data collection and for always ensuring there was a space available for me to conduct interviews.

To the principals, teachers, secretaries and therapists at the Mainstream and Special Schools, thank you for all your assistance with the data collection process and for allowing me to conduct interviews with your learners.

A very special thank you to all the incredible children with whom I interacted with throughout this project. Without your willingness and openness, this project would not have been possible. I would also like to thank their parents/legal guardians/caregivers for allowing their children to participate in this project.

Lastly, a huge thank you to my family and friends for their support, advice and encouragement throughout this project. You are all greatly appreciated.

Dedication

This research project is dedicated to the paediatric community of the Western Cape, South Africa in the hopes that all children are treated holistically by ensuring that their psychosocial wellbeing is not overshadowed by their physical wellbeing.

Abstract

Introduction

The interest in Health-Related Quality of Life (HRQoL) in the paediatric population has grown over the last decade as it allows for a more holistic approach which has the potential to positively influence treatment outcomes (1–4). With an increase in interest, the need for alternative modes of administration of HRQoL instruments has become more important to allow for self-report in younger age-groups. Despite their age and/or literacy levels, their ability to understand the concept of HRQoL would allow for accurate self-report if the correct instrument is used (5). Proxy-report is used often as a default in these younger age-groups as only two interviewer-administered instruments are currently available (6,7), neither of which have been validated for African populations. The newly developed EQ-5D-Y-3L Interviewer Administered (IA) instrument would allow for self-report in younger paediatric populations, therefore limiting the reliance on proxy-report which does not account for the subjectivity of HRQoL or allow for the inclusion of the child's view (2,8–10).

Aim

The first aim of this study was to determine the performance and preference of the EQ-5D-Y-3L-IA and self-complete (SC), in children aged 8-10-years. The second aim was to determine the psychometric performance of the EQ-5D-Y-3L-IA version in children aged 5-7-years compared to those aged 8-10-years.

Methods

A cross-sectional, descriptive observational, analytical design was used. Children were recruited in two age-groups, 5-7-years (n=177, 46%) and 8-10-years (n=211, 54%). Participants were drawn from the General Population (GenPop) attending a Mainstream School (n=109, 28%), Special Schools for learners with special educational needs (n=55, 14%) and healthcare facilities caring for children with orthopaedic conditions (n=161, 41%) or chronic respiratory illnesses (n=63, 16%). All children completed the EQ-5D-Y-3L-IA, Faces Pain Scale-Revised (FPS-R), Moods and Feelings Questionnaire (MFQ). The researcher completed the observational Functional Independence Measure (WeeFIM). In addition, children in the 8-10-year group completed the EQ-5D-Y-3L-SC. Dimension responses of the EQ-5D-Y-3L-IA and SC were analysed for floor and ceiling effects, inconsistent responses, missing

responses and differences in health states between age-groups and versions. Differences in reporting were determined by chi-square statistic (x^2). Known-group validity across age (years), sex and health conditions were analysed using Spearman's rank order coefficients (r_s) in addition to the median utility and Visual Analogue Scale (VAS) scores using Kruskal Wallis and Mann-Whitney U-test. Pearson's correlation was used to assess concurrent validity by comparing the utility and VAS scores between versions. Spearman's Rank Correlation was computed to assess the convergent validity of the EQ-5D-Y-3L-IA and SC compared to the FPS-R, MFQ and WeeFIM. Responses from structured cognitive debriefing interviews were grouped and coded by the researcher according to similar responses provided by participants. Cognitive debriefing was used to determine the acceptability, comprehensibility and where applicable, participants' preference between versions and the reasons for their preference. The researcher was aware of reflexivity and did not allow personal opinions to impact on participants' responses, nor the grouping and coding of responses. The EQ-5D-Y-3L-IA was retested 48 hours later only in children with a stable health condition, recruited from schools and analysed using weighted Cohen's kappa statistic (k) for dimension scores and the Intraclass Correlation Coefficient (ICC) for utility and VAS scores.

Results

There were no concerning differences in EQ-5D-Y-3L dimension responses, known-group validity, concurrent validity or correlation of VAS and utility scores between the IA and SC versions. The IA version had the advantage of no missing values and was preferred over the SC version by 8-10-year-olds (60%). When comparing the IA version between age-groups, the performance was similar. However, children aged 5-7-years reported significantly more problems with the *Looking After Myself* dimension ($x^2=31.021; p<.0001$) by which cognitive debriefing revealed developmental difficulty with advanced dressing tasks such as laces and buttons.

Conclusion

Validity and test-retest reliability of the EQ-5D-Y-3L-IA version was successfully assessed in children aged 5-10-years. As the results were comparable to the SC version in children aged 8-10-years, it therefore indicates that versions can be used interchangeably. In settings with low literacy levels, such as South Africa, the IA version is recommended for young children, most notably those 8-years of age. The performance of the IA version across age-groups showed that younger children can reliably report on their HRQoL therefore also proved useful in younger age-groups, however, adaptations to the

dimension of *Looking after myself* is suggested for improved developmental appropriateness. Therefore, it is recommended that EQ-5D-Y-3L-IA be included in children from 5-years in routine clinical practice and clinical trials.

Abbreviations

Activities of daily living	ADLs
Cerebral Palsy	CP
Child and Health Illness Profile - Child-report form/Child Edition	CHIP CRF/CE
Child Health Questionnaire	CHQ
Child Health Utility 9D	CHU-9D
Consensus-based Standards for the selection of health Measurement Instruments	COSMIN
Cystic Fibrosis	CF
EQ-5D-Y-3L Self-Complete	EQ-5D-Y-3L-SC
EQ-5D-Y-3L-IA	IA
EuroQoL Five Dimension	EQ-5D
EuroQoL Five Dimension Youth version – Three levels	EQ-5D-Y-3L
Faces Pain Scale-Revised	FPS-R
Functional Independence Measure	WeeFIM
Functional Status II (R)	FSIIR
General Population	GenPop
Health Utilities Index Mark 3	HUI_3
Health-Related Quality of Life	HRQoL
Human Research Ethics Committee	HREC
Intraclass Correlation Coefficient	ICC
Juvenile Arthritis Multidimensional Assessment Report	JAMAR
Juvenile Idiopathic Arthritis	JIA
KINDER Lemensqualitatsfragebogen: Children Quality of Life-Questionnaire Kiddy-KINDL version	Kiddy-KINDL
Learners with special educational needs	LSEN
Looking After Myself	LAM
Mobility	Mob
Moods and Feelings Questionnaire	MFQ
National Health Insurance	NHI
Netherlands by a programme started by the Netherlands Organisation for Applied Scientific Research Institute of	TNO-AZL TACQOL-PF

Prevention and Health and Leiden University Hospital (TNO-AZL) Questionnaire for Children's Health-Related Quality of Life – Parent/Child Form	
Pain or Discomfort	P/D
Patient Reported Outcome Measure	PROM
Patient-Reported Outcome Measure Information System Paediatric Global Health Measure	PROMIS-PGH-7
Pediatric Quality of Life Inventory 4.0 Generic Core Scales	PedsQL
Progress in International Reading Literacy Study	PIRLS
Quality Adjusted Life Years	QALY
Quality of Life	QoL
Self-complete	SC
The National Institute for Health and Care Excellence	NICE
University of Cape Town	UCT
Usual Activities	UA
Version Management Committee	VMC
Visual Analogue Scale	VAS
World Health Organization	WHO
Worried, Sad or Unhappy	WSU

Table of Contents

Declaration 1
Acknowledgements 2
Dedication 4
Abstract 5
Abbreviations 8
List of Tables 14
List of Figures 16
1. Introduction 17
1.1. Background 17
1.2. Rationale and significance 18
1.3. Aims 19
1.4. Specific objectives 19
1.5. Outline of the study 21
2. Literature Review 22
2.1. Introduction 22
2.2. Health status, Quality of Life and Health-Related Quality of Life 23
2.2.1. Health status 23
2.2.2. Quality of Life 23
2.2.3. Health-Related Quality of Life 24
2.3. Measuring health-related quality of life in the paediatric population 25
2.3.1. Importance of measuring health-related quality of life 25
2.3.2. Challenges in measuring health-related quality of life - proxy-report vs self-report 26
2.4. Health-related quality of life instruments for the paediatric population 27
2.4.1. Generic and disease-specific measures 28
2.4.2. Summary of generic health-related PROMs for children aged 5-7-years 29
2.4.3. Descriptions of generic health-related PROMs for children aged 5-7-years 35
2.4.4. Summary of generic health-related PROMs for children aged 5-7-years 39
2.5. Psychometric properties and appraisal of health-related patient reported outcome measures for children aged 5-7-years, according to COSMIN 40
2.5.1. Validity 43
2.5.2. Reliability 50
2.5.3. Feasibility 53
2.5.4. Acceptability 54
2.6. Summary of literature review 54
3. Comparison of the Performance of the EQ-5D-Y-3L-IA and EQ-5D-Y-3L-SC in children aged 8-10-years 56
3.1. Methodology 56
3.1.1. Introduction 56

3.1.2. Study settings 56
3.1.3. Participants 57
3.1.4. Sample size 58
3.1.5. Instruments 58
3.1.6. Procedure 60
3.1.7. Data management 61
3.1.8. Statistical analysis 62
3.1.9. Ethical considerations 64
3.1.10. COVID-19 considerations 65

3.2. Results 66
3.2.1. Descriptive statistics 68
3.2.2. General instrument performance and feasibility 70
3.2.3. Distribution of responses between the EQ-5D-Y-3L-IA and EQ-5D-Y-3L-SC 73
3.2.4. Known-group validity 78
3.2.5. Concurrent validity 88
3.2.6. Convergent validity 88
3.2.7. Preference of version 94
3.2.8. Summary of results 96

3.3. Discussion 97
3.3.1. Recruitment and descriptive statistics 97
3.3.2. General instrument performance and feasibility 98
3.3.3. Known-group validity 100
3.3.4. Concurrent validity 101
3.3.5. Convergent validity 102
3.3.6. Preference of version 103

3.4. Conclusion 104

4. Comparison of the Performance of the EQ-5D-Y-3L-IA in children aged 5-7-years and 8-10-years 105

4.1. Methodology 105
4.1.1. Introduction 105
4.1.2. Sample size 105
4.1.3. Instruments 105
4.1.4. Procedure 106
4.1.5. Statistical analysis 108

4.3. Results 111
4.2.1. Descriptive statistics 113
4.2.2. General instrument performance 115
4.2.3. Feasibility 116
4.2.4. Known-group validity 117
4.2.5. Concurrent validity 125
4.2.6. Convergent validity 125
4.2.7. Test-retest reliability 129
4.2.8. Cognitive debriefing 130
4.2.9. Summary of results 141

4.3. Discussion 142
4.3.1. Recruitment and descriptive statistics 142
4.3.2. Dimension performance between the age-groups 143
4.3.3. Feasibility 144
4.3.4. Known-group validity 145
4.3.5. Concurrent validity 145
4.3.6. Convergent validity 146
4.3.7. Test-retest reliability 147

4.3.8. Cognitive debriefing 148
4.4. Conclusion 149
5. Final Conclusion and Recommendations 150
5.1. Study limitations 150
5.2. Recommendations for practice 151
5.3. Recommendations for research 151
5.4. Recommendations for policies 152
5.5. Accessibility of research 153
6. References 154
7. Appendices 170
Appendix 1: COSMIN Risk of Bias Checklist – EQ-5D-Y-3L 170
Appendix 2: COSMIN Risk of Bias Checklist – PedsQL 199
Appendix 3: COSMIN Risk of Bias Checklist – HUI_3 225
Appendix 4: COSMIN Risk of Bias Checklist – Kiddy-KINDL 239
Appendix 5: COSMIN Risk of Bias Checklist – CHIP CRF/CE 252
Appendix 6: COSMIN Risk of Bias Checklist – PROMIS-PGH-7 266
Appendix 7: COSMIN Risk of Bias Checklist – FSIIR 285
Appendix 8: COSMIN Risk of Bias Checklist – CHU-9D 299
Appendix 9: COSMIN Risk of Bias Checklist – DISABKIDS-TAKE-6 317
Appendix 10: COSMIN Risk of Bias Checklist – TACQoL 337
Appendix 11: COSMIN Risk of Bias Checklist – CHQ 348
Appendix 12: EQ-5D-Y-3L-SC (English Version for South Africa) 364
Appendix 13: EQ-5D-Y-3L-IA (English Version for South Africa) 367
Appendix 14: Faces Pain Scale-R (FPS-R) 370
Appendix 15: Moods and Feelings Questionnaire (MFQ) 371
Appendix 16: WeeFIM 372
Appendix 17: Study specific demographic questionnaire 373
Appendix 18: Preference between SC and IA versions 375
Appendix 19: HREC approval 376
Appendix 20: Ministerial permission to conduct non-therapeutic research with minors (Form A) 380
Appendix 21: Permission from the Western Cape Educational Department to conduct research at schools 383
Appendix 22: Letter to schools 384
Appendix 23: Permission from healthcare facilities 388
Appendix 24: Informed consent (Parent/Legal Guardian) 389

Appendix 25: Assent form for child 393
Appendix 26: Telephonic consent form (Parent/Legal guardian) 396
Appendix 27: Data Management Plan 399
Appendix 28: Cognitive debriefing template 402
Appendix 29: Interviewer questionnaire 405
Appendix 30: Letter to healthcare facilities and schools outlining research findings 406
Appendix 31: Letter to participants outlining research findings 409

List of Tables

Table 1: Summary of generic health-related quality of life instruments for children aged 5-7-years . 30
Table 2: Summary of the all psychometric properties of health-related PROMs 42
Table 3: Descriptive statistics of participants 69
Table 4: Comparison of the EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA dimensions 71
Table 5: EQ-5D-Y-3L-SC missing values across age (years), sex and health conditions 72
Table 6: Inconsistent responses of the EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA across dimensions 74
Table 7: Inconsistent responses of the EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA by age-group 76
Table 8: Inconsistent responses of the EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA across dimensions and health conditions 77
Table 9: Spearman's rank correlation of EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA scores across health groups, age and sex 78
Table 10: Comparison of the EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA dimensions across age (years) 79
Table 11: Comparison of the EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA dimensions across sex 80
Table 12: Gamma Correlation Calculations of the EQ-5D-Y-3L-SC versus the EQ-5D-Y-3L-IA dimension responses across 8-, 9- and 10-year-olds 89
Table 13: Convergent validity of the EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA Mobility dimension and WeeFIM Mobility 90
Table 14: Convergent validity of the EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA Looking After Myself dimension and WeeFIM Self-care 91
Table 15: Convergent validity of the EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA Usual Activities dimension and WeeFIM Mobility and Social Interaction 92
Table 16: Convergent validity of the EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA Pain/Discomfort dimension and the Faces Pain Scale-Revised 92
Table 17: Convergent validity of the EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA Worried, Sad or Unhappy dimension and the Moods and Feelings Questionnaire 93
Table 18: Preference between EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA by age (years), sex and health conditions 94
Table 19: Reason for preference between EQ-5D-Y-3L-IA and EQ-5D-Y-3L-SC 95
Table 20: Descriptive statistics of participants across age-groups (5-7-years and 8-10-years) 114
Table 21: Known-group validity of EQ-5D-Y-3L-IA scores by age and health conditions with Spearman's rank order correlation 117
Table 22: Convergent validity of EQ-5D-Y-3L-IA Mobility dimension and WeeFIM Mobility 125

Table 23: Convergent validity of EQ-5D-Y-3L-IA Looking After Myself dimension and WeeFIM Self-care 126
Table 24: Convergent validity of EQ-5D-Y-3L-IA Usual Activities dimension and WeeFIM Mobility and Social Interaction 127
Table 25: Convergent validity of EQ-5D-Y-3L-IA Pain/Discomfort dimension and Faces-Pain Scale-Revised 127
Table 26: Convergent validity of EQ-5D-Y-3L-IA Worried, Sad or Unhappy dimension and Moods and Feelings Questionnaire 128
Table 27: Test-retest reliability of the EQ-5D-Y-3L-IA across age-groups 129
Table 28: Reasons for level reported in Mobility dimension 131
Table 29: Reasons for level reported in Looking After Myself dimension 133
Table 30: Reasons for level reported in Usual Activities dimension 135
Table 31: Reasons for level reported in Pain/Discomfort dimension 136
Table 32: Reasons for level reported in Worried, Sad or Unhappy dimension 138
Table 33: Reasons for difficulties reported with completion of the EQ-5D-Y-3L-IA across age-groups and dimensions 140

List of Figures

Figure 1: Inclusion process for health-related PROMs for children aged 5-7-years 28
Figure 2: COSMIN taxonomy of measurement properties adapted by researcher from Mokkink et al. (2010) (25) 41
Figure 3: Recruitment of sample 67
Figure 4: Comparison of the EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA Mobility dimension across health conditions 81
Figure 5: Comparison of the EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA Looking After Myself dimension across health conditions 82
Figure 6: Comparison of the EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA Usual Activities dimension across health conditions 83
Figure 7: Comparison of the EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA Pain/Discomfort dimension across health conditions 84
Figure 8: Comparison of the EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA Worried, Sad or Unhappy dimension across health conditions 85
Figure 9: EQ-5D-Y-3L-IA and EQ-5D-Y-3L-SC VAS scores 86
Figure 10: EQ-5D-Y-3L-IA and EQ-5D-Y-3L-SC utility scores 87
Figure 11: Scatterplot of utility scores versus VAS scores for the EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA versions in children 8-10-years (n=207) 88
Figure 12: Recruitment of children aged 5-7-years and 8-10-years 112
Figure 13: Comparison of the EQ-5D-Y-3L-IA dimensions across age-groups 115
Figure 14: Ceiling effects in children aged 5-7-years and 8-10-years across health conditions 116
Figure 15: Comparison of EQ-5D-Y-3L-IA Mobility dimension across condition groups 118
Figure 16: Comparison of EQ-5D-Y-3L-IA Looking After Myself dimension across condition groups 119
Figure 17: Comparison of EQ-5D-Y-3L-IA Usual Activities dimension across condition groups 120
Figure 18: Comparison of EQ-5D-Y-3L-IA Pain/Discomfort dimension across condition groups 121
Figure 19: Comparison of EQ-5D-Y-3L-IA Worried, Sad or Unhappy dimension across condition groups 122
Figure 20: EQ-5D-Y-3L-IA utility scores in 5-7-year-olds and 8-10-year-olds 123
Figure 21: EQ-5D-Y-3L-IA VAS scores in 5-7-year-olds and 8-10-year-olds 124
Figure 22: Scatterplot of EQ-5D-Y-3L-IA utility scores versus VAS scores for 5-7-year-olds and 8-10-year-olds 125

1. Introduction

1.1. Background

The term Health-related Quality of Life (HRQoL) has been described as a 'modern' approach to healthcare as it considers the physical, social, mental and emotional impact of a health condition on an individual's wellbeing and steers away from focusing solely on the health condition presented (11). HRQoL is a multidimensional subjective measure of these physical and psychosocial factors in the context of an individual's daily life (12). In the past, an emphasis had been placed on measuring HRQoL in the adult population, therefore limiting research in paediatric populations (2,8,9,13,14). A major contributor to this was the lack of paediatric HRQoL instruments, however, with more instruments being developed, a shift to paediatric health is now being seen (1). By introducing these multidimensional instruments, it encourages healthcare professionals to create more holistic treatment plans (2,3) which address all dimensions and provide a means of monitoring the effectiveness of their plans (4).

These multidimensional instruments are often referred to as preference-based HRQoL instruments which aim to "assess patient preference across broad areas including symptoms, physical functioning, work and social activities and mental wellbeing" (11 (p.37)). In 1990, the EuroQoL group developed a generic preference-based instrument, the EuroQoL Five Dimension (EQ-5D), which aimed to subjectively measure HRQoL in various adult populations (11). The group later adapted the EQ-5D to create a youth version, EQ-5D-Y-3L self-complete version (EQ-5D-Y-3L-SC) intended for children aged 8-15-years (3). The EQ-5D-Y-3L-SC assesses five dimensions: Mobility (Mob), Looking After Myself (LAM), doing Usual Activities (UA), having Pain/Discomfort (P/D) and Feeling Worried, Sad or Unhappy (WSU) based on three levels of report: 'no problems', 'some problems' and 'a lot of problems'. A Visual Analogue Scale (VAS) is included to assess overall health from 0 representing the worst health imaginable to 100 representing the best health imaginable. The five dimensions and VAS both assess HRQoL on the day of testing (3). The EQ-5D-Y-3L-SC has successfully been used in studies in South Africa to measure health and changes over time in children from the recommended age of 8-years (2,3,16–18).

The EQ-5D-Y-3L is one of approximately 30 generic HRQoL instruments that were developed over the last two decades specifically for the paediatric population (19). The EQ-5D-Y-3L has been translated into various languages including all 11 of South Africa's official languages (20) and has been used internationally in paediatric populations with and without health conditions (2,5,13,16). The National Institute for Health and Care Excellence (NICE) currently endorses the adult EQ-5D, from which the youth version was developed, for health technology appraisal in adults (21). However, the modes of administration remains limited especially in younger children who understand the concept of health (22) but lack the appropriate literacy levels to self-complete a questionnaire, and therefore have to rely on proxy-report which has been shown to be problematic with responses often mismatched between children and parents (2,8–10). Despite these mismatches, proxy-report remains an integral part of recording HRQoL especially in very young populations and children with cognitive impairments as both groups may not have the ability to provide any form of self-report (23). Therefore, the need for more interviewer-administered versions for younger children who are unable to read but able to understand and provide feedback, have become increasingly important to allow children the opportunity to self-report their HRQoL instead of defaulting to proxy-report.

1.2. Rationale and significance

Since 2001, the Progress in International Reading Literacy Study (PIRLS) assesses and monitors reading literacy at five-year intervals in over 60 countries to create international standards for reading literacy. In South Africa, as of 2016, PIRLS estimated that 78% of children between the ages of 9-10-years have not yet mastered basic reading by the end of their fourth year of formal schooling, compared to a mere 4% internationally (24). This would directly affect their ability to self-complete any HRQoL instrument despite their age or level of education suggesting otherwise. However, this may not affect their self-report ability if the concepts are understood.

Similarly, if younger children do not have the necessary literacy levels to self-complete an instrument due to their age and level of education but can also understand the concepts which HRQoL instruments address, these children would be able to self-report their HRQoL. Studies have suggested that children as young as 5-years-old, with varying health conditions, are able to reliably report their HRQoL by means of an interviewer-administered questionnaire (1,23). Considering the subjectivity of HRQoL, the recommendation for self-report and the recent literacy statistics in South Africa, the newly developed EQ-5D-Y-3L-IA (referred to as IA from here) could be useful in obtaining HRQoL information in older children with lower literacy levels or directly from the younger child who may not be able to

read yet, but who understands the concept of HRQoL. Although some studies have found proxy-report to be reliable, it is fair to say that it cannot be used in all dimensions, with a noticeable difference in the report when comparing psychosocial dimensions (16,23). The IA has the potential to allow younger and older children to subjectively report their HRQoL, therefore supporting the aim of this study.

The newly developed IA has yet to be evaluated for validity, reliability or feasibility, therefore warranting the need for this study. In order to evaluate the IA and its performance in various paediatric health conditions, it was tested in children from the general population (GenPop), children with functional disabilities, chronic respiratory illnesses and orthopaedic conditions. A variety of paediatric health conditions were included as self-reported HRQoL is a very important aspect in ensuring holistic medical and social care.

1.3. Aims

The two aims of this research study are as follows and will be presented in Chapter 3 and 4 respectively:

1. To compare the performance and preference of the IA to the SC version in children aged 8-10-years.
2. To compare the performance of the IA in children 5-7-years compared to children aged 8-10-years.

1.4. Specific objectives

Comparing the performance of the IA vs SC in children aged 8-10-years

- To assess the feasibility of the IA vs SC in children 8-10-years by:
    - Evaluating the ceiling effect by the proportion of respondents reporting level 1, indicating 'no problems', across all five dimensions. Ceiling effects occur when participants report the maximum score on a Patient Reported Outcome Measure (PROM) (25).
    - Evaluating the floor effect by the proportion of respondents reporting level 3, indicating 'a lot of problems', across all five dimensions. Similarly, floor effects occur when participants report the minimum score on a PROM (25).

    - Comparing the time taken to complete each instrument.
- To determine the proportion of inconsistent responses between the IA and SC versions and whether it differs by age, sex or health condition.
- To determine and compare the known-group validity of the IA and SC by comparing the HRQoL profiles of children known to belong to the GenPop or receive management for their medical conditions, including orthopaedic, chronic respiratory illness or functional disabilities.
- To determine and compare the concurrent validity of the IA and SC VAS scores and utility scores.
- To investigate the convergent validity of the IA and SC dimensions by comparing them to similar items on:
    - The Faces Pain Scale-Revised (FPS-R), a paediatric pain measure to compare to the IA and SC dimension of P/D.
    - The Mood and Feelings Questionnaire (MFQ), a measure of emotional wellbeing to compare to the IA and SC dimension of WSU.
    - The Functional Independence Measure (WeeFIM), a paediatric measure of functional independence, to compare to the IA and SC dimensions of Mob, LAM and UA.
- To determine whether children aged 8-10-years preferred using the SC or IA version by asking for their preference between versions. Participants were also asked to provide a reason for their preference

Comparing the EQ-5D-Y-3L-IA completion in children aged 5-7-years vs 8-10-years

- To assess the feasibility of the IA in children aged 5-7-years and 8-10-years by:
    - Evaluating the ceiling effect by the proportion of 5-7-year-olds reporting level 1, indicating ‘no problems’, across all five dimensions and comparing it to the ceiling effect found in 8-10-year-olds. Ceiling effects occur when participants report the maximum score on a PROM (25).
    - Evaluating the floor effect by the proportion of 5-7-year-olds reporting level 3, indicating ‘a lot of problems’, across all five dimensions and comparing it to the floor effect found in 8-10-year-olds.. Similarly, floor effects occur when participants report the minimum score on a PROM (25).
    - Comparing the time taken to complete the instrument in each age-group.
- To determine the known-group validity of the IA in 5-7-year-olds by comparing the HRQoL profiles of children known to belong to the GenPop or receive management for their medical

conditions including: orthopaedic, chronic respiratory illness or functional disabilities and comparing it to the known-group validity found in 8-10-year-olds.

- To determine the concurrent validity of the EQ-5D-Y-3L-IA VAS score and utility scores in 5-7-year-olds and comparing it to the concurrent validity found in the 8-10-year-olds. .
- To investigate the convergent validity of the IA dimensions in 5-7-year-olds and comparing it to the convergent validity found in the 8-10-year-olds by comparing them to similar items on:
    - The FPS-R, a paediatric pain measure to compare to the IA dimension of P/D.
    - The MFQ, a measure of emotional wellbeing to compare to the IA dimension of WSU.
    - The WeeFIM, a paediatric measure of functional independence to compare to the IA dimensions of Mob, LAM and UA.
- To assess test-retest reliability of the IA in children from the GenPop and those with stable functional disabilities, 48 hours after initial testing. The time interval of 48 hours was proved to be suitable as it is a long enough period for children with a stable health condition not to remember their initial score (26).
- To establish the perceived difficulty in answering the IA questions by age-group by assessing the comprehensibility by means of cognitive debriefing interviews with the children and a questionnaire for the interviewer on completion of each interview.

1.5. Outline of the study

In preparation for this study, a comprehensive literature review (Chapter 2) was conducted focusing on the history behind the term HRQoL, the increasing importance of measuring HRQoL specifically in the paediatric population and the evaluation of existing paediatric HRQoL instruments using the Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) Risk of Bias checklist (27). Following this review, the methodology, results, discussion and conclusion of each aim will be presented in Chapters 3 and 4, respectively. The dissertation will be concluded in Chapter 5, including a final conclusion, study limitations and recommendations for practice, future research and policies.

2. Literature Review

2.1. Introduction

This chapter provides a comprehensive review of existing literature addressing the various objectives of this study. Literature and data were sourced from the following online databases: Academic Search Premier, CINAHL, Health Source Nursing/Academic Edition via EBSCOhost, Google Scholar and PubMed. Keywords used in the search were as follows: Health-Related Quality of Life/HRQoL, Quality of Life/QoL, HRQoL paediatric/paediatric measures, EQ-5D-Y-3L Self-Complete, HRQoL proxy vs self-report, psychometric properties, validity, reliability and feasibility. Pearling was used to uncover additional information about a specific topic mentioned in an article and to ensure all relevant articles were included in the review. Wildcard symbols such as p*ediatric, reliabl*, valid* and feasibl* were used to maximize search results across databases. Boolean operators such as AND and/or OR were also used to combine terms in order to maximise search results. Examples of this include, HRQoL in young children OR paediatrics OR school-going children OR 5-10-year-olds or HRQoL in young children using self-report AND proxy-report.

Only full-text journal articles and web articles in the English language were included in the review. No time limit on research was set as research surrounding HRQoL has increased since the 1990's. However, literature from the last decade was highlighted as it was a very productive period regarding HRQoL research, specifically in the paediatric population. Generic HRQoL instruments in the English language, developed for children between the ages of 5-7-years were included to identify appropriate modes of administration of HRQoL instruments for this specific age-group based on their cognitive development and literacy skills.

The aim of this review was to define the many meanings associated with the term HRQoL, how they differ or relate and the factors which influence them. This was followed by the importance of measuring HRQoL, especially in paediatric populations with and without health conditions. Existing generic paediatric instruments developed for the paediatric population between the ages of 5-7-years were identified and described. Lastly, psychometric evaluation of these existing generic paediatric instruments were discussed using the COSMIN framework and appraised using the COSMIN Risk of Bias checklist.

2.2. Health status, Quality of Life and Health-Related Quality of Life

2.2.1. Health status

The WHO described health as "a state of complete physical, mental and social wellbeing, and not merely the absence of diseases and infirmity" (28 (p.1)). This definition highlights the importance of psychosocial factors rather than focusing on the physical impairments. It also suggests that health should not only be measured when a condition has been diagnosed, but throughout states of complete health as well. A limitation noted in this definition speaks to the fact that no individual is in a state of perfect health at any one given time (29). Additionally, with the advancement of medical interventions, chronic conditions are being better treated and managed therefore, an individual may not, by definition, be in a state of perfect health but when asked to subjectively report their health status, they might feel that their chronic illness has not affected their health status therefore reporting full health (29,30). Considering these limitations and the various definitions, Huber et al. (2011) recommended that the definition be altered in such a way that it includes adaptation by means of "the ability of an individual to adapt and self-manage in the face of social, physical and emotional challenges" (31 (p.1)). As many other health-related definitions, this one has also been challenged as the inclusion of social-wellbeing was not agreed upon by all (31). An alternative definition of health status includes, an individual's functional ability is assessed in relation to societal expectations of what physical and mental wellbeing looks like (32). Literature on health status is believed to have been the stepping stone in the development of the definition for HRQoL (33).

2.2.2. Quality of Life

Quality of Life (QoL) became increasingly important in healthcare as advancements in medical treatments either extend the length of life, sometimes at the expense of QoL, or enhance QoL with or without extending length of life. Providing one definition for QoL has been proven difficult as many variations and approaches have been discussed. The most commonly cited definition of QoL is that of the World Health Organization (WHO) which describes health as "an individual's perception of their position in life in the context of the culture and value systems in which they live in relation to their goals, expectations, standards and concerns" (27 (p.1405)). Ultimately, this definition emphasises the subjectivity of QoL all within the individual's daily environment, their ethnic/cultural beliefs/traditions, morals and values, all of which pertain to their desires/ambitions, the degree to which things are expected to be done and any relative anxieties and/or stressors. It is apparent that QoL is multi-

factorial and that each individual factor plays a significant role and should be carefully considered in order to accurately capture QoL in an adult or child (34). Therefore, QoL had also become more important to consider outcomes beyond morbidity and biological functioning (35,36). Many definitions emphasize the subjective influence in QoL, however, some authors explored the objectivity of QoL through the use of an objective instrument based on observations and measurement, such as physical ability, therefore, suggesting that both objective and subjective factors play a role in measuring QoL and could be measured simultaneously to ensure all aspects of an individual's health condition are addressed (37–39).

2.2.3. Health-Related Quality of Life

Karimi & Brazier (2016) identified four definitions that have been used to define HRQoL (33). The first definition describes HRQoL as "how well a person functions in their life and his or her perceived wellbeing in physical, mental and social domains of health" (36 (p.195)). This definition addresses a person's ability to perform activities of daily living (ADLs) while considering subjective feelings (40,41). The second definition suggests that non-health factors such as political or economic factors present in an individual's life does not affect their HRQoL (31). The third definition addresses an individual's own perception of their wellbeing in relation to an existing medical condition or intervention (42). Finally, the fourth definition refers to the "values assigned to different health states" (39 (p.83)) and can be calculated using Quality Adjusted Life Years (QALYs), "a measure of the value of health outcomes" (44 (p.1)), whereby zero is equivalent to death and conversely, one is equivalent to full health with values in between reflecting a loss in HRQoL (43). This implies that losses and gains can be aggregated which is useful in health economic decision-making (45).

Ferrans et al. (2005) argued that HRQoL distinguishes between factors which are largely health-related and those factors which are not health-related (39). Thus, as described by Ferrans et al. (2005), Spilker and Revicki (1996) developed a non-HRQoL taxonomy which comprised of "personal-internal", "personal-social", "external-natural" and "external-societal" (47 (p.194)). These domains address how an individual's intrinsic characteristics influences their experiences in an environment, their social circle and environment, their physical/natural environment and expectations created by society. Despite this distinction, factors which influence one's HRQoL in addition to non-health-related factors remain interrelated (47), as a disruption in an individual's health will undoubtedly have an impact on all factors of an individual's life (48).

Considering these many definitions, a consensus regarding HRQoL's defining characteristics has been reached, including that HRQoL is multidimensional, subjective and value-based therefore indicating that HRQoL is influenced by a several dimensions including physical, social, psychological and spiritual wellbeing rated by the individual (34). The subjectivity of HRQoL is constantly emphasized and stressed that it is the "perceived (health) status" of an individual (40 (p.21)). The next characteristic used is "dynamic" which indicates that HRQoL does not remain constant but will vary over time (44 (p.694)). Lastly, HRQoL encompasses both the negative and positive life experiences of an individual (50). Specific characteristics such as multidimensionality and subjectivity form the basis of the HRQoL definition used throughout this thesis.

2.3. Measuring health-related quality of life in the paediatric population

2.3.1. Importance of measuring health-related quality of life

According to the World Bank, as of 2019, approximately 25% of the world's population was between the ages of 0-14-years (51). Of the paediatric population, middle-income countries such as South Africa report high rates of under-five mortality with a rate of 37.5 deaths per 1000 live births which is a significant decrease in reporting since the mid-2000s which estimated 62 deaths per 1000 live births (52), with Sub-Saharan Africa as a whole having the highest rates of 76 deaths per 1000 live births (53). The decrease seen in South Africa is said to be partially due to the implementation of the Sustainable Development Goals set out by the United Nations (15). In line with goal number three, child and maternal health will be focused on by promoting and ensuring optimal health and wellbeing of these populations (15), therefore identifying an appropriate instrument to measure health is incredibly important as it would allow for the monitoring of both child and maternal health over time.

Due to their age, children are considered a vulnerable group. They are often dependent on adults to ensure that their rights are not overlooked as they lack empowerment and need assistance in identifying and addressing their needs (54). Children are commonly known as the future of society; however, this should not be the only contributing factor behind the assessment and monitoring of their health. Rather, children's health needs to be made a sole priority regardless of their societal role (54).

Due to the large emphasis placed on the adult population, paediatric HRQoL remains limited (2,9,10,14). Over the last decade, more HRQoL instruments have been developed and assessed specifically for the paediatric population (1) as their health has become of more interest compared to older populations (17). Recent literature has also highlighted the importance of measuring HRQoL in this population, especially in guiding healthcare professionals in developing more patient-centered and holistic treatment plans which identify and consider psychosocial factors in addition to physical impairments (3,8,16). It will also allow for monitoring of treatment plans over time to determine its effectiveness or need for modifications (16). By adopting a more holistic and patient-centered treatment plan, it allows children to feel as though their concerns are not dismissed but rather incorporated into their treatment plans where possible. This goes hand in hand with article 12 of the United Nations Convention on the Rights of the Child, which states that "children have a right to have their views taken into account in matters that affect them" (25 (p.132)). This type of approach has the ability to improve patient-clinician relationships and, as a result, may improve treatment outcomes (3,16).

2.3.2. Challenges in measuring health-related quality of life - proxy-report vs self-report

The development of health-related PROMs in the paediatric population remains a 'conceptual' and 'methodological' challenge as appropriate content, method of report, accessibility and target age-group needs to be considered (22). As described earlier, HRQoL is a subjective measure (12); therefore self-report is the preferred method of HRQoL evaluation and should be used as far as possible (55). However, it remains important to take age and cognitive ability into consideration as very young children or children with cognitive impairments may not be able to accurately report on their HRQoL, therefore, supporting the need for proxy-report (23).

Proxy-reporting has been used largely in the paediatric population despite parents and children often sharing different views (6) as children and adults prioritize their physical and emotional wellbeing differently, therefore they may report conflicting information when asked about each dimension (16). When comparing HRQoL in 7-12-year-olds from 567 GenPop children and 61 children with functional disabilities attending special schools, parents of children with functional disabilities reported poorer HRQoL compared to their child based on their VAS scores. While females and their female caregiver showed a low and insignificant correlation of r=0.16, males and their female caregiver also showed an insignificant but higher correlation of r=0.67 (17). Jelsma and Ramma (2010) attributed this to parents

viewing the functional disability as more of a limitation compared to how their child views their disability. As a result, proxy-report remains problematic and may not accurately capture the HRQoL from the child's perspective (56–60).

As discussed in chapter one, it has been suggested that children as young as 5-years may be able to reliably report on their HRQoL, although these young children may not have the necessary literacy skills (1), thereby affecting their ability to self-complete a measure. This may also be true for older children with lower literacy levels within low socioeconomic settings (5) or who have limited formal schooling due to ill health. However, if the concepts asked are understood by both groups, younger and older children will be given the opportunity to provide information about their health to an interviewer (61) without the pressure of self-completing an instrument. Thus, it is the opinion of the researcher, if the IA version proves valid and reliable in the younger age-group, it has the potential to extend the lower age-range for self-reported HRQoL, by alternate means of completion and therefore, eliminate the need to use proxy-report in the age-group between 5-7-years. Proxy-reporting remains problematic, with a noticeable difference between self-report and proxy-report (56–58,60) with mismatches often occurring in psychosocial dimensions suggesting that parents are more aware of their child's physical health rather than their mental health (62). Literature suggested that proxy-report may not accurately capture the health states of children (59) and reporting could be over- or underestimated based on bias, therefore self-report should be encouraged as far as possible (61). Proxy results may prove useful when comparing results between caregivers in relation to their child but not when compared to children's self-report and therefore should not be used interchangeably (17). Although proxy-reported HRQoL is not always accurate, caregivers' perceptions on their child's HRQoL remains important as they are often the decision-makers in their child's healthcare (57).

2.4. Health-related quality of life instruments for the paediatric population

Over the last 10 years, many health-related PROMs specifically for the paediatric population of varying ages and health conditions have been developed and assessed for validity, reliability and feasibility (1). The next section will provide a brief overview of these health-related PROMs and will later be appraised using the COSMIN Risk of Bias checklist.

2.4.1. Generic and disease-specific measures

Generic HRQoL measures are developed for a variety of health populations and are not limited to populations with specific health conditions (4). Although generic measures may not be sensitive enough to identify all changes in health, when these instruments are used in GenPop, results can be used comparatively or as a control in relation to populations with health conditions (63). Alternatively, disease-specific HRQoL measures were developed for populations with specific health conditions in order to measure disease-related symptoms or the effect of treatment methods which is not possible when using a generic HRQoL instrument due to its lack of sensitivity to health conditions (63). Generic and disease-specific measures are both able to measure change over time and can be scored with population norms, summary scores or utility values. The health state utility value obtained from these measures allow for QALYs to be calculated (64).

In keeping with the aim of this review to describe and evaluate generic HRQoL instruments, the inclusion process of identifying generic health-related PROMs for children aged 5-7-years is shown in Figure 1 below.

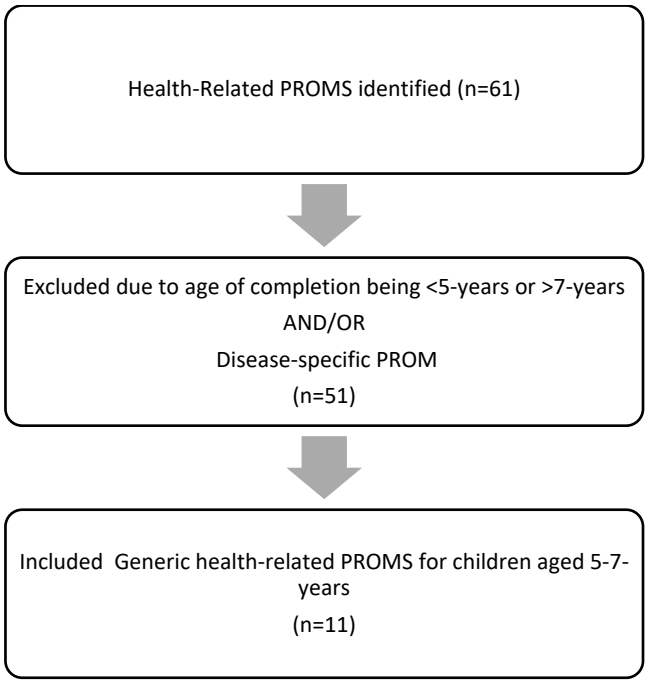


Figure 1: Inclusion process for health-related PROMs for children aged 5-7-years

As a result, 11 generic health-related PROMs (Figure 1) were included, described below and appraised further on in the review: the EQ-5D-Y-3L, Paediatric Quality of Life Inventory 4.0 Generic Core Scales

(PedsQL), Health Utilities Index Mark 3 (HUI_3), KINDER Lemensqualitatsfragebogen: Children Quality of Life-Questionnaire Kiddy-KINDL version (Kiddy-KINDL), the Child and Health Illness Profile-Child-Report Form/Child Edition (CHIP-CRF/CE), Patient-Reported Outcome Measure Information System-Paediatric Global Health Measure (PROMIS-PGH-7), Functional Status II (R) (FSIIR), Child Health Utility 9D (CHU-9D), DISABKIDS-TAKE-6, Netherlands Organisation for Applied Scientific Research Institute of Prevention and Health and Leiden University Hospital (TNO-AZL) Questionnaire for Children's Health-Related Quality of Life-Parent Form (TACQoL-PF) and Child Health Questionnaire (CHQ) (Table 1).

2.4.2. Summary of generic health-related PROMs for children aged 5-7-years

Table 1 shows a summary of the 11 generic health-related measures identified for children aged 5-7-years. The age of completion, method of completion, estimated time needed for completion and dimensions assessed, which vary between instruments with some overlap, are included in the table. The age-range for completion of these instruments range from a young age of 2-years, all the way to 18-years. Considering this wide age-range, the applicability of dimensions assessed may not be appropriate for all ages based on the developmental ability of children in different stages of their life. Therefore, instruments should aim to reduce their age-ranges, or include age specific questionnaires, in relation to the dimensions assessed so that applicability and appropriateness of dimensions are both ensured (22). Considering the country of development, all of the instruments were developed in the Northern Hemisphere in High Income Countries except for the EQ-5D-Y-3L which was developed in Europe and South Africa. Of the 11 instruments, seven of them assesses a large number of items covering various aspects of physical, social and emotional functioning. The more items assessed, the longer the time of completion becomes, which has a direct effect on feasibility especially in clinical settings where time is often limited (65). The scoring system of most instruments are based on a summated score with the HUI_3, CHU-9D and EQ-5D-Y-3L using preference-based scores which rely on societal-preference derived from elicitation techniques including 'standard gamble', 'time trade off' or 'discrete choice experiment'. Standard gamble is based on the utility theory by von Neumann and Morgenstern "to elicit the utility of a health state or health intervention" (67 (p.333)). A utility is a number that represents the strength of the individual's preference for a particular health outcome, in the face of uncertainty. It considers whether an individual is willing to accept a certain risk of death in order to avoid a certain health state. 'Time-trade off,' referring to an individual being prepared to essentially 'trade-off' of their current life expectancy to avoid prolonged time in a undesirable health state (66). In a 'discrete choice experiment' in the context of the EQ-5D instruments, participants are given a set of different health profiles and asked to choose the one which they prefer (67–69).

Table 1: Summary of generic health-related quality of life instruments for children aged 5-7-years

	Health-Related PROM	**County of origin**	**Methods of completion available**	**Age-range for completion (years)**	**Estimated time for completion (minutes)**	**Total items assessed**	**Description of dimensions**	**Scoring system**	**References**
1	EQ-5D-Y-3L	Multinational including Europe and South Africa	- SC - Proxy - IA	4-8 5-8/9	5	6	- Mobility - Self-care - Usual Activities - Pain or Discomfort - Worried, Sad or Unhappy - General health measured on a Visual Analogue Scale (VAS)	3-point Likert scale measuring severity VAS from 0-100	(3,10)
2	PedsQL	USA	- SC - Proxy	5-18	5-10	23	- Physical functioning - Emotional functioning - Social functioning - School functioning	5-point Likert scale rating frequency	(70)
3	HUI_3	Canada	- SC - Proxy	2-18	5-10	45	- Vision - Hearing - Speech - Ambulation	Scale ranging from 0-100	(67,71)

	Health-Related PROM	County of origin	Methods of completion available	Age-range for completion (years)	Estimated time for completion (minutes)	Total items assessed	Description of dimensions	Scoring system	References
							- Dexterity - Emotion - Cognition - Pain		
4	Kiddy-KINDL	Germany	- SC via IA	4-7	15	12	- Physical Health - Family functioning - Self-esteem - Social functioning - School functioning	3-point Likert scale Converted to a scale of 0-100	(6)
5	CHIP CRF/CE	USA	Proxy	6-11	15	76	- Satisfaction - Comfort - Resilience - Risk Avoidance - Achievement	5-point Likert scale rating frequency	(22,72)
6	PROMIS-PGH-7	USA	- SC - Proxy	5-17	1-2	7	- Physical health - Social health - Mental health	5-point score for individual items	(73)

	Health-Related PROM	County of origin	Methods of completion available	Age-range for completion (years)	Estimated time for completion (minutes)	Total items assessed	Description of dimensions	Scoring system	References
7	FSIIR	USA	- Proxy	>4	15-30	43 14 (short form available)	- General health - Interpersonal functioning	3-point Likert scale rating difficulty and the extent due to illness	(4,22)
8	CHU 9D	UK	- SC - Proxy	7-11	3-5	9	- Worried - Sad - Pain - Tired - Annoyed - Schoolwork - Sleep - Daily routine - Activities	5 ordinal levels	(22,74)
9	TAKE-6	Europe	- SC via IA - Proxy	4-7	2	6, 12 or 37	- Mental - Social - Physical	Ordinal scale from 1 to 5 Converted to a scale of 0 to 100	(7)
10	TACQOL-PF	Holland	- Proxy	5-15	10	56	- Physical complaints - Motor functioning	4-point scale	(75)

	Health-Related PROM	County of origin	Methods of completion available	Age-range for completion (years)	Estimated time for completion (minutes)	Total items assessed	Description of dimensions	Scoring system	References
							- Autonomous function - Social function - Cognitive function - Positive emotions - Negative emotions		
11	CHQ-PF 28/50	USA	Proxy	5-18	5-25	28	- Physical function - Pain - Role/social-physical - General health - Perception - Role/social emotional behaviour - Mental health - General behaviour - Self-esteem - Parental emotional impact	4 to 6 levels of reporting	(22,76)

	Health-Related PROM	County of origin	Methods of completion available	Age-range for completion (years)	Estimated time for completion (minutes)	Total items assessed	Description of dimensions	Scoring system	References
							- Parental time impact - Family impact		

2.4.3. Descriptions of generic health-related PROMs for children aged 5-7-years

2.4.3.1. EQ-5D-Y-3L

The EQ-5D-Y-3L, a youth-friendly version, was developed by an international task team from the EuroQoL group in 2010 to assess HRQoL in the paediatric population and was adapted from the adult version, the EQ-5D (3). The intended age for self-completion is 8-12-years. The proxy version of the instrument is recommended for use in children 4-8-years (10). The EQ-5D-Y-3L assesses a child's HRQoL on the day of data collection in five different dimensions: Mobility (Mob) (walking about), Looking After Myself (LAM) (washing or dressing), doing Usual Activities (UA) (for example, going to school, sports, hobbies, playing and doing things with friends or family), having Pain/Discomfort (P/D) and feeling Worried, Sad, or Unhappy (WSU). Each dimension has three levels of reporting in terms of having 'no problems', 'some problems' or 'a lot of problems'. A VAS is also included in the measure. The scale ranges from zero, which reflects the worst health ever imagined to 100 which reflects the best health ever imagined. The EQ-5D-Y-3L uses a preference-based scoring system. South Africa participated in the development and initial validity and reliability testing of the EQ-5D-Y-3L (2). It has been translated into various South African languages, including South African English, Afrikaans, Northern Sotho, Sesotho, Setswana, isiXhosa and Zulu (20). The EQ-5D-Y-3L is valid and reliable for use in South African children aged 8-15-years from the GenPop and with a health condition (2,5,13,16).

The EQ-5D-Y-3L has recently been adapted to allow for an interviewer-administered version and could be useful in obtaining HRQoL information directly from children younger than 8-years who may not be able to read, but who understand the concepts surrounding their health (22,32). Considering that this version is recently adapted, it needs to be tested for validity, reliability and feasibility in the target population of young children.

2.4.3.2. Paediatric Quality of Life Inventory 4.0 Generic Core Scales

The PedsQL Generic was developed to assess HRQoL in the paediatric population in four dimensions: physical functioning, emotional functioning, social functioning and school functioning (70). A proxy and self-complete version were developed for children from the age of 5-18-years and can be used in the GenPop and in children with a chronic or acute condition. The PedsQL uses a summated score with a rating scale of zero to four, the higher the score, the more problems reported (77). The PedsQL has

been used in both local and international populations and has successfully tested for construct validity, reliability, responsiveness and feasibility in both GenPop and disease-specific populations (70,77). The PedsQL has also been translated into various South African languages including, South African English, Afrikaans, isiXhosa, Setswana, Sesotho and Zulu (78).

2.4.3.3. Health Utilities Index (Mark 3)

The HUI was the first version created which was used in assessing outcomes for very-low birth weight infants. After which, the HUI_2 was developed for children with childhood cancer. The HUI_3 was later developed to create a version that could be used in children with and without health conditions and assesses physical and psychological functioning by looking at 'vision', 'hearing', 'speech', 'ambulation', 'dexterity', 'emotion' and 'cognition' (71). A preference-based score is used by which the higher the score for each dimension and overall, the better the HRQoL reported (79). The results of the literature search did not find any reference to the HUI instruments being used in the South African context in children. The HUI_3 however, has been validated in Sub-Saharan Africa, more specifically Uganda (15).

2.4.3.4. Kiddy-KINDL version

The KINDL offers a range of HRQoL instruments for three different age range of the paediatric population, including young children and adolescents. These versions can be used in children/adolescent with and without health conditions. The Kiddy-KINDL was developed for children in the age-range of 4-6-years and is completed by means of an interviewer which allows for self-report. This instrument assesses physical and psychosocial wellbeing, relationships with family and friends, self-esteem and everyday functioning. Scoring is based on a summated score, the higher the score reported on each dimension, the better the HRQoL (6).

2.4.3.5. The Child Health and Illness Profile –Child Report Form/Child Edition

The development of the CHIP CRF/CE began in 1995 at the same time literature, although limited, suggested that young school-going children were able to self-report on some aspects of their health (72). The dimensions assessed on this version are based off of the adolescent edition and parent-report form. These dimensions include: satisfaction with themselves and their health, comfort relating to physical and emotional symptoms/limitations in ADLs caused by illness, inter- and intra-personal

relationships in relation to resilience and enhancement of health, avoidance of risky/dangerous behaviour which has the potential to negatively affect health and age- or developmentally-appropriate academic and social performance. The higher the score on dimensions, the better the health is believed to be and uses a summated scoring system (72).

2.4.3.6. Patient-Reported Outcome Measure Information System-Paediatric Global Health Measure

The PROMIS-PGH-7 was based off the concepts assessed in the adult version. The paediatric version aimed to be shorter, practical and offer a complete overview of children's physical and psychosocial wellbeing that can be used in research and clinical settings to improve and monitor HRQoL in children. The PROMIS-PGH-7 offers two versions, a self-report for children 8-17-years and a proxy-report for children 5-17-years (73). When scoring the instrument, summated scoring is used with higher scores associated with better HRQoL (80).

2.4.3.7. Functional Status II (R)

The FS I assesses age-appropriate functional ability and the effect an illness can have on an individual's function. The FS I was later revised to create the FSIIR with the purpose of clarifying instrument items, testing in a wider age-range of children from two-weeks-old to 16-years-old, to determine the most appropriate cut-off ranges and finally, to assess the psychometric properties of the newly revised version. Thus, a 14-item instrument was created to assess age-appropriate communication, mobility, mood, eating, energy and sleeping by means of proxy-report. When/if an impairment is noted, parents are asked to indicate whether the impairment is a result of an illness. A summated scoring system is used for the FSIIR. Two scores are calculated from the 14-items: a 'Total Score' calculated by adding scores from all 14-items therefore indicating functional ability with or without an illness and the 'Illness Score' by which items affected by an illness are subtracted from the total score. The higher the score, the better functional ability. The FSIIR, has been made available in English and French (81,82).

2.4.3.8. Child Health Utility-9D

The CHU-9D is a nine-dimensional measure that may be used in children and adolescents aged 7-17-years by assessing their feelings towards schoolwork, sleep, daily routine, ability to join in on activities

and a sense of feeling worried, pain, sad, tired or annoyed by rating each dimension using a numerical scale of one to five. The higher the score, the poorer the result. Scoring is done using preference-based scores (83,84). The CHU-9D has not been translated into any South African languages nor has it been tested for reliability or validity in a South African context.

2.4.3.9. DISABKIDS-TAKE-6

The DISABKIDS-group identified the need for more research in younger paediatric populations as a large amount of research has been done in children aged 13-18-years. This group also emphasised the need for self-report measures for these younger populations due to the subjectivity of HRQoL. As a result, the TAKE-6 was developed for children between the ages of 4-7-years with and without health conditions. The DISABKIDS-TAKE 6 assesses physical, mental and social health using five smiley faces for children to rate each dimension. Summated scores are used by converting responses to a score ranging from 0-100, with a higher number indicating better overall wellbeing. This instrument has been successfully translated into six additional languages including: Dutch, English, French, German, Greek and Swedish (7).

2.4.3.10. Netherlands Organisation for Applied Scientific Research Institute of Prevention and Health and Leiden University Hospital Questionnaire for Children's Health-Related Quality of Life-Parent Form

The TACQoL, developed in the Netherlands by the TNO-AZL, aimed to create a generic HRQoL instrument for children aged 5-16-years. At the time, young children were not believed to be able to report on their health reliably therefore, the focus shifted to creating a parent-form, as it was assumed that parents are to be one of the best sources regarding their child's wellbeing. Thus, the TACQoL-PF was developed by assessing seven dimensions of HRQoL including: 'pain and symptoms', 'basic motor functioning', 'social functioning', 'global negative and positive emotional functioning'. Scoring of the TACQoL is done by summated scores (85).

2.4.3.11. Child Health Questionnaire

The CHQ was developed after a Child Health Assessment Project was created in 1990, which searched for methods to measure HRQoL in children. The aim of the CHQ was, therefore, to assess all aspects

of HRQoL, which included: physical functioning, social and mental functioning, pain/discomfort, relationships with parents and family, self-esteem and overall health in children/adolescents (5-18-years) with and without health conditions. Overall, HRQoL is determined by summated scores, by converting individual dimensions scores to a score ranging from 0-100 whereby a higher score is associated with better HRQoL (86). The CHQ has been translated from American-English to Canadian-French, German and United Kingdom-English (76).

2.4.4. Summary of generic health-related PROMs for children aged 5-7-years

As seen from Table 1, it is evident that the use of self-report via interviewer-administration is scarce amongst health-related PROMs with only two instruments, the DISABKIDS-TAKE-6 and the Kiddy-KINDL, offering it as a mode of administration despite this mode being the most subjective manner to gather information, other than self-complete, as it is obtained directly from the child thus, making it the most objective measure of HRQoL. If an instrument cannot be self-completed, the default seems to be proxy-report in most cases in spite of the mismatches found between proxy- and self-report (56–58,60). Both the CHU-9D and CHIP-CRF/CE have been administered using an interviewer but are yet to publish official interviewer-administered versions, therefore, limiting its standardisation and implementation in research and clinical settings (72,87). Both of the instruments with interviewer-administration were developed in higher income countries and neither offer translations into local South African languages therefore, limiting the use in a South African context. The EQ-5D-Y-3L and PedsQL SC versions are the only two generic PROMs that have been translated into South African languages, therefore, would be more commonly used in a South African setting. Although these versions are available, further consideration for the South African context is the comparatively lower literacy levels, possibly associated with socioeconomic status or inability to attend school due to a medical condition, which could exclude children from self-complete (24). In these instances, one would have to rely on proxy-report from children who may have the cognitive capacity to self-report but who are excluded due to their literacy.

Considering the dimensions assessed, all instruments show a holistic approach by assessing aspects of both physical and psychosocial health, although some instruments assess this with a limited number of items whereas others include multiple items per dimension. For example, the PedsQL assesses school and social functioning as independent dimensions with multiple items for each, while the EQ-5D-Y-3L assesses both aspects broadly in the UA dimension in addition to other activities such as hobbies and sports. Regardless of number of dimensions on the instruments, all 11 are consistent with

what the definition of HRQoL encompasses. To be able to use these HRQoL instruments in daily practice, they need to be psychometrically evaluated first.

2.5. Psychometric properties and appraisal of health-related patient reported outcome measures for children aged 5-7-years, according to COSMIN

Validity and reliability form the basis of psychometric evaluation and are equally important to quantitative research as they enhance the quality of said research (88). Validity and reliability ensures accurate measurement by an instrument and its reproducibility on different occasions by the same or different users (89). These psychometric factors will be discussed using COSMIN framework. In addition, The COSMIN Risk of Bias checklist will be used to appraise the generic health-related PROMs identified in the review (90). The COSMIN initiative includes an international multi-disciplinary team of researchers with a background in epidemiology, psychometrics, medicine, qualitative research and healthcare who specialize in the development and review of PROMs. COSMIN aims to improve the selection available and the suitability of health-related PROMs in research and clinical practice by using more transparent methodologies and practice tools (91).

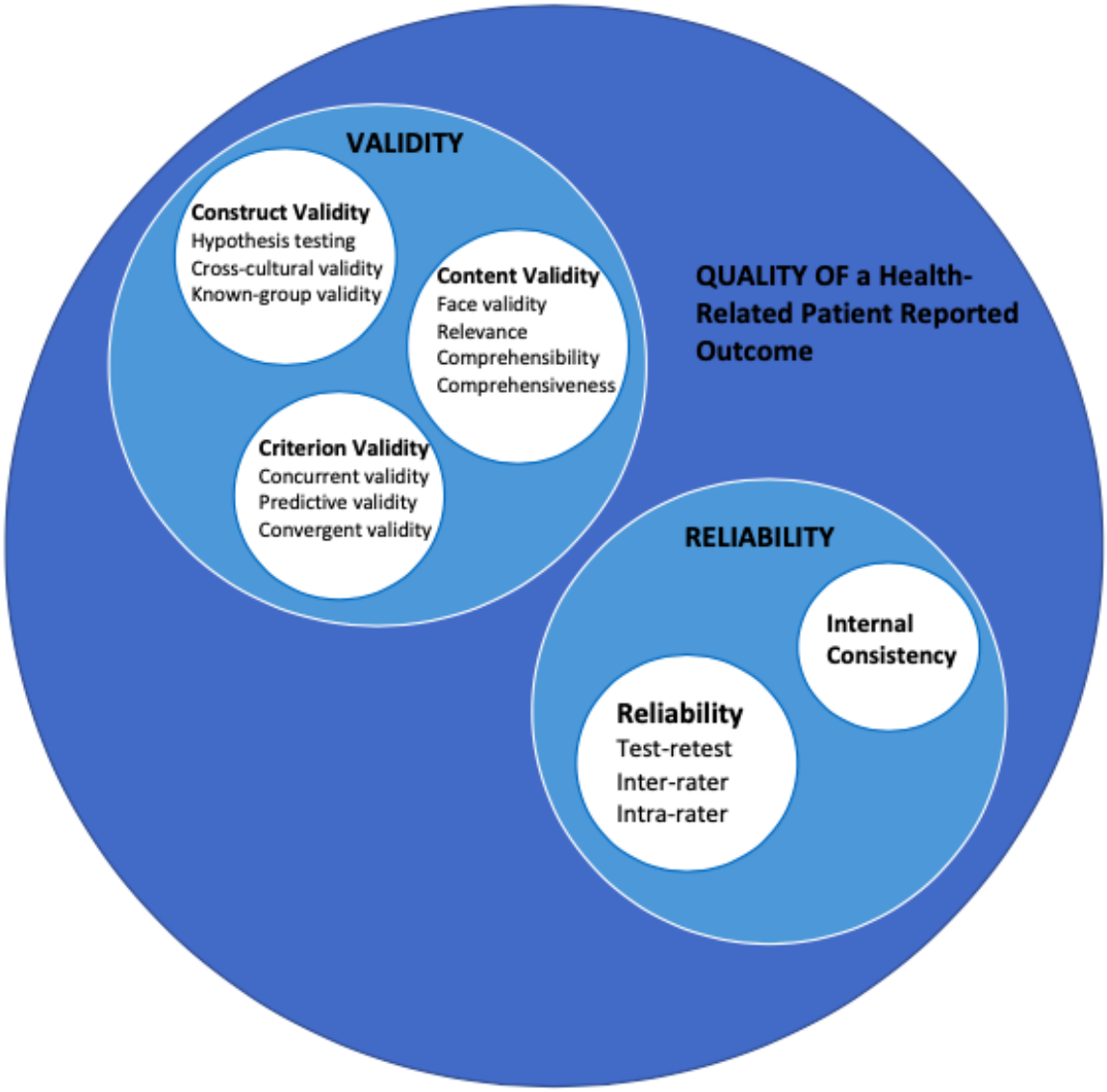


Figure 2: COSMIN taxonomy of measurement properties adapted by researcher from Mokkink et al. (2010) (25)

The health-related PROMs were appraised according to the COSMIN Risk of Bias Checklist (Appendices 1-11) and a summary of the psychometric properties are presented in Table 2 below.

Table 2: Summary of the all psychometric properties of health-related PROMs

	Validity					Reliability		Appendix
	Content	**Structural**	**Hypothesis testing and Known-group**	**Cross-cultural**	**Criterion**	**Internal Consistency**	**Test-retest Reliability**	
EQ-5D-Y-3L								1
PedsQL								2
HUI_3								3
Kiddy-KINDL								4
CHIP CRF/CE								5
PROMIS-PGH-7								6
FSIIR								7
CHU-9D								8
DISABKIDS-TAKE-6								9
TACQoL								10
CHQ								11

LEGEND	Tested 'very good' for >80% of criteria	Tested 'adequately' for >80% of criteria	Tested 'inadequately' for >80% of criteria	Tested 'doubtful' for >80% of criteria	Partially tested 'very good' for >80% of criteria	Partially tested 'inadequately' for >80% of criteria	Not tested

2.5.1. Validity

Validity is the degree in which a health-related PROM accurately measures what it intended to (89,90). According to COSMIN taxonomy there are three main types of validity: content validity, construct validity and criterion-related validity (Figure 2) (90).

2.5.1.1. Content validity

Content validity considers whether the health-related PROM accurately measures all aspects of HRQoL (89). This includes face validity which requires the respondent to offer an opinion, typically in qualitative interviews (92), on whether they think the instrument measures HRQoL (93). COSMIN suggests three criteria should be assessed when considering content validity: relevance, comprehensiveness and comprehensibility (94). Relevance ensures that the content being assessed remains relevant to the context in which it is being assessed (94). Comprehensiveness refers to the content of the instrument and whereby no relevant items are left out, while comprehensibility refers to the ability to understand the instrument (95). Literature stresses the importance of not assuming a HRQoL instrument developed for the adult population may accurately measure a child's HRQoL as its content may not be suitable or relevant to younger populations therefore child-centered instruments need to be newly developed or adapted from an adult instrument (61).

Content validity was tested in five out of the 11 instruments by means of cognitive debriefing with participants (children/parents) and professionals involved in the administration of the instrument (Table 2) However, two of the five instruments did not involve either patients or professionals in the cognitive debriefing therefore only partially tested content validity (PROMIS-PGH-7 and FSIIR). While the EQ-5D-Y-3L (Appendix 1), PedsQL (Appendix 2) and the DISABKIDS-TAKE-6 (Appendix 9) tested all aspects of content validity (Table 2).

The content validity of the EQ-5D-Y-3L was established in the development of the instrument by adapting the original adult version, the EQ-5D, by conducting cognitive interviews once participants had self-completed the measure (3). In order to obtain maximum feedback, various methods were included in the cognitive interviews. The 'paraphrasing' method included participants repeating the item in their own words to determine if they understood what was being asked, the 'general probing' technique allowed participants to suggest alternative words or phrases and lastly, a 'understanding numerical scale method' was used whereby participants rated their level of understanding from zero

to 10 (3). The relevance and comprehensiveness of the EQ-5D-Y-3L was tested in children 8-18-years using cognitive interviews, in a multinational study and found that almost all of the participants were able to self-complete the measure without assistance therefore indicating good comprehension of the measure with only minor cultural and language modifications needed (61). When using the EQ-5D-Y-3L in a multinational study, therapists found the measure to be "easy and quick" (6 (p.12)) and reported that the measure made them aware of aspects of a child's HRQoL that they were previously unaware of, especially with regards to the psychological dimensions which addressed P/D and feeling WSU (2).

The DISABKIDS-TAKE-6 was used in a European sample of 4-7-years with various chronic conditions. Content validity was assessed by means of cognitive debriefing during the administration of the instrument by identifying any ambiguous words or phrases reported by the caregiver and was recorded by the trained interviewer for future development of the instrument (7).

Similarly, when developing and refining the PedsQL, the first phase of development was met with extensive research and cognitive debriefing by means of open-ended questions and discussions with patients, parents and healthcare professionals to determine the most appropriate items that should be included. The aim of the second phase was to refine the items assessed on the instrument and was tested in a completely new but smaller disease-specific sample of patients, parents and healthcare professionals who also underwent cognitive debriefing after administration. Later, the third phase took place in a larger sample and similarly underwent cognitive debriefing post-administration. This process was repeated many times over a few years to ensure appropriateness of all items of the PedsQL for both cognitive and developmental ability of younger children (93).

2.5.1.2. Construct validity

Construct validity refers to whether a health-related PROM accurately measures what it claims to measure by evaluating the results and determining whether they are an accurate representation of the construct being measured (90). Construct validity can be divided into: structural validity hypothesis testing, cross-cultural validity and known-group validity.

2.5.1.2.1. Structural validity

Structural validity refers to the whether the outcome of a health-related PROM accurately reflects the "dimensionality" of the instrument and what was intended to be measured (90).

The PedsQL (Appendix 2), Kiddy-KINDL (Appendix 4), PROMIS-PGH-7 (Appendix 6), CHU-9D (Appendix 8) and CHQ (Appendix 11) successfully tested for structural validity while the remaining six instruments, including the EQ-5D-Y-3L (Appendix 1) did not.

2.5.1.2.2. Hypothesis testing

Hypothesis testing refers to the degree to which scores of a health-related PROM are consistent with what has been hypothesized based on relationships or differences between measures or within a measure (90). The Latin term '*a priori*' is often used when describing hypothesis testing and simply refers to a hypothesis being introduced before testing has occurred and results have been discussed (97).

The TACQoL (Appendix 10) was the only instrument out of the 11 that did not assess construct validity (hypothesis testing) while all other ten instruments successfully tested this (Table 2).

Mayoral et al. (2019) set hypothesis *a priori* based on literature findings that children with Type 1 diabetes would report similar HRQoL to the GenPop but with slightly worse physical wellbeing and emotional state. This was confirmed in the study with high ceiling effects on all EQ-5D-Y-3L domains except for having P/D and WSU (98). A set of *a priori* expectations were also set when comparing dimensions on the CHQ and HUI_3. Results obtained from this study found a strong correlation between pain dimensions, a moderate correlation between mobility and mental health dimensions, and lastly, a weak correlation for the general health perceptions (86). When using the FSIIR in a sample of parents of children diagnosed with asthma, it was hypothesised that poorer QoL would be associated with increased absenteeism at school and visits to healthcare facilities, social issues and physiological involvement. All expected hypotheses were proven except low QoL in relation to pulmonary function (82). The PedsQL also similarly used social influences such as poor access to healthcare facilities as a basis for a hypothesis whereby a lower score was expected to correlate to poorer access (70). A study using the PROMIS-PGH-7 also based part of their hypotheses on social influences such as low socioeconomic status along with presence of a medical condition and whether special educational

assistance is needed at school by which a poorer score is expected (99). The Kiddy-KINDL focused their hypotheses on psychosocial health by successfully hypothesising that a higher score of the Kiddy-KINDL would be associated with a lower score on an anxiety instrument (100). The EQ-5D-Y-3L and CHU-9D were also used to test hypotheses in children with cerebral palsy (CP) compared to children from the GenPop by stating that similar dimensions would show a high correlation. However, a weak correlation was found for all dimensions except the LAM which showed a strong correlation with the 'daily routine' dimension (83).

2.5.1.2.3. Known-group validity

Known-group validity addresses a health-related PROM's ability to discriminate between groups known to be different (101). As a result, scores from different known-groups including, the GenPop, acute health condition groups and chronic health condition groups, are expected to be different when compared (1,70).

Similarly to hypothesis testing, The TACQoL (Appendix 10) was the only instrument (Table 2) which did not test for known-group validity while all other instruments successfully tested for known-group validity.

The EQ-5D-Y-3L was able to differentiate between different severities of disease in children with Juvenile Idiopathic Arthritis (JIA), as measured on the Juvenile Arthritis Multidimensional Assessment Report (JAMAR) (102) and chronic kidney disease by which an increased severity was associated with a lower VAS score (103). The HUI_3 was also used to differentiate between different severities of chronic kidney disease, but in an adult population (104). The CHU-9D, CHIP-CRF/CE and PedsQL in addition to the EQ-5D-Y-3L have also been successfully used to differentiate between similar paediatric populations including children with respiratory illnesses, orthopaedic conditions, acute illnesses and functional disabilities often in comparison to children from the GenPop (2,5,105,9,10,16,59,83,87,102,103). The FSIIR was also used to differentiate between children with and without health conditions (81) while the PROMIS-PGH-7 and DISABKIDS-TAKE-6 were used in children with chronic health conditions (7,99).

2.5.1.2.4. Cross-cultural validity

This concept refers to the ability of a health-related PROM to be translated into various languages or adapted to different cultures in such a manner so it does not change what the original instrument intended to measure (90), therefore promoting inclusivity for people of all languages and cultures. In a South African context where a total of 11 official languages have been recognized (16), cross-cultural validity is exceptionally important to accurately measure HRQoL in all South African populations. To accurately measure their HRQoL, instruments need to be successfully validated cross-culturally.

All instruments except the Kiddy-KINDL (Appendix 4), PROMIS-PGH-7 (Appendix 6) and TACQoL (Appendix 10) assessed cross-cultural validity. The CHU-9D (Appendix 8) and HUI_3 (Appendix 3) performed slightly lower on the COSMIN checklist for cross-cultural validity compared to the EQ-5D-Y-3L (Appendix 1), PedsQL (Appendix 2), CHIP-CRF/CE (Appendix 5), FSIIR (Appendix 7), DISABKIDS-TAKE-6 (Appendix 9) and CHQ (Appendix 11) as they were 'adequately' and 'inadequately' tested compared the other six instruments being categorised as tested 'very good' based on the colour-coded ratings from the COSMIN Risk of Bias Checklist legend (Table 2).

Results of cross-cultural validity showed that most health-related instruments had been tested in developed countries such as United Kingdom, Europe, United States, Canada and Australia. A systemic review of HRQoL instruments that have been validated in sub-Saharan Africa, found that only two generic health-related PROMs were assessed for cross-cultural validity, which included the HUI_3 used in Uganda and the EQ-5D-Y-3L used in South Africa (15). The HUI_3 was forward-translated from its original English version to various languages spoken in Uganda however, the EQ-5D-Y-3L was not translated into South African English but the original UK version was used instead. Of the other eight generic instruments identified, none were developed in sub-Saharan Africa nor in a county with similar resource availability or restraints (15).

The EQ-5D-Y-3L has been successfully translated and used in various other languages and cultures including German, Italian, Korean, Swedish, Japanese and Spanish, etc. (9,14,16,30,59,106) with a total of about 40 versions available since 2019 (61). Both the CHIP-CRF/CE and FSIIR have been translated and used in Spanish populations (81,82,107) while the CHQ and DISABKIDS have been translated and used in a number of populations including Canadian-French, United Kingdom English, German, Greek, etc. (7,76).

2.5.1.3. Criterion validity

Criterion validity refers to how well a newly developed health-related PROM performs compared to an already established instrument, often known as the 'gold standard' (101) and can further be categorized into concurrent validity, convergent validity and predictive validity.

2.5.1.3.1. Concurrent validity

Concurrent validity refers to the relationship between responses of corresponding items from multiple health-related PROMs when tested simultaneously in order to predict outcomes on a different instrument (93). One of the limitations in assessing the concurrent validity of generic HRQoL is that there is no one 'gold standard' health-related PROM (108). Furthermore, each instrument assesses different and sometimes additional items or dimensions within the broader HRQoL framework, such as school functioning, autonomy, relationships with parents/peer and energy levels, therefore, limiting direct comparison between health-related PROMs (2,109). Although concurrent validity is not assessed in the COSMIN Risk of Bias checklist, it remains an important psychometric property to consider.

2.5.1.3.2. Convergent validity and divergent validity

Convergent validity and discriminant validity do not necessarily allow for predictions on different instruments, but rather convergent validity relates to the correlation among items on instruments with similar constructs (89). Conversely, divergent/discriminant validity tests whether instruments known to measure different constructs are in fact, not related (93).

The FSIIR (Appendix 7) and the TACQoL (Appendix 10) were the only two instruments (Table 2) which did not test for convergent validity while all other instruments successfully tested for convergent validity.

Convergent validity was proven in GenPop children with high correlations between WSU dimension on the EQ-5D-Y-3L and psychological functioning on the PedsQL and the Kidscreen-27 (2). When comparing dimensions of the EQ-5D-Y-3L to the Kidscreen-27 in children with a health condition, convergent validity was found with psychological dimensions but not for items of Mob and physical functioning (98). Convergent validity of the EQ-5D-Y-3L was also established by Scott et al. (2017) by

comparing similar dimensions from the EQ-5D-Y-3L to the WeeFIM, PedsQL and the FPS-R in children with and without a health condition (16). The WeeFim, an observational measure of functional independence, showed significant correlations with EQ-5D-Y-3L Mob in acutely-ill children ($p<0.001$, $H=21.75$), chronically ill children (p-0.01, $H=9.19$) and children with a functional disability attending a special school ($p<0.001$, $H=22.12$) (16). The FPS-R, a self-report measure of pain, was significantly correlated with the P/D dimension of the EQ-5D-Y-3L in acutely ill children (16). Good convergent validity was also found between the EQ-5D-Y-3L and all nine dimensions of the CHU-9D in GenPop children with the highest correlation found in the dimensions assessing pain (84). The EQ-5D-Y-3L was also compared to the CHU-9D and the PedsQL in children aged 6-7-years attending a mainstream school, hypotheses developed suggested that all similar dimensions will correlate, therefore demonstrating convergent validity. Of these hypotheses, a surprising result found the Mob dimension on the EQ-5D-Y-3L not statistically significant compared to the PedsQL physical dimension. However, all other hypotheses were either moderately or fairly statistically significant (87).

The DISABKIDS-TAKE-6 used the CHQ, General Health Profile and the KINDL-Revised to assess convergent validity in 4-7-year-olds with chronic health conditions. However, correlations were lower than expected and were attributed to the different age-groups these instruments target and suggested a better comparison would be the Kiddy-KINDL which similarly targets 4-7-years (7). The CHQ and HUI_3 successfully assessed in children/adolescents younger than 16-years by proving moderate to strong correlations between similar dimensions of pain, physical, mental and emotional health and overall health. The PROMIS-PGH-7 compared similar dimensions to the KIDSCREEN-10 and PedsQL in children 8-17-years using self- and proxy-report and found strong correlations with the KIDSCREEN-10 while moderate correlations with the PedsQL was found. Despite the objectivity of physical health and the subjectivity of psychosocial health, stronger correlations were found in the psychosocial dimensions (99). The CHIP-CRF-CE assessed convergent validity across limited dimensions which included ‘satisfaction’, ‘comfort’, ‘resilience’, ‘risks’ and ‘achievement’ which were compared to seven different instruments assessing similar constructs. Moderate to strong correlations between instruments were found with the ‘risk’ dimension having the highest correlation (72).

Divergent validity was proven with lower coefficients on unrelated dimensions when comparing the EQ-5D-Y-3L to the PedsQL e.g. Mob and school function (9). Similar correlations were found when comparing the EQ-5D-Y-3L LAM dimension to ‘autonomy and relationships with parents’, ‘social support and relationship with friends’ and ‘school environment’ from the KIDSCREEN-27 (98). The PROMIS-PGH-7 also successfully tested for discriminant validity by showing negative associations with

'physical health' and 'emotional distress' (99). The Kiddy-KINDL displayed discriminant validity by showing a negative correlation when compared to the Preschool Anxiety Scale where a higher score on the Kiddy-KINDL was associated with a lower score on the Preschool Anxiety Scale (100).

2.5.1.3.3. Predictive validity

Predictive validity relates to the ability of an instrument to predict future outcomes (94) and can be used to measure change to treatment outcomes (110). However, this sub-type of validity is not often measured in HRQoL research due to the unpredictability of health over time.

2.5.2. Reliability

Reliability is related to a health-related PROM's ability to consistently measure what it intended to (88) and when an instrument is used multiple times on the same participant, under similar conditions, the results should also be similar (89).

2.5.2.1. Internal consistency

Internal consistency relates to the interrelatedness of items on a health-related PROM (109). Internal consistency can be measured using the Cronbach coefficient where $\alpha>0.70$ is considered acceptable (111). Although internal consistency is an important aspect of reliability when assessing a PROM, it is not considered important for preference-based instruments as not all psychometric properties are equally important when assessing a PROM versus a preference-based instrument (21).

All 11 instruments assessed internal consistency. However, the quality of assessment differed across instruments (Table 2). The PedsQL (Appendix 2) and CHU-9D (Appendix 8) showed the most successful testing, while others showed poorer results (Table 2).

The PedsQL was successfully assessed for internal consistency when tested in children and adolescents aged 5-16-years with an acute or chronic health condition (1). In a Columbian study, the EQ-5D-Y-3L proxy version failed to meet the minimum Cronbach coefficient of 0.70 when assessed in each of the five dimensions with values ranging between $\alpha=0.43$-0.69 as well as $\alpha=0.64$ for the overall score (112). This was the only study which assessed internal consistency for the EQ-5D-Y-3L despite showing poor results which can be expected, as each dimension assesses different constructs.

The CHQ was successfully tested for internal consistency with Cronbach values ranging from 0.74 to 0.97 for sub-scores of both the CHQ-Child-Form-97 and CHQ-Parent-Form-50 with only one exception on the CHQ-Child-Form-87 dimension of general health perceptions, where a score of 0.69 was recorded (56). The Kiddy-KINDL was also successfully tested (α=0.70-0.80) in children between the ages of 3-7-years from GenPop or diagnosed with a cleft lip/palate (113). The DIASABKIDS-TAKE-6 showed moderate to acceptable Cronbach values ranging from 0.64 (interview-administered) to 0.71 (proxy-report) when assessed in children aged 4-7-years with various chronic conditions, including asthma, CP and Cystic Fibrosis (CF). (7). CHIP CRF/CE showed similar results ranging from 0.70 to 0.82 when assessed in American GenPop children between 6- and 11-years (72). While the PROMIS-PGH7 showed higher values of 0.84 for child-reporting (interview-administered) and 0.88 parent-report (proxy-report) when tested in GenPop children aged 5-17-years (73). The TACQOL-PF was assessed in children aged 6-11-years from outpatient clinics in the Netherlands and showed good internal consistency with values ranging from 0.71-0.89 for proxy-report and lower values for self-report (α=0.59-0.86). FSIIR showed good to excellent internal consistency (α=0.83-0.94) when assessed in children from 0-16-years with varying health conditions (81).

2.5.2.2. Inter-rater reliability

Inter-rater reliability refers to when an individual's health is rated by two different individuals on the same occasion and both raters score similarly (90).

2.5.2.3. Test-retest reliability

Test-retest reliability refers to the same individual rating their health on two or more occasions so long as their health remains stable (90). The interval between tests should be long enough to ensure that participants do not remember their initial responses but that their knowledge and attitudes remain unchanged (114). A randomized control trial using the 36-item Short Form survey, the Modified Lysholm scale, the Cincinnati Knee Rating System, the Activity of Daily Life of the Knee Outcome Survey and the American Academy of Orthopaedic Surgeons sport knee rating scale, found test-retesting done at a two-day or two-week interval was adequate for subjects in a stable health state (26). Test-retest reliability can be measured by calculating the intraclass correlation coefficient (ICC). ICC results between initial testing and retesting are interpreted as follows: scores <0.20 indicates a poor agreement, scores from 0.20 to 0.40 indicates a fair agreement, scores from 0.41 to 0.60

indicates a moderate agreement, scores from 0.61 to 0.80 indicates a good agreement and scores >0.81 indicates an excellent agreement (160).

Similarly to internal consistency, all of the 11 instruments assessed test-retest reliability, the EQ-5D-Y-3L (Appendix 1), PedsQL (Appendix 2) and CHU-9D (Appendix 8) showed the most successful testing compared to the other eight instruments (Table 2).

Scalone et al. (2011), Scott et al. (2017) and Ravens-Sieberer et al. (2010) are amongst the many authors who have assessed test-retest reliability of the EQ-5D-Y-3L in children and adolescents ranging from 8-18-years with acute/chronic health conditions and from the GenPop. The period between initial testing and retesting ranged from 24-hours later to ten days later based on previous literature or the researcher's aims/objectives (2,9,16). However, the longer the period between testing, the higher the possibility of a change in health state, therefore resulting in a poor test-retest reliability. Both Scalone et al. (2011) and Ravens-Sieberer et. al (2010) found fair to excellent test-retest agreement across all five dimensions (ICC=69.8-99.7%) and the VAS (ICC=0.82-0.83) (2,9). While Scott et al. (2017) found agreements ranging from poor to good (k=0.199-0.653) across the five dimensions and a good agreement for the VAS (ICC=0.77) (16).

Test-retest reliability of the PedsQL proxy and self-complete version was done in children 8-12-years with varying severities of congenital heart disease and from the GenPop two-weeks after initial testing. The self-report performed slightly better than the proxy-report when considering dimensions and the overall total. The self-complete total had an ICC=0.66, while the proxy-report had an ICC=0.63. The dimensions for self-complete ranged from 0.50-0.60 compared to a slightly lower range in the proxy-report (ICC=0.49-0.65) (115).

The CHU-9D was retested in children aged 7-15-years from the GenPop in Northern Sweden 7-15-days after initial testing (116). Since accurate retesting required participants to be in a stable health condition (26), 13 children were excluded from retesting due to major life events occurring during the interim period. Results across the nine dimensions ranged from fair to moderate agreement (k=0.32-0.54) with the dimension of pain showing the lowest agreement (116). The CHU-9D was also retested by Canaway and Frew (2012) in a younger sample of children aged 6-7-years from the GenPop. However, the retesting period was a lot shorter than recommended as it took place in the afternoon on the same day as the initial testing, which took place during the morning (87). Agreement across the nine dimensions ranged from 76-86.5% which was lower than what was expected since the retest

period was so short and lead authors to question the reliability of self-report in such young children warranting further research for the reason for the change in answers (87).

2.5.3. Feasibility

Feasibility, although not included by Mokkink et al. (25) in Figure 2 or on the COSMIN Risk of Bias Checklist, has been included as an additional psychometric concept and relates to the degree to which a newly developed instrument, in this case a health-related PROM, has been successfully used within a given setting (117). It is most commonly measured by assessing time taken to complete an instrument, number of missing responses after completion and ceiling/floor effects (2,16).

Table 1 provides a summary of the number of items assessed and includes the estimated time of completion for all 11 health-related PROMs whereby the more items assessed, the longer time of completion seems to be. Feasibility was not included in Table 2 as COSMIN describes feasibility as a description rather than a measurement as no official tests for feasibility have been developed (109).

Feasibility of the EQ-5D-Y-3L was assessed by determining the percentage of missing values, inappropriate responses on the VAS and the mean time taken to complete the measure (2). Little to no missing values were observed in children aged 8-17-years with and without health conditions, including chronic and acute illnesses (2,9,16,30,102,105,118). It was the quickest measure compared to both the PedsQL and JAMAR as it took just over one-minute to complete (102). The EQ-5D-Y-3L was also compared to the CHU-9D in a study measuring HRQoL in GenPop children aged 6-7-years attending a mainstream school. Neither health-related PROM had missing values and both measures were completed in less than three-minutes (87). When used in children aged 5-17-years, the PROMIS-PGH-7 also had no missing values across items (73) and was estimated to take 1-2 minutes to complete (73). Similarly, the DISABKIDS-TAKE-6 is estimated to take approximately two minutes to complete while missing values are not stated (7). The PedsQL showed less than 3% of missing values when tested in a large paediatric sample (1), with the school functioning dimension having the most missing values attributed to age-appropriateness of the dimension, especially in younger children aged 5-7-years (1,70). The CHIP-CRF/CE also had relatively low missing values (<4%) (72). However, completion time is slightly longer than other instruments due to the number of items assessed, with the interviewer-administered version taking an average of 21.4 minutes and self-complete an average of 22.6 minutes (22,72). The TACQoL-PF had slightly lower missing values ranging from 0.4-1.5% across its seven items (85). The CHQ and Kiddy-KINDL both had slightly higher missing values when tested in various paediatric populations. The Kiddy-KINDL had 8.1% missing values when administered in children 4-6-

years (6) while missing values on the CHQ ranged from >1.5%-9% (76,119,120) and took approximately 10-20 minutes to complete when tested in different cultural groups (121). The FSIIR is estimated to take the longest to complete compared to the other ten instruments described (Table 1) with an estimated time of completion ranging from 15-30 minutes (4,22) while the HUI_3 is estimated to take an average of 5-10 minutes despite it having a relatively large number of items (67,71). Therefore, In fast-paced clinical settings, HRQoL instruments are not always routinely used due to the time of completion which may result in shortening assessment/treatment time. After conducting a study exploring the reasons for as to why therapists do not use HRQoL instruments, 80% of clinicians used time constraints as their reasoning while others mentioned the difficulty in scoring and interpreting the results (65). Therefore, to encourage clinicians to use these instruments, they need to show that they are quick and easy to administer/interpret. Despite no information on the feasibility of the IA yet, if this newly developed instrument can be completed in a short period of time, it may prove useful in these fast-paced settings as its only requirement is a health professional to read the script to the patient.

2.5.4. Acceptability

Acceptability of a PROM can be determined by the individual using the PROM as it is based on their assessment on whether it accurately addresses the needs of the population intended to utilise the PROM (117,122). Despite acceptability not being included by Mokkink et al. (25) (Figure 2) or on the COSMIN Risk of Bias Checklist, it remains an important psychometric property to consider.

2.6. Summary of literature review

According to the COSMIN Risk of Bias checklist (summarised in Table 2), overall, the PedsQL (Appendix 2) showed the highest psychometric performance in terms of validity, as they explored all five subtypes of validity listed on the checklist, which included: content, structural, cross-cultural, criterion and construct validity. The EQ-5D-Y-3L (Appendix 1) fell slightly short by having not tested structural validity. The HUI_3 (Appendix 3), the Kiddy-KINDL (Appendix 4), FSIIR (Appendix 7) CHU-9D (Appendix 8), DISABKIDS-TAKE-6 (Appendix 9) and TACQoL-PF (Appendix 10) all showed the poorest validity by only assessing one or two of the five validity subtypes listed, with all similarly lacking in the criterion validity sections. The CHIP CRF/CE (Appendix 5), the PROMIS-PGH-7 (Appendix 6) and the CHQ (Appendix 11) all showed fair validity by testing three subtypes. The CHIP CRF/CE lacked evidence for content and structural validity, while the PROMIS-PGH-7 lacked in the cross-cultural and criterion validity sections.

All 11 health-related PROMs (Table 2) were tested for reliability and internal consistency with the FSIIR (Appendix 7) having shown the highest range of internal consistency. The time period between initial and retesting was not stated by all, including studies using the CHU-9D (Appendix 8), DISABKIDS-TAKE-6 (Appendix 9) and TACQoL-PF (Appendix 10), while the CHQ (Appendix 11) exceeded the recommended interval of two-days to two-weeks (26) and was retested three-weeks later or between two- and 16-days. The FSIIR did not assess test-retest but rather tested the reliability between an English and Spanish version of the FSIIR.

Currently, the EQ-5D-Y-3L and PedsQL, both of which have a proxy-report version, performed the best on the COSMIN Risk of Bias Checklist. While the only two PROMs which offers interviewer-administered administration, the DISABKIDS-TAKE-6 and Kiddy-KINDL, did not perform as well despite the growing need for this mode of administration in young children to allow them to self-report on their health. In order to accurately record HRQoL in young children, more valid and reliable health-related PROMs with minimal bias and various modes of administration are needed to allow for better inclusivity and accurate results.

The EQ-5D-Y-3L and PedsQL are both valid and reliable measures in South Africa and are available in many of the local South African languages. As the IA version of the EQ-5D-Y-3L is newly developed, it is prudent to determine whether it is valid and reliable in our context and could negate the use of the proxy version in children aged 5-7-years. It may further prove useful in older children, within the recommended age band of 8-12-years, with lower literacy levels due to their health condition, low socioeconomic setting or other contributing factors (5). This would allow greater inclusivity of children for self-report and potentially lower the age-range for children to subjectively report on their HRQoL, thus decreasing the reliance on proxy-report. Based on the issues and concerns surrounding proxy-report which have been discussed, this instrument will be an indispensable addition to the paediatric population.

3. Comparison of the Performance of the EQ-5D-Y-3L-IA and EQ-5D-Y-3L-SC in children aged 8-10-years

3.1. Methodology

3.1.1. Introduction

This chapter provides a detailed description of the methods implemented to compare the EQ-5D-Y-3L IA and SC versions in children aged 8-10-years. The study design and setting, sample size, instruments utilized, management of data, statistical analysis and ethical considerations have been described below.

A cross-sectional, descriptive observational, analytical design was conducted to determine performance of the EQ-5D-Y-3L-IA compared to the SC version in children aged 8-10-years with a known medical condition and from the GenPop. The cross-sectional aspect of the study design refers to data being collected from multiple individuals within the same period of time. The analytical aspect of the study design refers to the statistical analyses and psychometric testing of the IA compared to the SC. The descriptive aspect of the study design refers to the qualitative data or feedback obtained from participants when asked for their preference between the IA and SC version and the reason for their preference.

3.1.2. Study settings

Children with chronic respiratory illnesses or receiving orthopaedic intervention, were recruited from outpatient clinics at a paediatric tertiary hospital, and an inpatient paediatric orthopaedic hospital both of which are located in the Western Cape, South Africa. The paediatric tertiary hospital has inpatient and outpatient sub-speciality services, treating over 250 000 patients per year (123). The orthopaedic hospital offers orthopaedic surgery, inpatient care and rehabilitation services for children needing corrective surgeries or fracture management (123).

Children with functional disabilities were recruited from schools for learners with special educational needs (LSEN), who follow a mainstream curriculum, in the Western Cape, South Africa. These schools cater for children who require high levels of support and provide additional resources, which may

include access to adaptive equipment and learning materials as well as services from nurses, rehabilitation professionals, psychologists and social workers (124).

Children from the GenPop were recruited from mainstream schools. These children were not expected to have any serious health conditions which precluded their attendance at school. They may however have manageable or well controlled health conditions such as asthma or eczema (125). The schools are within the same geographical catchment area as the LSEN schools and hospitals and likely share similar socioeconomic circumstances. Groups were not matched for sex or socioeconomic status as they were not comparator groups but rather grouped according to health classification to allow us to better hypothesize how they would report their health state on the two EQ-5D-Y-3L versions.

3.1.3. Participants

3.1.3.1. Inclusion criteria

Children aged 8-10-years fluent in English were included as it is the source language of newly developed EQ-5D-Y-3L-IA and has yet to be adapted and translated into different languages. Fluency in English was determined if it was their self-classified home language or the language of instruction at school. Children with multi-morbidities were included and allocated to a known-group according to the condition that they were seeking care for on the day of recruitment and any additional health conditions were noted.

3.1.3.2. Exclusion criteria

Children who were medically diagnosed as unable to hear with assistive technology were excluded as the primary outcome measure relied on interviewer-administration. Children attending the LSEN schools with moderate to severe intellectual disability, diagnosed by a psychologist and typically educated in a unit class, were excluded from the study as their level of understanding of the questions asked could have been limited. Teachers and school psychologists at the LSEN schools identified children with moderate to severe intellectual disabilities.

Children who required admission to the intensive care or high care unit, with continuous monitoring, were considered critically ill and therefore excluded due to additional associated emotional stress that would have occurred if they were to participate.

3.1.4. Sample size

The sample size was powered to detect a difference in correlations between the IA and SC versions with a small effect size 0.2, slightly smaller than previous South African studies with an effect size of 0.3 and 0.4 (16,126). A sample of 211 was required to ensure a power of 90% and a significance of 0.05, similar to previous research done in South African children aged 8-15-years (126).

3.1.5. Instruments

EQ-5D-Y-3L-SC (Appendix 12) measures HRQoL on the day of testing on five dimensions: Mobility (Mob) (walking about), Looking After Myself (LAM) (washing or dressing), doing Usual Activities (UA) (for example, going to school, sports, hobbies, playing and doing things with friends or family), having Pain/Discomfort (P/D) and feeling Worried, Sad, or Unhappy (WSU). Each dimension has three levels of report categorized as level 1 indicating 'no problems', level 2 indicating 'some problems' or level 3 indicating 'a lot of problems' which results in 243 (3^5) health states (127). The SC includes a VAS which is a vertical, graduated number scale from worst imagined health state (0) to best imagined health state (100) on which the participant rates their overall health status also on the day of testing (17,18). A ceiling or floor effect may be present if 'no problems' are indicated for all dimensions (11111) or if 'a lot of problems' are reported for all dimensions (33333). The SC has been successfully tested for validity, reliability and responsiveness in South African children aged 8-15-years (2,5,16,102).

The newly developed IA version, **EQ-5D-Y-3L-IA** (Appendix 13) underwent standardized forward-backward translation from United Kingdom English to South African English by the EuroQoL Research Foundations Version Management Committee (VMC) (128). The process ensured that the forward translation from English for the United Kingdom to English for South Africa, was done by two independent translators. These two versions were compared and discussed, and a consensus version was compiled. Backward translation of the consensus version was done by two blinded translators. The translation process was overseen for quality assurance by members of the VMC who ensured that the meaning of the translated version matched the source and if any further modifications were needed. Once forward-backward translation had been completed, the translated version of the instrument was tested in face-to-face interviews with ten children in South Africa to ensure that the translated version retained its meaning, and that it was culturally accepted and relevant (129). As there is no utility value set available for South Africa, the recently published value set produced for Slovenia was used (130) as an indication of composite dimension performance. To ensure that the

societal preference-based score did not influence performance; comparison was made to the Japanese value set (131). The Slovenian and Japanese value sets were the only two available at the time of data analysis.

The **Faces Pain Scale-Revised (FPS-R)** (Appendix 14) is a self-report measure intended to determine the intensity of pain felt by children on the day of testing. It was developed using a series of six facial expressions depicting an increase in pain intensity from left to right. The scoring ranges from 0-10 and increases by increments of two. It can be used to self-rate pain intensity in children aged 4-years or older (105). The FPS-R was successfully used to determine concurrent validity for the dimension of P/D on the EQ-5D-Y-3L in South Africa at baseline (3).

The **Moods and Feelings Questionnaire** (**MFQ)** (Appendix 15) consists of 13 questions about the child's psychological wellbeing in the two-weeks prior to testing. Participants were asked by the student researcher to answer questions on a scale of 'not true', 'sometimes' or 'true' which was converted to a numerical value of zero, one and two, respectively. The measure has been found valid and reliable in an international study in children from age five years (132).

The **WeeFIM** (Appendix 16) is an observational instrument used to assess functional independence in children (133,134). Functional performance was measured, by the student researcher, in three dimensions, namely self-care, mobility and cognition. The self-care dimension includes eating, grooming, dressing, bathing, toileting, level of assistance needed to control bladder and bowel and frequency of bladder and/or bowel accidents. The mobility dimension includes transfers in/out of a chair or mobility assisted device, toilet transfers, tub and/or shower transfers and locomotion (including walking, wheelchair use, crawling and stairs). The cognition dimension includes comprehension of auditory or visual instructions, ability to vocally or non-vocally express basic needs, social interaction with peers, ability to solve everyday problems and memory of daily routines, however, social interaction was the only item of the cognition dimension assessed in this study. There is a total of 18 items, each rated on an ordinal scale from 1-7 whereby one indicates 'total assistance' and seven indicates 'complete independence' (Appendix 16). The scale gives scores for sub-scales (mobility, cognition and self-care) or a total score for functional performance, the higher score, the more independent the child is in that dimension.

The WeeFIM allows for a functional assessment for a range of children including those who are confined to bed or use a wheelchair as well as those who are more functionally advanced, making it

suitable for assessment of children across health conditions. The WeeFIM sub-scale of mobility and self-care was previously used to determine concurrent validity in the corresponding dimensions of Mob and LAM on the EQ-5D-Y-3L in South Africa (3) and was similarly used in this study.

Study specific medical and demographic questionnaire (Appendix 17) was completed by the parent/caregiver, or from the medical folder, with caregiver consent, and included details regarding the child's date of birth, sex, and diagnosis of a health condition. These details were deemed important as it allowed the researcher to accurately allocate children to the correct age-group and health condition group. The use of an assistive device and/or orthotic device was included in the event that children did not have their device with them on the day of data collection.

Preference for IA or SC version (Appendix 18): Children aged 8-10-years were asked, by the researcher, whether they preferred the IA version or the SC version.

3.1.6. Procedure

Necessary approvals were granted by the Faculty of Health Sciences, Human Research Ethics Committee (HREC), University of Cape Town (UCT) (HREC 369/2020) (Appendix 19), ministerial permission for non-therapeutic research with minors (Form A) (Appendix 20), Western Cape Education Department (Appendix 21), the respective school principals (Appendix 22) and the children's hospital management (Appendix 23). Children who fulfilled the inclusion criteria were recruited from schools and healthcare institutions.

Envelopes containing study information, an informed consent form (Appendix 24) and a demographic questionnaire (Appendix 17) were distributed to participating schools for learners to take home. Parents/legal guardians were invited to return the signed consent form and demographic questionnaire if they agreed to their child participating in the study. All children whose parents/legal guardians granted consent, were invited individually to a private room where they were given a detailed description of the study. Assent (Appendix 25) was obtained from those willing to participate. The research packs, including the EQ-5D-Y-3L-IA, FPS-R, MFQ, WeeFIM and EQ-5D-Y-3L-SC were completed in a random order (with the IA or SC versions always presented first and last) to minimize question order bias. Participants were also asked for their preference between the IA and SC version and a reason for the preference.

Children admitted to the orthopaedic hospital were recruited after admission. Parents/legal guardians who were not at the bedside, were contacted telephonically to describe the study and to ask for consent (Appendix 26). The procedure for data collection was as described for children attending school.

Children attending the sub-specialty outpatient clinics for either orthopaedics or chronic respiratory illness, were screened for inclusion according to the date of birth in the clinic diary and recruited systematically in the order of arrival. The eligible children and their parent/legal guardian were approached in the waiting room and invited to participate in the study by means of a brief verbal description of the study accompanied by an informed consent form. If consent was granted, the parent/legal guardian and their child were invited to sit in a private consultation room where assent was obtained, followed by the completion of the research packs. If the parent/legal guardian chose to accompany their child during the completion of the research packs, they were invited to sit slightly behind their child and asked not to influence their answers verbally or non-verbally. Children did not lose their place in the queue to see the medical professional nor were other medical investigations put on hold due to their participation. If the research pack was not completed by the time they were called in for their consultation with the medical professional, the research was stopped, and the child and parent/legal guardian were invited to complete the study after their consultation.

All screening, enrolment and data collection was done by the student researcher to ensure standardization and to minimize bias.

Participating schools received an education hamper to the value of R2000 for their assistance in identifying learners who met the inclusion criteria, communication with parents/legal guardians, handling of envelopes and arranging appropriate times for children to leave the classroom during school hours to conduct the interviews. These schools received the educational hamper regardless of how many children from their facility were enrolled. The participants nor their parents/legal guardians were reimbursed as they did not incur any costs to participate in the research.

3.1.7. Data management

All information obtained was entered into a password protected Excel spreadsheet under the code allocated to each participant thus ensuring confidentiality and anonymity. All hard copies were stored

in a locked cupboard in a secure office and will be destroyed ten years after publication. No identifying information of participants was recorded for data analysis or dissemination of results.

A data management plan (Appendix 27) was developed in line with the UCT Data Management Policy (135).

3.1.8. Statistical analysis

The Shapiro-Wilk test was used to test the normality of the data. Level of statistical significance was set at $p<0.05$.

3.1.8.1. General instrument performance and feasibility

The IA and SC responses and descriptive data were summarised in terms of frequency of responses. More problems were expected to be reported by the orthopaedic and respiratory illness groups. The ceiling effect of the two versions was defined as the proportion of children scoring no problems across all five dimensions (11111) or for each individual dimension. The floor effect of the two versions was assessed by the number of children reporting a lot of problems across all dimensions (33333) or for each individual dimension. High ceiling effects were expected in children from the GenPop but no differences between versions were expected. Differences in reporting was determined by chi-square statistic (x^2). The feasibility was assessed by comparing the number of missing values for the SC and IA versions. More missing values were expected for the SC version compared to the IA version.

3.1.8.2. Inconsistent responses

Paired dimension responses on the IA and SC were assessed for the respondents who had no missing responses and the proportion of inconsistencies was recorded by age, sex and health condition. Differences across age, sex and health condition were determined by chi-square statistic (x^2). Inconsistencies were not expected based on health condition, age, sex or between versions of the EQ-5D-Y-3L.

3.1.8.3. Known-group validity

Known-group validity was tested for the dimensions of the SC and IA versions for age, sex and by health condition by Spearman rank order coefficients (r_s). It was expected that children with an orthopaedic condition and those with a functional disability would report more problems in the Mob dimension compared to other groups (16,17,126). It was also anticipated that children with an orthopaedic condition (being more acutely ill), would report more problems with UA and P/D (16,136). Lastly, it was expected that all children with a health condition (orthopaedic, respiratory and functional disability) would report greater feelings of WSU than children from the GenPop (16,136). Effect sizes were interpreted according to Cohen's d interpretation whereby 0.2 = small, 0.5 = medium, 0.8 = large and 1.3 = very large (137).

The known-group validity across health condition was assessed for the median utility and VAS score across age, sex and health condition by Kruskal Wallis H-test (H) and Mann-Whitney U-test. It was anticipated that the VAS and utility scores would be higher for those from the GenPop, functional disability, respiratory condition, and orthopaedic condition in that order.

3.1.8.4. Concurrent validity

The Pearson's correlation of the utility score and VAS score was computed for the SC and IA versions and compared using the Fisher r-to-z transformation (http://vassarstats.net) (138). It was expected that there would be no difference in concurrent validity between the IA and SC.

3.1.8.5. Convergent validity

Convergent validity between the IA and SC was evaluated by individual dimension response-pairs, using Gamma Correlations statistics. Utility scores were compared with Pearson Correlation co-efficient. Correlation coefficients were interpreted according to Cohen: 0.1–0.29 low association, 0.3–0.49 moderate association and ≥0.5 high association (139). It was expected that similar dimensions would show similar correlations (2,9,16,87).

The convergent validity of the dimension scores of the SC and IA versions were compared to similar items on the MFQ, FPS-R and WeeFIM using Spearman correlations (r_s). Correlation coefficients were

compared between the SC and the IA versions of the EQ-5D-Y-3L using the Fisher r-to-z transformation (http://vassarstats.net) (138).

3.1.8.6. Preference between the EQ-5D-Y-3L-IA and SC

The preference of the EQ-5D-Y-3L-IA and SC were determined by frequency of responses and compared with chi-square statistics (x^2). It was expected that participants would prefer the IA as the respondent burden was reduced (140). Qualitative data was coded and grouped according to preference ie. SC or IA and similar reasons for preference were grouped together. The researcher was aware of reflexivity and did not allow personal opinions to impact on participants' responses, nor the grouping and coding of responses.

3.1.9. Ethical considerations

Ethical principles of autonomy, confidentiality, beneficence/non-maleficence and justice were applied according to the Helsinki Declaration (141). Every child who met the inclusion criteria and was eligible to participate was recruited. Since the IA is a newly developed English measure, only children fluent in English were recruited. Once the English source version is validated, translations into other languages can be done for further validity and reliability testing. No child was excluded based on ethnicity, sex, gender, religion, or other reason. None of the participants showed any signs of emotional distress during the interview.

The children's vulnerability was protected by ensuring that the following procedures were adhered to: informed consent and assent were obtained before data collection began, privacy during interviews was ensured, participants were not disadvantaged if they did not provide assent or chose to withdraw during the interview even after assent was granted. No information shared during interviews was of a concern, requiring follow-up or referral therefore no study information was shared. There were no added risks to this vulnerable group participating in the study as it did not affect the medical treatment or the schooling which they received. The benefit of including this vulnerable group included that the results of the study may benefit future HRQoL measurement in South Africa for those with lower literacy levels or where a child's medical condition does not allow for self-complete.

3.1.10. COVID-19 considerations

COVID-19 protocols were adhered to throughout this study by implementing safety precautions according to the National Health Department (142), provincial lockdown regulations, and the respective school or health institution guidelines.

3.2. Results

The recruitment of children aged 8-10-years is shown in Figure 3. A total of 211 children were recruited however, only 207 were included in Chapter 3 as four children did not complete the SC. There was a high proportion of non-responders in this age-group (n=207, 64%) due to the high number of parents/legal guardians from GenPop and special schools not returning consent forms. The reason for not wanting to participate was not recorded. A large number of children with orthopaedic problems withdrew (n=21, 20%) during interviews due to personal reasons, multiple medical appointments, time constraints and transport issues which include scheduled patient transport to/from healthcare facilities to patient's residential area which may be located in a different city within the Western Cape province, price of public transport and fear of missing public transport which do not follow fixed schedules.

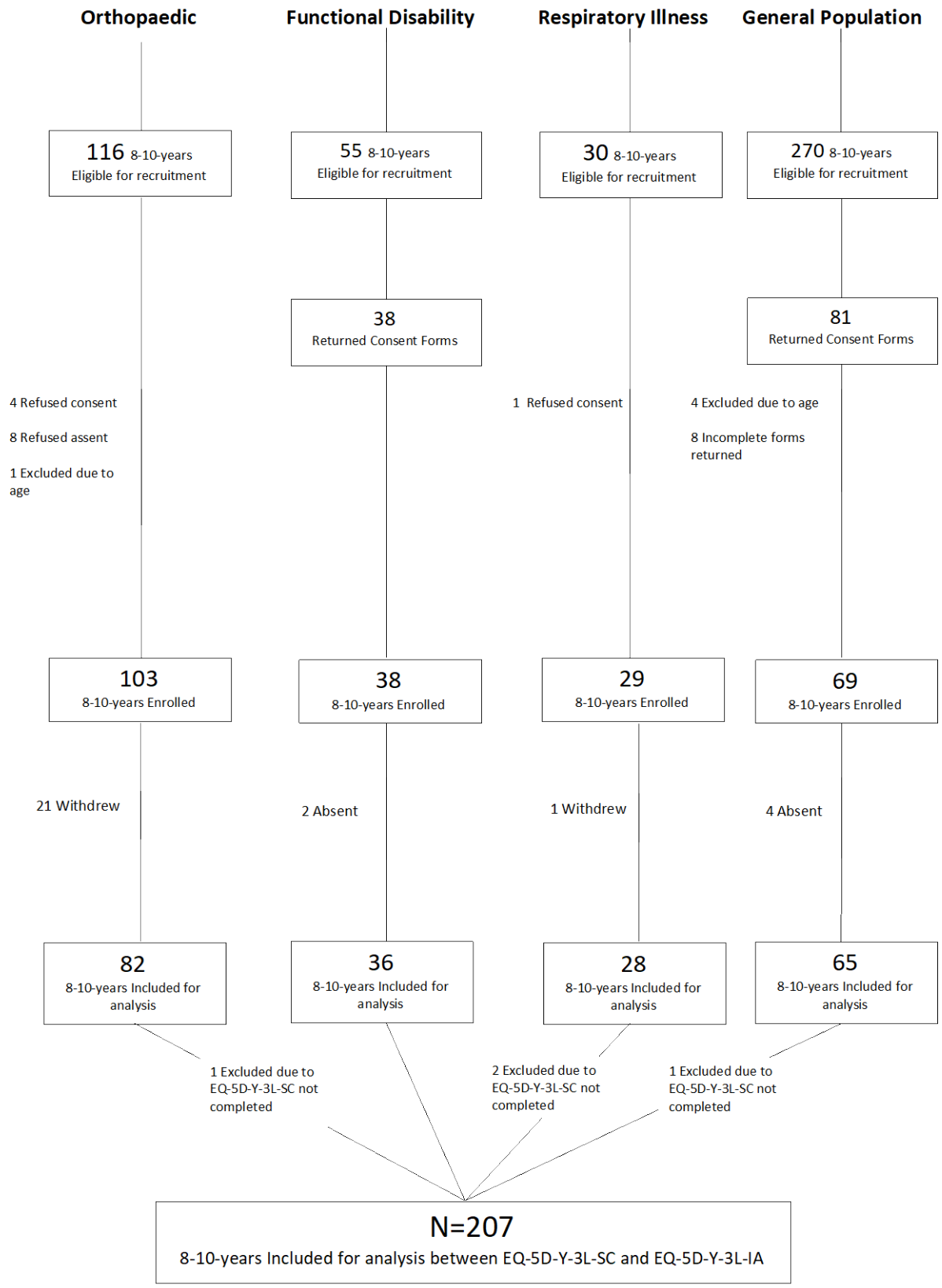


Figure 3: Recruitment of sample

3.2.1. Descriptive statistics

There was no significant difference between se[x] across age (8-, 9- and 10-year-olds) (x^2=0.03, df=2, p=0.985) (Table 3). In total, there were more children with orthopaedic conditions (n=81, 39%) and from the GenPop (n=64, 31%) than children with functional disabilities (n=36, 13%) and respiratory illnesses (n=26, 13%). Furthermore, there was no difference between health condition across age (x^2=3.61, p=0.729). The specific conditions included in these disease groups are shown in Table 3.

Table 3: Descriptive statistics of participants

	Age (years)						Total	
	8-years		9-years		10-years			
	(n=65)		(n=70)		(n=72)		(n=207)	
Sex	n	%	n	%	n	%	n	%
Female	30	46%	32	46%	34	47%	96	46%
Male	35	54%	38	54%	38	53%	111	54%
Orthopaedic	(n=29)		(n=25)		(n=27)		(n=81)	
Upper Limb Fracture	13	45%	9	36%	9	33%	31	38%
Lower Limb Fracture	6	21%	10	40%	6	22%	22	27%
Surgical correction of acquired or congenital orthopaedic condition[#]	5	17%	4	16%	10	37%	19	23%
Other*	5	17%	2	8%	2	7%	9	11%
Functional Disability	(n=11)		(n=12)		(n=13)		(n=36)	
Developmental Co-ordination Disorder^	6	55%	8	7%	7	4%	21	33%
Cerebral Palsy	1	9%	2	3%	3	4%	6	17%
Spina Bifida	2	18%	1	1%	2	3%	5	14%
Developmental Delay	1	9%	1	1%	1	1%	3	8%
Traumatic Brain Injury	1	9%	0	0%	0	0%	1	3%
Respiratory	(n=7)		(n=7)		(n=12)		(n=26)	
Atopy	3	43%%	2	29%	7	58%	12	46%
Cystic Fibrosis	2	29%	2	29%	1	8%	5	19%
Bronchiectasis	0	0%	0	0%	2	17%	2	8%
Other[¥]	2	29%	3	43%	2	17%	7	27%
GenPoP	(n=18)		(n=26)		(n=20)		(n=64)	
None	16	89%	21	81%	16	80%	53	83%
Atopy	1	6%	5	19%	2	10%	8	13%
Other[§]	1	6%	0	0%	2	10%	3	5%

*[#]Includes Blount's disease, osteogenesis imperfecta, developmental dysplasia of the hip, leg-length discrepancy and spinal deformity *includes osteitis; septic arthritis and a traumatic amputation. ^Includes learning disability and Human Immunodeficiency Virus. [¥]Includes damage to the lungs post-acute viral infection, congenital abnormalities of the respiratory system and idiopathic pulmonary haemorrhage. [§]Includes osteogenesis imperfecta and a congenital cardiac defect.*

3.2.2. General instrument performance and feasibility

3.2.2.1. General instrument performance

Table 4 shows that there were no significant differences in reporting between the IA and SC versions across all five dimensions. The utility score[1] and VAS score were similarly higher on the IA than the SC version, although not significantly so. However, a higher ceiling effect was found in the SC (n=80, 39%) compared to the IA version (n=62, 30%). One participant had reported 33333 on the IA only.

[1] Analysis with the Slovenian utility score is presented. There was no significant difference between results using the Slovenian or Japanese utility scores.

Table 4: Comparison of the EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA dimensions

		SC (n=207)		IA (n=207)				
		n	%	n	%	X^2	df	p-value
Mob	No	151	73%	146	71%	3.11	2	0.211
	Some	30	14%	43	21%			
	A lot	12	6%	18	9%			
	Missing	14	7%	0	0%			
LAM	No	143	69%	150	72%	0.98	2	0.613
	Some	37	18%	47	23%			
	A lot	12	6%	10	5%			
	Missing	15	7%	0	0%			
UA	No	145	70%	141	68%	3.79	2	0.150
	Some	29	14%	47	23%			
	A lot	18	9%	19	9%			
	Missing	15	7%	0	0%			
P/D	No	122	59%	133	64%	0.81	2	0.667
	Some	56	27%	60	29%			
	A lot	18	9%	14	7%			
	Missing	11	5%	0	0%			
WSU	No	135	65%	147	71%	0.40	2	0.819
	Some	52	25%	49	24%			
	A lot	10	5%	11	5%			
	Missing	10	5%	0	0%			
11111		80	39%	62	30%	3.1		0.078
33333		0	0%	1	0%			
Utility score	Median (IQR)	0.883 (0.608,1.00)		0.870 (0.614,1.00)		z=1.262		0.207
VAS*	Median (IQR)	95 (68,100)		100 (70,100)		z=0.496		0.62
	Missing	3	1%	0	0%			

*Mob=mobility, LAM=looking after myself, UA=usual activities, P/D= pain or discomfort, WSU=worried, sad or unhappy. SC=EQ-5DY-3L-SC, IA=EQ-5D-Y-3L-IA *Difference in continuous variables were calculated with Wilcox sign test.*

3.2.2.2. Feasibility

Both versions were statistically different, however, the IA took less time to complete (median=110s, IQR=98, 124s, p<0.001) compared to the SC version (median=157s, IQR=123s, 209s, p<0.001) and was not significant across versions (X^2=2.51, p=0.113). When comparing the time taken across ages, 8-year-olds took the longest to complete both versions but were able to complete the IA quicker than the SC.

The IA had no missing values compared to the SC which had a total of 32% (n=65) across all dimensions. The total missing values were made by 11% of the sample (n=22) (Table 4).

Table 5 shows comparison of the number of children who contributed to the missing values on the SC was not significant across age, sex or health condition, despite the 8-year-olds and those with an orthopaedic condition showing a higher proportion. Comparison of missing values on dimension scores across the total sample showed that the 8-year-olds had significantly more missing values than the 9-year-olds and the 10-year-olds. Children with an orthopaedic condition had significantly lower missing values across all dimensions than the other health conditions.

Table 5: EQ-5D-Y-3L-SC missing values across age (years), sex and health conditions

	Children with missing values				Missing values across five dimensions*			
Age	n	%	X^2	p-value	n	%	X^2	p-value
8-years (n=65)	11	17%	4.01	0.135	34	10%	14.23	<0.001
9-years- (n=70)	5	7%			14	4%		
10-years (n=72)	6	8%			17	5%		
Sex								
Female (n=96)	8	8%	0.56	0.454	24	5%	2.1	0.147
Male (n=111)	14	13%			41	7%		
Health condition								
Orthopaedic (n=81)	11	14%	2.12	0.548	37	9%	21.93	<0.001
Functional Disabilities (n=36)	4	11%			36	20%		
Respiratory (n=26)	1	4%			26	20%		
GenPop (n=64)	6	9%			64	20%		

**The denominator is the total number of answers per age/sex or health condition. Calculated as the number of children per category multiplied by the number of dimensions (five).*

3.2.3. Distribution of responses between the EQ-5D-Y-3L-IA and EQ-5D-Y-3L-SC

Table 6 shows the highest report of inconsistent responses on both versions were found in the P/D dimension (31%). The highest inconsistency across dimensions is moving from reporting no problems on the SC version and some problems on the IA version. An exception to this was the dimension of WSU where the highest inconsistency was reporting no problems on the IA version but some problems on the SC version.

When inconsistency was analysed by sex there were no significant differences between the number of inconsistent dimension responses between males and females (x^2=0.43, p=0.980). For males the inconsistencies were in keeping with those shown in Table 6. However, females showed the highest inconsistency with reporting some problems on the SC version and no problems on the IA version for dimensions of LAM, P/D and WSU.

Table 6: Inconsistent responses of the EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA across dimensions

	EQ-5D-Y-3L-IA							
EQ-5D-Y-3L-SC	**No**		**Some**		**A lot**		**Inconsistent Responses**	
Mob	n	%	n	%	n	%	n	%
No	122	66%	15	8%	7	4%	41	22%
Some	8	4%	17	9%	3	2%		
A lot	4	2%	4	2%	4	2%		
LAM	**No**		**Some**		**A lot**			
No	119	65%	16	9%	3	2%	42	23%
Some	13	7%	20	11%	2	1%		
A lot	4	2%	4	2%	3	2%		
UA	**No**		**Some**		**A lot**			
No	113	61%	18	10%	7	4%	45	24%
Some	7	4%	18	10%	3	2%		
A lot	5	3%	5	3%	8	4%		
P/D	**No**		**Some**		**A lot**			
No	96	52%	20	11%	3	2%	57	31%
Some	15	8%	29	16%	6	3%		
A lot	9	5%	4	2%	2	1%		
WSU	**No**		**Some**		**A lot**			
No	110	60%	15	8%	4	2%	48	26%
Some	21	11%	24	13%	3	2%		
A lot	3	2%	2	1%	2	1%		

n=184, shaded cells indicate consistent responses. Mob=mobility, LAM=looking after myself, UA=usual activities, P/D= pain or discomfort, WSU=worried, sad or unhappy

Table 7 shows that inconsistencies were lowest in the 10-year-olds when the total inconsistencies across all five dimensions are considered, however, there were differences in inconsistencies at an individual dimension level. The dimension of LAM had the highest number of inconsistencies for 8-year-olds compared to P/D for the 9- and 10-year-olds. For the dimension of Mob, the 8-year-olds had a high proportion of reporting no problems with Mob on the SC version and a lot of problems on the IA version whereas the 9- and 10-year-olds reported no problems on the SC version and some problems on the IA version. In the dimension of LAM, the highest inconsistency from the 8- and 10-year-olds was reporting some problems on the SC and no problems on the IA whereas the 9-year-olds reported more problems on the IA version. Across all three age-groups children moved from reporting no problems with UA on the SC version to some problems on the IA version. The youngest children reported greater P/D on the SC version and changed to no P/D on the IA version whereas the 9-10-year-olds both had the greatest shift with reporting no problems with P/D on the SC version and some

P/D on the IA version. WSU showed an under-reporting of problems on the IA version compared to the SC version across all age-groups.

Table 8 shows that the GenPop group has the least inconsistent responses across dimensions while the functional disability group had the highest inconsistent responses across dimensions when compared to other health condition groups. For those with an orthopaedic condition, the highest number of inconsistencies was reported for WSU whereas P/D had the highest inconsistencies for those with a functional disability or respiratory condition. For those with an orthopaedic condition, the highest inconsistencies were from reporting no problems on the SC and some problems on the IA version for all dimensions with the highest inconsistencies found in the UA dimension. Those with a functional disability had the highest number of inconsistencies with moving from reporting problems on the SC version to no problems on the IA version except for the dimension of Mob where they reported some problems on the SC and a lot of problems on IA. In the respiratory group, the inconsistencies were highest for reporting no problems on SC and some problems on the IA version except for the dimension of LAM. In the GenPop the inconsistencies were greatest in reporting some problems on the SC version and then changing to no problems on the IA version.

There were significant differences in inconsistent responses across health conditions in the dimensions of LAM (x^2=8.36, p=0.039), UA (x^2=7.69, p=0.053), and P/D (x^2=10.05, p=0.018).

Table 7: Inconsistent responses of the EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA by age-group

	EQ-5D-Y-3L-IA											
	8-years				9-years				10-years			
SC	(n=54)				(n=65)				(n=65)			
Mob	**no**	**some**	**a lot**	**Inconsistent Responses**	**no**	**some**	**a lot**	**Inconsistent**	**no**	**some**	**a lot**	**Inconsistent**
No	65%	6%	7%	24%	66%	11%	5%	25%	68%	8%	0%	18%
Some	4%	6%	2%		6%	9%	2%		3%	12%	2%	
A lot	4%	2%	6%		2%	0%	0%		2%	5%	2%	
LAM	**no**	**some**	**a lot**	**Inconsistent**	**no**	**some**	**a lot**	**Inconsistent**	**no**	**some**	**a lot**	**Inconsistent**
No	56%	13%	2%	28%	72%	6%	3%	22%	65%	8%	0%	23%
Some	9%	11%	2%		3%	9%	2%		9%	12%	0%	
A lot	2%	2%	4%		2%	2%	2%		3%	3%	0%	
UA	**no**	**some**	**a lot**	**Inconsistent**	**no**	**some**	**a lot**	**Inconsistent**	**no**	**some**	**a lot**	**Inconsistent**
No	57%	9%	4%	26%	65%	6%	6%	26%	62%	14%	2%	22%
Some	4%	13%	2%		5%	5%	2%		3%	12%	2%	
A lot	6%	2%	4%		3%	5%	5%		0%	2%	5%	
P/D	**no**	**some**	**a lot**	**Inconsistent**	**no**	**some**	**a lot**	**Inconsistent**	**no**	**some**	**a lot**	**Inconsistent**
No	54%	6%	4%	26%	49%	15%	0%	38%	54%	11%	2%	28%
Some	7%	19%	0%		12%	11%	3%		5%	18%	6%	
A lot	7%	2%	4%		5%	3%	2%		3%	2%	0%	
WSU	**no**	**some**	**a lot**	**Inconsistent**	**no**	**some**	**a lot**	**Inconsistent**	**no**	**some**	**a lot**	**Inconsistent**
No	65%	6%	0%	22%	55%	11%	5%	31%	60%	8%	2%	25%
Some	11%	13%	4%		11%	12%	0%		12%	14%	2%	
A lot	0%	2%	0%		3%	2%	2%		2%	0%	2%	

Mob=mobility, LAM=looking after myself, UA=usual activities, P/D= pain or discomfort, WSU=worried, sad or unhappy. Consistent responses are shaded grey

Table 8: Inconsistent responses of the EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA across dimensions and health conditions

	EQ-5D-Y-3L-IA															
	Orthopaedic				Functional Disabilities				Respiratory				GenPop			
SC	(n=70)				(n=32)				(n=24)				(n=58)			
Mob	**No**	**Some**	**A lot**	**Inconsistent**	**No**	**Some**	**A lot**	**Inconsistent**	**No**	**Some**	**A lot**	**Inconsistent**	**No**	**Some**	**A lot**	**Inconsistent**
No	57%	13%	6%	27%	56%	3%	3%	22%	63%	13%	4%	33%	84%	3%	2%	12%
Some	3%	13%	1%		3%	16%	6%		8%	4%	0%		5%	3%	0%	
A lot	1%	3%	3%		3%	3%	6%		4%	4%	0%		2%	0%	0%	
LAM	**No**	**Some**	**A lot**	**Inconsistent**	**No**	**Some**	**A lot**	**Inconsistent**	**No**	**Some**	**A lot**	**Inconsistent**	**No**	**Some**	**A lot**	**Inconsistent**
No	49%	14%	3%	27%	53%	9%	3%	34%	71%	0%	0%	25%	88%	5%	0%	10%
Some	4%	21%	3%		13%	13%	0%		13%	4%	0%		5%	0%	0%	
A lot	1%	1%	3%		3%	6%	0%		8%	4%	0%		0%	0%	2%	
UA	**No**	**Some**	**A lot**	**Inconsistent**	**No**	**Some**	**A lot**	**Inconsistent**	**No**	**Some**	**A lot**	**Inconsistent**	**No**	**Some**	**A lot**	**Inconsistent**
No	47%	16%	6%	30%	53%	6%	3%	34%	71%	13%	4%	25%	79%	3%	2%	12%
Some	0%	16%	1%		3%	6%	6%		8%	4%	0%		7%	7%	0%	
A lot	1%	6%	7%		13%	3%	6%		0%	0%	0%		0%	0%	2%	
P/D	**No**	**Some**	**A lot**	**Inconsistent**	**No**	**Some**	**A lot**	**Inconsistent**	**No**	**Some**	**A lot**	**Inconsistent**	**No**	**Some**	**A lot**	**Inconsistent**
No	43%	14%	1%	27%	44%	13%	0%	50%	50%	4%	8%	42%	69%	9%	0%	21%
Some	3%	27%	1%		16%	9%	9%		8%	8%	8%		10%	10%	0%	
A lot	4%	3%	3%		9%	3%	0%		8%	4%	0%		2%	0%	0%	
WSU	**No**	**Some**	**A lot**	**Inconsistent**	**No**	**Some**	**A lot**	**Inconsistent**	**No**	**Some**	**A lot**	**Inconsistent**	**No**	**Some**	**A lot**	**Inconsistent**
No	56%	10%	4%	29%	47%	6%	3%	34%	67%	8%	0%	17%	69%	7%	0%	22%
Some	10%	13%	3%		19%	19%	3%		4%	17%	0%		12%	9%	0%	
A lot	1%	0%	3%		3%	0%	0%		4%	0%	0%		0%	3%	0%	

Mob=mobility, LAM=looking after myself, UA=usual activities, P/D= pain or discomfort, WSU=worried, sad or unhappy. Consistent responses shaded in grey

3.2.4. Known-group validity

Although there were no significant differences in dimension rank order correlations for either the SC or IA (Table 9) by age, sex or health condition there were some differences to note as detailed below.

Table 9: Spearman's rank correlation of EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA scores across health groups, age and sex

	Age (years)*		Sex		Health Condition	
	SC	IA	SC	IA	SC	IA
Mob	0.03	-0.01	-0.01	0.09	0.04	0.00
LAM	-0.02	-0.04	0.07	0.06	0.02	-0.09
UA	-0.07	0.01	-0.07	0.10	0.01	0.03
P/D	-0.08	0.05	0.04	0.13	0.04	-0.10
WSU	0.02	-0.02	-0.04	0.05	0.13	0.00
Utility score	0.08	0.01	-0.02	-0.10	-0.07	0.03
VAS score	-0.04	-0.05	0.09	0.10	-0.02	0.09

**Age was computed as a continuous variable. Health condition was compared by those with an orthopaedic condition, functional disability, respiratory illness and GenPop. Mob=mobility, LAM=looking after myself, UA=usual activities, P/D= pain or discomfort, WSU=worried, sad or unhappy, SC=EQ-5D-Y-3L-SC, IA=EQ-5D-Y-3L-IA*

There was no significant difference between reporting of problems between any of the EQ-5D-Y-3L dimensions, utility or VAS score by age (Table 9). However, as seen in Table 10, 9-year-olds reported less problems on the SC than on the IA version for Mob. For the dimension of UA the 10-year-olds reported less problems on SC than on the IA. Conversely the 8-year-olds reported more problems with P/D on the IA than the SC. The 9-year-olds reported more problems with WSU on the IA than on the SC.

Nine-year-olds reported the highest utility score[2] and VAS score on the SC compared to the IA, however, there were no significant differences between utility scores[2] and VAS scores on the IA and SC versions across ages.

[2] Analysis with the Slovenian utility score is presented. There was no significant difference between results using the Slovenian or Japanese utility scores.

Table 10: Comparison of the EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA dimensions across age (years)

		SC (n=207)						IA (n=207)					
		8-year-olds (n=65)		9-year-olds (n=70)		10-year-olds (n=72)		8-year-olds (n=65)		9-year-olds (n=70)		10-year-olds (n=72)	
		n	%	n	%	n	%	n	%	n	%	n	%
Mob	No	45	69%	55	79%	51	71%	46	71%	50	71%	50	69%
	Some	6	9%	12	17%	12	17%	9	14%	16	23%	18	25%
	A lot	6	9%	1	1%	5	7%	10	15%	4	6%	4	6%
	Missing	8	12%	2	3%	4	6%	0	0%	0	0%	0	0%
LAM	No	39	60%	55	79%	49	68%	44	68%	54	77%	52	72%
	Some	14	22%	9	13%	14	19%	17	26%	12	17%	18	25%
	A lot	4	6%	3	4%	5	7%	4	6%	4	6%	2	3%
	Missing	8	12%	3	4%	4	6%	0	0%	0	0%	0	0%
UA	No	41	63%	52	74%	52	72%	43	66%	52	74%	46	64%
	Some	11	17%	7	10%	11	15%	16	25%	10	14%	21	29%
	A lot	6	9%	8	11%	4	6%	6	9%	8	11%	5	7%
	Missing	7	11%	3	4%	5	7%	0	0%	0	0%	0	0%
P/D	No	34	52%	43	61%	45	63%	44	68%	45	64%	44	61%
	Some	16	25%	18	26%	22	31%	17	26%	20	29%	23	32%
	A lot	9	14%	6	9%	3	4%	4	6%	5	7%	5	7%
	Missing	6	9%	2	3%	2	3%	0	0%	0	0%	0	0%
WSU	No	41	63%	47	67%	47	65%	46	71%	49	70%	52	72%
	Some	17	26%	16	23%	19	26%	15	23%	17	24%	17	24%
	A lot	2	3%	4	6%	4	6%	4	6%	4	6%	3	4%
	Missing	9	14%	3	4%	2%	0%	0	0%	0	0%	0	0%
Utility score	Median (IQR)	0.811 (0.59,1.00)		0.976 (0.57,1.00)		0.894 (0.69, 1.00)		0.848 (0.61,1.00)		0.883 (0.65,1.00)		0.848 (0.61,1.00)	
VAS	Median (IQR)	98 (58,100)		100 (85,100)		90 (65,100)		100 (60,100)		100 (85,100)		95 (68,100)	

Mob=mobility, LAM=looking after myself, UA=usual activities, P/D= pain or discomfort, WSU=worried, sad or unhappy, SC=EQ-5D-Y-3L-SC, IA=EQ-5D-Y-3L-IA

There were further no significant differences in dimensions, SC utility[3] (H=1.887, p=0.389), IA utility[3] (H=0.571, p=0.751), SC VAS (H=3.633, p=0.163) and IA VAS (H=4.411, p=0.110) scores between males and females. Differences in reporting across dimensions in males and females are highlighted in Table 11. Males reported less problems on the SC for WSU than females. Whereas on the IA version, males reported more problems than females for dimensions of LAM and P/D compared to females.

[3] Analysis with the Slovenian utility score is presented. There was no significant difference between results using the Slovenian or Japanese utility scores.

When dimension scores were evaluated by sex, females did not report any significant differences in dimension, utility[4] (z=0.15, p=0.880) or VAS scores (z=0.15, p=0.880) for either version. As seen in Table 11, females had greater reporting of problems on the SC version for dimensions of LAM, WSU and P/D. The reporting of problems in males was marginally higher on the IA version for Mob, LAM, P/D and the same for UA when compared to the SC version, although not significant. There was greater report of problems on the SC for WSU than the IA, although not significantly so. Males similarly did not show a significant difference in the utility[4] (z=1.39, p=0.162) nor VAS scores (z=0.00, p=1.00).

Table 11: Comparison of the EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA dimensions across sex

		SC (n=207)				IA (n=207)			
		Female (n=96)		Male (n=111)		Female (n=96)		Male (n=111)	
		n	%	n	%	n	%	n	%
Mob	No	70	73%	81	73%	68	71%	78	70%
	Some	14	15%	16	14%	18	19%	25	23%
	A lot	7	7%	5	5%	10	10%	8	7%
	Missing	5	5%	9	8%	0	0%	0	0%
LAM	No	66	69%	77	69%	74	77%	76	68%
	Some	19	20%	18	16%	18	19%	29	26%
	A lot	5	5%	7	6%	4	4%	6	5%
	Missing	6	6%	9	8%	0	0%	0	0%
UA	No	68	71%	77	69%	64	67%	77	69%
	Some	15	16%	14	13%	23	24%	24	22%
	A lot	8	8%	10	9%	9	9%	10	9%
	Missing	5	5%	10	9%	0	0%	0	0%
P/D	No	55	57%	67	60%	67	70%	66	59%
	Some	26	27%	30	27%	22	23%	38	34%
	A lot	10	10%	8	7%	7	7%	7	6%
	Missing	5	5%	6	5%	0	0%	0	0%
WSU	No	58	60%	77	69%	68	71%	79	71%
	Some	29	30%	23	21%	23	24%	26	23%
	A lot	6	6%	4	4%	5	5%	6	5%
	Missing	3	3%	7	6%	0	0%	0	0%
Utility score	Median (IQR)	0.883 (0.590,1.000)		0.917 (0.674,1.00)		0.883 (0.595,1.00)		0.838 (0.637,1.000)	
VAS	Median (IQR)	100 (60,100)		95 (75,100)		100 (68,100)		95 (80,100)	

Mob=mobility, LAM=looking after myself, UA=usual activities, P/D= pain or discomfort, WSU=worried, sad or unhappy, SC=EQ-5D-Y-3L-SC, IA=EQ-5D-Y-3L-IA

[4] Analysis with the Slovenian utility score is presented. There was no significant difference between results using the Slovenian or Japanese utility scores.

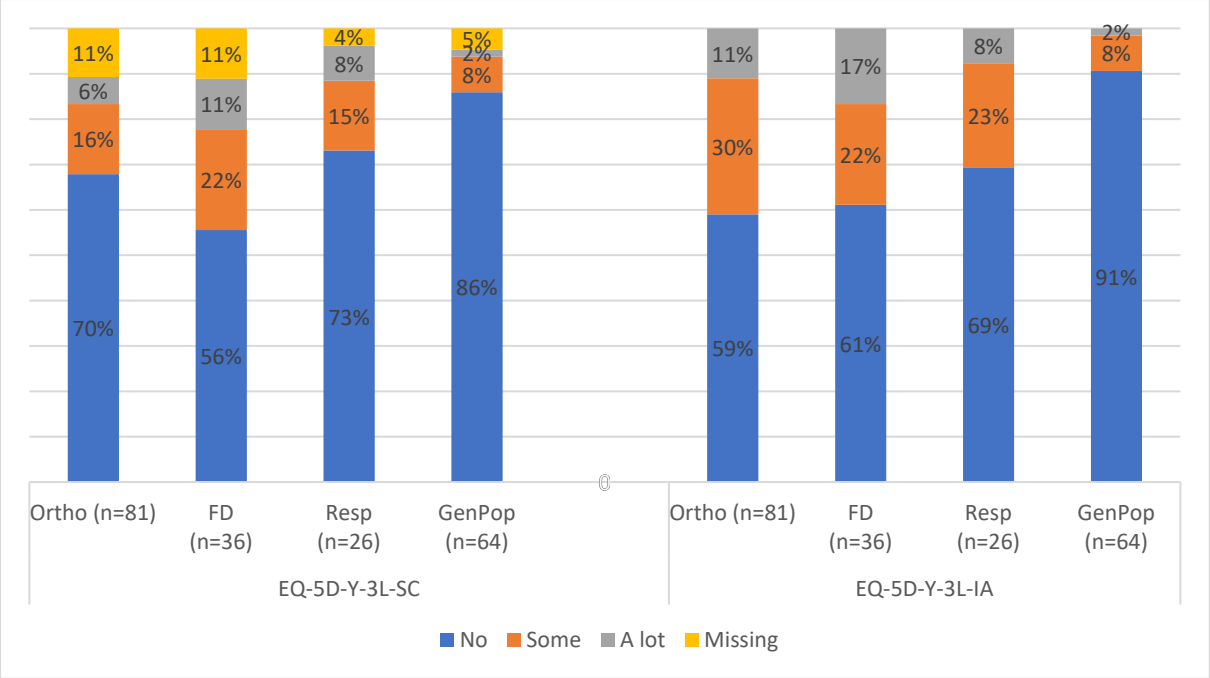


Figure 4: Comparison of the EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA Mobility dimension across health conditions

Figure 4 shows that for the SC, the dimension of Mob showed greater problems for those with functional disabilities, respiratory illnesses, orthopaedic conditions and the GenPop respectively, whereas on the IA, the order of problems was higher for those with orthopaedic conditions compared to all other health conditions.

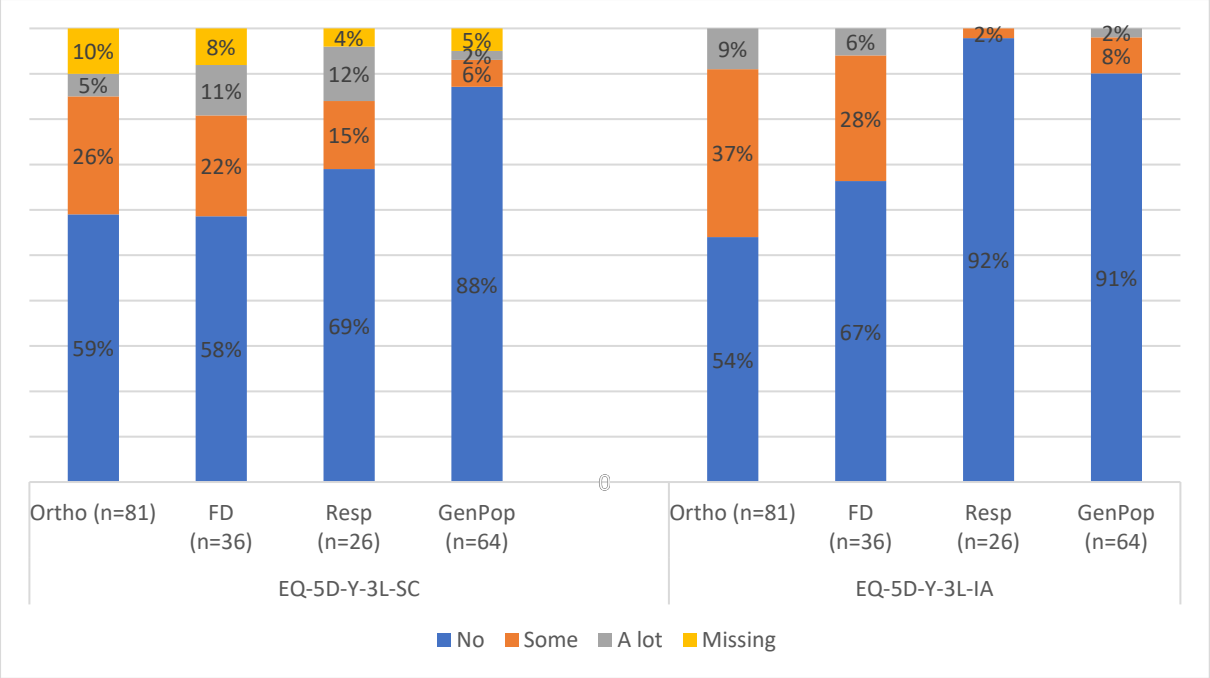


Figure 5: Comparison of the EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA Looking After Myself dimension across health conditions

Figure 5 shows that the orthopaedic group reported greater problems for the SC LAM dimension, compared to those with respiratory illnesses, functional disabilities and from the GenPop. A similar pattern was seen for the IA version regarding the orthopaedic and functional disability groups, however, those with respiratory illnesses and from the GenPop showed similar problems for LAM.

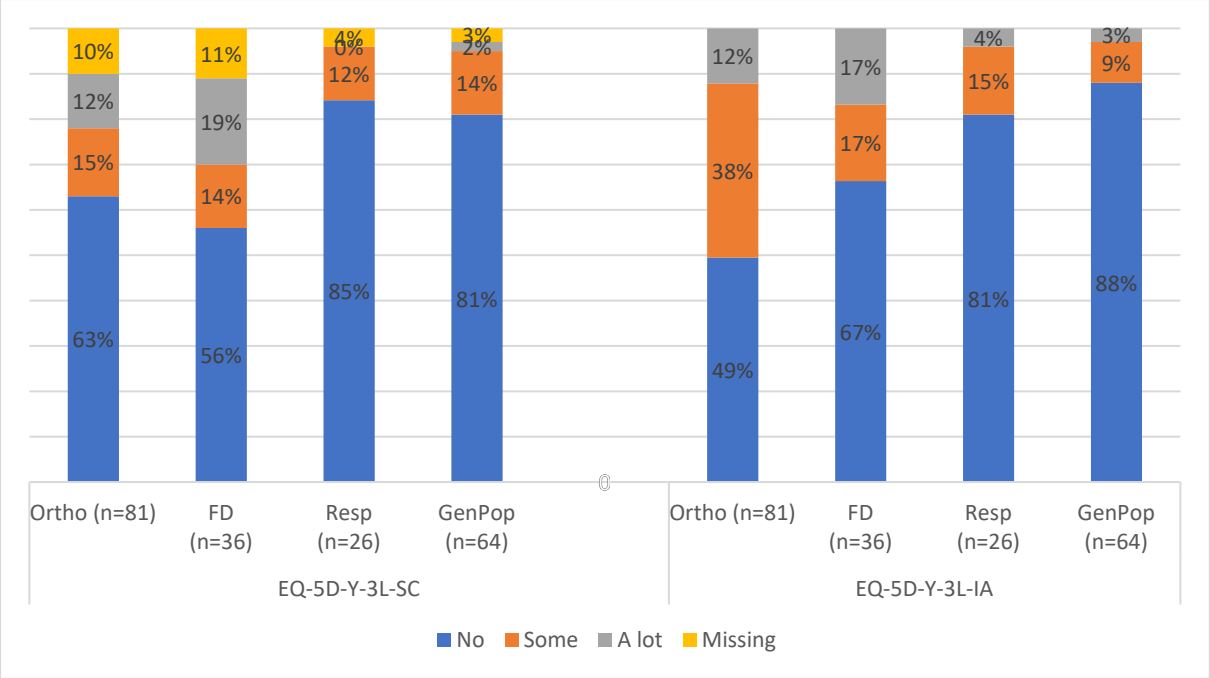


Figure 6: Comparison of the EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA Usual Activities dimension across health conditions

Figure 6 shows that children with functional disabilities reported the greatest proportion of problems with UA on the SC however, the orthopaedic group reported the greatest proportion of problems on the IA version which was followed by those with functional disabilities.

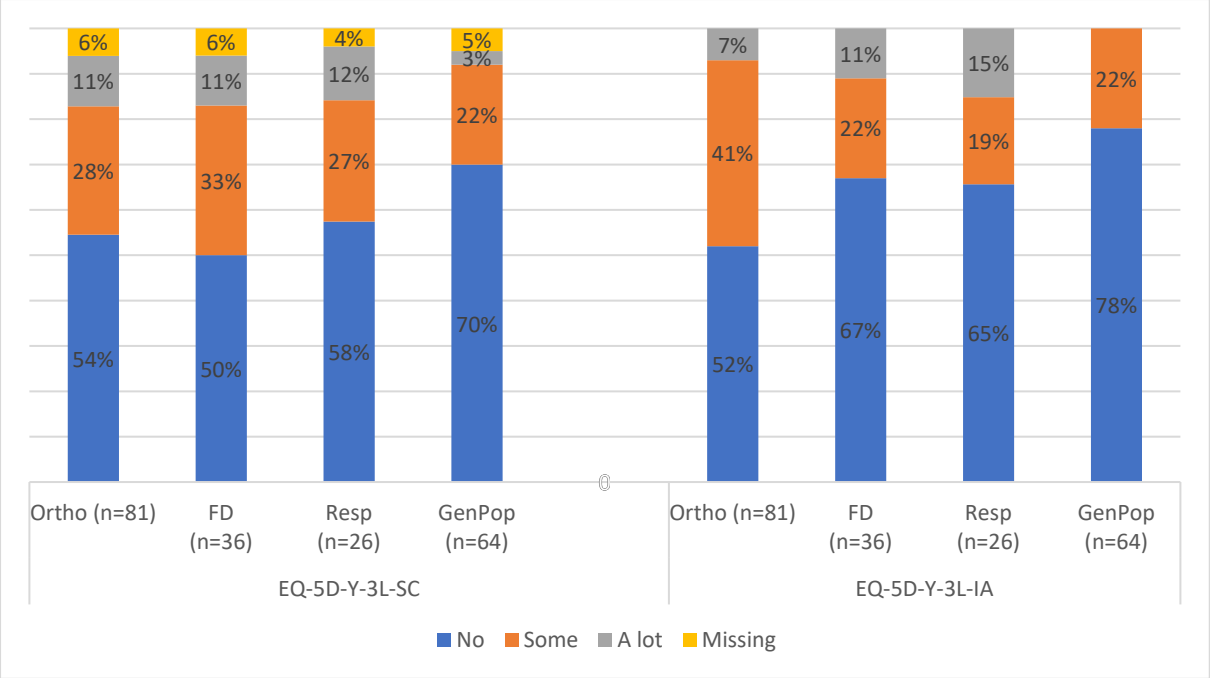


Figure 7: Comparison of the EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA Pain/Discomfort dimension across health conditions

Figure 7 shows that those with functional disabilities reported the greatest proportion of P/D. However, on the IA version, the orthopaedic group reported the greatest proportion of P/D.

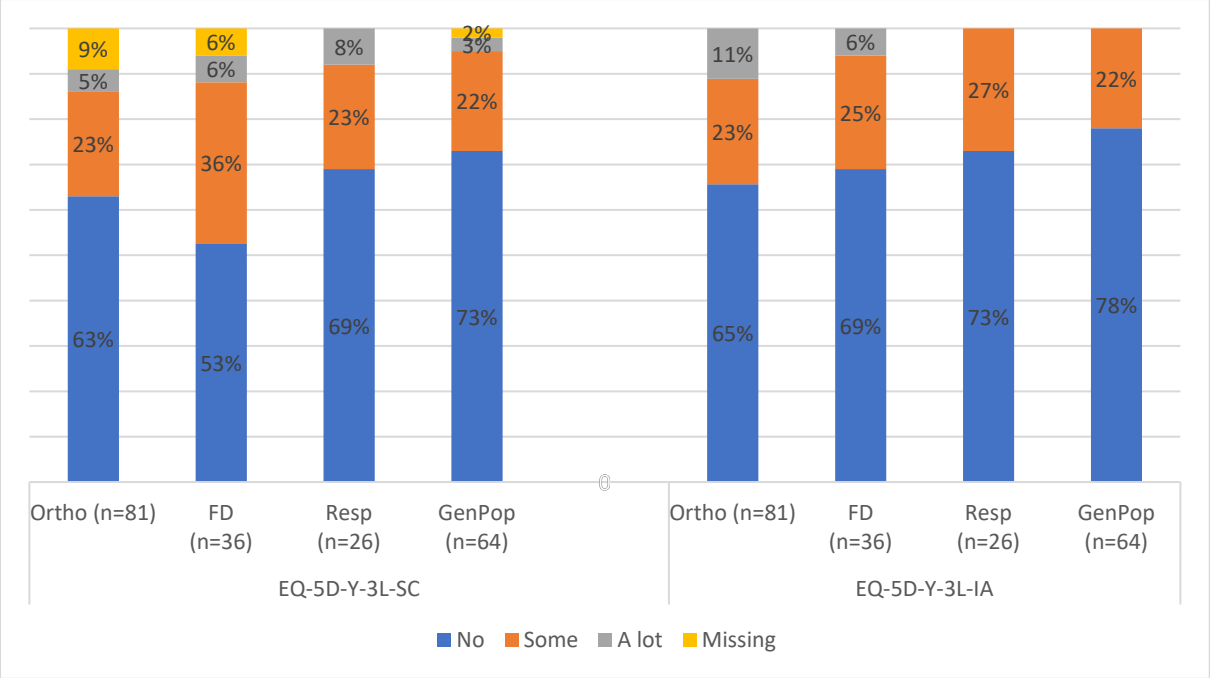
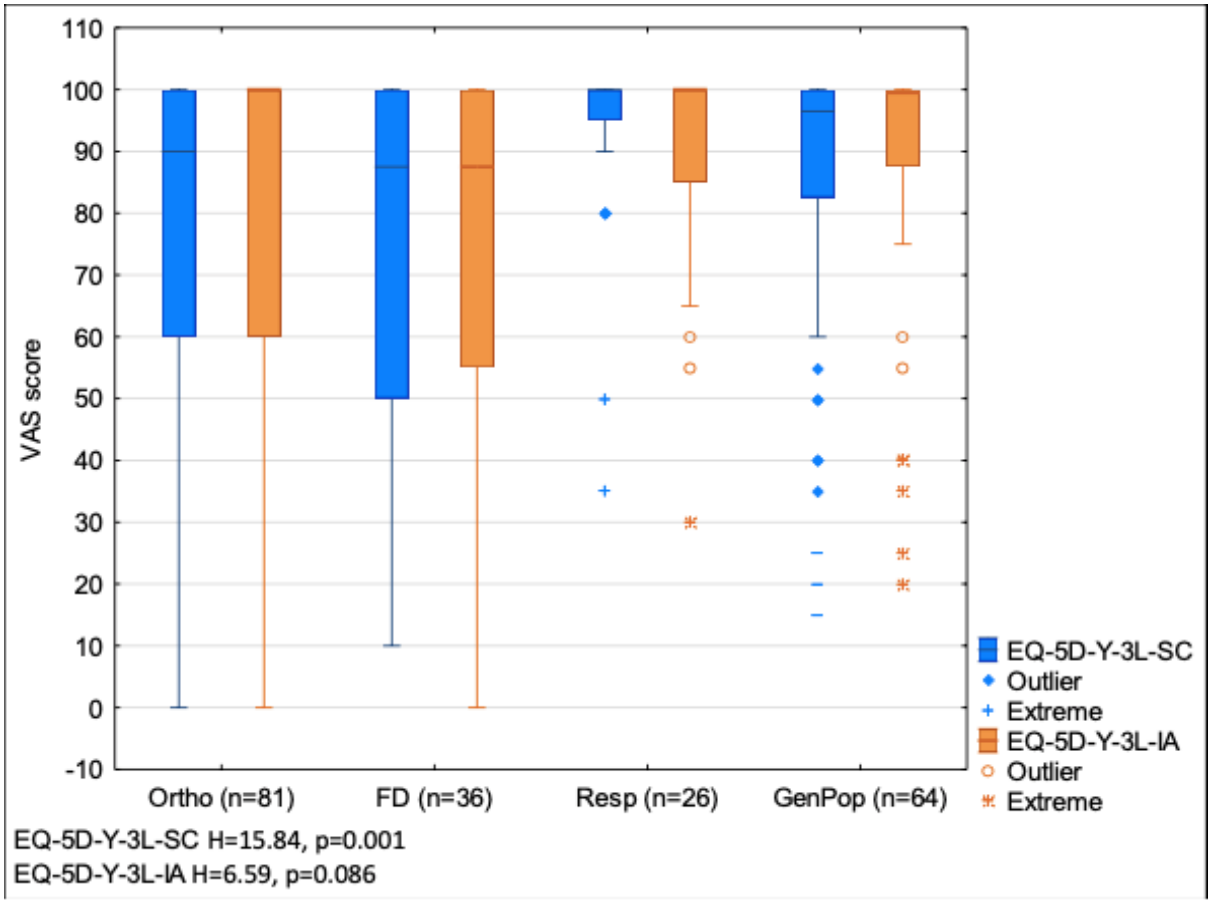


Figure 8: Comparison of the EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA Worried, Sad or Unhappy dimension across health conditions

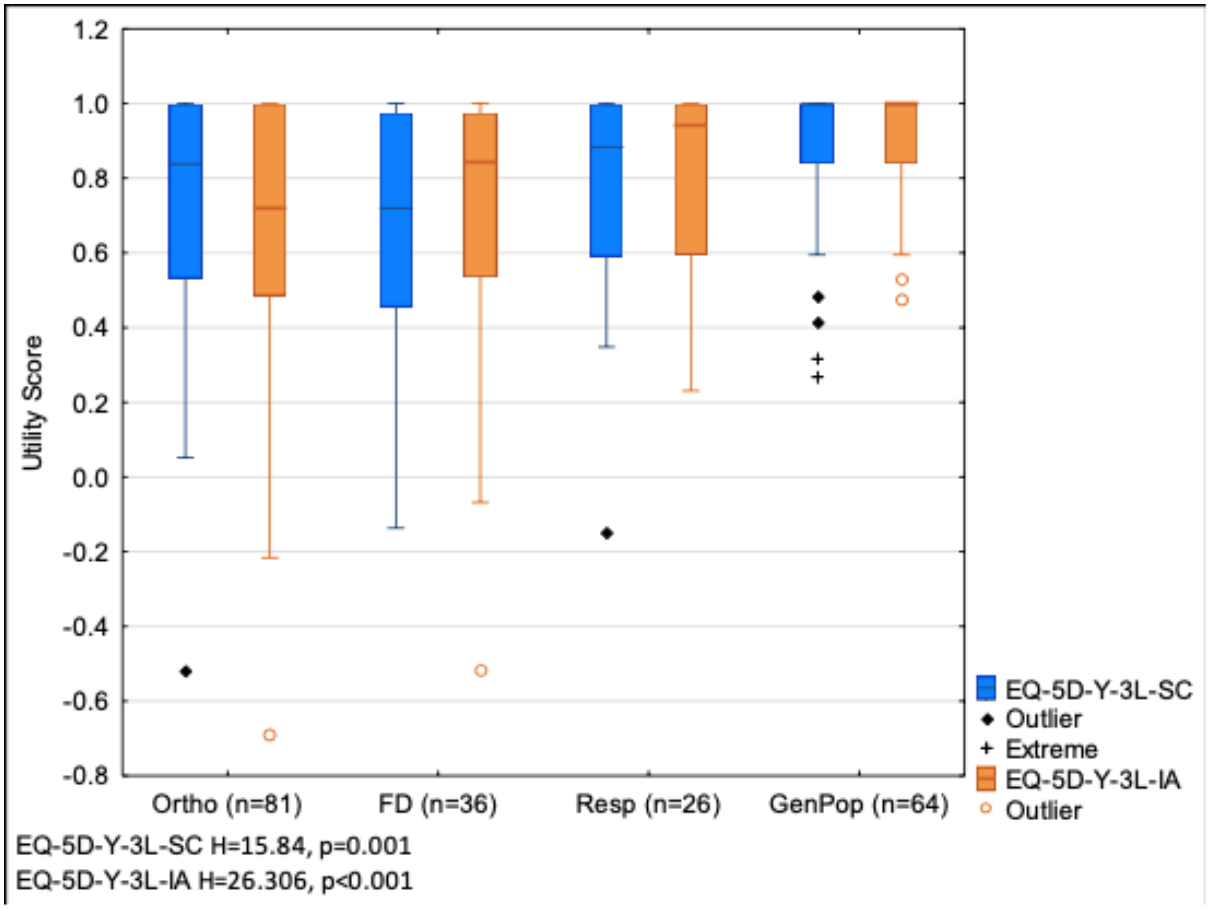
Figure 8 shows that those with functional disabilities reported the greatest proportion of problems in the WSU dimension on the SC version. However, on the IA version, the orthopaedic group reported the greatest proportion of problems.



Orthopaedic (Ortho), Functional Disability (FD), Respiratory illness (Resp) and General Population (GenPop)

Figure 9: EQ-5D-Y-3L-IA and EQ-5D-Y-3L-SC VAS scores

As seen in Figure 9, The VAS score was significantly different between groups on the SC version (H=15.84, p=0.001) but not the IA version (H=6.59, p=0.086). Post hoc analysis showed differences on the SC version between children with a respiratory illness and functional disability (H=-2.54, p=0.011) and orthopaedic condition (H=2.626, p=0.009).



Orthopaedic (Ortho), Functional Disability (FD), Respiratory illness (Resp) and General Population (GenPop)

Figure 10: EQ-5D-Y-3L-IA and EQ-5D-Y-3L-SC utility scores

As seen in Figure 10, the utility scores[5] were significantly different between groups on SC (H=15.84, p=0.001) and IA (H=26.306, p<0.001). Post-hoc analysis showed that SC differences were between the GenPop and children with an acute orthopaedic condition and (H=-3.59, p=0.001) and functional disability (H=-3.135, p=0.002). The IA similarly found differences between the GenPop and an acute orthopaedic condition (H=4.939, p<0.001), functional disability (H=3.252, p<0.001) and additionally those with a respiratory illness (H=-2.124, p<0.001).

[5] Analysis with the Slovenian utility score is presented. There was no significant difference between results using the Slovenian or Japanese utility scores.

3.2.5. Concurrent validity

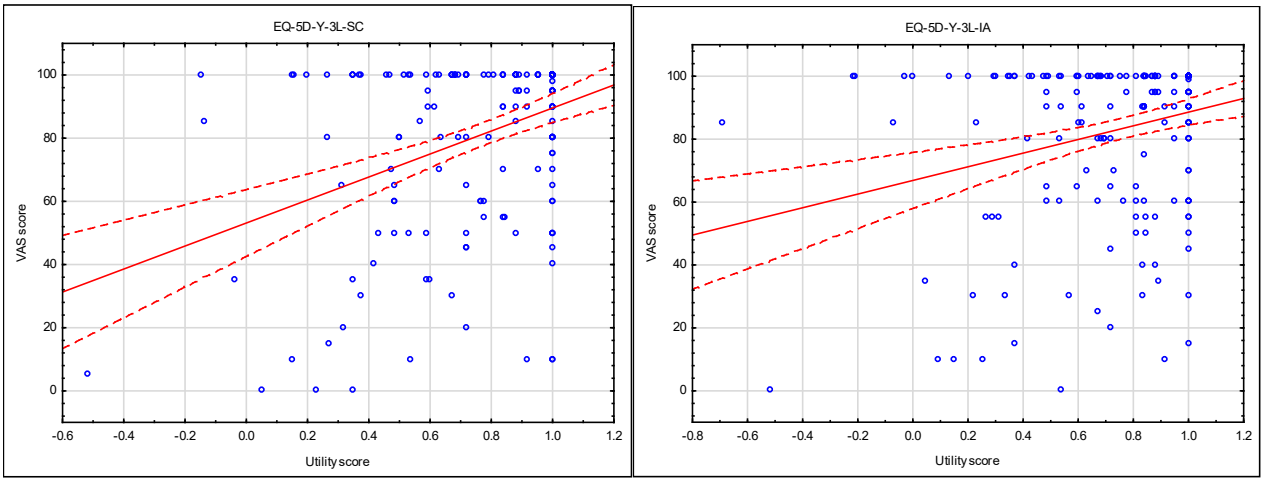


Figure 11: Scatterplot of utility scores versus VAS scores for the EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA versions in children 8-10-years (n=207)

The concurrent validity was assessed by the correlation of the VAS and utility score[6] which was significant and moderate for the SC version (r=0.38 p<0.001) and significant and low for the IA version (r=0.27 p<0.001). There was however, no significant difference between IA and SC versions when comparing their individual concurrent validity to one another (z=1.34, p=0.090) (Figure 11).

3.2.6. Convergent validity

The gamma correlation for the physical dimensions of Mob, LAM and UA showed similar high correlations with P/D and WSU showing significantly lower correlations when considering all children (Table 12). The dimension of Mob showed a significantly higher correlation than P/D (z=2.28, p=0.011) and WSU (z =1.59, p=0.05).

The 8-year-olds showed a significantly lower correlations than the 10-year-olds in the dimensions of Mob (z=-2.88, p=0.002), UA (z=-4.08, p<0.001) and P/D (z=-3.75, p<0.001). While the 9-year-olds similarly showed significantly lower correlations than the 10-year-olds for dimensions of Mob (z=-2.88, p=0.002), UA (z=-3.17, p<0.001), P/D (-2.88, p=0.002) and WSU (z=-1.97, p=0.020). However, the correlation for LAM was significantly higher in the 9-year-olds when compared to the 10-year-olds (z=1.71, p=0.04). When comparing the 8-year-olds to the 9-year-olds, there was a significantly lower

[6] Analysis with the Slovenian utility score is presented. There was no significant difference between results using the Slovenian or Japanese utility scores.

correlation for LAM (z=-2.01, p=0.022), while WSU showed a significantly higher correlation (z=3,34, p<0.001).

Table 12: Gamma Correlation Calculations of the EQ-5D-Y-3L-SC versus the EQ-5D-Y-3L-IA dimension responses across 8-, 9- and 10-year-olds.

SC	**IA**				
Total (n=207)	Mob	LAM	UA	P/D	WSU
Mob	0.74*	0.18	0.52*	0.55*	0.19
LAM	0.58*	0.76*	0.59*	0.20	0.32*
UA	0.51*	0.46*	0.75*	0.31*	0.21*
P/D	0.44*	0.41*	0.43*	0.62*	0.47*
WSU	0.44*	0.39*	0.28*	0.52*	0.66*
8-years (n=65)					
Mob	0.68*	0.04	0.54*	0.47*	-0.34
LAM	0.91*	0.70*	0.69*	0.36*	0.26
UA	0.53*	-0.01	0.64*	0.10	-0.13
P/D	0.52*	0.32*	0.18	0.59*	0.58*
WSU	0.50*	-0.09	0.12	0.32	0.80*
9-years (n=70)					
Mob	0.62*	0.24	0.60*	0.73*	0.44*
LAM	0.17	0.84*	0.39*	-0.10	0.42*
UA	0.61*	0.70*	0.73*	0.61*	0.43*
P/D	0.48*	0.45*	0.53*	0.51*	0.19
WSU	0.36*	0.49*	0.29	0.56*	0.47*
10-years (n=72)					
Mob	0.87*	0.30	0.46*	0.46*	0.30
LAM	0.42*	0.73*	0.67*	0.20	0.30
UA	0.38*	0.59*	0.90*	0.14	0.27
P/D	0.30	0.44*	0.58*	0.74*	0.61*
WSU	0.47*	0.65*	0.40*	0.61*	0.69*

**p<0.05. Mob=mobility, LAM=looking after myself, UA=usual activities, P/D= pain or discomfort, WSU=worried, sad or unhappy, SC=EQ-5D-Y-3L-SC, IA=EQ-5D-Y-3L-IA*

Table 13 shows that IA and SC Mob dimension had a significant low to moderate association with all WeeFIM items of mobility as well as the motor total score. Comparison of the correlations between the two versions showed that the correlations were significantly higher for all WeeFIM items and the IA version.

Table 13: Convergent validity of the EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA Mobility dimension and WeeFIM Mobility

	Mobility			
	SC	IA	SC vs IA	
WeeFIM Mobility			z-score	p-value
Sit to stand transfer	-0.17*	-0.33**	1.73	0.042
Toilet transfer	-0.21**	-0.36**	1.65	0.050
Tub or shower transfer	-0.24**	-0.40**	1.81	0.035
Locomotion (walk/wheelchair for ≥45m or crawl ≥15m)	-0.31**	-0.45**	1.66	0.049
Stairs climbing (ascend and descend 12-14 stairs)	-0.23**	-0.47**	2.79	0.003
Motor Total	-0.23**	-0.40**	1.91	0.028

*N=207 *Spearman's correlation p<0.05, **Spearman's correlation p<0.001, A higher WeeFIM score indicates greater independence, a higher EQ-5D-Y-3L-IA score indicates greater problems. Shaded cells indicate comparison of correlations on the IA and SC versions. SC=EQ-5D-Y-3L-SC, IA=EQ-5D-Y-3L-IA*

Table 14 shows significant moderate to strong associations for the SC and IA LAM dimensions and WeeFIM self-care items, self-care total excluding grooming, bladder and bowel continence. There were significantly higher correlations on the IA with all WeeFIM items except toileting, bladder and bowel continence.

Table 14: Convergent validity of the EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA Looking After Myself dimension and WeeFIM Self-care

	Looking After Myself			
WeeFIM Self-Care	SC	IA	SC vs IA	
			z-score	p-value
Grooming	-0.25**	-0.39**	1.58	0.057
Bathing (washing body excluding back)	-0.45**	-0.68**	3.48	0.000
Dressing Upper Body	-0.38**	-0.59**	2.69	0.004
Dressing Lower Body	-0.43**	-0.62**	2.68	0.004
Toileting (perineal hygiene and adjusting clothing before and after toilet use)	-0.30**	-0.31**	0.11	0.456
Bladder (continence)	-0.11	-0.15*	0.41	0.341
Bowel (continence)	-0.12	-0.16*	0.41	0.341
Self-Care Total	-0.44**	-0.66**	3.24	0.001

N=207 **Spearman's correlation p<0.05, **Spearman's correlation p<0.001, A higher WeeFIM score indicates greater independence, a higher EQ-5D-Y-3L-IA score indicates greater problems. The WeeFIM item of eating was excluded from analysis as all children scored total independence with eating and there was thus no variation. Shaded cells indicate comparison of correlations on the IA and SC versions. SC=EQ-5D-Y-3L-SC, IA=EQ-5D-Y-3L-IA*

Table 15 shows no significance between the SC and IA UA dimension and the WeeFIM social interaction item however, the IA and SC UA dimension showed significant low to moderate associations for the WeeFIM mobility items and mobility total. There were significantly higher correlations on all of the WeeFIM items of mobility and the UA dimension on the IA version.

Table 15: Convergent validity of the EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA Usual Activities dimension and WeeFIM Mobility and Social Interaction

	Usual Activities			
	SC	IA	SC vs IA	
WeeFIM Mobility			z-score	p-value
Sit to Stand Transfer	-0.17*	-0.33**	1.73	0.042
Toilet transfer	-0.21**	-0.36**	1.65	0.050
Tub or shower transfer	-0.24**	-0.40**	1.81	0.035
Locomotion (walk/wheelchair ≥45m OR crawl ≥15m)	-0.31**	-0.45**	1.66	0.049
Stairs climbing (ascend and descend 12-14 stairs)	-0.23**	-0.47**	2.79	0.003
Mobility Total	-0.29**	-0.48**	2.27	0.012
Motor Total[§]	-0.23**	-0.40**	1.91	0.028
WeeFIM Cognition				
Social Interaction (interaction with other children)	0.04	0.05	-0.1	0.4602

*N=207 *Spearman's correlation p<0.05, **Spearman's correlation p<0.001, [†§]Motor Total = Mobility Total + Self-care Total. A higher WeeFIM score indicates greater independent, a higher EQ-5D-Y-3L-IA score indicates greater problems. Shaded cells indicate comparison of correlations on the IA and SC versions. SC=EQ-5D-Y-3L-SC, IA=EQ-5D-Y-3L-IA*

Table 16 shows a moderate association between the SC and IA P/D dimension and the FPS-R with no significant difference between the SC and IA versions.

Table 16: Convergent validity of the EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA Pain/Discomfort dimension and the Faces Pain Scale-Revised

	Pain/Discomfort			
	SC	IA	SC vs IA	
			z-score	p-value
Faces Pain Scale-Revised	0.33**	0.38**	-0.58	0.281

*N=207 *Spearman's correlation p<0.05, **Spearman's correlation p<0.001, A higher Faces Pain Scale and EQ-5D-Y-3L-IA score both indicate greater pain or discomfort. Shaded cells indicate comparison of correlations on the IA and SC versions. SC=EQ-5D-Y-3L-SC, IA=EQ-5D-Y-3L-IA*

Table 17 shows the SC and IA WSU dimension both had a moderate association with the MFQ total score. Both versions showed low and similar associations between MFQ items, except the IA and the bad person item which showed a moderate and significant association.

Table 17: Convergent validity of the EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA Worried, Sad or Unhappy dimension and the Moods and Feelings Questionnaire

	Worried, Sad or Unhappy			
	SC	IA	SC vs IA	
Moods and Feelings Questionnaire			z-score	p-value
Unhappy	0.26**	0.18**	-0.58	0.281
Enjoyment	0.21**	0.19**	0.21	0.417
Tired	0.12	0.16*	-0.41	0.341
Restless	0.21**	0.19**	0.21	0.417
No good	0.21**	0.16*	0.52	0.302
Crying	0.17*	0.22**	-0.53	0.298
Concentration	0.14	0.22**	-0.84	0.201
Hate	0.14*	0.12	0.21	0.417
Bad person	0.08	0.30**	-2.32	0.010
Lonely	0.17*	0.20**	-0.31	0.378
Love	0.15*	0.16*	-0.1	0.460
Comparison	0.09	0.17*	-0.82	0.206
Did things wrong	0.15*	0.20**	-0.52	0.302
Total	0.33**	0.34**	-0.11	0.456

*N=207 *Spearman's correlation p<0.05, **Spearman's correlation p<0.001, A higher Moods and Feelings score and EQ-5D-Y-3L-IA score both indicate greater problems. Shaded cells indicate comparison of correlations on the IA and SC versions. SC=EQ-5D-Y-3L-SC, IA=EQ-5D-Y-3L-IA*

3.2.7. Preference of version

As seen in Table 18 below, overall, there were more children who preferred the IA (n=125, 60%) compared to those who preferred the SC (n=77, 37%) or had no preference (n=5, 2%) (x^2*=21.87, p<0.001).* There was no significant difference between preferences for age, sex or health conditions.

Table 18: Preference between EQ-5D-Y-3L-SC and EQ-5D-Y-3L-IA by age (years), sex and health conditions

	IA		SC		No preference			
Age	n	%	n	%	n	%	x^2	p-value
8-years (n=65)	39	60%	26	40%	0	0%	5.12	0.275
9-years (n=70)	44	63%	25	36%	1	1%		
10-years (n=72)	42	58%	26	36%	4	6%		
Sex								
Female (n=96)	65	68%	28	29%	3	3%	5.07	0.079
Male (n=111)	60	54%	49	44%	2	2%		
Health condition								
Orthopaedic (n=81)	49	60%	29	36%	3	4%	3.72	0.715
Functional Disability (n=36)	21	58%	14	39%	1	3%		
Respiratory (n=26)	13	50%	13	50%	0	0%		
GenPop (n=64)	42	66%	21	33%	1	2%		

SC=EQ-5D-Y-3L-SC, IA=EQ-5D-Y-3L-IA

As seen in Table 19, the IA version was preferred across all age-groups as they reported that they did not yet have the literacy skills for self-completion "*I can't read yet, I am still learning to read".* This was notably higher in those aged 8-9-years. However, the 10-year-olds did report that they preferred it to the SC version as it was easier, quicker, more understandable and factors associated with the interviewer "*you read it nice and slow"* which could all indicate some difficulty with literacy. The general preference included children stating that it was "*better"* or "*nicer."*

The preference for the SC version across the age-groups was related to an independence on completion of the SC version with children stating, "*I liked to do it on my own"*. General preference for the measure was not specific and included "*I liked it more, it was better".*

The reason for no preferences included: *"both were fine", "both were easy" or "I liked both".*

Table 19: Reason for preference between EQ-5D-Y-3L-IA and EQ-5D-Y-3L-SC

	8-years		9-years		10-years		Total	
	(n=65)		(n=70)		(n=72)		(n=207)	
Reason for IA preference	n	%	n	%	n	%	n	%
Associated with literacy skills	16	25%	16	23%	13	18%	45	22%
Easier	5	8%	6	9%	4	6%	15	7%
More understandable	4	6%	8	11%	7	10%	19	9%
Listening/answering preferred over reading	4	6%	2	3%	6	8%	12	6%
Associated with interviewer	3	5%	3	4%	8	11%	14	7%
General preference	3	5%	4	6%	3	4%	10	5%
Other[#]	3	5%	5	7%	2	3%	10	5%
Reason for SC preference								
Associated with independence	20	31%	16	23%	19	26%	55	27%
General preference	4	6%	3	4%	1	1%	8	4%
More understandable	1	2%	2	3%	1	1%	4	2%
Easier	1	2%	1	1%	1	1%	3	1%
No talking required	0	0%	1	1%	1	1%	2	1%
More time to think about answers	0	0%	0	0%	2	3%	2	1%
Other*	0	0%	2	3%	1	1%	3	1%
No preference	0	0%	1	1%	4	6%	5	2%

*[#]Included the ability to complete the IA and provide the correct answers, the IA being quicker, enjoyed the conversation, feeling nervous to complete the IA version and the lack of enjoyment associated with reading. *Included the preference of reading over listening, the SC version was quicker and the instructions were clear.*

3.2.8. Summary of results

A total of 207 8-10-years-olds were recruited and allocated to four know-groups: orthopaedic condition (n=81, 39%), GenPop (n=64, 31%), functional disability (n=36, 13%) and respiratory illnesses (n=26, 13%). The 8-year-olds had significantly higher missing responses (x^2=14,23, p<0.001) on the SC version compared to no missing values on the IA version. The most inconsistencies between versions were noted in the psychosocial dimensions with a higher report of problems on the SC in the P/D dimension (31%) and WSU (26%) had the largest proportion of inconsistent responses with a higher report of problems on the SC version. Known-group and concurrent validity were comparable across dimensions, utility and VAS scores for the two versions. The dimensions showed low to moderate convergent validity with similar items on the, MFQ, FPS-R and WeeFIM with significantly higher correlations between the IA dimension of Mob and WeeFIM mobility total (z=1.91, p=0.028) and LAM and WeeFIM self-care total (z=3.24, p=0.001). Children preferred the IA version (60%) (x^2=21.87, P<0.001) with 22% of the reasons attributed to literacy level.

3.3. Discussion

The aim of this chapter was to compare the performance of the newly developed EQ-5D-Y-3L-IA in children aged 8-10-years, to the already validated SC version. Due to its recent development, this is the first study to assess the performance of the IA version. The results of this chapter will be discussed in the order by which they were presented above.

3.3.1. Recruitment and descriptive statistics

This sample of 207 participants was recruited from similar socioeconomic settings in the Western Cape, South Africa across four known condition groups. These groups included children with acute orthopaedic conditions, chronic respiratory illnesses, functional disabilities and from the GenPop who presented with minor atopic conditions. Analyses of known-groups were done using non-parametric tests but were not matched for age and sex. Although the number of participants recruited in each known-group was different, there was no significant difference across health conditions across age (x^2=3.61, p=0.729) or across 8-, 9- and 10-year-olds (x^2=0.03, p=0.985).

Previous studies using the EQ-5D-Y-3L have mostly assessed larger samples of children from the GenPop in countries including Canada (n=3421), South Africa (n=521), Spain (n=620) and the United Kingdom (n=160) (18,87,106,143), in comparison this study which used a smaller sample of 207 children but included children with various health conditions and differing severities. A large variety of paediatric health conditions have also been assessed in various studies some of which overlap with those assessed in this study and included chronic conditions such as CF, diabetes mellitus, JIA and acute conditions such as traumatic or orthopaedic injuries, acute lymphoblastic leukaemia, in addition to the GenPop (9,13,16,30,83,98). Some studies have used the GenPop groups as a comparator group to a single disease group, (9,17) while others, including this study, used the GenPop as a comparator group to multiple disease groups (10,16).

In this study, there was a high report of non-responders (64%) in the GenPop to which a very similar result of 63% was found by Jelsma et al. (2012) which looked closely at the process of obtaining consent from parents of GenPop children in a South African population. Apart from refusing consent, another reason suggested for poor response may also be related to the steps leading up to providing consent which often relies on the child to deliver the form to the parent. This is often where the issues arise as parents sometimes never receive the forms and if they do receive them, they are then expected to read the form, decide to either grant or refuse consent and finally, ensure that the consent

form is safely returned by the child (144). It may be useful to align research studies at schools with parent-teacher conferences so that information and forms can be handed directly to parents, therefore, children are only expected to return the forms to their school teacher. Considering that this study was conducted during the Covid-19 pandemic, the response rate may have been further impacted due to the additional educational responsibilities placed on staff, children and caregivers with reduced in-person time or alternate teaching days/weeks at the schools. This may have affected teachers' ability to remind learners and increased the chances of forgetting to complete and return forms by children/parents as they might not have been at school every day. Consideration also needs to be given to the difference in methods of recruitment between school-going children and in-patient children. Face-to-face contact with parents allowed for direct communication between parents and the researcher, therefore allowing concerns or queries to be addressed immediately and being more willing to allow their child to participate compared to receiving an envelope from school.

3.3.2. General instrument performance and feasibility

When comparing the IA and SC versions, there were no statistical differences between dimensions, utility scores or VAS scores.

The missing values on the SC version were high (11%) in comparison to other studies, in children aged 8-16-years, where missing values ranged from none (9,16,87) to a maximum of 2% (2,14,145). Overall, both LAM and UA had the most missing values (n=15). The missing values in the UA dimension may have been due to children struggling to consider the large number of activities in the UA dimension, while the LAM missing values were most common in orthopaedic group (10%) and may have been due to the inability to complete washing/dressing tasks due to their injury and therefore skipped the question. The number of missing responses was significantly higher in 8-year-olds and thus, it may be beneficial to recommend the use of the IA in settings where the literacy levels may negatively influence the ability to self-complete. This is supported by the fact that the children, notably children aged 8-years, who preferred the IA reported difficulty with literacy skills. Difficulty with literacy skills may be unique to the South African sample recruited, which is reported to have lower literacy levels in this age-group compared to international levels (24). As of 2004, the reading ability of Grade 3 learners in South Africa, which comprises mostly of 8- and 9-year-olds, was 44.2%, while overall literacy score was only 35.9% (146).

Although the feasibility of the IA is improved with lower missing values, this is at the cost of a longer time for completion and the added resource of an interviewer. Considering the administration in a

clinical setting, the IA was however still feasible with a relatively low completion time of under three minutes (median=110s, IQR=98s,124s). This is lower than the times reported for self-complete on other generic measures of HRQoL i.e. EQ-5D-Y-3L (5 minutes), CHU-9D (3-5 minutes), HUI_3 (8-10 minutes), KIDSCREEN (5-20 minutes) and PedsQL (10-15 minutes) (22). Therefore suggesting that the IA may be more feasible in a clinical setting compared to other instruments, however, no additional evidence regarding the EQ-5D IA versions is currently available. Unfortunately, direct comparisons cannot be made to the Kiddy-KINDL and DISABKIDS-TAKE-6 either, as time was not recorded during administration, however in relation to time, the DISAKKIDS-TAKE-6 was described as 'short' therefore suggesting it may be feasible option but cannot be known for certain (7).

At a composite level, there was a higher ceiling effect on the SC version across all dimensions (11111). This was notable at an individual dimension level only for Mob and UA. As seen by the ceiling effect and confirmed with the distribution of responses between the IA and SC versions there was a greater reporting of no problems for LAM, P/D and WSU on the IA version. The inconsistency with WSU and P/D was similarly noted on the lower correlation of these dimensions. The inconsistency in responses between the versions, although not significant, may be attributed to social-desirability bias (147) as face-to-face interviews have been shown to produce more socially desirable responses compared to self-complete versions as participants feel as though they need to present themselves in the best way (140). This was similarly noted when comparing the SC and IA versions of the EQ-5D in adult cardiac patients whereby less problems were reported when interviewed, with a significant difference in the P/D dimension (148).

Conversely, there was a higher report of problems on the IA for physical dimensions which may be attributed to observation bias as children may have been more aware of the interviewer's ability to observe their current ability. This may have been further strengthened by the interviewer being a physiotherapist and assessing functional ability on the WeeFIM. A similar observation was seen in a study by Scott et al. (2017) whereby 14% of children reported problems with Mob which was not observed by the researcher on completion of the WeeFIM. It was found that the report of problems were not only associated with physical impairments but also environmental barriers linked to safety in the areas in which they live (16). The influence of the interviewer may further be contributing to the significantly higher convergent validity noted with the IA dimensions of Mob and LAM and the corresponding interviewer-rated WeeFIM items, as the physical ability of the child could objectively be determined and not influenced by subjective aspects such as pain or lack of enjoyment during the activity, which in a child's mind might influence their physical ability. Ultimately, due the subjectivity

of HRQoL, if problems are reported but not objectively observed, it may be useful to note the reason for reporting problems to understand the origin of the problem so that health professionals are able to address and monitor the problem appropriately. Conversely, problems with Mob might be objectively observed, but not subjectively reported by a child which may affect the ability of a health professional to adequately treat the observed problem if it is not identified by the child.

3.3.3. Known-group validity

Known-group validity was compared across age (in years), sex and health conditions with no significant differences in dimensions, utility scores or VAS scores across any known-groups.

There were slightly more males enrolled in this study however, no selection/recruitment bias was identified. According to the 2016 South African census, there are slightly more males in the 5-9-year and 10-14-year age-groups (149) which may have affected enrolment or may have been due to chance that more males were present during the enrolment period. Both the chronic respiratory and acute orthopaedic groups included more females, whereas the functional disability group was equally distributed while previous studies found males to present with more functional disabilities than females (9,17,150). Although not significant, females reported more problems in the WSU dimension compared to males on both the IA and SC versions. Similar findings, also not significant, were seen in a Korean study including children from the GenPop aged 7-12-years (14) and a Swedish study including children aged 13-18-years with various self-reported health conditions such as asthma, allergic eyes or nasal symptoms, food allergy, nickel allergy, eczema, other skin disease, diabetes, and/or epilepsy (105) both of which used the SC version. While a Columbian study including children/adolescents (7-17.9-years) from the GenPop used the proxy version but yielded the same results (112). Despite these studies not exploring the reasons for the higher report of feeling WSU amongst females, in a European study using the KIDSCREEN-52 in children/adolescents aged 8-18-years, authors linked this observation to hormonal imbalances, females generally being more sensitive and concerned about their health and as a result, may be more susceptible to psychological troubles (151).

At a dimension level, there was no difference in the ranking by age, sex or health condition, but at a composite level, there were differences in the utility scores between those with and without a health condition between both versions, by which children with a health condition reported poorer HRQoL. Similarly, when compared to the GenPop of the same age and sex but without a health condition, children with acute lymphoblastic leukaemia reported more problems on the EQ-5D-Y-3L and overall

lower VAS scores, which was expected (9). Of note, the difference between those with a respiratory illness and the GenPop was only recorded on the IA therefore suggesting children with functional disabilities and acute orthopaedic conditions did not associate their conditions with a poorer HRQoL. The Mob dimension followed a similar trend as suggested by previous studies with the acute orthopaedic and functional disability groups reporting the most problems (16,17,126). This was an expected difference and could not be attributed to any single factor but likely multi-factorial with a difference in reporting of health with improved understanding on the IA and/or observation and social-desirability bias affecting their level of report. It was unlikely to be influenced by the use of the Slovenian utility index as it was consistently used across the versions for comparison of composite performance. Repeated analysis was done using the Japanese value set which showed to have no effect on the results in any way. At the time of data analysis only the Slovenian and Japanese value sets were published.

3.3.4. Concurrent validity

In this study, concurrent validity between the utility and VAS scores were significant for both versions (p=<0.001) but ranged from low to moderate associations (r=0.27-0.38). One would expect that the dimensions on the EQ-5D-Y-3L would account for the measure of general health as scored on the VAS and there would be no difference between the IA and SC descriptive systems. The association between the scores was lower in this study than a previous comparison between the VAS and composite score in children with acute illness (r=-0.786, p=<0.001). Composite scores are a summary of the EQ-5D-Y-3L dimensions using QALY weightings as suggested by Craig et al. (2016), therefore provides a total score for all five dimensions (152). To note though that Scott et al. (2017) did not find any association between the composite score and VAS in children with chronic illness or the GenPop. As this study analysed a heterogenous group of children including those with acute and chronic illness and from the GenPop it could account for the lower correlation. This could be due to the disability paradox reported in previous studies where children with chronic health conditions, such as CF and functional disabilities, found that those children did not necessarily report poorer HRQoL as one would have expected, as children with long-term conditions often find ways to adapt to their environment or the manner in which they complete a task so that it suits their abilities (17,29,30). Importantly, there was no difference between the scores on the IA and SC versions.

3.3.5. Convergent validity

Convergent validity between the IA and SC physical dimensions of Mob, LAM and UA were highly associated between versions compared to the psychosocial dimensions of P/D and WSU which showed lower significant associations. Correlations between versions were expected as both assessed the same dimensions with three levels of the report; however, as mentioned earlier, social-desirability bias may account for the low association in psychosocial dimensions (147).

The physical dimensions of Mob and LAM were previously compared to the WeeFIM items of mobility and self-care whereby, Spearman's correlation was high (r_s=-0.60) in all condition groups including chronically ill, acutely ill and children with functional disabilities (16). Although, in this study, low to moderate associations between instruments were seen in the Mob dimensions with the IA version showing higher correlations. LAM also showed higher associations with the WeeFIM items of self-care on the IA version. This may suggest that children were more aware of their physical ability when interviewed or may have not understood the question when self-completing and as a result answered incorrectly in relation to their functional ability.

As this is the first study to use the MFQ as a comparison to the WSU dimension on either version of the EQ-5D-Y-3L, comparisons to other studies were unfortunately not possible. However, in this study, the SC showed slightly stronger associations between WSU and the MFQ compared to the IA version, with both versions showing moderate associations with the MFQ total. Previous studies have tested convergent validity of the WSU dimension against psychosocial dimensions on the generic HRQoL instruments such as the KIDSCREEN, PedsQL and CHU-9D and found strong associations between instruments (2,84). The WSU dimension was also tested against a disease-specific instrument, the JAMAR to assess the association between feeling WSU and feeling anxious/nervous which, to the authors surprise, showed no association and attributed this to the difference in words used and not understanding the concept of anxiety/nervousness compared to feeling WSU (13). It may be helpful to divide the WSU dimension by separating the terms 'sad/unhappy' which may be more associated with mental wellbeing while, 'worried' may relate more to anxiety. This change could allow for a wider range of information to be obtained and for more direct comparisons between instruments, therefore could be beneficial to explore with further research.

The P/D dimension has previously been compared to the FPS-R which showed a significant correlation between instruments for acutely ill children only (p=<0.001) (16). Similarly, significant and moderate

associations were found between the FPS-R and the IA (r_s=0.33, p <0.001) and SC versions (r_s=0.38, p<0.001) with no significant difference between versions (p=0.281). As a result, this may suggest that the P/D dimension was accurately able to reflect children's experience of feeling P/D using either version.

Assessing psychosocial dimensions remains a challenge due to its subjectivity compared to dimensions such as Mob, LAM and UA which may be objectively observed (153), therefore physical dimensions were expected to present with better convergent validity between instruments than psychosocial dimensions.

3.3.6. Preference of version

In this sample, the IA version was highly preferred (60%) compared to the SC version (37%) with literacy levels largely influencing the reason for their preference (22%) especially in 8-year-olds (25%). While 27% of children preferred the SC version as it gave them a sense of independence by being able to complete the instrument by themselves. However, a preference for interviewer-based completion has been expressed in a South African context which was not surprising, as reading ability and overall literacy levels amongst 8-year-olds are both below 50% (146), therefore allowing us to understand why so few reported this sense of independence. Ultimately, reading difficulties have the potential to exclude children from self-completing (154) if alternative modes are not available even though the SC is valid and reliable in children from 8-years (2,3,9). The implementation of the IA version in populations with lower literacy levels is recommended in order to record their HRQoL accurately and not exclude them based on literacy levels, and subsequently socioeconomic status (5).

Additional reasons for preferring the IA may be associated with acquiescence bias (155) which is mostly associated with interviewer-based instruments rather than SC versions. This may be as a result of participants finding it easier to respond with a positive response option rather than considering their own 'true' preference or the simplest answer, (140) or the first answer presented to them (62). In the context of the EQ-5D-Y-3L, this would translate to reporting level one (no problems) for all dimensions thus, presenting with a better HRQoL on the IA version than SC version. This was evident in the comparison between interviewer-based and self-administered versions of a disease-specific instrument developed for asthma, which found similar results with a higher HRQoL being reported when using interviewer-based versions (156). Conversely, a longitudinal study assessing HRQoL in adults with Acquired Immunodeficiency Syndrome, across clinics in the United States, did not find a

meaningful effect between modes of administration which was converted to a quantitative measure using adjusted differences in standard deviation of ≥0.2 (157). In addition, a study measuring HRQoL, in relation to oral health, in children aged 9-16-years also found the IA and SC versions of the same instrument to perform similarly when compared (154). Similarly in this study, there was an equal preference of 50% between versions n the respiratory group.

3.4. Conclusion

The IA proved to be valid and performed as well as the SC version in children aged 8-10-years with no significant differences in reporting across versions. The feasibility of the instrument is supported by the fewer missing responses with the IA compared to the SC. The burden of IA, with regards to the need of an interviewer in a clinical setting (148,154), may outweigh the benefit of reduction in missing responses. The results of the SC and IA versions were comparable which would further allow researchers to use the results interchangeably in a study and select a version most appropriate to the child's literacy level and/or medical condition or based on the child's preference, therefore also granting the child a sense of autonomy.

Due to the literacy levels in South Africa, children fall within the recommended age-range of 8-years and older for self-completion, but it should not be assumed that these children have the ability to self-complete despite their age and understanding of health. Each child should be assessed individually to determine which mode of administration is most appropriate, especially in 8-year-olds who proved, in this setting, to struggle the most with self-completing.

Further studies are recommended to assess whether the social-desirability bias significantly impacts the reporting of WSU and P/D in children with conditions that are hypothesised to impact these dimensions i.e. children experiencing anxiety and/or depression and children with acute pain. Additional research to assess response bias is also recommended by determining the effect of order of questionnaires on responses.

4. Comparison of the Performance of the EQ-5D-Y-3L-IA in children aged 5-7-years and 8-10-years

4.1. Methodology

4.1.1. Introduction

This chapter provides a detailed description of the methods implemented to carry out the research project comparing the performance of the EQ-5D-Y-3L-IA in children aged 5-7-years compared to those aged 8-10-years with a known medical condition and from the GenPop. This information was compared to the FPS-R, MFQ and the WeeFIM, all of which measure similar constructs to those measured in each dimension on the IA. Longitudinal data, for test-retest reliability, was collected 48 hours after initial testing from a smaller sample of children with stable health conditions including from the GenPop attending mainstream schools or who have a functional disability and are attending a LSEN school.

The procedure followed that described in 3.1 with differences in methodology highlighted below.

4.1.2. Sample size

The sample size was adequately powered to detect a difference in correlations between 5-7-year-olds and 8-10-year-olds with a small effect size 0.2, slightly smaller than previous South African studies with an effect size of 0.3 and 0.4 (16,126). A sample of 177 children was needed for a power of 85% and a significance of 0.05. The difference in power calculations were done to accommodate for the difference in sample sizes between the two age-groups."

4.1.3. Instruments

The **EQ-5D-Y-3L-IA** (Appendix 13) as described in 3.1.5.

The **Faces Pain Scale-Revised (FPS-R)** (Appendix 14) as described in 3.1.5.

The **Moods and Feelings Questionnaire** (**MFQ)** (Appendix 15) as described in 3.1.5.

The **WeeFIM** (Appendix 16) as described in 3.1.5.

Study specific medical and demographic questionnaire (Appendix 17) as described in 3.1.5.

Cognitive debriefing template (Appendix 28) was used as a guide for the interviews with the children after completion of the IA to determine the comprehensibility of the instrument. The structured script allowed the student researcher to probe the child into the reason behind their answer for each of the dimension scores and the VAS scores '*why did you say you have a lot of problems with mobility?*'. If there were any apparent inconsistencies this was also probed e.g. Level 3 (a lot of problems) on Mob but level 1 (no problems) on UA or Level 1 (no problems) or Mob but apparently confined to bed. The cognitive debriefing further aimed to identify any potentially difficult or confusing words (158).

Interviewer Questionnaire (Appendix 29) recorded the interviewer's opinion on whether the child understood the IA or if any indecisiveness was noted. This was done through a series of questions (Appendix 18) and recording whether there was indecisiveness or clarification of questions and/or terms needed on the IA.

4.1.4. Procedure

Necessary approvals were granted by the Faculty of Health Sciences, Human Research Ethics Committee (HREC), University of Cape Town (UCT) (HREC 369/2020) (Appendix 19), ministerial permission for non-therapeutic research with minors (Form A) (Appendix 20), Western Cape Education Department (Appendix 21), the respective school principals (Appendix 22) and the children's hospital management (Appendix 23). Children who fulfilled the inclusion criteria were recruited from schools and healthcare institutions.

Envelopes containing study information, an informed consent form (Appendix 24) and a demographic questionnaire (Appendix 17) were distributed to participating schools for learners to take home. Parents/legal guardians were invited to return the signed consent form and demographic questionnaire if they agreed to their child participating in the study. All children whose parents/legal guardians granted consent, were invited individually to a private room where they were given a detailed description of the study. Assent (Appendix 25) was obtained from those willing to participate.

The research packs, including the EQ-5D-Y-3L-IA (timed), cognitive debriefing questionnaire, FPS-R, MFQ and WeeFIM to minimize question order bias.

Children admitted to the orthopaedic hospital were recruited after admission. Parents/legal guardians who were not at the bedside, were contacted telephonically to describe the study and to ask for consent (Appendix 26). The procedure for data collection was as described for children attending school.

Children attending the sub-specialty outpatient clinics for either orthopaedics or chronic respiratory illness, were screened for inclusion according to the date of birth in the clinic diary and recruited systematically in the order of arrival. The eligible children and their parent/legal guardian were approached in the waiting room and invited to participate in the study by means of a brief verbal description of the study accompanied by an informed consent form. If consent was granted, the parent/legal guardian and their child were invited to sit in a private consultation room where assent was obtained, followed by the completion of the research packs. If the parent/legal guardian chose to accompany their child during the completion of the research packs, they were invited to sit slightly behind their child and asked not to influence their answers verbally or non-verbally. Children did not lose their place in the queue to see the medical professional nor were other medical investigations put on hold due to their participation. If the research pack was not completed by the time they were called in for their consultation with the medical professional, the research was stopped, and the child and parent/legal guardian were invited to complete the study after their consultation.

All screening, enrolment and data collection was done by the student researcher to ensure standardization and to minimize bias.

Participating schools received an education hamper to the value of R2000 for their assistance in identifying learners who met the inclusion criteria, communication with parents/legal guardians, handling of envelopes and arranging appropriate times for children to leave the classroom during school hours to conduct the interviews. These schools received the educational hamper regardless of how many children from their facility were enrolled. The participants nor their parents/legal guardians were reimbursed as they did not incur any costs to participate in the research.

Children with a stable functional disability and those from the GenPop completed the IA again after 48 hours to determine test-retest reliability.

4.1.5. Statistical analysis

The Shapiro-Wilk test was used to test the normality of the data. Level of statistical significance was set at $p < 0.05$.

4.1.5.1. General instrument performance and feasibility

The IA responses and descriptive data were summarised in terms of frequency of responses. The feasibility was assessed by comparing the number of missing values for two age-groups. The ceiling effect of the IA was defined as the proportion of children scoring no problems in all five dimensions (11111) or for each individual dimension. Higher ceiling effects were expected in children from the GenPop but differences between age-groups were not expected to be significant (2,9,14,16,87). The floor effect is the proportion of children scoring a lot of problems for all five dimensions (33333) or for each individual dimension. The number of unique health states (as recorded by the 5-digit code recorded from the IA) was computed across age-groups and condition groups. Differences in reporting was determined by chi-square statistic (x^2). The median time taken to complete the IA version was compared with the Mann-Whitney U-test. It was expected that the 5-7-year-old group would take longer to complete the measure than the 8-10-year-olds due to the time needed to comprehend questions (159).

4.1.5.2. Known-group validity

The frequency of responses was compared across health conditions (orthopaedic, respiratory illness, functional disability and GenPop) for the two age-groups and compared with Chi-square statistics (x^2). The dimension responses, utility scores and VAS scores were assessed with Spearman rank order coefficients (r_s) across age (continuous variable) and for the two age-groups by health condition. Furthermore, median utility scores and VAS scores were compared with Kruskall-Wallis H-test (H) across health condition for those aged 5-7-years-old and 8-10-years-old. No differences were expected across age-groups. Effect sizes were interpreted according to Cohen's d interpretation whereby 0.2 = small, 0.5 = medium, 0.8 = large and 1.3 = very large (137).

4.1.5.3. Concurrent validity

The Pearson's correlation of the utility score and VAS score was computed for the age-groups and compared using the Fisher r-to-z transformation (http://vassarstats.net) (138). More problems were expected in the 5-7-year group compared to the 8-10-year group (3,10,87).

4.1.5.4. Convergent validity

The convergent validity of the dimension scores of the IA across the age-groups were compared to the corresponding scores from the WeeFIM, FSP-R and MFQ using Spearman correlations (r_s). Correlation coefficients were compared between age-groups using the Fisher r-to-z transformation (http://vassarstats.net) (138). Spearman's rank order and Pearson's correlations coefficients were interpreted according to Cohen: 0.1–0.29 low association, 0.3–0.49 moderate association and ≥0.5 high association (139). It was expected that there would be no differences in age-groups and similar dimensions would show stronger correlations (2,9,16,87).

4.1.5.5. Test-retest reliability

Test-retest reliability was assessed using weighted Cohen's kappa statistic (k) for dimension scores and the ICC for utility and VAS scores across the two age-groups. Kappa values were interpreted according to Landis and Koch's guidelines: <0.2 poor agreement, 0.21–0.40 fair agreement, 0.41–0.60 moderate agreement, 0.61–0.80 substantial agreement and 0.81-1.00 almost perfect (160). An ICC of >0.7 was considered reliable (161). It was expected that there would be no significant differences in age-groups (2,9,16).

4.1.5.6. Cognitive debriefing

Qualitative data collected from participants regarding reasons for level reported, understanding and inconsistencies were grouped according to similar responses per age-group and tabulated. Similarly, the responses from the interviewer questionnaire were also grouped according to similar reasons per age-group and tabulated. More inconsistencies were expected in the 5-7-year group due to level of understanding. It was expected that the 5-7-year-olds would report more difficulties in answering the IA.

All data analyses were conducted using SPSS Windows 27.0 (IBM SPSS Inc., Chicago, IL, USA) and Statistica Windows Version 13.0 (TIBCO Software Inc., Palo Alto, CA, USA).

4.3. Results

The recruitment of children aged 5-7-years and those aged 8-10-years is shown in Figure 12. The same group of 8-10-year-olds included in Chapter 3 was included in Chapter 4. There was a high proportion of non-responders in the 8-10-year-olds (n=207, 64%) than 5-7-year-olds (n=78, 44%) due to the high number of parents/legal guardians from GenPop and special schools not returning consent forms. Reasons for refusing participation were not recorded. There were more children with orthopaedic problems who refused consent or assent in the 5-7-year-old group (n=33, 29%) than the 8-10-year-old group (n=12, 10%). More 8-10-year-olds (n=21, 20%) than 5-7-year-olds (n=11, 12%) withdrew. All participants who withdrew, did so for personal reasons, multiple medical appointments, time constraints and transport issues which include scheduled patient transport to/from healthcare facilities to patient's residential area which may be located in a different city within the Western Cape province, price of public transport and fear of missing public transport which do not follow fixed schedules.

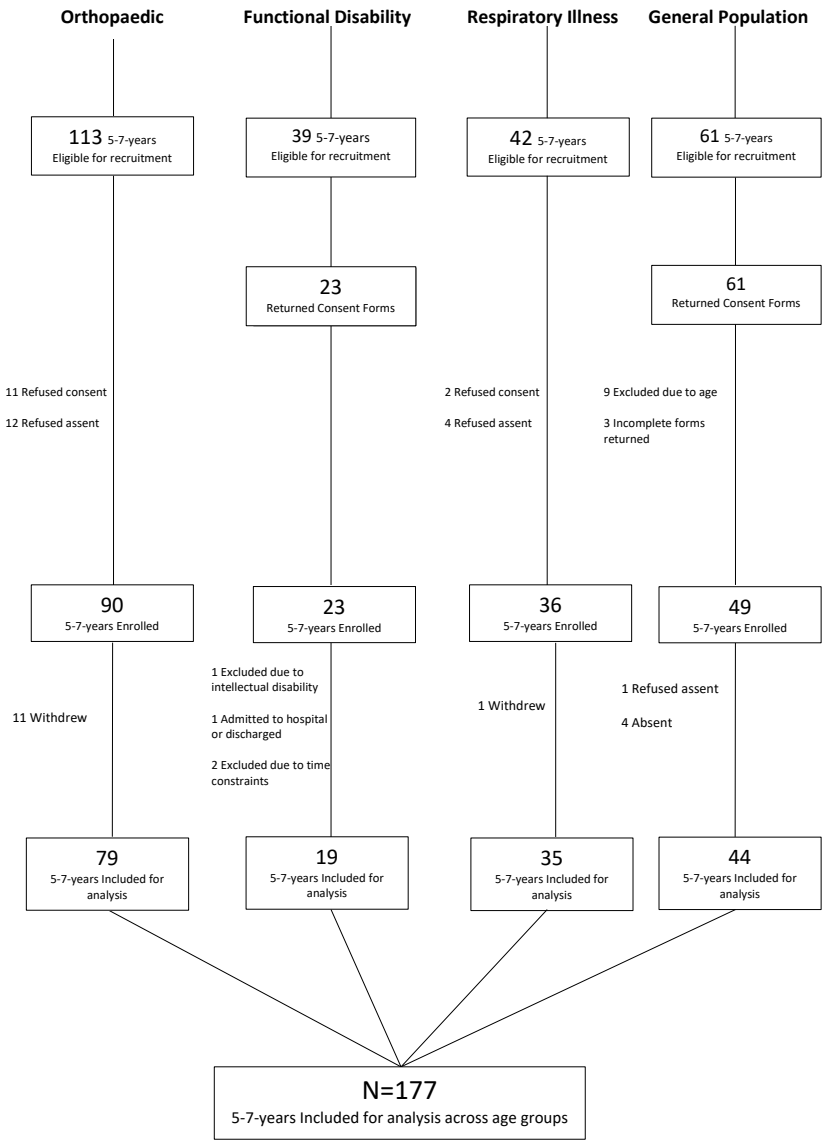


Figure 12: Recruitment of children aged 5-7-years and 8-10-years

4.2.1. Descriptive statistics

A total of 388 children were recruited across 5-7-years (n=177, 46%) and 8-10-years (n=211, 54%). There was no difference between sex (x^2=2.34, p=0.126) or health condition (x^2=7.21, p=0.065) across the age-groups. In the 5-7-year-old group there was a higher number of children with an orthopaedic condition (n=79, 45%), followed by GenPop (n=44, 25%), respiratory illness (n=35, n=20%) and functional disabilities (n=19, 11%). Similarly, the 8-10-year group had the highest proportion of children with an orthopaedic condition (n=82, 39%) followed by GenPop (n=65, 31%), functional disabilities (n=36, 17%) and respiratory illness (n=28, 13%). The disease groups under these conditions are shown in Table 20.

Table 20: Descriptive statistics of participants across age-groups (5-7-years and 8-10-years)

	Age Category			
	5-7-years		8-10-years	
	n	%	n	%
Sex	**(n=177)**		**(n=211)**	
Female	96	54%	98	46%
Male	81	46%	113	54%
Orthopaedic	**(n=79)**		**(n=82)**	
Upper Limb Fracture	34	43%	31	38%
Lower Limb Fracture	24	30%	21	26%
Surgical correction of acquired or congenital orthopaedic condition[#]	17	22%	21	26%
Other*	4	5%	9	11%
Functional Disabilities	**(n=19)**		**(n=36)**	
Cerebral Palsy	8	42%	6	17%
Spina Bifida	3	16%	5	14%
Development Co-ordination Disorderˆ	8	42%	23	63%
Developmental Delay	0	0%	2	6%
Other[¥]	5	26%	11	29%
Respiratory	**(n=35)**		**(n=28)**	
Atopy	6	17%	12	43%
Cystic Fibrosis	13	37%	5	18%
Bronchiectasis	4	11%	3	11%
Acute Respiratory Illness	3	9%	0	0%
Other[¥]	9	26%	8	29%
GenPop	**(n=44)**		**(n=65)**	
None	40	91%	54	83%
Atopy	3	7%	9	14%
Other[§]	1	2%	2	3%

*[#]Includes Blount's disease, osteogenesis imperfecta, developmental dysplasia of the hip, leg-length discrepancy and spinal deformity. *Includes osteitis; septic arthritis and a traumatic amputation. Includes learning disability and Human Immunodeficiency Virus. [¥]Includes damage to the lungs post-acute viral infection, congenital abnormalities of the respiratory system and idiopathic pulmonary haemorrhage. [§]Includes Osteogenesis imperfecta and a congenital cardiac defect.*

4.2.2. General instrument performance

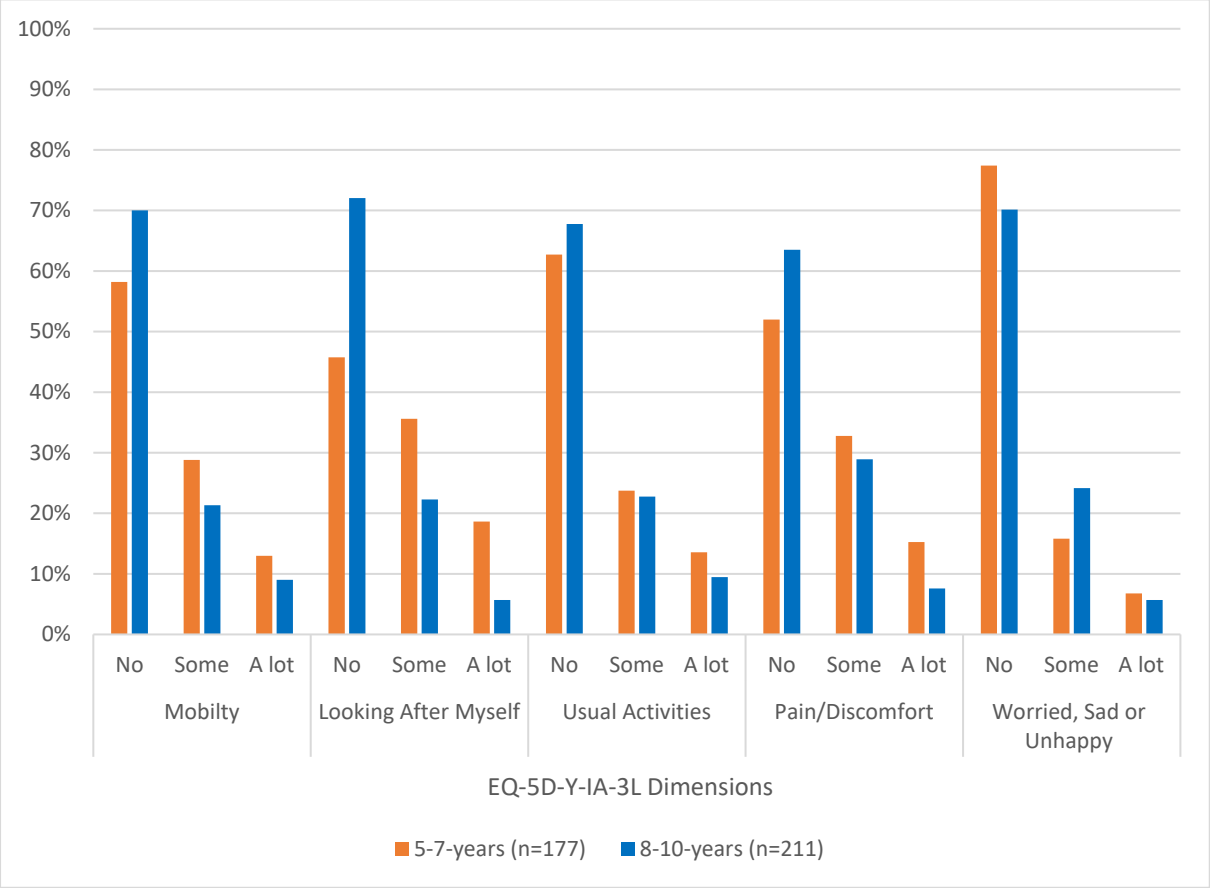


Figure 13: Comparison of the EQ-5D-Y-3L-IA dimensions across age-groups

As seen in Figure 13, there was a higher percentage of no problems for Mob for those aged 8-10-years compared to 5-7-year-olds, however, there was no significant difference between groups (x^2=5.563, p=0.062). There were significantly higher reports of some problems and a lot of problems in LAM in the 5-7-year-olds compared to the 8-10-year-olds (x^2=31.021, p<.0001). There were also significantly higher reports of some and a lot of problems in P/D across age-groups (x^2=7.775, p=0.020) with the 33% of 5-7-years reporting some P/D and 15% reporting a lot of P/D compared to 29% of 8-10-year-olds reporting some P/D and 8% reporting a lot of P/D. There were no significant differences in UA (x^2=1.830, p=0.401), and WSU (x^2=4.173, p=0.124), across age-groups.

4.2.3. Feasibility

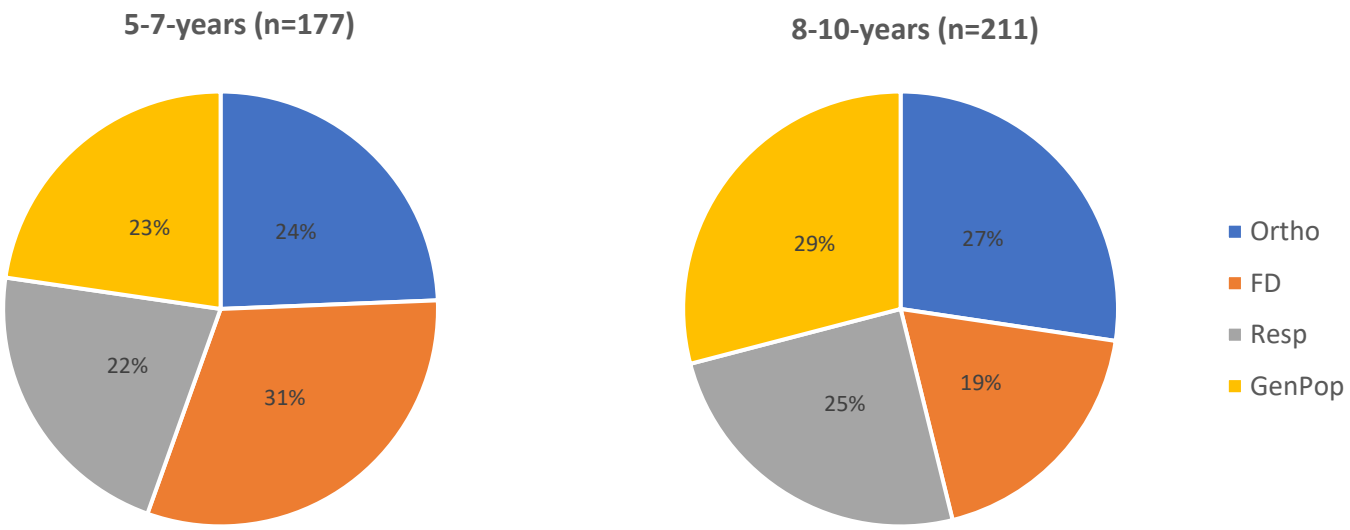


Figure 14: Ceiling effects in children aged 5-7-years and 8-10-years across health conditions

As seen in Figure 14 children with functional disabilities reported a higher proportion of no problems across all dimensions in the younger group however, there was no significant difference between the age-groups (x^2=1.8, p=0,615). There was no difference between ceiling effect across health condition in the 5-7-year-olds (x^2=0.82, p=0.845) nor the 8-10-year-olds (x^2=1.61, p=0.845). Considering the total ceiling effect by age-group, there was no difference between those aged 5-7-years (n=51, 29%) and 8-10-years (n=64, 30%) (x^2=0.05, p=0.82). Reports of floor effects were low with only one participant reporting 33333 in each age-group, respectively.

The total reporting of unique health states was significantly higher in the 8-10-year group (n=111, 53%) than the 5-7-year group (n=66, 37%) (x^2=8.5, p=0.004).

The frequency of time taken to complete the IA per age-group differed as the 5-7-year group took significantly longer to complete the measure (median= 134s, IQR= 118, 157) compared to the 8-10-year group (median= 110s, IQR= 98, 125) (Mann-Whitney U= 8389.5, p=<0.001).

4.2.4. Known-group validity

The known-group validity of the IA scores is shown in Table 21 across the ages and health conditions. When considering the age of the children, younger children with a health condition reported significantly more problems with Mob, LAM, P/D. Conversely, older children reported more problems with WSU, but this was not significant. Although UA did not show any significant differences across the ages, younger children tended to score less problems in this dimension. For the dimension of P/D and the VAS score, however, younger children reported more problems.

When considering health conditions within age-groups, children 5-7-years with orthopaedic conditions reported significantly more problems in all dimensions and on the utility score[7] compared to those with functional disabilities, respiratory illnesses and from the GenPop. In the 8-10-year group, children with orthopaedic conditions reported significantly more problems in Mob, LAM, UA and on the utility score[7] compared to those with functional disabilities, respiratory illnesses and from the GenPop.

Table 21: Known-group validity of EQ-5D-Y-3L-IA scores by age and health conditions with Spearman's rank order correlation

	Age (years)	**Health Condition**	
		5-7-years	**8-10-years**
		(n=177)	**(n=211)**
Mobility	-0.111*	-0.18*	-0.27*
LAM	-0.322*	-0.26*	-0.35*
UA	-0.055	-0.25*	-0.35*
P/D	-0.116*	-0.18*	-0.22*
WSU	0.061	-0.17*	-0.13
VAS score	-0.077	-0.07	0.08
Utility score	0.171*	0.32*	0.34*

*Age in years was computed as a continuous variable, *p<0.05. Mob=mobility, LAM=looking after myself, UA=usual activities, P/D= pain or discomfort, WSU=worried, sad or unhappy.*

[7] Analysis with the Slovenian utility score is presented. There was no significant difference between results using the Slovenian or Japanese utility scores.

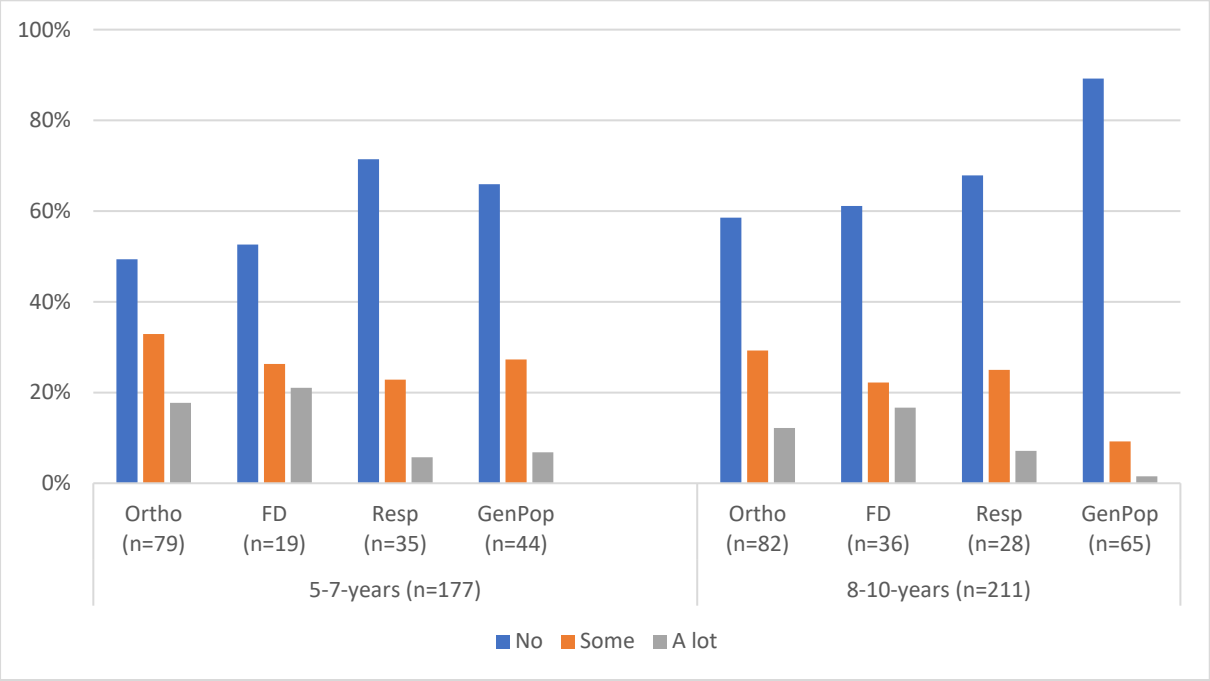


Figure 15: Comparison of EQ-5D-Y-3L-IA Mobility dimension across condition groups

As seen in Figure 15 there was a higher reporting of problems in the Mob dimension in the GenPop for children 5-7-year-old compared to the 8-10-years-olds (x^2= 8.95, p=0.011).

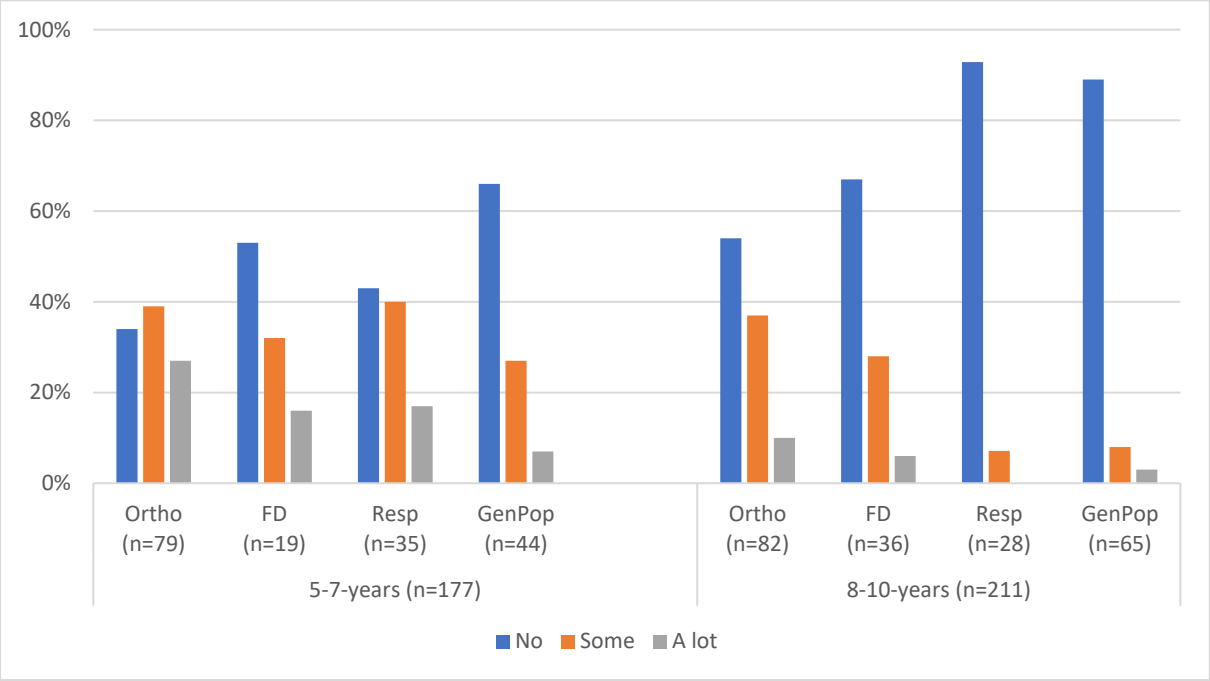


Figure 16: Comparison of EQ-5D-Y-3L-IA Looking After Myself dimension across condition groups

As seen in Figure 16 there was a higher reporting of problems in the LAM dimension in the GenPop and respiratory illness group for children 5-7-years compared to the 8-10-years-olds (x^2= 9.04, p=0.0109 and x^2= 17,39, p<0.001, respectively).

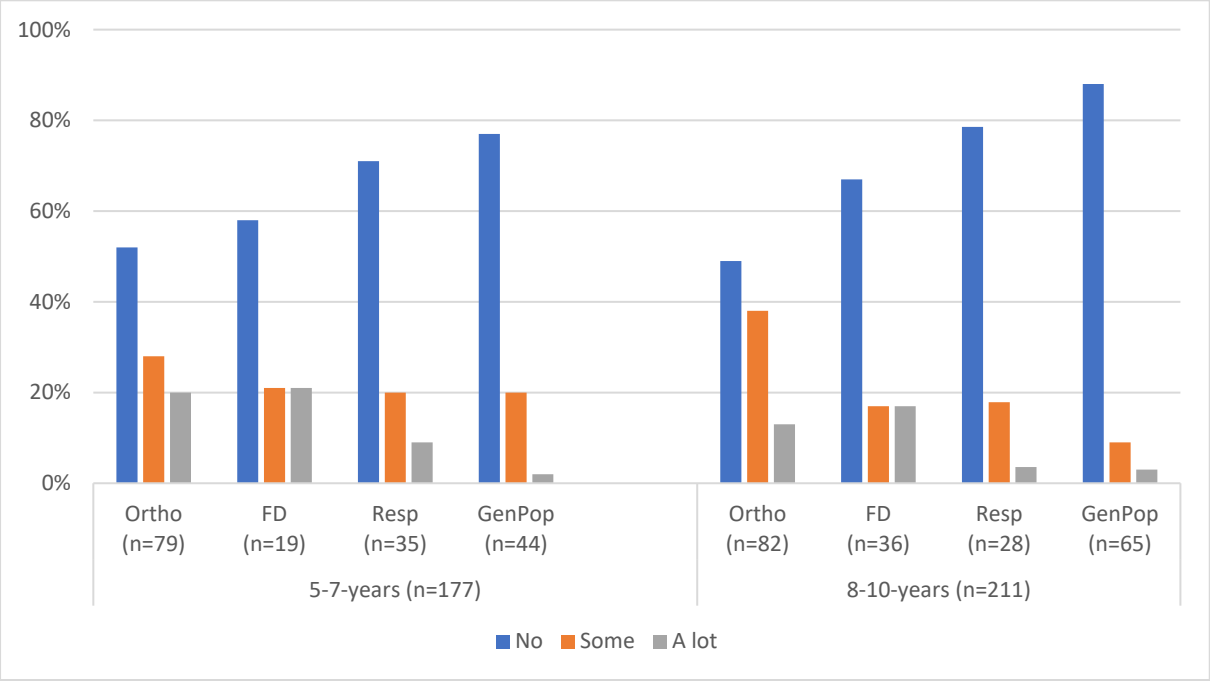


Figure 17: Comparison of EQ-5D-Y-3L-IA Usual Activities dimension across condition groups

Figure 17 shows that in the dimension of UA, there was no significant differences in reporting across age-groups and health conditions.

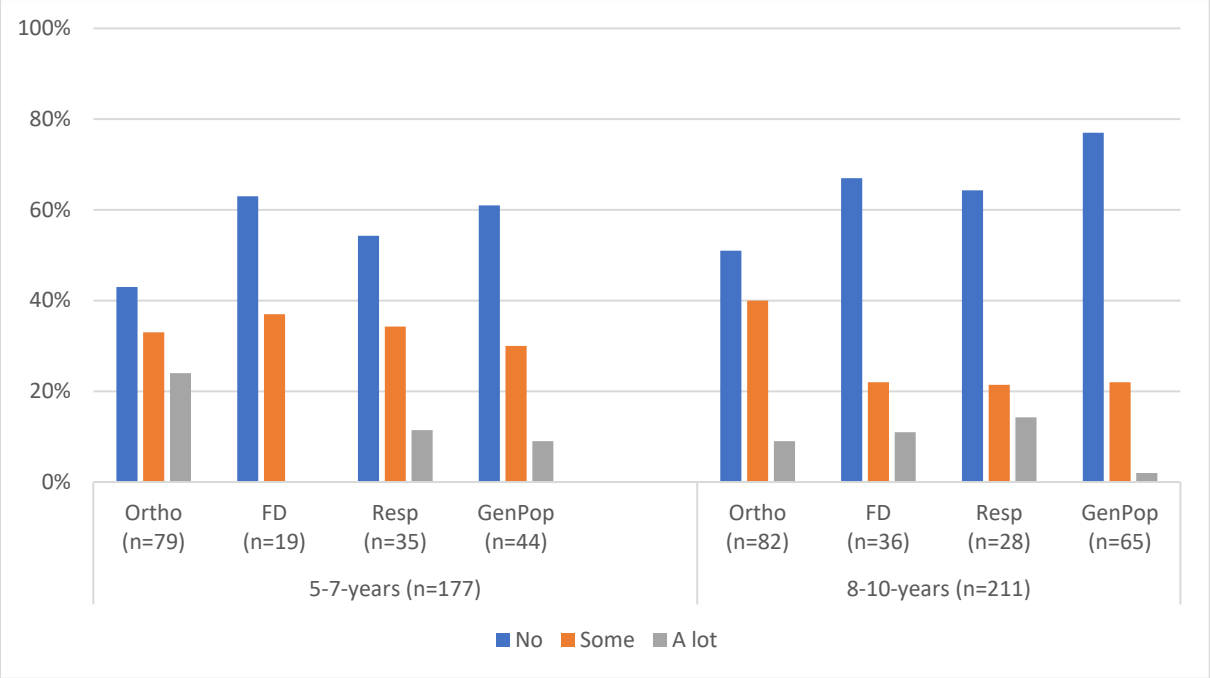


Figure 18: Comparison of EQ-5D-Y-3L-IA Pain/Discomfort dimension across condition groups

As seen Figure 18 there was a higher report of P/D in the orthopaedic group for children 5-7-years compared to 8-10-year-olds (x^2= 7.16, p=0.0279).

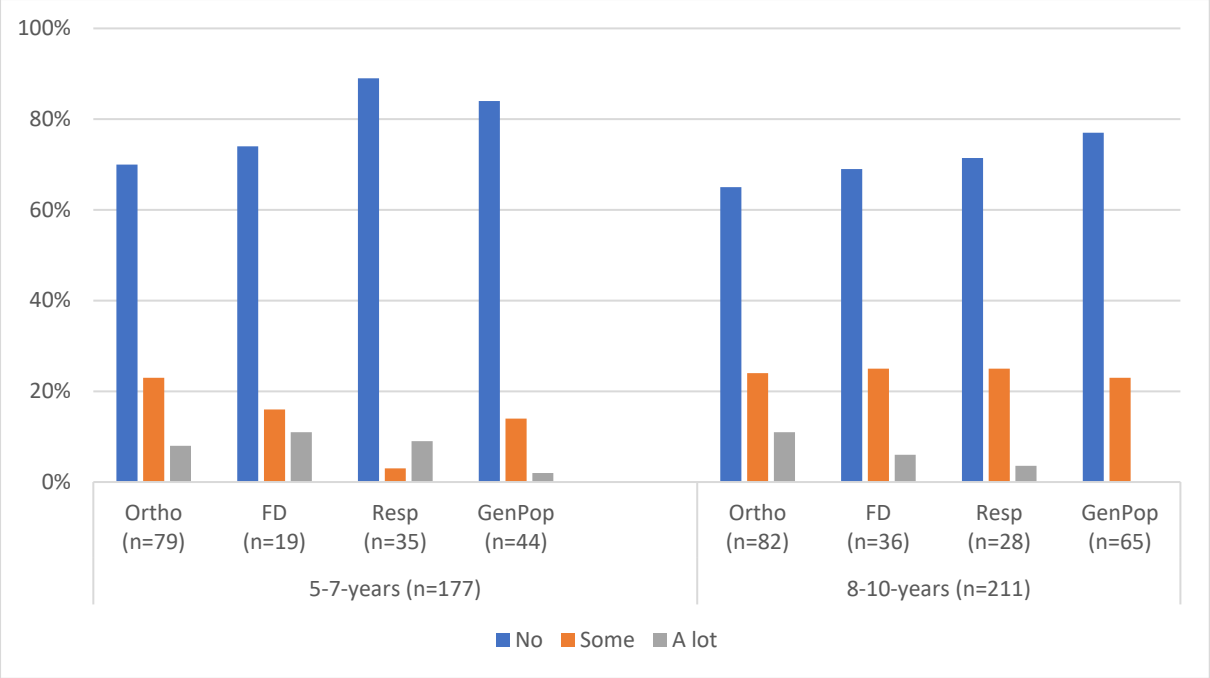
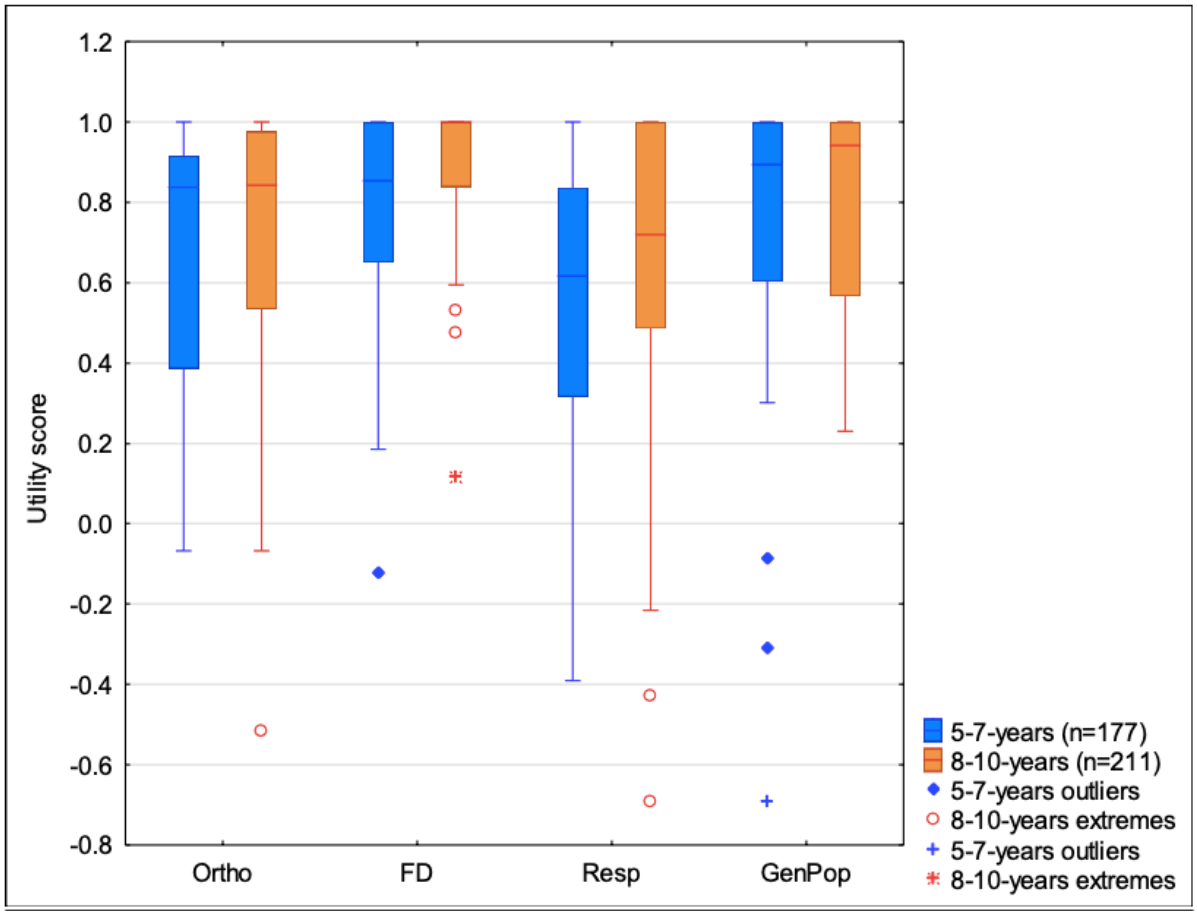


Figure 19: Comparison of EQ-5D-Y-3L-IA Worried, Sad or Unhappy dimension across condition groups

Figure 19 shows that for the dimension of WSU there was a higher report of feeling WSU in the chronic respiratory group for children 8-10-years compared to 5-7-year-olds (x^2= 7.18, p=0.0276).

In summary, Figures 16-20 show that those with orthopaedic conditions and respiratory illnesses reported more problems with LAM than those with functional disabilities and from the GenPop. Whereas, for UA, there was a higher reporting of problems in the functional disability and orthopaedic groups. All dimensions were significantly different for the older group except the WSU dimension.



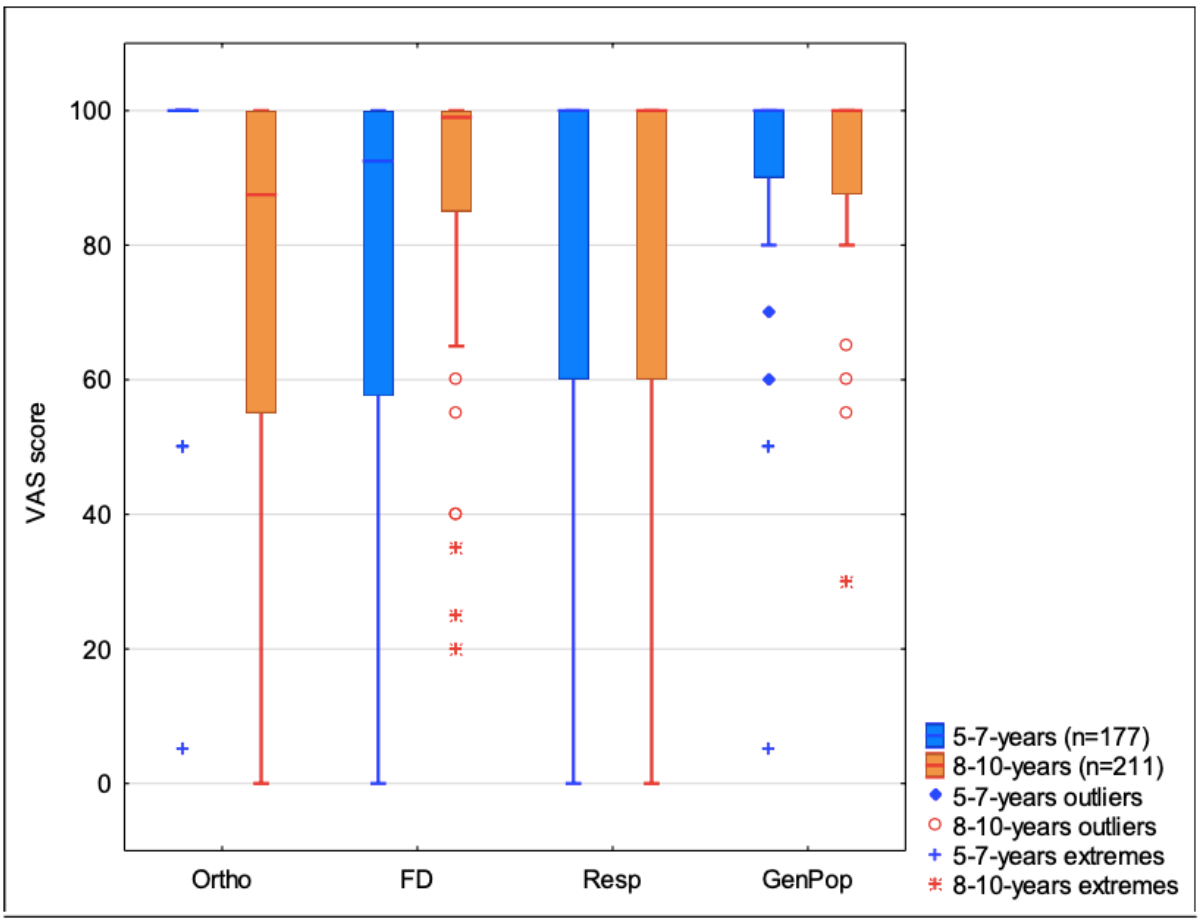
5-7-years: Orthopaedic (Ortho) (n=79), Functional Disability (FD) (n=19), Respiratory illness (Resp) (n=35) and General Population (GenPop) (n=44). 8-10-years Orthopaedic (Ortho) (n=82), Functional Disability (FD) (n=36), Respiratory illness (Resp) (n=28) and General Population (GenPop) (n=65)

Figure 20: EQ-5D-Y-3L-IA utility scores in 5-7-year-olds and 8-10-year-olds

As seen in Figure 20, there was a significant difference in utility scores[8] between health conditions in the 5-7-year group (H=18.57, p<0.001). Post-hoc analysis in the 5-7-year-olds show that there was a significant difference in the utility scores[8] for the orthopaedic group compared to the respiratory group (H=-31.64, p=0.013) and GenPop (H=37.16, p=0.001).

There were similarly significant differences between the utility score[8] across health condition in the 8-10-year-olds (H=24.89, p<0.001). Post-hoc analysis showed that there was a significant difference between GenPop and those with functional disabilities (H=-37.91, p=0.013) and the orthopaedic group (H=47.61, p<0.001).

[8] Analysis with the Slovenian utility score is presented. There was no significant difference between results using the Slovenian or Japanese utility scores.



5-7-years: Orthopaedic (Ortho) (n=79), Functional Disability (FD) (n=19), Respiratory illness (Resp) (n=35) and General Population (GenPop) (n=44). 8-10-years Orthopaedic (Ortho) (n=82), Functional Disability (FD) (n=36), Respiratory illness (Resp) (n=28) and General Population (GenPop) (n=65)

Figure 21: EQ-5D-Y-3L-IA VAS scores in 5-7-year-olds and 8-10-year-olds

As seen in Figure 21, in the 5-7-year group there was a significant difference in VAS score across health condition (H=8.032, p=0.045). Although post-hoc analysis shows that there was no significant difference between health conditions. The GenPop had the lowest median (IQR) VAS score [93 (58,100)], indicating a worse HRQoL.

There was no significant difference between any of the health conditions in children aged 8-10-years (H=6.549, p=0.088) although those with functional disabilities reported the lowest median (IQR) VAS score [88 (55,100)].

4.2.5. Concurrent validity

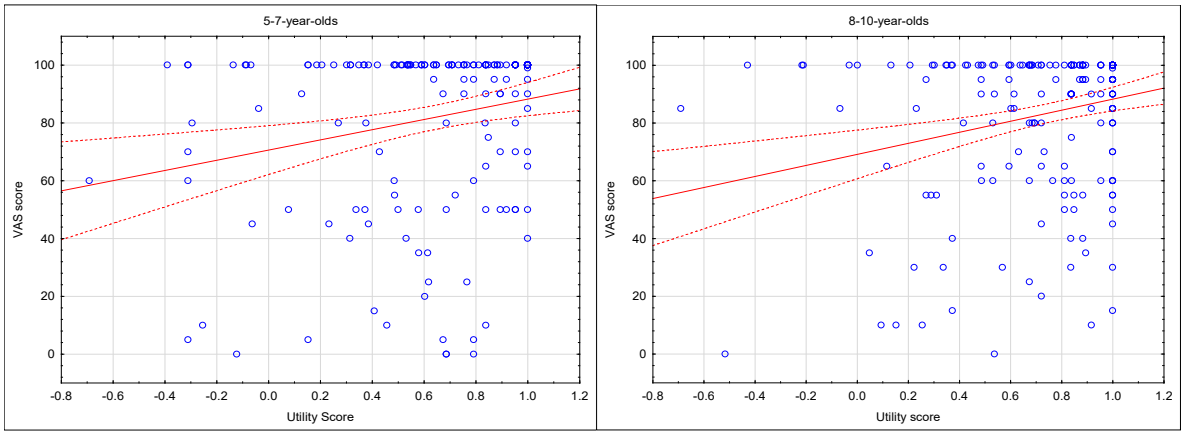


Figure 22: Scatterplot of EQ-5D-Y-3L-IA utility scores versus VAS scores for 5-7-year-olds and 8-10-year-olds

Figure 22 shows that the Pearson's correlation of the VAS and utility scores[9] for the 5-7-year-olds (r=0.225, p=0.003) was similar to the 8-10-year-olds (r=0.246, p<0.001). There was no significant difference between these correlations (z=-0.22, p=0.413).

4.2.6. Convergent validity

Table 22 shows that Mob had a moderate to high correlation with WeeFIM items of mobility as well as the mobility and motor total scores. There was however, a significantly higher correlation with toilet transfers and sit to stand transfer in the 5-7-year-olds than the 8-10-year-olds.

Table 22: Convergent validity of EQ-5D-Y-3L-IA Mobility dimension and WeeFIM Mobility

	Mobility			
	5-7-years	8-10-years	5-7 vs 8-10-years	
WeeFIM Mobility	(n=177)	(n=211)	z-score	p-value
Sit to stand transfer	-0.48**	-0.32**	-1.86	0.031
Toilet transfer	-0.49**	-0.35**	-1.66	0.049
Tub or shower transfer	-0.38**	-0.38**	0.00	0.500
Locomotion (walk/wheelchair for ≥45m or crawl ≥15m)	-0.51**	-0.43**	-1.03	0.151
Stairs climbing (ascend/descend 12-14 stairs)	-0.43**	-0.47**	0.49	0.312
Mobility Total	-0.38**	-0.47**	1.07	0.142

**Spearman's correlation p<0.05, **Spearman's correlation p<0.001, A higher WeeFIM score indicates greater independence, a higher EQ-5D-Y-3L-IA score indicates greater problems. Shaded cells indicate comparison of correlations between 5-7-year-olds and 8-10-year-olds.*

[9] Analysis with the Slovenian utility score is presented. There was no significant difference between results using the Slovenian or Japanese utility scores.

Table 23 shows that LAM had moderate to high association with items of bathing, and dressing the upper and lower body, self-care total and motor total scores in both age-groups. Associations with dressing upper and lower body and the self-care total were significantly higher in the 8-10-year-olds compared to the 5-7-year-olds.

Table 23: Convergent validity of EQ-5D-Y-3L-IA Looking After Myself dimension and WeeFIM Self-care

	Looking After Myself			
	5-7-years	8-10-years	5-7 vs 8-10-years	
WeeFIM Self-care	(n=177)	(n=211)	z-score	(n=177)
Grooming	-0.27**	-0.38**	1.20	0.115
Bathing (washing body excluding back)	-0.61**	-0.69**	1.35	0.089
Dressing Upper Body	-0.42**	-0.59**	2.24	0.013
Dressing Lower Body	-0.40**	-0.61**	2.78	0.003
Toileting (including perineal hygiene and adjusting clothing before and after toilet use)	-0.20**	-0.30**	1.04	0.149
Bladder (continence)	-0.134	-0.14*	0.06	0.476
Bowel (continence)	-0.13	-0.16*	0.30	0.382
Self-Care Total	-0.54**	-0.67**	2.01	0.022

**Spearman's correlation p<0.05, **Spearman's correlation p<0.001, A higher WeeFIM score indicates greater independence, a higher EQ-5D-Y-3L-IA score indicates greater problems. The WeeFIM item of eating was excluded from analysis as all children aged 8-10 years scored total independence with eating and there was thus no variation. Shaded cells indicate comparison of correlations between 5-7-year-olds and 8-10-year-olds.*

Table 24 shows that UA was not associated with WeeFIM social interaction. There was however, a low to moderate association between WeeFIM items of mobility and the total scores. The association with transfers was significantly higher in children aged 5-7-years than those aged 8-10-years.

Table 24: Convergent validity of EQ-5D-Y-3L-IA Usual Activities dimension and WeeFIM Mobility and Social Interaction

	Usual Activities			
	5-7-years	8-10-years	5-7 vs 8-10-years	
WeeFIM Mobility	(n=177)	(n=211)	z-score	(n=177)
Sit to Stand Transfer	-0.35**	-0.19**	-1.68	0.047
Toilet transfer	-0.34**	-0.22**	-1.27	0.102
Tub or shower transfer	-0.24**	-0.27**	0.31	0.378
Locomotion (walk/wheelchair ≥45m or crawl ≥15m)	-0.35**	-0.25**	-1.07	0.142
Stairs climbing (ascend and descend 12-14 stairs)	-0.37**	-0.28**	-0.98	0.164
Mobility Total	-0.28**	-0.33**	0.54	0.295
Motor Total[§]	-0.36**	-0.39**	0.34	0.369
WeeFIM Cognition				
Social Interaction (interaction with other children)	-0.01	0.05	-0.58	0.281

**Spearman's correlation p<0.05, **Spearman's correlation p<0.001, † §Motor Total = Mobility Total + Self-care Total. A higher WeeFIM score indicates greater independent, a higher EQ-5D-Y-3L-IA score indicates greater problems. Shaded cells indicate comparison of correlations between 5-7-year-olds and 8-10-year-olds.*

The FPS-R and P/D showed moderate association with no difference between age-groups (Table 25).

Table 25: Convergent validity of EQ-5D-Y-3L-IA Pain/Discomfort dimension and Faces-Pain Scale-Revised

	Pain/Discomfort			
	5-7-years	8-10-years	5-7 vs 8-10-years	
	(n=177)	(n=211)	z-score	p-value
Faces Pain Scale-Revised	0.48**	0.38**	1.20	0.115

**Spearman's correlation p<0.05, **Spearman's correlation p<0.001, A higher Faces Pain Scale and EQ-5D-Y-3L-IA score both indicate greater pain or discomfort. Shaded cells indicate comparison of correlations between 5-7-year-olds and 8-10-year-olds.*

Table 26 shows that there were significant, low correlations with the MFQ and WSU, with a moderate association with the total score for the 8-10-year-olds. There were fewer significant correlations in the 5-7-year-olds with lonely, love, compassion and the total score showing low associations. This was further evident from the significant difference in correlations of no good, bad person and total score between the age-groups.

Table 26: Convergent validity of EQ-5D-Y-3L-IA Worried, Sad or Unhappy dimension and Moods and Feelings Questionnaire

	Worried, Sad or Unhappy			
	5-7-years	8-10-years	5-7 vs 8-10-years	
Moods and Feelings Questionnaire	(n=177)	(n=211)	z-score	p-value
Unhappy	0.12	0.18**	-0.60	0.274
Enjoyment	0.09	0.18**	-0.89	0.187
Tired	0.11	0.17*	-0.60	0.274
Restless	0.12	0.19**	-0.70	0.242
No good	-0.05	0.18**	-2.26	0.012
Crying	0.08	0.21**	-1.29	0.099
Concentration	0.12	0.23**	-1.11	0.134
Hate	0.05	0.11	-0.59	0.278
Bad person	-0.04	0.28**	-3.19	0.001
Lonely	0.21**	0.22**	-0.10	0.460
Love	0.15*	0.14*	0.10	0.460
Comparison amongst peers	0.18*	0.18*	0.00	0.500
Did things wrong	0.12	0.19**	-0.70	0.242
Total	0.19*	0.34**	-1.57	0.058

**Spearman's correlation p<0.05, **Spearman's correlation p<0.001, A higher Moods and Feelings score and EQ-5D-Y-3L-IA score both indicate greater problems. Shaded cells indicate comparison of correlations between 5-7-year-olds and 8-10-year-olds.*

4.2.7. Test-retest reliability

In the younger group, Mob, LAM and P/D showed moderate test-retest reliability while UA and WSU showed fair reliability (Table 27). Utility score[10] and VAS score were both reliable. In the older group, a moderate reliability was found in Mob with all other dimensions showing fair reliability. VAS scores were reliable whereas the utility scores[10] showed an ICC=0.60.

Table 27: Test-retest reliability of the EQ-5D-Y-3L-IA across age-groups

		5-7-years	**8-10-years**
		(n=177)	(n=211)
Mob	k	0.46**	0.50**
LAM	k	0.47**	0.38**
UA	k	0.27**	0.34**
P/D	k	0.42*	0.29*
WSU	k	0.27*	0.30*
Utility score	ICC	0.73**	0.60**
VAS score	ICC	0.77**	0.70**

*k: Cohen's weighted Kappa, ICC: Intra-class correlation coefficient. **p>0.00, *p<0.05, Mob=mobility, LAM=looking after myself, UA=usual activities, P/D= pain or discomfort, WSU=worried, sad or unhappy.*

[10] Analysis with the Slovenian utility score is presented. There was no significant difference between results using the Slovenian or Japanese utility scores.

4.2.8. Cognitive debriefing

4.2.8.1. Cognitive debriefing template

4.2.8.1.1. Reasons for level reported on EQ-5D-Y-3L-IA dimensions

Table 28 shows that children across age-groups who reported no problems on the IA Mob dimension did so as they were able to mobilise independently. Some participants reported no problems despite their presenting medical condition by reporting "*my legs get tired, but I can still walk by myself*", "*I can walk fine even on my sore leg"*.

Across age-groups, those reporting some problems referred to their presenting medical conditions as a reason for their problems. Children from the orthopaedic group reported "*when I walk or run, my arm pains", "because of my sore foot".* Children with a respiratory condition reported "*my chest gets tight and I get tired", "I get tired when I walk a lot".* Additional reasons for some problems across age-groups included not feeling safe in their environment/community *"it is dangerous to walk alone so I always walk with my mommy", "I might get stolen", "I can't walk alone because they steal children".*

For those reporting a lot of problems, many children similarly associated their problems with their presenting medical condition and reported "*because of the POP [Plater of Paris Cast] on my leg", "I can't walk because of my CP"* while others associated their problems with the use of an assistive device *"because of my wheelchair".*

Table 28: Reasons for level reported in Mobility dimension

	Age Category			
	5-7-years		8-10-years	
	(n=177)		(n=211)	
	n	%	n	%
No problems	**(n=103)**		**(n=147)**	
Independent walking	99	96%	140	95%
Independent walking despite crutch use	1	1%	1	1%
Pain	1	1%	0	0%
Falls	0	0%	1	1%
Environment enables independence	0	0%	2	1%
Other*	2	2%	3	2%
Some problems	**(n=51)**		**(n=45)**	
Presenting medical condition	18	35%	23	51%
Pain unrelated to presenting medical condition	10	20%	4	9%
Falls unrelated to presenting medical condition	5	10%	3	7%
Fatigue unrelated to a presenting medical condition	4	8%	2	4%
Crutch use	4	8%	4	9%
Independent walking	3	6%	1	2%
Unsafe Environment	3	6%	3	7%
Other†	4	8%	5	11%
A lot of problems	**(n=23)**		**(n=19)**	
Presenting medical condition	15	65%	8	42%
Independent walking	2	9%	0	0%
Wheelchair use	2	9%	1	5%
Pain unrelated to presenting medical condition	1	4%	2	11%
Crutch use	1	4%	2	11%
Falls unrelated to presenting medical condition	0	0%	3	16%
Fatigue unrelated to presenting medical condition	0	0%	0	0%
Walking frame or rollator use	0	0%	2	11%
Other§	2	9%	1	5%

**Other includes independent mobility before injury, walking slowly so affected arm does not hurt, feeling lazy to walk and independent mobility due to medication use. †Other includes not enjoying walking, foot going skew while walking, interference by younger siblings when walking, bones feeling lazy, struggling to walk, friends walk too fast and shoes affecting walking. §Other includes not enjoying walking and resting affected leg.*

Table 29 shows that children who reported no problems on the IA LAM dimension did so as they were able to perform both washing and dressing and included reasons such as "*I do it by myself", "it's easy".* Other reasons for independent washing and dressing included being taught by their parents as they get older. Inconsistencies within this dimension were noted as participants reported no problems; however, mentioned that they need physical assistance to wash and/or dress, that they were still learning and needed help to set up the bath or shower.

Children aged 5-7-years who reported some problems, reported needing physical assistance from a helper unrelated to a medical condition and reported *"mommy helps me", "mommy has to help me".* More participants in this age-group reported problems with washing (n=4) compared to dressing (n=2) while most reported problems with both (n=9). More of the younger participants also reasoned with "*I am still learning"* to wash and dress and needing help with setting up the bath/shower area by reporting "*my mom helps me with the taps"* compared to the older participants. However, physical assistance with washing and dressing was also needed in the 8-10-year group, but due to the presenting medical condition as reported "*because of my arm in the cast", "because my arm is broken".*

Similarly, those aged 5-7-years who reported a lot of problems with washing and dressing, reported reasons unrelated to their health. While others reported needing assistance due to a presenting medical condition by saying "*I can't do it with my arm", "because my leg is bandaged and tied up".*

Table 29: Reasons for level reported in Looking After Myself dimension

	5-7-years		8-10-years	
	(n=177)		(n=211)	
	n	%	n	%
No problems	**(n=81)**		**(n=152)**	
Independent washing and dressing	71	88%	146	96%
Assistance with washing and dressing due to developmental age	4	5%	0	0%
Physical assistance to wash and dress - due to developmental capacity	4	5%	2	1%
Still learning to wash and/or dress	1	1%	0	0%
Minimal assistance with shower/bath set up	1	1%	3	2%
Independent washing and dressing despite lower limb plaster of Paris	0	0%	1	1%
Some problems	**(n=63)**		**(n=47)**	
Physical assistance to wash and dress -due to developmental capacity	33	52%	16	34%
Assistance with washing and dressing due to presenting medical condition	15	24%	19	40%
Minimal assistance with shower/bath set up	5	8%	6	13%
Still learning to wash and dress	5	8%	2	4%
Independent washing and dressing	3	5%	2	4%
Assistance needed with buttons and laces	1	2%	0	0%
Other*	1	2%	2	4%
A lot of problems	**(n=33)**		**(n=12)**	
Assistance required to dress – due to developmental capacity	17	52%	5	42%
Help with washing and dressing due to presenting medical condition	10	30%	5	42%
Still learning to wash and/or dress	2	6%	2	17%
Independent washing and dressing	2	6%	0	0%
Minimal assistance with shower/bath set up	1	3%	0	0%
Other[ŧ]	1	3%	0	0%

**Other includes not knowing which clothes fit, doesn't enjoy doing it, feeling lazy and the long time it takes to bath independently. [ŧ]Other includes a parent being injured therefore unable to assist participant with bathing and not wanting to do it sometimes.*

Table 30 shows that children who reported no problems on the IA UA dimension did so as they were able to perform all UA independently and associated their independent UA with enjoyment of the activities included in the dimension. Some participants reported no problems with UA despite their medical condition by reporting "*I am playing and doing schoolwork at hospital".*

Participants who reported some problems associated their problems with their injury or medical condition and reported "*because of my crutches and my broken leg", "I can't use my sore arm".* While most participants referred to the play aspect of UA, one participant in the older group reported "*I cannot play soccer because of my leg"* therefore referring to the sport aspect of UA while only two participants referred to a specific activity "*I get hurt when I run*" or "*I get tired when I run*". Some inconsistencies were noted in the older group as they reported some problems with UA but reasoned with independent UA by reporting "*I am used to it and doing activities", "it's easy to do those things", "I like playing and walking at school".*

Those who reported a lot of problems also associated their problems with their presenting medical condition and reported *"It is difficult to do those things with a cast", "I have asthma, so it is difficult to play".* Some participants from the orthopaedic and functional disability groups associated their problems with the use of an assistive device by reporting "*it is difficult to play with crutches" or "because of my wheelchair".* As with the reporting of some problems, most participants referred to the play aspect of UA. However, three participants in the 5-7-year group who referred to other aspects such as family, school and running by reporting "*I missing them and because of the pins", "I never went to school because of my broken arm"* and *"I get tired when I run"*.

Table 30: Reasons for level reported in Usual Activities dimension

	5-7-years		8-10-years	
	(n=177)		(n=211)	
	n	%	n	%
No problems	**(n=111)**		**(n=143)**	
Independent usual activities	95	86%	124	87%
Independent usual activities with friends	8	7%	9	6%
Independent usual activities despite presenting medical condition or plaster of Paris	4	4%	4	3%
COVID-19 related	0	0%	1	1%
Other*	4	4%	5	3%
Some problems	**(n=42)**		**(n=48)**	
Presenting medical condition	25	60%	31	65%
Bullying, rude/rough children	5	12%	3	6%
Crutches use	2	5%	4	8%
Disagreements with siblings	2	5%	1	2%
I don't like usual activities	2	5%	0	0%
Pain	0	0%	4	8%
Unrelated to presenting medical condition	0	0%	0	0%
Independent usual activities	0	0%	3	6%
Other†	6	14%	2	4%
A lot of problems	**(n=24)**		**(n=20)**	
Presenting medical condition	16	67%	10	50%
Bullying, rude/rough children	2	8%	0	0%
Wheelchair use	2	8%	0	0%
Crutches use	2	8%	0	0%
Disagreements with siblings	1	4%	2	10%
Walker use	0	0%	1	5%
Pain	0	0%	2	10%
Other§	1	4%	5	25%

**Other includes all usual activities are part of daily routines, COVID-19 protocols are followed when playing, assisting siblings with schoolwork, playing with siblings, willingness to learn, being friendly, having fun while young. †Other includes waking up early, struggling to play, difficulties playing with other children and making friends, falling a lot, getting hurt when running, busy parents therefore no one to play with and getting confused when swopping stationery with peers. §Other includes waking up early, difficulties with schoolwork and games, preference to be alone, being late for school, feeling lonely and COVID-related issues such as social distancing.*

Table 31 shows that children who reported no problems on the P/D dimension did so as they had no pain on the day of data collection. Other reasons for no pain included that they had healthy lifestyles and their due to their strong faith.

Those who reported some P/D, mostly associated it with their presenting medical condition and reported pain in the affected body part by reporting "*my chest is a tiny bit sore today", "my leg feels still and pains".* Other reasons for reporting some problems included P/D unrelated to a medical condition such as "*my stomach is cramping", "I have a headache".* One participant associated some problems with missing family by reporting "*I miss my mommy".*

Both age-groups reported a lot of problems due to their presenting medical condition and reported "*my chest is sore", because the cast came off".* While others associated their problems with an unrelated medical condition that occurred on the day of data collection by reporting *"my stomach is sore", "my tooth is sore".*

Table 31: Reasons for level reported in Pain/Discomfort dimension

	5-7-years (n=177)		**8-10-years** (n=211)	
	n	%	n	%
No problems	**(n=92)**		**(n=134)**	
No Pain	79	86%	124	93%
Healthy lifestyle habits	3	3%	2	3%
Faith and religion	0	0%	2	1%
Ability to play all day	0	0%	1	1%
Other*	1	1%	0	0%
Some problems	**(n=58)**		**(n=61)**	
Related to presenting medical condition	28	48%	37	61%
Unrelated to presenting medical condition	26	45%	17	28%
Related to falls/clumsy	3	5%	0	0%
Emotional pain	0	0%	1	2%
Other[ǂ]	0	0%	1	2%
A lot of problems	**(n=27)**		**(n=16)**	
Related to presenting medical condition	18	67%	7	44%
Unrelated to presenting medical condition	5	19%	6	38%

**Other includes participant's bones not cracking on day of data collection. [ǂ]Other includes feeling uncomfortable with multiple layers of clothing on*

Table 32 shows that children who reported no problems on the WSU dimension mainly associated their feelings with school by commenting, "*I like being at school", "I get to see my friends", "my teacher is so nice".* Participants also reported general feelings of happiness and excitement some of which were associated with getting to visit the doctor/healthcare facility.

Across age-groups, those who reported some problems mostly associated those problems with their presenting medical condition and expressed that "*I'm worried what the doctor is going to say"* and *"I am worried for when they take the cast off".* In the older group, more participants also associated their feelings with missing their family due to hospital admissions such as "*I miss my mom*", *"I am away from my mom and I miss her".* Both age-groups reported a lot of problems due to their presenting medical conditions.

Table 32: Reasons for level reported in Worried, Sad or Unhappy dimension

	5-7-years (n=177)		8-10-years (n=211)	
	n	%	n	%
No problems	**(n=137)**		**(n=148)**	
Related to school	39	28%	34	23%
Related of family	27	20%	15	10%
Feeling happy	23	17%	32	47%
Related to presenting medical condition and/or hospital visits	15	11%	12	8%
Ability to play	4	3%	7	5%
Not sad	3	2%	4	3%
Related to missing a day at school due to hospital visit	3	3%	3	2%
Not worried, sad or unhappy	2	1%	2	1%
Not worried	1	1%	7	5%
Related to faith/religion	0	0%	1	1%
Other*	16	12%	23	16%
Some problems due to	**(n=28)**		**(n=51)**	
Related to presenting medical condition	7	25%	11	22%
Missing home or family due to hospital visit/admission	5	18%	7	14%
Family-related issues	3	11%	5	10%
Difficulties at school	2	7%	7	14%
Feeling bad/sick/tired	0	0%	4	8%
COVID-related	0	0%	3	6%
Other†	5	18%	8	16%
A lot of problems	**(n=12)**		**(n=12)**	
Related to presenting medical condition	3	25%	5	42%
Family-related issues	2	17%	2	17%
Missing home or family due to hospital visit/admission	2	8%	2	17%
Feeling bad/sick/tired	1	8%	0	0%
Difficulties at school	1	8%	1	8%
Other§	1	8%	2	17%

**Other includes condition improving, not getting hurt again, ability to go out, witnessing a nice/beautiful day, experiencing kind people, eating yummy food, missed school. †Other includes inability to play, missing school, woken up early, not sure of feelings, not having fun. §Other includes feeling sad due to it not being participant's birthday and feeling worried something bad might happen.*

4.2.8.1.2. Inconsistencies in EQ-5D-Y-3L-IA dimensions as noted by the interviewer

Inconsistencies were noted in relation to functional disability and in consideration to other dimensions. Overall, more inconsistencies were noted in the 5-7-year-olds (19%) compared to the 8-10-year-olds (12%).

In the 5-7-year-olds, the most inconsistencies across dimensions were noted in the Mob and UA dimensions as participants with a presenting medical condition reported some problems with Mob but no problems with UA and justified their answers by reporting "*my foot is fine when I play", "I don't think about my pain when I play".* When looking at the LAM dimension, participants presenting with a medical condition reported no problems with LAM but had an apparent difficulty e.g. with orthopaedic conditions such as upper limb fractures and casts therefore required some assistance but justified their answers by saying "*I can do somethings with one hand but not everything so my mom helps me", "mommy is always there to help me".*

In the 8-10-year-old group, inconsistencies were noted in relation to participants' functional ability whereby no problems were reported on Mob but an assistive device was used due to an orthopaedic condition or functional disability and was justified by "*I can go everywhere with my wheelchair", "I can go everywhere with my crutches".* When inconsistencies were noted across dimensions of Mob and UA also in participants with a presenting medical condition which required an assistive device, the levels of report were justified with "*I can play games with my friends in bed"* or "*I can sit in my chair and play games. I don't have to be standing or walking".*

4.2.8.1.3. Understanding of EQ-5D-Y-3L-IA dimensions

More 5-7-year-olds reported difficulty understanding the questionnaire (n=17, 10%) than the 8-10-year-olds (n=8, 4%) (x^2=4.47, p=0.035). Of the children aged 5-7 years reporting difficulty, one reported that he/she did not understand any of the questions asked. Reasons for difficulties experienced per age-group are shown in Table 33, most of which were associated with difficulty with certain words or comprehension of items.

Table 33: Reasons for difficulties reported with completion of the EQ-5D-Y-3L-IA across age-groups and dimensions

		5-7-years (n=17)		**8-10-years** (n=8)	
Dimension/s	Reason	n	%	n	%
Mobility	I didn't understand "about"	1	6%	1	13%
	I didn't understand the question	3	18%	0	0%
	I had to think a lot	1	6%	0	0%
Looking After Myself	The question was difficult	1	6%	0	0%
	I didn't understand the question	0	0%	1	13%
Usual Activities	There was a lot to think about	5	29%	0	0%
	I don't enjoy those activities so I didn't know what to say	0	0%	1	13%
	I don't understand the question	1	6%	0	0%
Pain/Discomfort	I didn't understand the words	1	6%	2	25%
	I didn't understand "discomfort"	2	12%	1	13%
Worried, Sad or Unhappy	I wasn't sure/can't explain how I was feeling	1	6%	1	13%
	I didn't understand the words	1	6%	0	0%
	I didn't know how to answer	0	0%	1	13%
	I had to think about it	1	6%	0	0%

Some children reported more than one difficulty

4.2.8.2. Interviewer assessment of difficulty with understanding and uncertainty

The interviewer recorded any observed difficulty with understanding and uncertainty on the IA. Children 5-7-years showed less understanding of the IA questions (n=26, 15%) compared to 8-10-years (n=18, 9%) although not significantly different (x^2*=3.04, p=0.081)*. Similarly, children aged 5-7-years showed more uncertainty (n=15, 8%) when answering compared to children 8-10-years (n=8, 4%) (x^2*=2.99, p=0.084)*.

One of the reasons given for misunderstanding the IA was due to the timeframe. Respondents would report based on previous ability instead of their ability for 'Today'. This was commonly seen on the Mob, LAM and UA dimensions whereby a participant reported *"I usually did it before the car accident".* In the 5-7-year group, participants also seemed to misunderstand the Mob dimension as they did not understand the term "walking about" (12%). The LAM dimension (16%) was also misunderstood in this age-group as participants would report no problems but still required assistance (16%).

In the younger group, Mob, UA, P/D and WSU all created equal uncertainty (13%) compared to Mob in the 8-10-year-olds (25%). Uncertainty was determined by the interviewer when participants asked for questions to be repeated and/or explained multiple times, if participants changed their answers multiple times or if a long pause was taken between questions and answers.

4.2.9. Summary of results

A total of 388 5-10-year-olds were recruited across two age-groups, 5-7-years (n=177, 46) and 8-10-years (n=211, 54%) and were allocated to four known-groups. The 5-7-year-olds included orthopaedic conditions (n=79, 45%), from the GenPop (n=44, 25%), respiratory illnesses (n=35, n=20%) and functional disabilities (n=19, 11%). Similarly, the 8-10-year-olds also included, orthopaedic condition (n=82, 39%), from the GenPop (n=65, 31%), functional disabilities (n=36, 17%) and respiratory illnesses (n=28, 13%). There were significantly higher reports of problems in the LAM dimension in the 5-7-year-olds (55%) compared to the 8-10-year-olds (28%) (x^2=31.021 ; p<.0001). The younger children took significantly longer to complete the measure (Mann-Whitney U=8389.5, p<0.001). Known-group validity was found at dimension level with children receiving orthopaedic management reporting more problems on physical dimensions and a significantly lower utility score across both age groups. There was no difference between correlation of VAS and utility scores between age groups (z=0.22, p=0.413). Convergent validity between LAM and WeeFIM items of self-care showed moderate to high correlations for both age groups with a significantly higher correlation in the 8-10-year-olds for dressing upper (z=2.24; p=0.013) and lower body (z= 2.78; p=0.003) and self-care total (z=2.01; p=0.022). There were low to moderate correlations between the other dimensions and corresponding items of the MFQ, FPS-R and WeeFIM. There were fair to moderate levels of test-retest reliability across age groups.

4.3. Discussion

The aim of this chapter was to determine the performance of the newly developed EQ-5D-Y-3L-IA in children aged 5-7-years and compare the results to children aged 8-10-years. Based on previous research, we know children aged 8-years and older understand the concept of health and can reliably report on the HRQoL, whereas; it was suggested by Varni et al. (2007) and Riley (2004) that children as young as 5- and 6-years are also able to report on their HRQoL reliably (1,162).

4.3.1. Recruitment and descriptive statistics

This sample was made up of two age-groups, 5-7-years (n=177) and 8-10-years (n=211) across four known condition groups which included children with acute orthopaedic conditions, chronic respiratory illnesses, functional disabilities and from the GenPop. The condition groups included, ranged in severity and dimensions that were hypothesised to be affected to indicate the performance of the measure across diverse health conditions. There was a large number of non-responders in the 8-10-year-olds (64%), with similar response rates to that of Jelsma et al. (2012) (144). Although reasons for poor response by caregivers is yet to be extensively covered in research, parental understanding and/or knowledge regarding research and consent may be lacking and may contribute to poor response rates (144). The response rate in the 5-7-year-olds was slightly better than 8-10-year-olds with only 44% of potential participants/caregivers not responding. As previously discussed, the difference in methods of recruitment between school-going children and in-patient children may have also impacted on the non-response rates as face-to-face contact may have been more favourable my parents compared to being sent an envelope.

Similar to previous research, also conducted in South Africa, there were no significant difference between health condition or sex (10,16). There were more females amongst the 5-7-year-olds compared to 8-10-year-olds, while previous studies found similar results in their entire sample size or within a specific health condition group (10,17), a different study found the opposite for their 6-7-year-old group when evaluating the EQ-5D-Y-3L proxy-version in Spanish children and adolescents, however, the reason for this was not discussed (106). Conversely, there were more males amongst the 8-10-year-olds which had been found in previous studies conducted in Italy and South Africa, especially amongst children with health conditions (9,17,150).

4.3.2. Dimension performance between the age-groups

The differences in the levels reported across dimensions between age-groups could be largely associated with understanding and age of children. All dimensions except LAM had showed no significant differences between age-groups. To highlight the performance across dimensions, each dimension will be discussed in further detail.

The dimension of Mob had more problems, although not significant, reported in the 5-7-year-olds by which cognitive debriefing revealed an association with problems and a presenting medical condition which would have been expected. Both age-groups equally reported problems due to the environment in which they live despite being able to physically mobilise. A similar finding was seen by Scott et al. (2017) whereby children associated problems in the Mob dimension with safety when walking in their community (16).

When addressing LAM, the younger group was expected to have more problems compared to the older group (118,163). As hypothesised, the dimension of LAM had a higher report of problems for the 5-7-year-olds than the older children, with similar high reports of problems reported on the EQ-5D-Y-3L proxy in young children (10). This is further highlighted by the significant difference between age-groups for convergent validity with WeeFIM items of dressing. The report of problems in younger children is due to their developmental age with still needing assistance with more advanced tasks of dressing, such as fastening buttons and tying shoelaces, as they have not yet learned how to perform all activities of washing and dressing independently (3). Adaptation of the wording of this question may make it more age-appropriate for younger children (10). This may be possible by including simpler activities such as removing their socks or washing their hands to prevent younger children from assuming that washing and dressing refers to more advanced tasks, therefore, reporting problems in this dimension without having any physical impairments which would prevent them from performing washing and dressing tasks independently.

Younger children reported more problems in the UA dimension compared to older children with a small proportion of children reporting problems due to bullying at school. This has yet to be explored in this age-group of children. However, assessing the link between HRQoL and bullying at school in older children aged 11-18-years found that poorer HRQoL was directly linked to experiencing some form of bullying at school (164–166).

In contrast to previous research assessing pain in children/adolescents, whereby pain is thought to increase with age (167), there was a significantly higher report of P/D by the younger group in this study. This finding was supported by Burstrom et al. (2014) and Kim et al. (2017) who both found more problems in the P/D dimension in younger children (14,168). This may suggest that younger children may have been more aware of their pain or chose to report their pain during the interview compared to older children.

In line with Kim et al. (2017) and Burstrom et al. (2014), older children reported more problems in the dimension of WSU (30%) compared to younger children (23%). The reasons for this were not explored previously but when children were asked for a reason in this study, majority of children associated their feelings with their presenting medical condition or missing family due to a hospital admission subsequent to their medical condition/injury. Studies exploring the effects of hospitalisation on children found that children admitted to healthcare facilities without a parent or familiar person often experience more emotional trauma in addition to the trauma caused by the reason for admission (169,170).

4.3.3. Feasibility

The time taken to complete the questionnaire was significantly longer for the younger children although, both questionnaires could be completed in under 2.5 minutes. This is not much longer than the 1 minute completion time reported for the EQ-5D-Y-3L-SC in a sample of children aged 8-12 years (102) and is still feasible for administration in a clinical setting. The feasibility of the measure in the younger children was further shown with a similar ceiling effect to older children. In accordance with other studies and as hypothesised, a higher ceiling effect was seen in the GenPop compared to other condition groups (2,9,14,30,87,106). The younger children did however, report significantly fewer unique health states. This may result in a concentration of select health profiles and may negatively impact the ability to detect a change in the distribution of profile data over time and to compare profiles between children with different health conditions (171). This may, however, be an accurate reflection on the younger children's perceived health as results from this study showed a significant and fair correlation between the utility score, being a composite measure of dimension performance and the VAS score measuring general health (10,16).

4.3.4. Known-group validity

Known-group validity was shown for the utility score across both groups but not the VAS score. This may have been due to dimensions being more specific, while the VAS score represented a more general representation of their health. Previous studies reported significant differences between children with an acute health condition and those from the GenPop or with a chronic health condition (10,16). This is possibly due to the difference in children recruited with an acute injury with this sample having an orthopaedic injury which they did not feel impacted their general health but did recognise the impact it had on individual dimensions. In the studies by Scott et al. (2017) and Verstraete et al. (2020), children with an acute injury were recruited from a tertiary hospital with acute illnesses where there was arguably a greater impact on health with conditions such as cancer, pneumonia, organ transplants, or surgery and seizures. All dimensions were significantly different except the dimension of WSU, more so in the 8-10-year-olds compared to the 5-7-year-olds, which shows that the children from the GenPop also experience feeling WSU despite not presenting with a major health condition but may be linked to social issues such as divorce or deaths. Similar results were seen in a multinational, Korean and Australian studies where by problems were reported in the P/D and WSU dimensions despite participants being a part of the GenPop (2,14,172).

4.3.5. Concurrent validity

Concurrent validity of the IA was assessed by comparing the VAS and utility scores in each age-group. Similar significant correlations were found in the 5-7-year ($r=0.225$, $p=0.003$) and 8-10-year age-group ($r=0.246$, $p<0.001$) with no significant difference between age-groups ($z=-0.22$, $p=0.413$). The correlations found in each age-group proved that utility scores were able to accurately record general health as children reported similarly on the VAS score and across the five dimensions which is used to determine utility scores. As discussed previously, when comparing composite scores to VAS scores, in children with various health conditions, associations were found in acutely ill children but not chronically ill children or children from the GenPop (16). Children with chronic conditions often do not view their physical immobility as a limitation which has been described as the disability paradox (17,29,30). This was evident during cognitive debriefing when children with functional disabilities reported no problems with Mob and/or UA but were confined to a wheelchair. These children then reasoned with being able to go anywhere with their wheelchair or even play in their wheelchair therefore did not report any problems in those dimensions.

4.3.6. Convergent validity

The convergent validity with the WeeFIM and FPS-R was comparable across both age-groups to previous South African results for children aged 8-12-years which showed a significance of p=<0.001 for LAM and WeeFIM self-care total. (16). Similarly, in this study, significant differences across age-groups were found between the LAM dimension and WeeFIM items of self-care, more specifically upper limb dressing (p=0.013), lower limb dressing (p=0.003) and self-care total (p=0.022). The LAM dimension has also been assessed against the CHU-9D item of 'daily routine' which also demonstrated a significant correlation (p=<0.001) (83).

The dimension of Mob had previously been assessed for convergent validity against the CHU-9D item of 'ability to join in on activities' (83,84), PedsQL physical functioning score (2,9), KIDSCREEN-27 physical wellbeing items (16,98,106) and the WeeFIM mobility dimension (16). Some studies have found significant associations with these instruments while others did not. Significant associations were found when comparing Mob to the WeeFIM mobility total in 8-12-year-olds with acute and chronic conditions (16), whereas in this study, no significant difference was seen in WeeFIM mobility total across age-groups (p=0.142) but rather for single mobility items labelled 'toilet transfers' (p=0.049) and 'sit to stand transfers' (p=0.031).

The UA dimension which addresses a wide range of daily activities such as going to school, hobbies, sports, playing and doing things with family or friends was compared to the WeeFIM item of social interaction which has some overlap by addressing interaction with peers/other children during play and social situations but does not necessarily emphasise activities such as school, hobbies and sport. As a result, the UA dimension did not show any significant associations with the WeeFIM item of social interaction for either age-group or across age-groups. Previous studies have used other generic HRQoL instruments such as the CHU-9D items of 'daily routine' and 'able to join in activities' (83), PedsQL total (16) and KIDSCREEN (2) to assess the convergent validity of the UA in children aged 8-19-years. Significant correlations were only found for the CHU-9D items of 'daily routine' (p=0,006) and 'able to join in' (p=<0.001) (83), PedsQL total (p=0.007-0.002) (16) but not the KIDSREEN (2). The JAMAR item of 'school activities and playing with friends' was also compared to UA in children with varying severities of JIA, which found a significant correlation (p=0.011) (13).

The MFQ showed low significant correlations in the 8-10-year group while showing no significant correlations in the 5-7-year group or across age-groups. This could be associated with the two-week

recall period of the MFQ, compared to the IA which refers to 'Today'. Despite young children understanding the concept of time, their ability to recall physical and psychological functioning lessens over time therefore, the recall period becomes incredibly important in this younger age-group (62). Furthermore, this could be attributed to the great variation in emotions experienced in a two-week period. Future research may consider comparing results to the CHU-9D as it assesses similar items as those assessed in the WSU dimension within the same timeframe i.e. day of testing, therefore it might show stronger associations with WSU (87). There were however significant correlations across age-groups for MFQ items of 'I felt like I was no good anymore' (p=0.012) and 'I am a bad person' (p=0.001) with a higher correlation in older children. As children get older and transition to primary school, the expectations and pressure placed on these children especially in academic settings tend to increase (173) which may result in children being compared to one another. Whether this is done by teachers, parents or even amongst peers, it may cause children to feel as though they are not good enough and has the potential to have a negative effect on their overall mental health (174).

The FPS-R, which similarly assesses pain or hurt on the same day as testing, showed significant and moderate associations with the P/D dimension in both age-groups with no significant difference between age-groups (p=0.115). As a result, this may suggest the same recall period allows for better associations between instruments.

4.3.7. Test-retest reliability

Test-retest reliability of the IA was done in school-going children only as their health condition was not expected to change while children from healthcare facilities were expected to experience changes in their conditions and were not seen regularly enough to follow-up 48 hours later. The period between initial testing and retesting was 48 hours as Marx et al. proved no significant difference between two-days or two-weeks, as long as children are not able to remember their initial answers but still understand the concept (26).

Children aged 5-7-years in this sample showed no systematic differences in test-retest reliability with similar reliability reported by Canaway and Frew (2013) when using the EQ-5D-Y-3L-SC in children 6-7-years (87) and in children from a similar South African population aged 8-12-years (102). At a dimension level, when ranked in order of strongest to weakest associations compared to children aged 8-15-years, Mob ranked higher in this study, LAM and UA ranked similarly between studies while WSU was ranked lower in this study (9), thus suggesting that physical health was more likely to remain

stable over time compared to emotional health. In relation to other studies in children 8-19-years from the GenPop the dimension of WSU showed a significant agreement between initial testing and retesting (p=<0.001) (2,16) but when assessed in children 8-19-years with acute and chronic conditions, a significant agreement was only demonstrated in the group of children acute conditions (9,98). While in this study, a stronger agreement was seen in the older group with a moderate test-retest reliability (k=0.30, p<0.05) compared to a fair test-retest in the younger group (k=0.27, p<0.05). P/D seemed to be more stable with moderate agreement on test-retest reliability in the 5-7-year-olds and fair agreement in the 8-10-year-olds. Reasons for level of reporting was not taken at retesting therefore only assumptions can be made regarding the instability in older children.

4.3.8. Cognitive debriefing

Difficulty with understanding the IA was low across both age-groups, although higher in the 5-7-year group (15%) and higher when compared to a total of 7.1% of 6- and 7-year-olds who showed poor or very poor understanding in a pilot study by Canaway and Frew (2013) who used the EQ-5D-Y-3L-SC and CHU-9D (87). Of the total sample, there was only one child in the 5-7-year group who did not understand any of the dimensions on the IA. All other children reported difficulty with certain words or with certain dimensions. The most frequent reason for the difficulty in the 5-7-year group was that it required a lot of thinking. This was most notable for the dimension of UA which may have been too complex for the younger group due to the number of activities within the dimension and their recall ability. A possible solution could be to identify the more common activities in which the child participates in regularly to ensure better recall, therefore also ensuring the instrument is patient-specific. The UA dimension was similarly seen as problematic by therapists assessing the EQ-5D-Y-3L in children aged 8-15-years with acute illnesses (16). It may be helpful to identify activities which children with acute illnesses may be able to participate in, for example reading a book, colouring, interacting with other children in the ward, to ensure all health conditions are considered.

When considering inconsistencies across age-groups, the 5-7--year-old group showed more inconsistencies (19%) compared to the 8-10-year-old group (10%) with the most reported in the Mob and UA dimensions across both age-groups. Despite this, It is important to consider that Mob and UA are not necessarily related ie. if Mob is affected, it should not be assumed that UA will be affected. This was evident in the reasons reported for the inconsistency whereby most children reported that they were still able to do their UA without necessarily being mobile for example, being wheelchair bound due to a functional disability but still being able to interact and play with friends. As mentioned

earlier, this is often referred to the disability paradox whereby children with chronic illness do not always report problems with ADLs as they have often modified activities to suit them (17,29,30).

4.4. Conclusion

The EQ-5D-Y-3L-IA is valid and reliable for measuring health in children aged 5-7-years. The performance of the measure was similar to children aged 8-10-years although there was more report of problems with the dimension of LAM due developmental difficulty as younger children required more help with dressing, including buttons and shoelaces. Further, there was some reported difficulty with thinking about the dimensions in the younger age-group, most notably for UA which includes a large number of examples and therefore, may be too complex for younger children to report on. Adaptations to the dimensions of LAM and UA could improve the suitability of the EQ-5D-Y-3L to IA in younger children.

The IA version is shown to be reliable and valid for measuring HRQoL in children aged 5-7-years. The performance of the measure showed similar validity and reliability as demonstrated for children aged 8-10-years with similar health conditions. Many of the differences noted between the age-groups can be attributed to the developmental age of the child rather than a poor understanding of the concept or an inability to rate their health. With appropriate adaptations to the LAM and UA dimensions, the IA is recommended for routine use in younger children with and without health conditions to accurately record their HRQoL therefore providing a valid and reliable alternative to proxy-report in this age-group. In settings where safety in communities is in question, further explanations of the Mob dimension referring to physical ability may be needed, especially in settings where safety is a concern. While verbal explanations and prompts may be possible with the IA version, written explanations may be warranted for the SC version to avoid this during self-complete as well.

Further studies are recommended to assess the IA in various geographical locations, from different socioeconomic statuses and disease groups. Additionally, further research into the responsiveness of the IA is recommended to determine its ability to detect change in paediatric health status over time.

5. Final Conclusion and Recommendations

To conclude, the newly developed EQ-5D-Y-3L-IA version proved more suitable and preferrable amongst children based on their literacy and showed no systematic differences when compared to the EQ-5D-Y-3L SC version. This proves that the IA version can be used in various age-groups however, keeping in mind the developmental ability of the child when administering the instrument. Furthermore, the IA allows for the subjective report on HRQoL despite age and/or literacy skills but rather, takes understanding of health into consideration to determine ability to self-report. Lastly, this version, in addition to the Kiddy-KINDL and DISABKIDS-TAKE-6, aims to bridge the gap between the number of proxy- and self-report instruments by expanding the pool of IA versions available to the paediatric population.

5.1. Study limitations

The GenPop group was from the same geographical catchment area as the tertiary paediatric hospital from where those with a health condition were recruited. The issues found, seemed to be reflective of the GenPop, the results cannot be generalised to the greater Western Cape region as no data on race, home language, or socioeconomic status were collected for comparison to the GenPop of the Western Cape. Recruitment from the same geographical area attempted to ensure that the socioeconomic background of the groups was similar.

In addition, COVID-19 and local government restrictions impacted recruitment of participants across health conditions and recruitment was stopped earlier than anticipated due to the second wave of COVID-19 in South Africa. More specifically, outpatient clinics were downscaled to prevent overcrowding in waiting areas and to ensure social distancing. As a result, there was a smaller number of children recruited in the chronic respiratory illness group. In addition, recruitment of inpatient orthopaedic patients was affected as selective/non-emergency orthopaedic surgeries were reduced or stopped according to level of restrictions in the Western Cape. Recruitment of children with functional disabilities and GenPop was also largely affected by COVID-19 as special permission was sought after to allow external persons onto school premises due to their own COVID-19 policies despite the level of local government restrictions.

As there is no published utility value set available for South Africa, the recently published value set for Slovenia was used and comparison was made to results using the Japanese value set, showing no

difference in results between the two value sets. At the time of data collection only the Slovenian and Japanese value sets were published. The utility value was used in this study as an indication of composite performance of the EQ-5D-Y-3L-IA and SC descriptive system. However, these value sets take into account societal-preferences and neither the Slovenian nor Japanese utility values are reflective of the preferences in South Africa. For decision-making or application of these utility scores the author would not recommend the use of either of these value sets until standardised methods are recommended to determine which value set is more appropriate.

5.2. Recommendations for practice

Measuring HRQoL in the paediatric population has been emphasised throughout this study with a large focus on the increase in interest in HRQoL in the paediatric population, the subjectivity of HRQoL and how it allows health professionals to monitor the effectiveness of treatment plans/outcomes. With this increase in interest, its subjectivity and ability to assist healthcare professionals, careful consideration should be taken before deciding on which instrument should be utilised to ensure its appropriateness and effectiveness in a given setting. In a clinical setting, where time is of the essence, a quick and easy instrument should be chosen. Although the IA requires an interviewer, due to the nature of its administration by means of a script, this instrument may be quick and easy enough for most clinical settings. In the event of COVID-19, we have further seen an increase in the use of telemedicine, thus the IA would be appropriate for completion during a telehealth consultation. It is therefore recommended that HRQoL be routinely measured in all clinical settings where deemed appropriate. For older children (8-years and older), it should not be assumed that they are able to self-complete an instrument despite their age or the age-range stated by an instrument. On the other hand, for younger children (5-7-years) who are not yet able to read, it should not be assumed that they do not understand the concept of health, therefore proxy-report of HRQoL should not be the default for all young children. It is imperative to assess children individually to determine the most appropriate mode of administration when measuring HRQoL and thus ensuring that information gathered, is done as subjectively as possible.

5.3. Recommendations for research

The IA version proved to be valid and reliable in young children and also interchangeable with the SC version in older children. In research involving younger children, the option for self-report should be encouraged and included in the form of either self-complete or interviewer-administration. The

appropriate version of the instrument should be selected based on the literacy levels of the population with further consideration to school policies in the setting.

The validity and reliability of the EQ-5D-Y-3L-IA in children 5-7-years is an important conclusion for the research community as protocols can now be designed to include self-reported HRQoL in younger children with less reliance on proxy-report.

Other researchers are encouraged to extend the age-range for self-report in future studies which includes the EQ-5D-Y-3L-IA to allow further evidence generation on the performance of the measure in this young age-group across different geographic locations, cultural groups and disease groups. Furthermore, the responsiveness and inter-rater reliability of the EQ-5D-Y-3L-IA needs to be tested. To allow for use in different cultural settings, cross-cultural validity is encouraged. Adaptation to the wording of questions and activities included in dimensions should be considered to ensure developmental appropriateness. Research into the recall ability using the newly developed five level version should also be encouraged. Although, it is hypothesised that this may be more challenging for the 5-7-year olds who struggled with the recall of all of the examples given for the dimension of UA.

Further research is also needed into the development of a South African value set for the EQ-5D-Y as only two have been published thus far, a Slovenian and Japanese value set.

5.4. Recommendations for policies

The South African healthcare system is proposed to change to that of a National Health Insurance (NHI) Fund which aims to ensure greater accessibility and equity of healthcare services for all South Africans (175,176). The proposed NHI would be managed similarly to the National Health scheme used in the United Kingdom which is guided by evidence-based guidelines from the NICE which, as mentioned earlier, currently endorses the adult EQ-5D to be included in appraisals for health technologies for decision-making (21). HRQoL is further used as an outcome measure in clinical practice in this setting to motivate the need for clinical services. Currently, NICE is considering the recommendation for a HRQoL measure for children (21). The Pharmaceutical Benefits Advisory Committee in Australia is similarly contemplating the recommendation for a routine HRQoL instrument for children within clinical trials between the ages of 2-18-years with children over the age of 7-years expected to self-complete the instrument (177). As South Africa navigates the systems within the NHI and considers evidence-based guidelines for decision-making and clinical outcome

measure, it may be prudent to consider the systems used in other national health schemes. This study provides further evidence on the validity and reliability of the EQ-5D-Y-3L-IA and SC versions in young South African children which should be considered for recommendation for policy. However, mode of administration should be determined individually based on ability rather than age.

5.5. Accessibility of research

A summary of the research findings was circulated to all healthcare facilities and schools involved in the study (Appendix 30).

A child-friendly information sheet was created so that participants were able to access and understand the conclusions reached, based on their responses on the questionnaires (Appendix 31).

6. References

1. Varni JW, Limbers CA, Burwinkle TM. How young can children reliably and validly self-report their health-related quality of life? An analysis of 8,591 children across age subgroups with the PedsQL™ 4.0 Generic Core Scales. Health and Quality of Life Outcomes. 2007;5:1–13.
2. Ravens-Sieberer U, Wille N, Badia X, Bonsel G, Burström K, Cavrini G, Devlin N, Egmar AC, Gusi N, Herdman M, Jelsma J, Kind P, Olivares PR, Scalone L, Greiner W. Feasibility, reliability, and validity of the EQ-5D-Y: Results from a multinational study. Quality of Life Research. 2010;19(6):887–97.
3. Wille N, Badia X, Bonsel G, Burström K, Cavrini G, Devlin N, Egmar AC, Greiner W, Gusi N, Herdman M, Jelsma J, Kind P, Scalone L, Ravens-Sieberer U. Development of the EQ-5D-Y: A child-friendly version of the EQ-5D. Quality of Life Research. 2010;19(6):875–86.
4. Solans M, Pane S, Estrada MD, Serra-Sutton V, Berra S, Herdman M, Alonso J, Rajmil L. Health-related quality of life measurement in children and adolescents: A systematic review of generic and disease-specific instruments. Value in Health. 2008;11(4):742–64.
5. Ferguson GD, Jelsma J, Derrett S. The use of the Visual Analogue Scale in the European Quality of Life -5 Dimension Scale- Youth Version (EQ5DY). 1st EuroQol African Regional Meeting, Cape Town, South Africa. Cape Town; 2020.
6. Villalonga-Olives E, Kiese-Himmel C, Witte C, Almansa J, Dusilova I, Hacker K, von Steinbuechel N. Self-reported health-related quality of life in kindergarten children: Psychometric properties of the Kiddy-KINDL. Public Health. 2015;129(7):889–95.
7. Chaplin J, Koopman H, Schmidt S. DISABKIDS Smiley questionnaire: The TAKE 6 assisted health-related quality of life measure for 4 to 7-year-olds. Clinical psychology & psychotherapy. 2008;15:173–80.
8. Varni JW, Burwinkle TM, Lane MM. Health-related quality of life measurement in pediatric clinical practice: An appraisal and precept for future research and application. Health and Quality of Life Outcomes. 2005;3:1–9.
9. Scalone L, Tomasetto C, Matteucci MC, Selleri P, Broccoli S, Pacelli B, Cavrini G. 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. 2011;8(4):331–41.
10. Verstraete J, Lloyd A, Scott D, Jelsma J. How does the EQ-5D-Y Proxy version 1 perform in 3, 4 and 5-year-old children? Health and Quality of Life Outcomes. 2020;18(1):1–10.
11. Zhang Y, Zhou Z, Gao J, Wang D, Zhang Q, Zhou Z, Su M, Li D. Health-related quality of life and its influencing factors for patients with hypertension: Evidence from the urban and rural areas of Shaanxi Province, China. BMC Health Services Research. 2016;16(1):1–9.

12. Khanna D, Tsevat J. Health-related Quality of Life—An Introduction. The American Journal of Managed Care. 2007;13(9):218–23.
13. Scott D, Scott C, Jelsma J, Abraham D, Verstraete J. Validity and feasibility of the self-report EQ-5D-Y generic Health-related Quality of Life outcome measure in children and adolescents with Juvenile Idiopathic Arthritis in Western Cape, South Africa. South African Journal of Physiotherapy. 2019;75(1):1–9.
14. Kim SKSK, Jo MW, Kim SH. A cross sectional survey on health-related quality of life of elementary school students using the Korean version of the EQ-5D-Y. Peer-reviewed Journal. 2017;5(e3115):1–13.
15. Ngwira LG, Khan K, Maheswaran H, Sande L, Nyondo-Mipando L, Smith SC, Petrou S, Niessen L. A Systematic Literature Review of Preference-Based Health-Related Quality-of-Life Measures Applied and Validated for Use in Childhood and Adolescent Populations in Sub-Saharan Africa. Value in Health Regional Issues. 2021;25:37–47.
16. Scott D, Ferguson GD, Jelsma J. The use of the EQ-5D-Y health related quality of life outcome measure in children in the Western Cape, South Africa: Psychometric properties, feasibility and usefulness - a longitudinal, analytical study. Health and Quality of Life Outcomes. 2017;15(1):1–14.
17. Jelsma J, Ramma L. How do children at special schools and their parents perceive their HRQoL compared to children at open schools? Health and Quality of Life Outcomes. 2010;8:2–8.
18. Jelsma J. A comparison of the performance of the EQ-5D and the EQ-5D-Y Health-Related Quality of Life instruments in South African children. International Journal of Rehabilitation Research. 2010;33:172–7.
19. Kenzik KM, Tuli SY, Revicki DA, Shenkman EA, Huang IC. Comparison of 4 pediatric health-related quality-of-life instruments: A study on a Medicaid population. Medical Decision Making. 2014;34(5):590–602.
20. EuroQoL. EQ-5D [Internet]. 2021 [cited 2021 Apr 30]. Available from: https://euroqol.org
21. Rowen D, Keetharuth AD, Poku E, Wong R, Pennington B, Wailoo A. A Review of the Psychometric Performance of Selected Child and Adolescent Preference-Based Measures Used to Produce Utilities for Child and Adolescent Health. Value in Health. 2021;24(3):443–60.
22. Janssens A, Thompson Coon J, Rogers M, Allen K, Green C, Jenkinson C, Tennant A, Logan S, Morris C. A systematic review of generic multidimensional patient-reported outcome measures for children, part I: Descriptive characteristics. Value in Health. 2015;18(2):315–33.
23. Germain N, Aballéa S, Toumi M. Measuring health-related quality of life in young children: how far have we come? Journal of Market Access & Health Policy. 2019;7(1):1618661.

24. Howie SJ, Combrinck C, Roux K, Tshele M, Mokoena GM, N MP. PIRLS Literacy 2016 : South African Highlights Report What is PIRLS? South African Highlights Report. 2017;(December):1–12.
25. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, De Vet HCW. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study. Quality of Life Research. 2010;19(4):539–49.
26. Marx RG, Menezes A, Horovitz L, Jones EC, Warren RF. A comparison of two time intervals for test-retest reliability of health status instruments. Journal of Clinical Epidemiology. 2003;56(8):730–5.
27. Mokkink LB. COSMIN Risk of Bias checklist [PDF File] [Internet]. 2018 [cited 2021 Apr 25]. p. 1–37. Available from: www.cosmin.nl
28. World Health Organization. Constitution of the World Health Organization. 48th ed. Basic documents of the World Health Organization. Geneva. 2014.
29. Huber M, André Knottnerus J, Green L, Van Der Horst H, Jadad AR, Kromhout D, Leonard B, Lorig K, Loureiro MI, Van Der Meer JWM, Schnabel P, Smith R, Van Weel C, Smid H. How should we define health? British Medical Journal. 2011;343(7817).
30. Eidt-Koch D, Mittendorf T, Greiner W. Cross-sectional validity of the EQ-5D-Y as a generic health outcome instrument in children and adolescents with cystic fibrosis in Germany. BMC Pediatrics. 2009;9(55).
31. Torrance GW. Utility approach to measuring health-related quality of life. Journal of Chronic Diseases. 1987;40(6):593–600.
32. Patrick D., Bush J., Chen M. Toward an Operational Definition of Health. Journal of Health and Social Behavior. 1982;14(6):6–23.
33. Karimi M, Brazier J. Health, Health-Related Quality of Life, and Quality of Life: What is the Difference? PharmacoEconomics. 2016;34(7):645–9.
34. Kuyken W, Group T. The World Health Organization Quality of Life assessment (WHOQOL): position paper from the World Health Organization. Social science & medicine. 1995;41(10):1403–9.
35. Ware JE. The status of health assessment 1994. Annual Review of Public Health. 1995;16(7):327–54.
36. Wenger NK, Mattson ME, Furberg CD, Elinson J. Assessment of quality of life in clinical trials of cardiovascular therapies. The American journal of cardiology. 1984 Oct;54(7):908—913.
37. Cummins RA. Moving from the quality of life concept to a theory. Journal of Intellectual

Disability Research. 2005;49(10):699–706.

38. Felce D, Perry J. Quality of life: Its definition and measurement. Research in Developmental Disabilities. 1995;16(1):51–74.
39. Ferrans CE. Definitions and conceptual models of quality of life. In: Outcomes assessment in cancer: Measures, methods, and applications. New York, NY, US: Cambridge University Press; 2005. p. 14–30.
40. Hays R., Reeve B. Measurement and Modeling of Health-Related Quality of Life. Epidemiology and Demography in Public Health San Diego: Academic Press. 2010;195–205.
41. Wilson IB, Cleary P. Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. The Journal of the American Medical Association. 1995;273(1):59–65.
42. Ebrahim S. Clinical and public health perspectives and applications of health-related quality of life measurement. Social Science and Medicine. 1995;41(10):1383–94.
43. Weinstein MC. Recommendations of the Panel on Cost-Effectiveness in Health and Medicine. The Journal of the American Medical Association. 1996;276(15):1253.
44. Prieto L, Sacristán JA. Problems and solutions in calculating quality-adjusted life years (QALYs). Health and Quality of Life Outcomes. 2003;1:1–8.
45. Whitehead SJ, Ali S. Health outcomes in economic evaluation: The QALY and utilities. British Medical Bulletin. 2010;96(1):5–21.
46. Gurková E. Issues in the definitions of HRQoL. Journal of Nursing, Social Studies, Public Health and Rehabilitation. 2011;34(2009):190–7.
47. Spilker B, Revicki B. Quality of life and clinical trials. The Lancet. 1995;346(8966):1–2.
48. Guyatt G, Mitchell A, Irvine EJ, Singer J, Williams N, Goodacre R, Tompkins C. A New Measure of Health Status for Clinical Trials in Inflammatory Bowel Disease. Gastroenterology. 1989;96(2):804–10.
49. Speight J, Shaw JAM. Does one size really fit all? Only by considering individual preferences and priorities will the true impact of insulin pump therapy on quality of life be determined. Diabetic Medicine. 2007;24(7):693–5.
50. Hagerty M, Cummins R, Ferris A, Land K, Michalos A, Peterson M, Sharpe A, Sirgy M, Vogel J. Quality of life indexed for national policy: review and agenda for research. Social Indicattos Research. 2001. 55:1-96.
51. World Bank. Population ages 0-14 (% of total population). 2019;19. Available from: https://data.worldbank.org/indicator/SP.POP.0014.TO.ZS?locations=SN
52. UNICEF. Key demographic indicators [Internet]. UNICEF. 2019 [cited 2021 Apr 30]. Available

from: https://data.unicef.org/country/zaf/

53. Van Malderen C, Amouzou A, Barros AJD, Masquelier B, Van Oyen H, Speybroeck N. Socioeconomic factors contributing to under-five mortality in sub-Saharan Africa: A decomposition analysis. BMC Public Health. 2019;19(1):1–19.
54. Wallander JL, Koot HM. Quality of life in children: A critical examination of concepts, approaches, issues, and future directions. Clinical Psychology Review. 2016;45:131–43.
55. Matza LS, Patrick DL, Riley AW, Alexander JJ, Rajmil L, Pleil AM, Bullinger M. Pediatric patient-reported outcome instruments for research to support medical product labeling: Report of the ISPOR PRO good research practices for the assessment of children and adolescents task force. Value in Health. 2013;16(4):461–79.
56. Baca CB, Vickrey BG, Hays RD, Vassar SD, Berg AT. Differences in child versus parent reports of the child's health-related quality of life in children with epilepsy and healthy siblings. Value in Health. 2010;13(6):778–86.
57. Kaartina S, Chin YS, Fara Wahida R, Woon FC, Hiew CC, Zalilah MS, Mohd Nasir MT. Adolescent self-report and parent proxy-report of health-related quality of life: An analysis of validity and reliability of PedsQL^{TM} 4.0 among a sample of Malaysian adolescents and their parents. Health and Quality of Life Outcomes. 2015;13(1):0–9.
58. Bjornson KF, Belza B, Kartin D, Logsdon RG, McLaughlin J. Self-Reported Health Status and Quality of Life in Youth With Cerebral Palsy and Typically Developing Youth. Archives of Physical Medicine and Rehabilitation. 2008;89(1):121–7.
59. Shiroiwa T, Fukuda T, Shimozuma K. Psychometric properties of the Japanese version of the EQ-5D-Y by self-report and proxy-report: reliability and construct validity. Quality of Life Research. 2019;28(11):3093–105.
60. Frøisland DH, Graue M, Markestad T, Skrivarhaug T, Wentzel-Larsen T, Dahl-Jørgensen K. Health-related quality of life among Norwegian children and adolescents with type 1 diabetes on intensive insulin treatment: A population-based study. Acta Paediatrica, International Journal of Paediatrics. 2013;102(9):889–95.
61. Kreimeier S, Greiner W. EQ-5D-Y as a Health-Related Quality of Life Instrument for Children and Adolescents: The Instrument's Characteristics, Development, Current Use, and Challenges of Developing Its Value Set. Value in Health. 2019;22(1):31–7.
62. Petrou S. Methodological issues raised by preference-based approaches to measuring the health status of children. Health Economics. 2003;12(8):697–702.
63. Wells GA, Russell AS, Haraoui B, Bissonnette R, Ware CF. Validity of quality of life measurement tools - From generic to disease-specific. Journal of Rheumatology. 2011;38(SUPPL. 88):2–6.

64. Bray N, Spencer LH, Edwards RT. Preference-based measures of health-related quality of life in congenital mobility impairment: A systematic review of validity and responsiveness. Health Economics Review. 2020;10(1).
65. Hefford C, Abbott JH, Baxter GD, Arnold R. Outcome measurement in clinical practice: practical and theoretical issues for health related quality of life (HRQOL) questionnaires. Physical Therapy Reviews. 2011;16(3):155–67.
66. Lugnér AK, Krabbe PFM. An overview of the time trade-off method: concept, foundation, and the evaluation of distorting factors in putting a value on health. Expert Review of Pharmacoeconomics and Outcomes Research. 2020;20(4):331–42.
67. Furlong WJ, Feeny DH, Torrance GW, Barr RD. The Health Utilities Index (HUI®) system for assessing health-related quality of life in clinical studies. Annals of Medicine. 2001;33(5):375–84.
68. McCabe CJ, Stevens KJ, Brazier JE. Utility scores for the Health Utilities Index Mark 2: An empirical assessment of alternative mapping functions. Medical Care. 2005;43(6):627–35.
69. Mulhern B, Norman R, De Abreu Lourenco R, Malley J, Street D, Viney R. Investigating the relative value of health and social care related quality of life using a discrete choice experiment. Social Science and Medicine. 2019;233:28–37.
70. Varni JW, Burwinkle TM, Seid M. The PedsQL™ 4.0 as a school population health measure: Feasibility, reliability, and validity. Quality of Life Research. 2003;15(2):203–15.
71. Horsman J, Furlong W, Feeny D, Torrance G. The Health Utilities Index (HUI®): concepts , measurement properties and applications. Health and quality of life outcomes. 2003;13(54):1–13.
72. Riley AW, Forrest CB, Rebok GW, Starfield B, Green BF, Robertson JA, Friello P, Rebok GW, Robertson JA, Green BF. The child report form of the CHIP-child edition reliability and validity. Medical Care. 2004;42(3):221–31.
73. Forrest CB, Bevans KB, Pratiwadi R, Moon J, Teneralli RE, Minton JM, Tucker CA. Development of the PROMIS® pediatric global health (PGH-7) measure. Quality of Life Research. 2014;23(4):1221–31.
74. Furber G, Segal L. The validity of the Child Health Utility instrument (CHU9D) as a routine outcome measure for use in child and adolescent mental health services. Health and Quality of Life Outcomes. 2015;13(1):1–14.
75. Verrips EGH, Vogels TGC, Koopman HM, Theunissen NCM, Kamphuis RP, Fekkes M, Wit JM, Verloove-Vanhorick SP. Measuring health-related quality of life in a child population. European Journal of Public Health. 1999;9(3):188–93.

76. Landgraf JM, Maunsell E, Nixon Speechley K, Bullinger M, Campbell S, Abetz L, Ware JE. Canadian-French, German and UK versions of the child health questionnaire: Methodology and preliminary item scaling results. Quality of Life Research. 1998;7(5):433–45.
77. Desai AD, Zhou C, Stanford S, Haaland W, Varni JW, Mangione-Smith RM. Validity and responsiveness of the Pediatric Quality of Life Inventory (PedsQL) 4.0 generic core scales in the pediatric inpatient setting. The Journal of the American Medical Association Pediatrics. 2014;168(12):1114–21.
78. PedsQL-Translation-Tables [Internet]. Available from: https://www.pedsql.org
79. Busija L, Pausenberger E, Haines TP, Haymes S, Buchbinder R, Osborne RH. Adult measures of general health and health-related quality of life: Medical Outcomes Study Short Form 36-Item (SF-36) and Short Form 12-Item (SF-12) Health Surveys, Nottingham Health Profile (NHP), Sickness Impact Profile (SIP), Medical Outcomes Study Sh. Arthritis Care and Research. 2011;63(SUPPL. 11).
80. Forrest CB, Zorc JJ, Moon JH, Pratiwadi R, Becker BD, Maltenfort MG, Guevara JP. Evaluation of the PROMIS pediatric global health scale (PGH-7) in children with asthma. Journal of Asthma. 2019;56(5):534–42.
81. Stein REK, Jessop DJ. Functional Status II(R): A measure of child health status. Medical Care. 1990;28:1041–55.
82. Kromer ME, Prihoda TJ, Hidalgo HA, Wood PR. Assessing quality of life in Mexican-American children with asthma: Impact-on-family and functional status. Journal of Pediatric Psychology. 2000;25(6):415–26.
83. Ryan JM, McKay E, Anokye N, Noorkoiv M, Theis N, Lavelle G. Comparison of the CHU-9D and the EQ-5D-Y instruments in children and young people with cerebral palsy: A cross-sectional study. BMJ Open. 2020;10(9).
84. Chen G, Flynn T, Stevens K, Brazier J, Huynh E, Sawyer M, Roberts R, Ratcliffe J. Assessing the Health-Related Quality of Life of Australian Adolescents: An Empirical Comparison of the Child Health Utility 9D and EQ-5D-Y Instruments. Value in Health. 2015;18(4):432–8.
85. Vogels TGC, Verloove-Vanhorick SP, Verrips EGH, Fekkes M. Measuring health-related quality of life in children: The development of the TACQOL parent form. Quality of Life Research. 1998;7:457–65.
86. Speechley KN, Maunsell E, Desmeules M, Schanzer D, Landgraf JM, Feeny DH, Barrera ME. Mutual concurrent validity of the child health questionnaire and the health utilities index: An exploratory analysis using survivors of childhood cancer. International Journal of Cancer. 1999;83(SUPPL. 12):95–105.

87. Canaway AG, Frew EJ. Measuring preference-based quality of life in children aged 6 – 7 years : a comparison of the performance of the CHU-9D and EQ-5D-Y — the WAVES Pilot Study. Quality of life research. 2013;22(1):173–83.
88. Noble H, Smith J. Issues of validity and reliability in qualitative research. Evidence-Based Nursing. 2015;18(2):34–5.
89. Heale R, Twycross A. Validity and reliability in quantitative studies. Evidence-Based Nursing. 2015;18(3):66–7.
90. COSMIN. COSMIN-definitions of domains, measurement properties, and aspects of measurement properties. 2018.
91. Amsterdam Public Health. About the COSMIN initiative. Cosmin [Internet]. 2019;1. Available from: https://www.cosmin.nl/about/
92. Connell J, Carlton J, Grundy A, Taylor Buck E, Keetharuth AD, Ricketts T, Barkham M, Robotham D, Rose D, Brazier J. The importance of content and face validity in instrument development: lessons learnt from service users when developing the Recovering Quality of Life measure (ReQoL). Quality of Life Research. 2018;27(7):1893–902.
93. Taherdoost H. Validity and Reliability of the Research Instrument; How to Test the Validation of a Questionnaire/Survey in a Research. SSRN Electronic Journal. 2018.
94. Terwee CB, Prinsen CAC, Chiarotto A, De Vet HCW, Westerman MJ, Patrick DL, Alonso J, Bouter LM, Mokkink LB. COSMIN standards and criteria for evaluating the content validity of health-related Patient-Reported Outcome Measures: a Delphi study. Quality of Life Research. 2018;27:1159–1170.
95. Terwee CB, Prinsen CA, Chiarotto A, Cw De Vet H, Bouter LM, Marjan JA, Donald W, Patrick L, Mokkink LB, Terwee CB. COSMIN methodology for assessing the content validity of PROMs: User manual. Circulation. 2018;120(9):0–70.
96. Varni JW, Seid M, Rode CA. The PedsQL ™ : Measurement Model for the Pediatric Quality of Life Inventory. Medical Care. 1998;37(2):126–39.
97. Erren TC. The case for a posteriori hypotheses to fuel scientific progress. Medical Hypotheses. 2007;69(2):448–53.
98. Mayoral K, Rajmil L, Murillo M, Garin O, Pont A, Alonso J, Bel J, Perez J, Corripio R, Carreras G, Herrero J, Mengibar JM, Rodriguez-Arjona D, Ravens-Sieberer U, Raat H, Serra-Sutton V, Ferrer M. Measurement properties of the online EuroQol-5D-youth instrument in children and adolescents with type 1 diabetes mellitus: Questionnaire study. Journal of Medical Internet Research. 2019;21(11).
99. Forrest CB, Tucker CA, Ravens-Sieberer U, Pratiwadi R, Moon JH, Teneralli RE, Becker B, Bevans

KB. Concurrent validity of the PROMIS® pediatric global health measure. Quality of Life Research. 2016;25(3):739–51.

100. Orgilés M, Melero S, Penosa P, Espada JP, Morales A. Parent-reported health-related quality of life in Spanish pre-schoolers: Psychometric properties of the Kiddy-KINDL-R. Anales de Pediatría (English Edition). 2019;90(5):263–71.

101. Guyatt GH. Measuring health-related quality of life in childhood cancer: lessons from the workshop (discussion). International Journal of Cancer. 1999;83(SUPPL. 12):143–6.

102. Scott, D., Scott, C., Jelsma, J., Abraham, D. & Verstraete J. Validity and feasibility of the self-report EQ-5D-Y generic Health-related Quality of Life outcome measure in children and adolescents with Juvenile Idiopathic Arthritis in Western Cape, South Africa. South African Journal of Physiotherapy. 2019;75(1).

103. Hsu CN, Lin HW, Pickard AS, Tain YL. EQ-5D-Y for the assessment of health-related quality of life among Taiwanese youth with mild-to-moderate chronic kidney disease. International journal for quality in health care : journal of the International Society for Quality in Health Care. 2018;30(4):298–305.

104. Davison SN, Jhangri GS, Feeny DH. Comparing the Health Utilities Index Mark 3 (HUI3) with the short form-36 preference-based SF-6D in chronic kidney disease. Value in Health. 2009;12(2):340–5.

105. Åström M, Persson C, Lindén-Boström M, Rolfson O, Burström K. Population health status based on the EQ-5D-Y-3L among adolescents in Sweden: Results by sociodemographic factors and self-reported comorbidity. Quality of Life Research. 2018;27(11):2859–71.

106. Gusi N, Perez-Sousa MA, Gozalo-Delgado M, Olivares PR. Validity and reliability of the Spanish EQ-5D-Y Proxy version. Anales de Pediatría (English Edition). 2014;81(4):212–9.

107. L . Rajmil , V . Serra-Sutton , J . Alonso , B . Starfield , A . W . Riley J. R. V. The Spanish Version of the Child Health and Illness Profile-Adolescent Edition (CHIP-AE). Quality of Life Research. 2003;12(3):303–13.

108. Feeny DH, Eckstrom E, Whitlock EP, Perdue LA. A Primer for Systematic Reviewers on the Measurement of Functional Status and Health-Related Quality of Life in Older Adults. A Primer for Systematic Reviewers on the Measurement of Functional Status and Health-Related Quality of Life in Older Adults. 2013; Available from:

109. Mokkink LB, Prinsen CAC, Patrick DL, Alonso J, Bouter LM, De Vet HCW, Terwee CB. COSMIN methodology for systematic reviews of Patient - Reported Outcome Measures (PROMs). User Manual. 2018;(February):1–78.

110. Harris AHS, Gupta S, Bowe T, Ellerbe LS, Phelps TE, Rubinsky AD, Finney JW, Asch SM,

Humphreys K, Trafton J. Predictive validity of two process-of-care quality measures for residential substance use disorder treatment. Addiction Science and Clinical Practice. 2015;10(1):1–8.

111. Taber KS. The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. Research in Science Education. 2018;48(6):1273–96.

112. Gaitán-López DF, Correa-Bautista JE, Vinaccia S, Ramírez-Vélez R. Self-report health-related quality of life among children and adolescents from Bogotá, Colombia. The FUPRECOL study. Colombia Medica. 2017;48(1):12–8.

113. Leopoldo-Rodado M, Pantoja-Pertegal F, Belmonte-Caro R, Garcia-Perla A, Gonzalez-Cardero E, Infante-Cossio P. Quality of life in early age Spanish children treated for cleft lip and/or palate: a case-control study approach. Clinical Oral Investigations. 2021;25(2):477–85.

114. Devon HA, Block ME, Moyle-Wright P, Ernst DM, Hayden SJ, Lazzara DJ, Savoy SM, Kostas-Polston E. A psychometric toolbox for testing validity and reliability. Journal of Nursing Scholarship. 2007;39(2):155–64.

115. Amedro P, Huguet H, Macioce V, Dorka R, Auer A, Guillaumont S, Auquier P, Abassi H, Picot MC. Psychometric validation of the French self and proxy versions of the PedsQL^{TM} 4.0 generic health-related quality of life questionnaire for 8–12 year-old children. Health and Quality of Life Outcomes. 2021;19(1):1–14.

116. Lindvall K, Vaezghasemi M, Feldman I, Ivarsson A, Stevens KJ, Petersen S. Feasibility, reliability and validity of the health-related quality of life instrument Child Health Utility 9D (CHU9D) among school-aged children and adolescents in Sweden. Health and Quality of Life Outcomes. 2021;19(1):1–12.

117. Weiner BJ, Lewis CC, Stanick C, Powell BJ, Dorsey CN, Clary AS, Boynton MH, Halko H. Psychometric assessment of three newly developed implementation outcome measures. Implementation Science. 2017;12(1):1–12.

118. Bergfors S, Åström M, Burström K, Egmar AC. Measuring health-related quality of life with the EQ-5D-Y instrument in children and adolescents with asthma. Acta Paediatrica, International Journal of Paediatrics. 2015;104(2):167–73.

119. Raat H, Botterweck AM, Landgraf JM, Hoogeveen WC. Reliability and validity of the short form of the child health questionnaire for parents (CHQ-PF28) in large random school based and general population samples. 2005;75–82.

120. Pistorio A, Ruperto N, Ravelli A, Pistorio A, Malattia C, Viola S, Cavuto S, Alessio M, Alpigiani MG, Buoncompagni A, Corona F, Cortis E, Falcini F, Gerloni V, Lepore L, Sardella ML, Medica I, Matteo IS, Pavia U, Ii UF. Cross-cultural adaptation and psychometric evaluation of the

Childhood Health Assessment Questionnaire (CHAQ) and the Child Health Questionnaire (CHQ) in 32 countries . Review of The Italian version of the Childhood Health Assessment Questionnaire. 2015;(July 2001):2–5.

121. Norrby U, Nordholm L, Fasth A. Reliability and validity of the Swedish version of Child Health Questionnaire Reliability and validity of the Swedish version of Child Health Questionnaire. 2009;9742.
122. Ayala GX, Elder JP. Qualitative methods to ensure acceptability of behavioral and social interventions to the target population. Journal of Public Health Dentistry. 2011;71(SUPPL. 1):1–17.
123. The Children's Hospital Trust. 2020;27(930004493):1121573. Available from: https://www.childrenshospitaltrust.org.za/the-hospital/
124. Department of Basic Education. Guidelines To Ensure Quality Education and Support in Special Schools and Special School. Government Gazette. 2014;1–25. A
125. Colice GL. Categorizing asthma severity: an overview of national guidelines. Clinical medicine & research. 2004;2(3):155–63.
126. Verstraete J, Marthinus Z, Dix-Peek S, Scott D. Measurement properties and responsiveness of the EQ-5D-Y-5L compared to the EQ-5D-Y-3L in children and adolescents receiving acute orthopaedic care. 2021.
127. EuroQol Research Foundation. EQ-5D-Y User Guide. EuroQol Research Foundation 2020 [Internet]. 2020;(September):1–20. Available from: www.impact-test.co.uk
128. Wild D, Grove A, Eremenco S, McElroy S, Verjee-Lorenz A, Erikson P. Value in Health. 2005;8(2):94–104.
129. Englund Dimitrova B. Translation process. 2010;(April):406–11. Available from: https://euroqol.org/support/translation-process/
130. Prevolnik Rupel V, Ogorevc M, Greiner W, Kreimeier S, Ludwig K, Ramos-Goni JM. EQ-5D-Y Value Set for Slovenia. PharmacoEconomics. 2021;39(4):463–71.
131. Shiroiwa T, Ikeda S, Noto S, Fukuda T, Stolk E. Valuation Survey of EQ-5D-Y Based on the International Common Protocol: Development of a Value Set in Japan. Medical Decision Making. 2021;41(5):597–606.
132. Angold A, Costello J, Van Kämmen W, Stouthamer-Loeber M. Development of a short questionnaire for use in epidemiological studies of depression in children and adolescents: factor composition and structure across development. International Journal of Methods in Psychiatric Research. 1996;5(4):251–62.
133. Graham JE, Granger C V., Karmarkar AM, Deutsch A, Niewczyk P, Divita MA, Ottenbacher KJ.

The uniform data system for medical rehabilitation: Report of follow-up information on patients discharged from inpatient rehabilitation programs in 2002-2010. American Journal of Physical Medicine and Rehabilitation. 2014;93(3):231–44.

134. Ottenbacher KJ, Msall ME, Lyon N, Duffy LC, Ziviani J, Granger C V., Braun S, Feidler RC. The WeeFIM instrument: Its utility in detecting change in children with developmental disabilities. Archives of Physical Medicine and Rehabilitation. 2000;81(10):1317–26.
135. University of Cape Town Research Office. University of Cape Town Research Data Management Policy. 2019;(February):2–5.
136. Verstraete J, Amien R, Jelsma J, Scott D. Comparing the English EQ-5D-Y Three-Level Version with the Five-Level Version in South Africa. 2021.
137. Sullivan GM, Feinn R. Using Effect Size—or Why the P Value Is Not Enough . Journal of Graduate Medical Education. 2012;4(3):279–82.
138. VassarStats: Website for statistical computation 2004 [Internet]. 2004 [cited 2021 Aug 17]. Available from: http://vassarstats.net/clin1.html
139. Cohen S., Percival A. Prolonged Peritoneal Dialysis in Patients Awaiting Renal Transplantation. British Medical Journal. 1968;1:409–13.
140. Bowling A. Mode of questionnaire administration can have serious effects on data quality. Journal of Public Health. 2005;27(3):281–91.
141. World Medical Association Declaration ofHelsinki Ethical Principles for Medical Research Involving Human Subjects. The Journal of the American Medical Association. 2013;53(9):739.
142. Africa ND of HR of S. COVID-19 Disease: Infection Prevention and Control Guidelines. 2020;2(April):1–25. Available from: https://j9z5g3w2.stackpathcdn.com/wp-content/uploads/2020/04/Covid-19-Infection-and-Prevention-Control-Guidelines-1-April-2020.pdf
143. Wu X, Ohinmaa A, Veugelers P. Sociodemographic and neighbourhood determinants of health-related quality of life among grade-five students in Canada. Quality of life research : An international journal of quality of life aspects of treatment, care and rehabilitation. 2010;19:969–76.
144. Jelsma J, Burgess T, Henley L. Does the requirement of getting Active consent from parents in school-based research result in a biased sample? an empirical study. Journal of Empirical Research on Human Research Ethics. 2012;7(5):56–62.
145. Pan CW, Zhong H, Li J, Suo C, Wang P. Measuring health-related quality of life in elementary and secondary school students using the Chinese version of the EQ-5D-Y in rural China. BMC Public Health. 2020;20(1):1–8.

146. Govender R, Hugo AJ. An analysis of the results of literacy assessments conducted in South African primary schools. South African Journal of Childhood Education. 2020;10(1).
147. Nederhof AJ. Methods of coping with social desirability bias : a review. European Journal of Social Psychology. 1985;15(April 1984):263–80.
148. Lozano F, Lobos JM, March JR, Carrasco E, Barros MB, González-Porras JR. Self-administered versus interview-based questionnaires among patients with intermittent claudication: Do they give different results? A cross-sectional study. Sao Paulo Medical Journal. 2016;134(1):63–9.
149. Statistics South Africa, Statistical release 2016 P0301. Statistics South Africa. 2016.
150. Ras M, Human AT. The Health Related Quality of Life of Children Living with a Physical Disability Attending Physiotherapy at Meerhof School in the North West Province. Journal of Community and Health Sciences. 2014;9(2):48–61.
151. Michel G, Bisegger C, Fuhr DC, Abel T. Age and gender differences in health-related quality of life of children and adolescents in Europe: A multilevel analysis. Quality of Life Research. 2009;18(9):1147–57.
152. Craig BM, Greiner W, Brown DS, Reeve BB. Valuation of child health-related quality of life in the United States. Health Economics. 2016;25:768–77.
153. Olsen JA, Misajon RA. A conceptual map of health-related quality of life dimensions: key lessons for a new instrument. Quality of Life Research. 2020;29(3):733–43.
154. Tsakos G, Bernabé E, O'Brien K, Sheiham A, de Oliveira C. Comparison of the self-administered and interviewer-administered modes of the child-OIDP. Health and Quality of Life Outcomes. 2008;6:1–8.
155. Rosenthal R, Fode K. Psychology of the Scientist: V. Three experiments in experimenter bias. Psychological Reports. 1963;12:491–511.
156. Cook DJ, Guyatt GH, Juniper E, Griffith L, McIlroy W, Willan A, Jaeschke R, Epstein R. Interviewer versus self-administered questionnaires in developing a disease-specific, health-related quality of life instrument for asthma. Journal of Clinical Epidemiology. 1993;46(6):529–34.
157. Puhan MA, Ahuja A, Van Natta ML, Ackatz LE, Meinert C. Interviewer versus self-administered health-related quality of life questionnaires - Does it matter? Health and Quality of Life Outcomes. 2011;9(1):30.
158. Farnik M. Instrument development and evaluation for patient-related outcomes assessments. Patient Related Outcome Measures. 2012;1.
159. Ponizovsky-Bergelson Y, Dayan Y, Wahle N, Roer-Strier D. A Qualitative Interview With Young Children: What Encourages or Inhibits Young Children's Participation? International Journal of Qualitative Methods. 2019;18:1–9.

160. Landis J, Koch G. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–74.
161. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. Journal of Chiropractic Medicine. 2016;15(2):155–63.
162. Riley AW. Evidence that school-age children can self-report on their health. Ambulatory Pediatrics. 2004;4(4 SUPPL.):371–6.
163. Burström K, Svartengren M, Egmar AC. Testing a Swedish child-friendly pilot version of the EQ-5D instrument - Initial results. European Journal of Public Health. 2011;21(2):178–83.
164. Chester KL, Spencer NH, Whiting L, Brooks FM. Bullying and Adolescent Health-Related. Journal of School Health. 2017;87(11):865–72.
165. Otto C, Haller AC, Klasen F, Hölling H, Bullinger M, Ravens-Sieberer U. Risk and protective factors of health-related quality of life in children and adolescents: Results of the longitudinal BELLA study. PLoS ONE. 2017;12(12):1–17.
166. Haraldstad K, Kvarme LG, Christophersen KA, Helseth S. Associations between self-efficacy, bullying and health-related quality of life in a school sample of adolescents: A cross-sectional study.BMC Public Health. 2019;19(1):1–9.
167. Haraldstad K, Christophersen KA, Helseth S. Health-related quality of life and pain in children and adolescents: A school survey. BMC Pediatrics. 2017;17(1):1–8.
168. Burström K, Bartonek A, Broström EW, Sun S, Egmar AC. 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, International Journal of Paediatrics. 2014;103(4):426–35.
169. Shields L, Mcn HZ, Practice C. Review title Review objective. 2(2):1–13.
170. Rokach A. Psychological, emotional and physical experiences of hospitalized children. Clinical Case Reports and Reviews. 2016;2(4):399–401.
171. Devlin N, Parkin D, Janssen B. Methods for Analysing and Reporting EQ-5D Data. Methods for Analysing and Reporting EQ-5D Data. Springer Nature Switzerland AG; 2020.
172. Wu XY, Ohinmaa A, Johnson JA, Veugelers PJ. Assessment of children's own health status using visual analogue scale and descriptive system of the EQ-5D-Y: Linkage between two systems. Quality of Life Research. 2014;23(2):393–402.
173. Chan WL. Expectations for the transition from kindergarten to primary school amongst teachers, parents and children. Early Child Development and Care. 2012;182(5):639–64.
174. Cushman P, Clelland T, Hornby G. Health-promoting schools and mental health issues: A survey of New Zealand schools. Pastoral Care in Education. 2011;29(4):247–60.
175. Passchier R V. Exploring the barriers to implementing national health insurance in South Africa:

The people's perspective. South African Medical Journal. 2017;107(10):836–8.

176. Murphy SD, Moosa S. The views of public service managers on the implementation of National Health Insurance in primary care: a case of Johannesburg Health District, Gauteng Province, Republic of South Africa. BMC Health Services Research. 2021;21(1):1–9.

177. Jones R, Mulhern B, McGregor K, Yip S, O'Loughlin R, Devlin N, Hiscock H, Dalziel K. Psychometric Performance of HRQoL Measures : An. Children. 2021;8(714).

178. Varni JW, Limbers C, Burwinkle TM. Literature review: Health-related quality of life measurement in pediatric oncology: Hearing the voices of the children. Journal of Pediatric Psychology. 2007;32(9):1151–63.

179. Hill CD, Edwards MC, Thissen D, Langer MM, Wirth RJ, Burwinkle TM, Varni JW. Practical issues in the application of item response theory: A demonstration using items from the Pediatric Quality of Life Inventory (PedsQL) 4.0 generic core scales. Medical Care. 2007;45(5 SUPPL. 1):39–47.

180. Limbers CA, Newman DA, Varni JW. Factorial Invariance of Child Self-report Across Healthy and Chronic Health Condition Groups: A Confirmatory Factor Analysis Utilizing the PedsQLTM 4.0 Generic Core Scales*. Journal of Pediatric Psychology. 2008;33(6):630–9.

181. Limbers CA, Newman DA, Varni JW. Factorial invariance of child self-report across age subgroups: A confirmatory factor analysis of ages 5 to 16 years utilizing the PedsQL 4.0 Generic Core Scales. Value in Health. 2008;11(4):659–68.

182. Varni JW, Limbers CA, Newman DA. Factorial invariance of the PedsQL^{TM} 4.0 generic core scales child self-report across gender: A multigroup confirmatory factor analysis with 11,356 children ages 5 to 18. Applied Research in Quality of Life. 2008;3(2):137–48.

183. Morrow AM, Hayen A, Quine S, Scheinberg A, Craig JC. 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. 2012;38(2):186–95.

184. Maddigan SL, Feeny DH, Majumdar SR, Farris KB, Johnson JA. Health Utilities Index mark 3 demonstrated construct validity in a population-based sample with type 2 diabetes. Journal of Clinical Epidemiology. 2006;59(5):472–7.

185. Castel LD, Williams KA, Bosworth HB, Eisen S V., Hahn EA, Irwin DE, Kelly MAR, Morse J, Stover A, DeWalt DA, DeVellis RF. Content validity in the PROMIS social-health domain: A qualitative analysis of focus-group data. Quality of Life Research. 2008;17(5):737–49.

186. Flynn KE, Dew MA, Lin L, Fawzy M, Graham FL, Hahn EA, Hays RD, Kormos RL, Kiu H, McNulty M, Weinfurt K. Reliability and Construct Validity of PROMIS® Measures for Patients With Heart Failure Who Undergo Heart Transplant. Quality of Life Research. 2015;24(11§2591-2599).

187. Wolf RT, Ratcliffe J, Chen G, Jeppesen P. The longitudinal validity of proxy-reported CHU9D. Quality of Life Research. 2021;30(6):1747–56.

188. Li J, Liu Y, Yu C, Cui B, Du M. Comparison of Incisions and Outcomes for Closure of Ventricular Septal Defects. Annals of Thoracic Surgery. 2008;85(1):199–203.

189. Byrne MW, Honig J. Psychometrics of Child Health Questionnaire Parent Short Form (CHQ-28) Used to Measure Quality of Life in HIV-Infected Children on Complex Anti-Retroviral Therapy. Quality of Life Research. 2021;14(7):1769–74.

190. Drotar D, Schwartz L, Palermo TM, Burant C. Factor Structure of the Child Health Questionnaire-Parent Form in Pediatric Populations. 2006;31(2):127–38.

191. Hepner KA, Sechrest L. Confirmatory factor analysis of the Child Health Questionnaire-Parent Form 50 in a predominantly minority sample Confirmatory factor analysis of the Child Health Questionnaire-Parent Form 50 in a predominantly minority sample. Quality of Life Research. 2014;11:763-773.

192. Ferro MA, Landgraf JM, Speechley KN, Speechley KN. Factor structure of the Child Health Questionnaire Parent Form-50 and predictors of health-related quality of life in children with epilepsy Factor structure of the Child Health Questionnaire Parent Form-50 and predictors of health-related quality of life. 2013;22(8):2201–11.

7. Appendices

Appendix 1: COSMIN Risk of Bias Checklist – EQ-5D-Y-3L

(level of performance highlighted in grey)

Box General requirement for studies that applied Item Response Theory (IRT) Models							
	Very good	**Adequate**	**Doubtful**	**Inadequate**	**Not applicable**	**Reference**	**Notes**
Box 1. PROM development							
1a. PROM design						(2,3,98,103,105,106,9,10,14,16,17,30,59,83)	
General design requirements							
Is a clear description provided of the construct to be measured?	Construct clearly described			Construct notclearly described	Not Applicable		
Is the origin of the construct clear: was a theory, conceptual framework or disease model used or clear rationale provided todefine the construct to be measured?	Origin of the construct clear		Origin of the construct not clear				From adult eq5d
Is a clear description provided of the target population for whichthe PROM was developed?	Target population clearly described			Target populationnot clearly			

				described			
Is a clear description provided of the context of use	Context of use clearly described		Context of use not clearly described				
Was the PROM development study performed in a sample representing the target population for which the PROM wasdeveloped?	Study performed in a sample representing the target population	Assumable that the study was performed in a sample representin g the target population, butnot clearly described	Doubtful whether the study was performed ina sample representing the target population	Study not performed in a sample representing the target population **(SKIP items 6-12)**			
Concept elicitation (relevance and comprehensiveness)							
Was an appropriate qualitative data collection method used toidentify relevant items for a new PROM?	Widely recognized or well justified qualitative methodused, suitable for the construct and study population	Assumable thatthe qualitative method was appropriate and suitable forthe construct and study population, butnot clearly described	Only quantitative (survey) method(s) used or doubtful whether the method was suitable for the construct and study population	Method used not appropriate or not suitable for the construct or study population			
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewers used	Group moderators/ interviewers had limited	Not clear if group moderators/		Not Applica ble		

		experience or were trained specifically for thestudy	interviewers weretrained or group moderators/ interviewers nottrained and no experience				
Were the group meetings or interviews based on an appropriatetopic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate, but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applica ble		
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews wererecorded and transcribed verbatim	Assumable that allgroup meetings or interviews were recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings or interviews were recorded and transcribed verbatimor recordings not transcribed verbatim or only notes were made during the group meetings/ interviews	No recording and no notes	Not applica ble		
Was an appropriate approach used to analyse the data?	A widely recognized or welljustified approachwas used	Assumable thatthe approach was appropriate, but not	Not clear what approach was used ordoubtful whether the approach was appropriate	Approach not appropriate			

		clearly described					
Was at least part of the data coded independently?	At least 50% of thedata was coded byat least two researchers independently	11-49% of the data was codedby at least two researchers independent ly	Doubtful if two researchers were involved in the codingor only 1-10% of the data was coded by at least two researchers independently	Only one researcher was involvedin coding or no coding			
Was data collection continued until saturation was reached?	Evidence providedthat saturation was reached	Assumable that saturation was reached	Doubtful whether saturation was reached	Evidence suggests that saturation was not reached	Not applica ble		
For quantitative studies (surveys): was the sample size appropriate?	≥100	50-99	30-49	<30	Not applica ble		
1b. <u>Cognitive interview study or other pilot test</u>							
Was a cognitive interview study or other pilot test conducted?	YES			**NO (SKIP items 15-35**			
General design requirements							
Was the cognitive interview study or other pilot test performed ina sample representing the target population?	Study performed in a sample representing the target population	Assumable that the study was performed in a sample representin g the target population, butnot clearly	Doubtful whether the study was performed in a sample representing the target population	Study not performed in a sample representing the target population			

		described					
Comprehensibility							
Were patients asked about the comprehensibility of the PROM?	YES	Not clear (SKIP standards 17-25)	No **(SKIP standards 17-25)**				
Were all items tested in their final form?	All items were tested in their finalform	Assumable thatall items were tested in their final form, but not clearly described	Not clear if all itemswere tested in theirfinal form	Items were not tested in their final form or itemswere not re- tested after substantial adjustments			
Was an appropriate qualitative method used for assessing the comprehensibility of the PROM instructions, items, response options, and recall period?	Widely recognized or well justified qualitative method used	Assumable that the method was appropriate butnot clearly described	Only quantitative (survey) method (s) used or doubtful whether the method was appropriate or not clear if patients were asked about the comprehensibility of the items, response options or recall period or patients not asked about the comprehensibility of the PROM instructions	Method used not appropriate or patients were not asked about the comprehensibility of the items, response options or recall period or patients			
Was each item tested in an appropriate number of patients? For qualitative studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear				

For quantitative (survey) studies							
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewers used	Group moderators/ interviewers had limited experience or were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers not trained and no experience		Not Applicable		
Were the group meetings or interviews based on an appropriatetopic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate, but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable		
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews wererecorded and transcribed verbatim	Assumable that allgroup meetings or interviews were recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings or interviews were recorded and transcribed verbatimor recordings not transcribed verbatim or only notes were made during the group meetings/ interviews	No recording and no notes	Not applicable		

Was an appropriate approach used to analyse the data?	A widely recognized or well justified approach was used	Assumable that the approach was appropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate			
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only oneresearcher involvedin the analysis				
Were problems regarding the comprehensibility of the PROMinstructions, items, response options, and recall period appropriately addressed by adapting the PROM?	No problems found or problems appropriately addressed and PROM was adapted and re- tested if necessary	Assumable thatthere were no problems or that problems were appropriatel y addressed, but not clearly described	Not clear if there were problems or doubtful if problems were appropriately addressed	Problems not appropriately addressed or PROM was adapted but items were not re-tested after substantial adjustments	Not applica ble		
Comprehensiveness							
Were patients asked about the comprehensiveness of the PROM?	YES		NO or not clear **(SKIP items 27-35)**				

Was the final set of items tested?	The final set of items was tested	Assumable that the final set of items was tested but not clearly described	Not clear if the final set of items was tested or the final set of items was not tested or the set of items were not re-tested after items were removed or added				
Was an appropriate method used for assessing the comprehensiveness of the PROM?	Widely recognized method used	Assumable that method was appropriate but not clearly described or only quantitative (survey) method (s) used	Doubtful whether the method was appropriate or methodused not appropriate				
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear				
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewers used	Group moderators/ interviewers had limited experience or were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers not trained		Not Applicable		

			and no experience				
Were the group meetings or interviews based on an appropriatetopic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate, but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable		
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews wererecorded and transcribed verbatim	Assumable that allgroup meetings or interviews were recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings or interviews were recorded and transcribed verbatimor recordings not transcribed verbatim or only notes were made during the group meetings/ interviews	No recording and no notes	Not applicable		
Was an appropriate approach used to analyse the data?	A widely recognized or well justified approach was used	Assumable that the approach was appropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate			
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers	Not clear if two researchers were included in the analysis or only				

		were involved in the analysis, but not clearly described	oneresearcher involvedin the analysis				
Were problems regarding the comprehensiveness of the PROM appropriately addressed by adapting the PROM?	No problems found or problems were appropriately addressed and PROM adapted and re-tested if necessary	Assumable that there were problems or that problems were appropriatel y addressed but not clearly described	Not clear if there were problems or doubtful if the problems were appropriately addressed or PROM was adapted but items were not re-tested after substantial adjustments	Problems not appropriately addressed	Not applica ble		
Box 2 Content validity							
2a. Asking patients about relevance							
Design requirements							
Was an appropriate method used to ask patients whether each item is relevant for their experience with the condition?	Widely recognized or well justified method used	Only quantitative (survey) method(s) used or assumable that the method was appropriate but not clearly described	Not clear if patients were asked whether each item is relevantor doubtful whether the method was appropriate	Method used notappropriate or patients not asked about the relevance of all items			

Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	≥7 ≥50	4-6 ≥30	<4 or not clear <30 or not clear				
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewers used	Group moderators/ interviewers had limited experience or were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers not trained and no experience		Not Applica ble		
Were the group meetings or interviews based on an appropriatetopic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate, but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applica ble		
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews wererecorded and transcribed verbatim	Assumable that allgroup meetings or interviews were recorded and transcribed verbatim, but not	Not clear if all group meetings or interviews were recorded and transcribed verbatimor recordings not transcribed verbatim or only notes were made	No recording and no notes	Not applica ble		

		clearly described	during the group meetings/ interviews				
Analyses							
Was an appropriate approach used to analyse the data?	A widely recognized or well justified approach was used	Assumable that the approach was appropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate			
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only oneresearcher involvedin the analysis				
2b. Asking patients about comprehensiveness							
Was an appropriate method used for assessing the comprehensiveness of the PROM?	Widely recognized or well justified method used	Only quantitative (survey) method(s) used or assumable that the method was appropriate but not clearly described	Doubtful whether the method was appropriate	Method used notappropriate			

Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	≥7 ≥50	4-6 ≥30	<4 or not clear <30 or not clear				
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewers used	Group moderators/ interviewers had limited experience or were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers nottrained and no experience		Not Applica ble		
Were the group meetings or interviews based on an appropriatetopic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate, but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applica ble		
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews wererecorded and transcribed verbatim	Assumable that allgroup meetings or interviews were recorded and transcribed verbatim, but not	Not clear if all group meetings or interviews were recorded and transcribed verbatimor recordings not transcribed verbatim or only notes were made	No recording and no notes	Not applica ble		

		clearly described	during the group meetings/ interviews				
Analyses							
Was an appropriate approach used to analyse the data?	A widely recognized or well justified approach was used	Assumable that the approach was appropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate			
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only oneresearcher involvedin the analysis				
2c. Asking patients about comprehensibility							
Was an appropriate qualitative method used for assessing the comprehensibility of the PROM	Widely recognized or well justified qualitative method used	Assumable that the method was appropriate butnot clearly described	Only quantitative (survey) method (s) used or doubtful whether the method was appropriate or not clear if patients were asked about the	Method used not appropriate or patients were not asked about the comprehensibility of the items,			

instructions, items, response options, and recall period?			comprehensibility of the items, response options or recall period or patients not asked about the comprehensibility of the PROM instructions	response options or recall period or patients			
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear				
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewers used	Group moderators/ interviewers had limited experience or were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers nottrained and no experience				
Were the group meetings or interviews based on an appropriatetopic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate, but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable		

Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews wererecorded and transcribed verbatim	Assumable that allgroup meetings or interviews were recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings or interviews were recorded and transcribed verbatimor recordings not transcribed verbatim or only notes were made during the group meetings/ interviews	No recording and no notes	Not applica ble		
Analyses							
Was an appropriate approach used to analyse the data?	A widely recognized or well justified approach was used	Assumable that the approach was appropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate			
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only oneresearcher involvedin the analysis				
2d. Asking professionals about relevance							

Was an appropriate method used to ask professionals whether each item is relevant for the construct of interest?	A widely recognized or well justified approach was used	Only quantitative (survey) method(s) used or assumable that the method was appropriate but not clearly described	Not clear if professionals were asked whether each item is relevant or doubtful if the method was appropriate	Method used not appropriate or professionals not asked about the relevance of all items			
Were professionals from all relevant disciplines included?	Professionals from all required disciplines were included	Assumable that professionals fromall required disciplines were included, but not clearly described	Doubtful whether professionals fromall required disciplines were included or relevant professionals werenot included				
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear				
Analyses							
Was an appropriate approach used to analyse the data?	A widely recognized or well justified approach was used	Assumable that the approach was appropriate, but not	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate			

		clearly described					
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only oneresearcher involvedin the analysis				
2e. Asking professionals about comprehensiveness							
Design requirements							
Was an appropriate method used to for assessing comprehensiveness of the PROM?	A widely recognized or well justified approach was used	Only quantitative (survey) method(s) used or assumable that the method was appropriate but not clearly described	Not clear if professionals were asked whether each item is relevant or doubtful if the method was appropriate	Method used not appropriate or professionals not asked about the relevance of all items			
Were professionals from all relevant disciplines included?	Professionals from all required disciplines were included	Assumable that professionals fromall required disciplines were included, but not	Doubtful whether professionals fromall required disciplines were included or relevant professionals werenot included				

		clearly described					
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	≥7 ≥50	4-6 ≥30	<4 or not clear <30 or not clear				
Analyses							
Was an appropriate approach used to analyse the data?	A widely recognized or well justified approach was used	Assumable that the approach was appropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate			
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only oneresearcher involvedin the analysis				
Box 3. Structural Validity – Not tested							
Box 4. Internal Consistency							
Does the scale consist of effect indicators i.e. is it based on a reflective model? Yes or No							
Design Requirements							
Was an internal consistency statistic calculated foreach	Internal consistency	Unclear whether	Internal consistency			(106)	Failed to meet

unidimensional scale or subscale separately?	statistic calculated for each unidimensional scale or subscale	scale or sub scale is unidimensional	statisticNOT calculated for each unidimensional scale or sub scale				acceptable 0.70
Statistical methods							
For continuous scores: Was Cronbach's alpha oromega calculated?	Cronbach's alpha, or Omega calculated		Only item-total correlations calculated	No Cronbach's alpha and no item-total correlations calculated	Not Applicable		
For dichotomous scores: Was Cronbach's alpha or KR-20 calculated?	Cronbach's alpha or KR-20 calculated		Only item-total correlations calculated	No Cronbach's alpha or KR-20 and no item-total correlations calculated	Not Applicable		
For IRT-based scores: Was standard error of the theta(SE (θ)) or reliability coefficient of estimated latent trait value (index of (subject or item) separation) calculated?	SE(θ) or reliability coefficient calculated			SE(θ) or reliability coefficient NOT calculated	Not Applicable		
Other							
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study			None identified but further research is suggested
Box 5: Cross-cultural validity/Measurement invariance							
Design requirements							

Were the samples similar for relevant characteristics except for thegroup variable?	Evidence provided that samples were similar for relevant characteristics except group variable	Stated (but no evidence provided) that samples were similar for relevant characteristi cs except group variable	Unclear whether samples were similar for relevant characteristics exceptgroup variable	Samples were NOT similar for relevant characteristics except group variable			Over 30 languages, VMC process described, used internationa lly
Statistical Methods							
Was an appropriate approach used to analyse the data?	A widely recognizedor well justified approach was used	Assumable that the approach was appropriate, but not clearly described	Not clear what approach was used ordoubtful whether the approach was appropriate	Approach not appropriate	Not Applica ble		
Was the sample size included in the analysis adequate?	Regression analysesor IRT/Rasch based analyses: 200 subjects per group	150 subjects pergroup	100 subjects per group	< 100 subjects pergroup			
	MGCFA*: 7 times the number of itemsand ≥100	5 times the numberof items and ≥100; OR 5-7 times the number of items but <100	5 times the numberof items but <100	<5 times the number of items			
Other							

Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study			
Box 6. Reliability							
Design Requirements							
Were patients stable in the interim period on the construct to be measured?	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable			7 studies
Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate			2d-2w
Were the test conditions similar for the measurements? e.g. typeof administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar			Not always stated
Statistical methods							
For Continuous scores: Was an Intraclass correlation coefficient (ICC) calculated?	ICC calculated and model or formula of the ICC described	ICC calculated but model or formula of the ICC not described or optimal. Pearson or Spearman correlation calculated with evidence provided	Pearson or Spearman correlation coefficient calculated WITHOUT evidence provided that no systematic change has occurred WITH evidence that systematic	No ICC or Pearson or Spearman correlations calculated	Not Applicable		

		that no systematic change has occurred	change has occurred				
For dichotomous/nominal/ordinal scores: Was kappa calculated?	Kappa calculated			No kappa calculated	Not Applicable		
For ordinal scores: Was a weighted kappa calculated?	Weighted Kappa Calculated		Unweighted Kappa calculated or not described		Not Applicable		
For ordinal scores: Was the weighting scheme described? E.g. linear, quadratic	Weighting scheme described	Weighting scheme NOT described					
Other							
Were there any other important flaws in the design or statisticalmethods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws			
Box 7. Measurement error: absolute measures							
Design requirements							
Were patients stable in the interim period on the construct to be measured?	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable			
Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate			
Were the test conditions similar for the measurements? e.g. typeof administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar			
Statistical methods							

For continuous scores: Was the Standard Error of Measurement (SEM), Smallest Detectable Change (SDC) or Limits of Agreement (LoA) calculated?	SEM, SDC, or LOA calculated	Possible to calculate LoA from the data presented		SEM calculated based on Cronbach's alpha, or on SDfrom another population	Not Applicable		
For dichotomous/ nominal/ordinal scores: Was the percentage(positive and negative) agreement calculated?	% positive and negative agreement calculated	% positive agreement calculated		% agreement not calculated	Not Applicable		
Other							
Were there any other important flaws in the design or statisticalmethods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws			
Box 8: Criterion Validity							
Statistical Methods							
For continuous scores: Were correlations, or the area under the receiver operating curve calculated?	Correlations or AUC calculated			Correlations or AUC NOT calculated	Not Applicable		
For dichotomous scores: Were sensitivity and specificity determined?	Sensitivity and specificity calculated			Sensitivity and specificity NOT calculated	Not Applicable		
Other							
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or			

	execution of the study			execution of the study			
Box 9: Hypothesis testing for construct validity							
9a. Comparison with other outcome measurement instruments (convergent validity)							7 studies
Design requirements							
Is it clear what the comparator instrument(s) measure(s)?	Constructs measured by the comparator instrument(s) is clear		Constructs measured by the comparator instrument(s) is not clear				
Were the measurement properties of the comparator instrument(s) sufficient?	Sufficient measurement properties of the comparator instrument(s) in a population similar tothe study population	Sufficient measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties of the comparator instrument(s) in any study population	No information on the measurement properties of the comparator instrument(s), OR evidence for insufficient measurement properties of the comparator instrument(s)			
Statistical methods							
Was the statistical method appropriate for the hypotheses to betested?	Statistical method was appropriate	Assumable that statistical methodwas appropriate	Statistical method applied NOT optimal	Statistical methodapplied NOT appropriate			

Other							
Were there any other important flaws in the design or statisticalmethods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws			
9b. Comparison between subgroups (discriminative or known-groups validity)							13 studies
Design requirements							
Was there adequate description provided of important characteristics of the subgroups?	Adequate description of the important characteristics of the subgroups	Adequate description of most of the important characteristics of the subgroups	Poor description of the important characteristics of the subgroups				
Statistical methods							
Was the statistical method appropriate for the hypotheses to betested?	Statistical method was appropriate	Assumable that statistical methodwas appropriate	Statistical method applied NOT optimal	Statistical methodapplied NOT appropriate			
Other							
Were there any other important flaws in the design or statisticalmethods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws			
Box 10: Responsiveness							
10a. Criterion approach (ie. comparison to a gold standard)							
Statistical Methods							

For continuous scores: Were correlations between change scores,or the area under the Receiver Operator Curve (ROC) curve calculated?	Correlations or Areaunder the ROC Curve (AUC) calculated			Correlations or AUC NOT calculated	Not Applicable	(13)	
For dichotomous scales: Were sensitivity and specificity (changedversus not changed) determined?	Sensitivity and specificity calculated			Sensitivity and specificity NOT calculated	Not Applicable	(13)	
Other							
Were there any other important flaws in the design or statisticalmethods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws		(13)	
10b. Construct approach (i.e. hypotheses testing; comparison with other outcome measurement instruments)							
Is it clear what the comparator instrument(s) measure(s)?	Constructs measured by the comparator instrument(s) is clear		Constructs measured by the comparator instrument(s) is not clear			(13)	Compared to JAMAR (disease-specific)
Were the measurement properties of the comparator instrument(s) sufficient?	Sufficient measurement properties of the comparator instrument(s) in a population similar tothe	Sufficient measurement properties of the comparator instrument(s) but not sure if these apply to the	Some information on measurement properties of the comparator instrument(s) in any study population	NO information onthe measurement properties of the comparator instrument(s) OR evidence of poor		(13)	

	study population	study population		quality of comparator instrument(s)			
Statistical Methods							
Was the statistical method appropriate for the hypotheses to betested?	Statistical methodwas appropriate	Assumable that statistical method were appropriate	Statistical method applied NOT optimal	Statistical method applied NOT appropriate		(13)	
Other							
Were there any other important flaws in the design or statisticalmethods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws		(13)	
10c. Construct approach: (i.e. hypotheses testing: comparison between subgroups)							
Design requirements							
Was an adequate description provided of important characteristicsof the subgroups?	Adequate description of the important characteristics of the subgroups	Adequate description of most of the important characteristics of the subgroups	Poor or no descriptionof the important characteristics of the subgroups			(13)	
Statistical Methods							
Was the statistical method appropriate for the hypotheses to betested?	Statistical methodwas appropriate	Assumable that statistical method was appropriate	Statistical method applied NOT optimal	Statistical method applied NOT appropriate		(13)	

Other							
Were there any other important flaws in the design or statisticalmethods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws		(13)	
10d. Construct approach: (i.e. hypotheses testing: before and after intervention)	No intervention administered therefore not tested						

Appendix 2: COSMIN Risk of Bias Checklist – PedsQL

(level of performance highlighted in grey)

Box General requirement for studies that applied Item Response Theory (IRT) Models							
	Very good	**Adequate**	**Doubtful**	**Inadequate**	**Not applicable**	**Reference**	**Notes**
Box 1. PROM development							
1a. PROM design						(1,19,96,178,179) (93)	
General design requirements							
Is a clear description provided of the construct to be measured?	Construct clearly described			Construct not clearly described	Not Applicable		
Is the origin of the construct clear: was a theory, conceptual framework or disease model used or clear rationale provided todefine the construct to be measured?	Origin of the construct clear		Origin of the construct not clear				
Is a clear description provided of the target population for whichthe PROM was developed?	Target population clearly described			Target population not clearly described			
Is a clear description provided of the context of use	Context of use clearly described		Context of use not clearly described				
Was the PROM development study performed in a sample representing the target population for which the PROM was developed?	Study performed in a sample representing the target population	Assumable that the study was performed in a sample	Doubtful whether the study was performed ina sample	Study not performed in a sample representingthe			

		representing the target population, butnot clearly described	representing the target population	target population **(SKIP items 6-12)**			
Concept elicitation (relevance and comprehensiveness)							
Was an appropriate qualitative data collection method used to identify relevant items for a new PROM?	Widely recognized or well justified qualitative methodused, suitable for the construct and study population	Assumable thatthe qualitative method was appropriate and suitable forthe construct and study population, butnot clearly described	Only quantitative (survey) method(s) used or doubtful whether the method was suitable for the construct and study population	Method used not appropriate or not suitable for the construct or study population			
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewersused	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers nottrained and no experience		Not Applicable		
Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate,	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable		

		but not clearly described					
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and transcribed verbatim	Assumable that allgroup meetings or interviews were recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings or interviews were recorded and transcribed verbatimor recordings not transcribed verbatim or only notes were made during the group meetings/ interviews	No recording and no notes	Not applicable		
Was an appropriate approach used to analyse the data?	A widely recognized or welljustified approachwas used	Assumable that the approach was appropriate, but not clearly described	Not clear what approach was used ordoubtful whether the approach was appropriate	Approach not appropriate			
Was at least part of the data coded independently?	At least 50% of thedata was coded byat least two researchers independently	11-49% of the data was codedby at least two researchers independently	Doubtful if two researchers were involved in the codingor only 1-10% of the data was coded by at least two researchers independently	Only one researcher was involvedin coding or no coding			
Was data collection continued until saturation was reached?	Evidence providedthat saturation was reached	Assumable that saturation was reached	Doubtful whether saturation was reached	Evidence suggests that saturation was not reached	Not applicable		
For quantitative studies (surveys): was the sample sizeappropriate?	≥100	50-99	30-49	<30	Not applicable		

1b. Cognitive interview study or other pilot test							
Was a cognitive interview study or other pilot test conducted?	YES			**NO (SKIP items 15-35**			
General design requirements							
Was the cognitive interview study or other pilot test performed ina sample representing the target population?	Study performed in a sample representing the target population	Assumable that the study was performed in a sample representing the target population, butnot clearly described	Doubtful whether the study was performed in a sample representing the target population	Study not performed in a sample representingthe target population			
Comprehensibility							
Were patients asked about the comprehensibility of the PROM?	YES	Not clear (SKIP standards 17-25)	No **(SKIP standards 17-25)**				
Were all items tested in their final form?	All items were tested in their finalform	Assumable that all items were tested in their final form, but not clearly described	Not clear if all itemswere tested in theirfinal form	Items were not tested in their final form or itemswere not re- tested after substantial adjustments			
Was an appropriate qualitative method used for assessing the comprehensibility of the PROM instructions, items, response options, and recall period?	Widely recognized or well justified qualitative method used	Assumable that the method was appropriate butnot clearly described	Only quantitative (survey) method (s) used or doubtful whether the method was appropriate or not clear if patients were asked about the	Method used not appropriate or patients were not asked about the comprehensibility of the items, response options or recall period or patients			

			comprehensibility of the items, response options or recall period or patients not asked about the comprehensibility of the PROM instructions				
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear				
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewersused	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers not trained and no experience		Not Applicable		
Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate, but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable		
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and	Assumable that allgroup meetings or interviews were recorded	Not clear if all group meetings or interviews were recorded and transcribed	No recording and no notes	Not applicable		

	transcribed verbatim	and transcribed verbatim, but not clearly described	verbatimor recordings not transcribed verbatim or only notes were made during the group meetings/ interviews				
Was an appropriate approach used to analyse the data?	A widely recognized or well justified approach was used	Assumable that the approach was appropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate			
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only oneresearcher involvedin the analysis				
Were problems regarding the comprehensibility of the PROM instructions, items, response options, and recall period appropriately addressed by adapting the PROM?	No problems found or problems appropriately addressed and PROM was adapted and re-tested if necessary	Assumable that there were no problems or that problems were appropriately addressed, but not clearly described	Not clear if there were problems or doubtful if problems were appropriately addressed	Problems not appropriately addressed or PROM was adapted but items were not re-tested after substantial adjustments	Not applicable		
Comprehensiveness							
Were patients asked about the comprehensiveness of the PROM?	YES		NO or not clear **(SKIP items 27-35)**				

Was the final set of items tested?	The final set of items was tested	Assumable that the final set of items was tested but not clearly described	Not clear if the final set of items was tested or the final set of items was not tested or the set of items were not re-tested after items were removed or added				
Was an appropriate method used for assessing the comprehensiveness of the PROM?	Widely recognized method used	Assumable that method was appropriate but not clearly described or only quantitative (survey) method (s) used	Doubtful whether the method was appropriate or methodused not appropriate				
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear				
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewersused	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers not trained and no experience		Not Applicable		
Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview	Not clear if a topic guide was used or doubtful if topic or		Not Applicable		

		guide was appropriate, but not clearly described	interview guide wasappropriate or no guide				
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and transcribed verbatim	Assumable that allgroup meetings or interviews were recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings or interviews were recorded and transcribed verbatimor recordings not transcribed verbatim or only notes were made during the group meetings/ interviews	No recording and no notes	Not applicable		
Was an appropriate approach used to analyse the data?	A widely recognized or well justified approach was used	Assumable that the approach was appropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate			
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only oneresearcher involvedin the analysis				
Were problems regarding the comprehensiveness of the PROM appropriately addressed by adapting the PROM?	No problems found or problems were appropriately addressed and	Assumable that there were problems or that problems were	Not clear if there were problems or doubtful if the problems were appropriately	Problems not appropriately addressed	Not applicable		

	PROM adapted and re-tested if necessary	appropriately addressed but not clearly described	addressed or PROM was adapted but items were not re-tested after substantial adjustments				
Box 2 Content validity							
2a. Asking patients about relevance						(6)	
Design requirements							
Was an appropriate method used to ask patients whether eachitem is relevant for their experience with the condition?	Widely recognized or well justified method used	Only quantitative (survey) method(s) used or assumable that the method was appropriate but not clearly described	Not clear if patients were asked whether each item is relevantor doubtful whether the method was appropriate	Method used not appropriate or patients not asked about the relevance of all items			
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear				
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewersused	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers not trained and no experience		Not Applicable		

Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate, but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable		
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and transcribed verbatim	Assumable that allgroup meetings or interviews were recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings or interviews were recorded and transcribed verbatimor recordings not transcribed verbatim or only notes were made during the group meetings/ interviews	No recording and no notes	Not applicable		
Analyses							
Was an appropriate approach used to analyse the data?	A widely recognized or well justified approach was used	Assumable that the approach was appropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate			
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only oneresearcher involvedin the analysis				

2b. Asking patients about comprehensiveness						(6)	
Was an appropriate method used for assessing the comprehensiveness of the PROM?	Widely recognized or well justified method used	Only quantitative (survey) method(s)used or assumable that the method was appropriate but not clearly described	Doubtful whether the method was appropriate	Method used not appropriate			
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear				
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewersused	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers nottrained and no experience		Not Applicable		
Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate, but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable		
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were	Assumable that allgroup meetings or	Not clear if all group meetings or interviews were	No recording and no notes	Not applicable		

	recorded and transcribed verbatim	interviews were recorded and transcribed verbatim, but not clearly described	recorded and transcribed verbatimor recordings not transcribed verbatim or only notes were made during the group meetings/ interviews				
Analyses							
Was an appropriate approach used to analyse the data?	A widely recognized or well justified approach was used	Assumable that the approach was appropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate			
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only oneresearcher involvedin the analysis				
2c. Asking patients about comprehensibility						(6)	
Was an appropriate qualitative method used for assessing the comprehensibility of the PROM instructions, items,	Widely recognized or well justified qualitative method used	Assumable that the method was appropriate butnot clearly described	Only quantitative (survey) method (s) used or doubtful whether the method was appropriate or not clear if patients were asked about	Method used not appropriate or patients were not asked about the comprehensibility of the items, response options			

response options, and recall period?			the comprehensibility of the items, response options or recall period or patients not asked about the comprehensibility of the PROM instructions	or recall period or patients			
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear				
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewersused	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers nottrained and no experience				
Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate, but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable		
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and	Assumable that allgroup meetings or interviews were recorded	Not clear if all group meetings or interviews were recorded and transcribed	No recording and no notes	Not applicable		

	transcribed verbatim	and transcribed verbatim, but not clearly described	verbatimor recordings not transcribed verbatim or only notes were made during the group meetings/ interviews				
Analyses							
Was an appropriate approach used to analyse the data?	A widely recognized or well justified approach was used	Assumable that the approach was appropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate			
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only oneresearcher involvedin the analysis				
2d. Asking professionals about relevance						(6)	
Was an appropriate method used to ask professionals whether each item is relevant for the construct of interest?	A widely recognized or well justified approach was used	Only quantitative (survey) method(s) used or assumable that the method was appropriate	Not clear if professionals were asked whether each item is relevant or doubtful if the method was appropriate	Method used not appropriate or professionals not asked about the relevance of all items			

		but not clearly described					
Were professionals from all relevant disciplines included?	Professionals from allrequired disciplines were included	Assumable that professionals fromall required disciplines were included, but not clearly described	Doubtful whether professionals from all required disciplines were included or relevant professionals were not included				
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear				
Analyses							
Was an appropriate approach used to analyse the data?	A widely recognized or well justified approach was used	Assumable that the approach was appropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate			
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only oneresearcher involvedin the analysis				
2e. Asking professionals about comprehensiveness						(6)	
Design requirements							
Was an appropriate method used to for assessing	A widely recognized or	Only quantitative	Not clear if professionals were	Method used not appropriate or			

comprehensiveness of the PROM?	well justified approach was used	(survey) method(s) used or assumable that the method was appropriate but not clearly described	asked whether each item is relevant or doubtful if the method was appropriate	professionals not asked about the relevance of all items			
Were professionals from all relevant disciplines included?	Professionals from allrequired disciplines were included	Assumable that professionals fromall required disciplines were included, but not clearly described	Doubtful whether professionals from all required disciplines were included or relevant professionals were not included				
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear				
Analyses							
Was an appropriate approach used to analyse the data?	A widely recognized or well justified approach was used	Assumable that the approach was appropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate			
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis,	Not clear if two researchers were included in the analysis or only oneresearcher involvedin the analysis				

		but not clearly described					
Box 3. Structural Validity							
Does the scale consist of effect indicators, i.e. is it based on a reflective model? Yes/No							
Does the study concern unidimensionality or structural validity?							
Statistical methods						(180–182)	
For CTT: Was exploratory or confirmatory factor analysis performed?	Confirmatory factor analysis performed	Exploratory factor analysis performed		No exploratory or confirmatory factoranalysis performed	Not Applicable		3 studies
For IRT: Were IRT/Rasch: does the chosen model fit to the research topic	Chosen model fits wellto the research question	Assumable that the chosen model fits wellto the research question	Doubtful if the chosen model fits well to the research question	Chosen model doesnot fit to the research question	Not Applicable		
	FA: 7 times the numberof items and ≥100	FA: at least 5 times the number of items and ≥100; OR at least6 times number of items but <100	FA: 5 times the number of items but <100	FA: < 5 times the number of items			
	Rasch/1PL models: ≥ 200 subjects	Rasch/1PL models:100-199 subjects	Rasch/1PL models: 50-99 subjects	Rasch/1PL models: <50 subjects			
	2PL parametric IRT models OR Mokken scale analysis: ≥ 1000 subjects	2PL parametric IRT models OR Mokkenscale analysis: 500-999 subjects	2PL parametric IRT models OR Mokkenscale analysis: 250- 499 subjects	2PL parametric IRT models OR Mokken scale analysis: < 250 subjects			

Other							
Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws			
Box 4. Internal Consistency							
Does the scale consist of effect indicators i.e. is it based on a reflective model? Yes or No						(1)	
Design Requirements							
Was an internal consistency statistic calculated foreach unidimensional scale or subscale separately?	Internal consistency statistic calculated for each unidimensional scale or subscale	Unclear whether scale or sub scale is unidimensional	Internal consistency statisticNOT calculated for each unidimensional scale or sub scale				
Statistical methods							
For continuous scores: Was Cronbach's alpha oromega calculated?	Cronbach's alpha, or Omega calculated		Only item-total correlations calculated	No Cronbach's alpha and noitem-total correlations calculated	Not Applicable		
For dichotomous scores: Was Cronbach's alpha or KR-20 calculated?	Cronbach's alpha or KR-20 calculated		Only item-total correlations calculated	No Cronbach's alpha or KR-20 and no item-total correlations calculated	Not Applicable		
For IRT-based scores: Was standard error of the theta(SE (θ)) or reliability coefficient of estimated latent trait value (index of (subject or item) separation) calculated?	SE(θ) or reliability coefficient calculated			SE(θ) or reliability coefficient NOT calculated	Not Applicable		
Other							
Were there any important flaws in the design or methods of the study?	No other important methodological		Other minor methodological flaws in the design	Other important methodological flaws in the design			

	flaws in the design or execution of the study		or execution of the study	or execution of the study			
Box 5: Cross-cultural validity/Measurement invariance							
Design requirements							
Were the samples similar for relevant characteristics except for thegroup variable?	Evidence provided that samples were similar for relevant characteristics except group variable	Stated (but no evidence provided)that samples were similar for relevant characteristics except group variable	Unclear whether samples were similar for relevant characteristics exceptgroup variable	Samples were NOT similar forrelevant characteristics except group variable			Over 30 languages, VMC process described
Statistical Methods							
Was an appropriate approach used to analyse the data?	A widely recognizedor well justified approach was used	Assumable that the approach was appropriate, but notclearly described	Not clear what approach was used ordoubtful whether the approach was appropriate	Approach not appropriate	Not Applicable		
Was the sample size included in the analysis adequate?	Regression analysesor IRT/Rasch based analyses: 200 subjects per group	150 subjects pergroup	100 subjects per group	< 100 subjects per group			
	MGCFA*: 7 times the number of itemsand ≥100	5 times the numberof items and ≥100; OR 5-7 times the number of items but <100	5 times the numberof items but <100	<5 times the number of items			
Other							

Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study			
Box 6. Reliability							
Design Requirements						(115)	
Were patients stable in the interim period on the construct to be measured?	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable			
Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate			
Were the test conditions similar for the measurements? e.g. typeof administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar			
Statistical methods							
For Continuous scores: Was an Intraclass correlation coefficient (ICC) calculated?	ICC calculated and model or formula of the ICC described	ICC calculated but model or formula of the ICC not described or optimal. Pearson or Spearman correlation calculated with evidence provided that no systematic change has occurred	Pearson or Spearman correlation coefficient calculated WITHOUT evidence provided that no systematic change has occurred WITH evidence that systematic change has occurred	No ICC or Pearson or Spearman correlations calculated	Not Applicable		

For dichotomous/nominal/ordinal scores: Was kappa calculated?	Kappa calculated			No kappa calculated	Not Applicable		
For ordinal scores: Was a weighted kappa calculated?	Weighted Kappa Calculated		Unweighted Kappa calculated or not described		Not Applicable		
For ordinal scores: Was the weighting scheme described? E.g. linear, quadratic	Weighting scheme described	Weighting scheme NOT described					
Other							
Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws			
Box 7. Measurement error: absolute measures							
Design requirements							
Were patients stable in the interim period on the construct to be measured?	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable			
Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate			
Were the test conditions similar for the measurements? e.g. typeof administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar			
Statistical methods							
For continuous scores: Was the Standard Error of Measurement (SEM), Smallest Detectable Change (SDC) or Limits of Agreement(LoA) calculated?	SEM, SDC, or LOA calculated	Possible to calculate LoA from the data presented		SEM calculated based on Cronbach's alpha, or on SDfrom another population	Not Applicable		

For dichotomous/ nominal/ordinal scores: Was the percentage (positive and negative) agreement calculated?	% positive and negative agreement calculated	% positive agreement calculated		% agreement not calculated	Not Applicable		
Other							
Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws			
Box 8: Criterion Validity							
Statistical Methods							
For continuous scores: Were correlations, or the area under the receiver operating curve calculated?	Correlations or AUC calculated			Correlations or AUC NOT calculated	Not Applicable		
For dichotomous scores: Were sensitivity and specificity determined?	Sensitivity and specificity calculated			Sensitivity and specificity NOT calculated	Not Applicable		
Other							
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study			
Box 9: Hypothesis testing for construct validity							
9a. Comparison with other outcome measurement instruments (convergent validity)						(70)	
Design requirements							
Is it clear what the comparator instrument(s) measure(s)?	Constructs measured by the comparator instrument(s) is clear		Constructs measured by the comparator instrument(s) is not clear				

Were the measurement properties of the comparatorinstrument(s) sufficient?	Sufficient measurement properties of the comparator instrument(s) in a population similar tothe study population	Sufficient measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information onmeasurement properties of the comparator instrument(s) in any study population	No information on the measurement properties of the comparator instrument(s), OR evidence for insufficient measurement properties of the comparator instrument(s)			
Statistical methods							
Was the statistical method appropriate for the hypotheses to betested?	Statistical method was appropriate	Assumable that statistical methodwas appropriate	Statistical method applied NOT optimal	Statistical method applied NOT appropriate			
Other							
Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws			
9b. Comparison between subgroups (discriminative or known-groups validity)							
Design requirements							
Was there adequate description provided of important characteristics of the subgroups?	Adequate description of the important characteristics of the subgroups	Adequate description of most of the important characteristics of the subgroups	Poor description of the important characteristics of the subgroups				
Statistical methods							

Was the statistical method appropriate for the hypotheses to betested?	Statistical method was appropriate	Assumable that statistical methodwas appropriate	Statistical method applied NOT optimal	Statistical method applied NOT appropriate			
Other							
Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws			
Box 10: Responsiveness							
10a. Criterion approach (ie. comparison to a gold standard)						(77)	1 study – no intervention, just initial testing and follow-up
Statistical Methods							
For continuous scores: Were correlations between change scores,or the area under the Receiver Operator Curve (ROC) curve calculated?	Correlations or Areaunder the ROC Curve (AUC) calculated			Correlations or AUC NOT calculated	Not Applicable		
For dichotomous scales: Were sensitivity and specificity (changed versus not changed) determined?	Sensitivity and specificity calculated			Sensitivity and specificity NOT calculated	Not Applicable		
Other							
Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws			
10b. Construct approach (i.e. hypotheses testing; comparison with other outcome measurement instruments)							
Is it clear what the comparator instrument(s) measure(s)?	Constructs measured by the comparator		Constructs measured by the comparator				

	instrument(s) is clear		instrument(s) is not clear				
Were the measurement properties of the comparatorinstrument(s) sufficient?	Sufficient measurement properties of the comparator instrument(s) in a population similar tothe study population	Sufficient measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information onmeasurement properties of the comparator instrument(s) in any study population	NO information onthe measurement properties of the comparator instrument(s) OR evidence of poor quality of comparator instrument(s)			
Statistical Methods							
Was the statistical method appropriate for the hypotheses to betested?	Statistical methodwas appropriate	Assumable that statistical method were appropriate	Statistical method applied NOT optimal	Statistical methodapplied NOT appropriate			
Other							
Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws			
10c. Construct approach: (i.e. hypotheses testing: comparison between subgroups)							
Design requirements							
Was an adequate description provided of important characteristicsof the subgroups?	Adequate description of the important characteristics of the subgroups	Adequate description of most of the important characteristics of the subgroups	Poor or no descriptionof the important characteristics of the subgroups				
Statistical Methods							

Was the statistical method appropriate for the hypotheses to betested?	Statistical methodwas appropriate	Assumable that statistical method was appropriate	Statistical method applied NOT optimal	Statistical methodapplied NOT appropriate			
Other							
Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws			
10d. Construct approach: (i.e. hypotheses testing: before and after intervention)	No intervention administered therefore not tested						

Appendix 3: COSMIN Risk of Bias Checklist – HUI_3

(level of performance highlighted in grey)

Box General requirement for studies that applied Item Response Theory (IRT) Models							
	Very good	**Adequate**	**Doubtful**	**Inadequate**	**Not applicable**	**Reference**	**Notes**
Box 1. PROM development							
1a. PROM design							
General design requirements							
Is a clear description provided of the construct to be measured?	Construct clearly described			Construct not clearly described	Not Applicable	(79,86,104,183,184)	
Is the origin of the construct clear: was a theory, conceptual framework or disease model used or clear rationale provided todefine the construct to be measured?	Origin of the construct clear		Origin of the construct not clear			(79,86,104,183,184)	From original HUI
Is a clear description provided of the target population for whichthe PROM was developed?	Target population clearly described			Target population not clearly described		(79,86,104,183,184)	Various paediatric populations
Is a clear description provided of the context of use	Context of use clearly described		Context of use not clearly described			(79,86,104,183,184)	Health-related quality of life
Was the PROM development study performed in a sample representing the target population for which the PROM was developed?	Study performed in a sample representing the target population	Assumable that the study was performed in a sample	Doubtful whether the study was performed ina sample	Study not performed in a sample representingthe		(79,86,104,183,184)	

		representing the target population, butnot clearly described	representing the target population	target population **(SKIP items 6-12)**			
Concept elicitation (relevance and comprehensiveness)							
Was an appropriate qualitative data collection method used to identify relevant items for a new PROM?	Widely recognized or well justified qualitative methodused, suitable for the construct and study population	Assumable thatthe qualitative method was appropriate and suitable forthe construct and study population, butnot clearly described	Only quantitative (survey) method(s) used or doubtful whether the method was suitable for the construct and study population	Method used not appropriate or not suitable for the construct or study population		(79,86,104,183,184)	
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewersused	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers nottrained and no experience		Not Applicable	(79,86,104,183,184)	Health professionals involved but not clear if they were trained for HUI SC version used therefore ?need for interviewer
Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview	Not clear if a topic guide was used or doubtful		Not Applicable	(79,86,104,183,184)	

		guide was appropriate, but not clearly described	if topic or interview guide wasappropriate or no guide				
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and transcribed verbatim	Assumable that allgroup meetings or interviews were recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings or interviews were recorded and transcribed verbatimor recordings not transcribed verbatim or only notes were made during the group meetings/ interviews	No recording and no notes	Not applicable	(79,86,104,183,184)	
Was an appropriate approach used to analyse the data?	A widely recognized or welljustified approachwas used	Assumable thatthe approach was appropriate, but not clearly described	Not clear what approach was used ordoubtful whether the approach was appropriate	Approach not appropriate		(79,86,104,183,184)	
Was at least part of the data coded independently?	At least 50% of thedata was coded byat least two researchers independently	11-49% of the data was codedby at least two researchers independently	Doubtful if two researchers were involved in the codingor only 1-10% of the data was coded by at least two researchers independently	Only one researcher was involvedin coding or no coding		(79,86,104,183,184)	
Was data collection continued until saturation was reached?	Evidence providedthat saturation was reached	Assumable thatsaturation was reached	Doubtful whether saturation was reached	Evidence suggests that saturation was not reached	Not applicable		

For quantitative studies (surveys): was the sample sizeappropriate?	≥100	50-99	30-49	<30	Not applicable	(79,86,104,183,184)	
1b. Cognitive interview study or other pilot test							
Was a cognitive interview study or other pilot test conducted?	YES			**NO (SKIP items 15-35**		(79,86,104,183,184)	
General design requirements							
Was the cognitive interview study or other pilot test performed ina sample representing the target population?	Study performed in a sample representing the target population	Assumable that the study was performed in a sample representing the target population, butnot clearly described	Doubtful whether the study was performed in a sample representing the target population	Study not performed in a sample representingthe target population			
Comprehensibility							
Were patients asked about the comprehensibility of the PROM?	YES	Not clear (SKIP standards 17-25)	No **(SKIP standards 17-25)**				
Were all items tested in their final form?	All items were tested in their finalform	Assumable thatall items were tested in their final form, but not clearly described	Not clear if all itemswere tested in theirfinal form	Items were not tested in their final form or itemswere not re- tested after substantial adjustments			
Was an appropriate qualitative method used for assessing the comprehensibility of the PROM	Widely recognized or well justified qualitative method used	Assumable that the method was appropriate butnot clearly described	Only quantitative (survey) method (s) used or doubtful whether the method was appropriate or	Method used not appropriate or patients were not asked about the comprehensibility of the items,			

instructions, items, response options, and recall period?			not clear if patients were asked about the comprehensibility of the items, response options or recall period or patients not asked about the comprehensibility of the PROM instructions	response options or recall period or patients			
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear				
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewersused	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers not trained and no experience		Not Applicable		
Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate, but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable		

Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and transcribed verbatim	Assumable that allgroup meetings or interviews were recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings or interviews were recorded and transcribed verbatimor recordings not transcribed verbatim or only notes were made during the group meetings/ interviews	No recording and no notes	Not applicable		
Was an appropriate approach used to analyse the data?	A widely recognized or well justified approach was used	Assumable that the approach was appropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate			
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only oneresearcher involvedin the analysis				
Were problems regarding the comprehensibility of the PROM instructions, items, response options, and recall period appropriately addressed by adapting the PROM?	No problems found or problems appropriately addressed and PROM was adapted and re-tested if necessary	Assumable thatthere were no problems or that problems were appropriately addressed, but not clearly described	Not clear if there were problems or doubtful if problems were appropriately addressed	Problems not appropriately addressed or PROM was adapted but items were not re-tested after substantial adjustments	Not applicable		

Comprehensiveness							
Were patients asked about the comprehensiveness of the PROM?	YES		NO or not clear **(SKIP items 27-35)**				
Was the final set of items tested?	The final set of items was tested	Assumable that the final set of items was tested but not clearly described	Not clear if the final set of items was tested or the final set of items was not tested or the set of items were not re-tested after items were removed or added				
Was an appropriate method used for assessing the comprehensiveness of the PROM?	Widely recognized method used	Assumable that method was appropriate but not clearly described or only quantitative (survey) method (s) used	Doubtful whether the method was appropriate or methodused not appropriate				
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear				
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewersused	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers		Not Applicable		

			not trained and no experience				
Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate, but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable		
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and transcribed verbatim	Assumable that allgroup meetings or interviews were recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings or interviews were recorded and transcribed verbatimor recordings not transcribed verbatim or only notes were made during the group meetings/ interviews	No recording and no notes	Not applicable		
Was an appropriate approach used to analyse the data?	A widely recognized or well justified approach was used	Assumable that the approach was appropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate			
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis,	Not clear if two researchers were included in the analysis or only oneresearcher involvedin the analysis				

		but not clearly described					
Were problems regarding the comprehensiveness of the PROM appropriately addressed by adapting the PROM?	No problems found or problems were appropriately addressed and PROM adapted and re-tested if necessary	Assumable that there were problems or that problems were appropriately addressed but not clearly described	Not clear if there were problems or doubtful if the problems were appropriately addressed or PROM was adapted but items were not re-tested after substantial adjustments	Problems not appropriately addressed	Not applicable		
Box 2 Content validity – Not tested							
Box 3. Structural Validity – Not tested							
Box 4. Internal Consistency							
Does the scale consist of effect indicators i.e. is it based on a reflective model? Yes or No							
Design Requirements							Systemic review
Was an internal consistency statistic calculated foreach unidimensional scale or subscale separately?	Internal consistency statistic calculated for each unidimensional scale or subscale	Unclear whether scale or sub scale is unidimensional	Internal consistency statisticNOT calculated for each unidimensional scale or sub scale			(79)	=0/71-0/79 =0.51
Statistical methods							
For continuous scores: Was Cronbach's alpha oromega calculated?	Cronbach's alpha, or Omega calculated		Only item-total correlations calculated	No Cronbach's alpha and no item-total correlations calculated	Not Applicable	(79)	

For dichotomous scores: Was Cronbach's alpha or KR-20 calculated?	Cronbach's alpha or KR-20 calculated		Only item-total correlations calculated	No Cronbach's alpha or KR-20 and no item-total correlations calculated	Not Applicable		
For IRT-based scores: Was standard error of the theta(SE (θ)) or reliability coefficient of estimated latent trait value (index of (subject or item) separation) calculated?	SE(θ) or reliability coefficient calculated			SE(θ) or reliability coefficient NOT calculated	Not Applicable		
Other							
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study		(79)	
Box 5: Cross-cultural validity/Measurement invariance							
Design requirements							
Were the samples similar for relevant characteristics except for thegroup variable?	Evidence provided that samples were similar for relevant characteristics except group variable	Stated (but no evidence provided)that samples were similar for relevant characteristics except group variable	Unclear whether samples were similar for relevant characteristics exceptgroup variable	Samples were NOT similar for relevant characteristics except group variable		(79)	Over 35 languages Origin - English
Statistical Methods							
Was an appropriate approach used to analyse the data?	A widely recognizedor well justified approach was used	Assumable that the approach was appropriate, but notclearly described	Not clear what approach was used ordoubtful whether the approach was appropriate	Approach not appropriate	Not Applicable	(79)	

Was the sample size included in the analysis adequate?	Regression analysesor IRT/Rasch based analyses: 200 subjects per group	150 subjects pergroup	100 subjects per group	< 100 subjects pergroup		(79)	
	MGCFA*: 7 times the number of itemsand ≥100	5 times the numberof items and ≥100; OR 5-7 times the number of items but <100	5 times the numberof items but <100	<5 times the number of items		(79)	
Other							
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study			
Box 6. Reliability							
Design Requirements							
Were patients stable in the interim period on the construct to be measured?	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable		(79)	
Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate		(79)	=2-week period =3-month retest period also found acceptable =1-month
Were the test conditions similar for the measurements? e.g. typeof	Test conditions were similar	Assumable that test	Unclear if test conditions were similar	Test conditions were NOT similar			

administration, environment, instructions	(evidence provided)	conditions were similar					
Statistical methods							
For Continuous scores: Was an Intraclass correlation coefficient (ICC) calculated?	ICC calculated and model or formula of the ICC described	ICC calculated but model or formula of the ICC not described or optimal. Pearson or Spearman correlation calculated with evidence provided that no systematic change has occurred	Pearson or Spearman correlation coefficient calculated WITHOUT evidence provided that no systematic change has occurred WITH evidence that systematic change has occurred	No ICC or Pearson or Spearman correlations calculated	Not Applicable	(79)	=0.75 (95% CI 0.65,0.83) =0.81 (95% CI 0.66, 0.90) =0.77
For dichotomous/nominal/ordinal scores: Was kappa calculated?	Kappa calculated			No kappa calculated	Not Applicable	(79)	
For ordinal scores: Was a weighted kappa calculated?	Weighted Kappa Calculated		Unweighted Kappa calculated or not described		Not Applicable	(79)	
For ordinal scores: Was the weighting scheme described? E.g. linear, quadratic	Weighting scheme described	Weighting scheme NOT described				(79)	
Other							
Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws		(79)	
Box 7. Measurement error: absolute measures – Not tested							
Box 8: Criterion Validity – Not tested							
Box 9: Hypothesis testing for construct validity							

9a. Comparison with other outcome measurement instruments (convergent validity)							
Design requirements							
Is it clear what the comparator instrument(s) measure(s)?	Constructs measured by the comparator instrument(s) is clear		Constructs measured by the comparator instrument(s) is not clear			(86,104,183,184)	David- SF6D, BDI-II, CCI Morrow- HUI2vs3 Maddigan- CHR, HUI2,HUI3 Speechley- healthcare resource, HUI3
Were the measurement properties of the comparatorinstrument(s) sufficient?	Sufficient measurement properties of the comparator instrument(s) in a population similar tothe study population	Sufficient measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties of the comparator instrument(s) in any study population	No information on the measurement properties of the comparator instrument(s), OR evidence for insufficient measurement properties of the comparator instrument(s)		(86,104,183,184)	
Statistical methods							
Was the statistical method appropriate for the hypotheses to betested?	Statistical method was appropriate	Assumable that statistical methodwas appropriate	Statistical method applied NOT optimal	Statistical methodapplied NOT appropriate		(86,104,183,184)	
Other							
Were there any other important flaws in the design or statistical methods of the study?	No other important methodological		Other minor important	Other important methodological flaws		(86,104,183,184)	

	flaws		methodological flaws				
9b. Comparison between subgroups (discriminative or known-groups validity)							
Design requirements							
Was there adequate description provided of important characteristics of the subgroups?	Adequate description of the important characteristics of the subgroups	Adequate description of most of the important characteristics of the subgroups	Poor description of the important characteristics of the subgroups			(86)	
Statistical methods							
Was the statistical method appropriate for the hypotheses to betested?	Statistical method was appropriate	Assumable that statistical methodwas appropriate	Statistical method applied NOT optimal	Statistical methodapplied NOT appropriate		(86)	
Other							
Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws		(86)	
Box 10: Responsiveness – Not tested							

Appendix 4: COSMIN Risk of Bias Checklist – Kiddy-KINDL

(level of performance highlighted in grey)

Box General requirement for studies that applied Item Response Theory (IRT) Models							
	Very good	**Adequate**	**Doubtful**	**Inadequate**	**Not applicable**	**Reference/s**	**Notes**
Box 1. PROM development							
1a. PROM design							
General design requirements							
Is a clear description provided of the construct to be measured?	Construct clearly described			Construct not clearly described		(100,113)	
Is the origin of the construct clear: was a theory, conceptual framework or disease model used or clear rationale provided todefine the construct to be measured?	Origin of the construct clear		Origin of the construct not clear			(100,113)	
Is a clear description provided of the target population for whichthe PROM was developed?	Target population clearly described			Target population not clearly described		(100,113)	
Is a clear description provided of the context of use	Context of use clearly described		Context of use not clearly described			(100,113)	Control vs condition group
Was the PROM development study performed in a sample representing the target population for which the PROM wasdeveloped?	Study performed in a sample representing the target population	Assumable that the study was performed in a sample representing	Doubtful whether the study was performed ina sample	Study not performed in a sample representingthe		(100,113)	

		the target population, but not clearly described	representing the target population	target population **(SKIP items 6-12)**			
Concept elicitation (relevance and comprehensiveness)							
Was an appropriate qualitative data collection method used to identify relevant items for a new PROM?	Widely recognized or well justified qualitative methodused, suitable for the construct and study population	Assumable that the qualitative method was appropriate and suitable for the construct and study population, but not clearly described	Only quantitative (survey) method(s) used or doubtful whether the method was suitable for the construct and study population	Method used not appropriate or not suitable for the construct or study population			
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewersused	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers nottrained and no experience		Not Applicable	(113)	
Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate,but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable	(113)	

Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and transcribed verbatim	Assumable that allgroup meetings or interviews were recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings or interviews were recorded and transcribed verbatimor recordings not transcribed verbatim or only notes were made during the group meetings/ interviews	No recording and no notes	Not applicable	(113)	Not mentioned
Was an appropriate approach used to analyse the data?	A widely recognized or well justified approach was used	Assumable that the approach was appropriate, but not clearly described	Not clear what approach was used ordoubtful whether the approach was appropriate	Approach not appropriate		(100,113)	Cronbach's alpha Mann-Whitney U Kruskal-Wallis The Wilcoxon test Spearman correlation
Was at least part of the data coded independently?	At least 50% of thedata was coded byat least two researchers independently	11-49% of the data was coded by at least two researchers independently	Doubtful if two researchers were involved in the codingor only 1-10% of the data was coded by at least two researchers independently	Only one researcher was involvedin coding or no coding		(100,113)	
Was data collection continued until saturation was reached?	Evidence provided that saturation was reached	Assumable that saturation was reached	Doubtful whether saturation was reached	Evidence suggests that saturation was not reached	Not applicable	(100,113)	

For quantitative studies (surveys): was the sample sizeappropriate?	≥100	50-99	30-49	<30	Not applicable	(100,113)	
1b. Cognitive interview study or other pilot test							
Was a cognitive interview study or other pilot test conducted?	YES			**NO (SKIP items 15-35**			
General design requirements							
Was the cognitive interview study or other pilot test performed ina sample representing the target population?	Study performed in a sample representing the target population	Assumable that the study was performed in a sample representing the target population, but not clearly described	Doubtful whether the study was performed in a sample representing the target population	Study not performed in a sample representingthe target population			
Comprehensibility							
Were patients asked about the comprehensibility of the PROM?	YES	Not clear (SKIP standards 17-25)	No **(SKIP standards 17-25)**				
Were all items tested in their final form?	All items were tested in their finalform	Assumable that all items were tested in their final form, but not clearly described	Not clear if all itemswere tested in theirfinal form	Items were not tested in their final form or itemswere not re- tested after substantial adjustments			
Was an appropriate qualitative method used for assessing the comprehensibility of the PROM instructions, items, response options,	Widely recognized orwell justified qualitative method used	Assumable that the method was appropriate but not clearly described	Only quantitative (survey) method (s) used or doubtful whether the method was appropriate or not clear if patients were asked about	Method used not appropriate or patients were not asked about the comprehensibility of the items, response options			

and recall period?			the comprehensibility of the items, response options or recall period or patients not asked about the comprehensibility of the PROM instructions	or recall period or patients			
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear				
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewersused	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers not trained and no experience		Not Applicable		
Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate,but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable		
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and	Assumable that allgroup meetings or interviews were	Not clear if all group meetings or interviews were recorded and	No recording and no notes	Not applicable		

	transcribed verbatim	recorded and transcribed verbatim, but not clearly described	transcribed verbatimor recordings not transcribed verbatim or only notes were made during the group meetings/ interviews				
Was an appropriate approach used to analyse the data?	A widely recognized orwell justified approach wasused	Assumable that the approach wasappropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate			
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only oneresearcher involvedin the analysis				
Were problems regarding the comprehensibility of the PROM instructions, items, response options, and recall period appropriately addressed by adapting the PROM?	No problems found or problems appropriately addressed and PROM was adapted and re-tested if necessary	Assumable that there were no problems or that problems were appropriately addressed, but not clearly described	Not clear if there were problems or doubtful if problems were appropriately addressed	Problems not appropriately addressed or PROM was adapted but items were not re-tested after substantial adjustments	Not applicable		
Comprehensiveness							
Were patients asked about the comprehensiveness of the PROM?	YES		NO or not clear **(SKIP items 27-35)**				

Was the final set of items tested?	The final set of items was tested	Assumable that the final set of items was tested but not clearly described	Not clear if the final set of items was tested or the final set of items was not tested or the set of items were not re-tested after items were removed or added				
Was an appropriate method used for assessing the comprehensiveness of the PROM?	Widely recognized method used	Assumable that method was appropriate but not clearly described or only quantitative (survey) method (s) used	Doubtful whether the method was appropriate or methodused not appropriate				
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear				
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewersused	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers not trained and no experience		Not Applicable		
Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide	Not clear if a topic guide was used or doubtful if topic or		Not Applicable		

		was appropriate,but not clearly described	interview guide wasappropriate or no guide				
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and transcribed verbatim	Assumable that allgroup meetings or interviews were recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings or interviews were recorded and transcribed verbatimor recordings not transcribed verbatim or only notes were made during the group meetings/ interviews	No recording and no notes	Not applicable		
Was an appropriate approach used to analyse the data?	A widely recognized orwell justified approach wasused	Assumable that the approach wasappropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate			
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only oneresearcher involvedin the analysis				
Were problems regarding the comprehensiveness of the PROM appropriately addressed by adapting the PROM?	No problems found or problems were appropriately addressed and PROM adapted	Assumable that there were problems or that problems were appropriately	Not clear if there were problems or doubtful if the problems were appropriately addressed or	Problems not appropriately addressed	Not applicable		

	and re-tested if necessary	addressed but not clearly described	PROM was adapted but items were not re-tested after substantial adjustments				
Box 2 Content validity – Not tested							
Box 3. Structural Validity							
Does the scale consist of effect indicators, i.e. is it based on a reflective model? Yes/No							
Does the study concern unidimensionality or structural validity?							
Statistical methods							
For CTT: Was exploratory or confirmatory factor analysis performed?	Confirmatory factor analysis performed	Exploratory factor analysis performed		No exploratory or confirmatory factoranalysis performed	Not Applicable	(6,100)	
For IRT: Were IRT/Rasch: does the chosen model fit to the research topic	Chosen model fits wellto the research question	Assumable that the chosen model fits well to the research question	Doubtful if the chosen model fits well to the research question	Chosen model doesnot fit to the research question	Not Applicable	(6,100)	
	FA: 7 times the numberof items and ≥100	FA: at least 5 times the number of items and ≥100; OR at least6 times number of items but <100	FA: 5 times the number of items but <100	FA: < 5 times the number of items		(6,100)	
	Rasch/1PL models: ≥ 200 subjects	Rasch/1PL models:100-199 subjects	Rasch/1PL models: 50-99 subjects	Rasch/1PL models: <50 subjects			
	2PL parametric IRT models OR	2PL parametric IRT models OR	2PL parametric IRT models OR	2PL parametric IRT models OR			

	Mokken scale analysis: ≥ 1000 subjects	Mokkenscale analysis: 500-999 subjects	Mokkenscale analysis: 250- 499 subjects	Mokken scale analysis: < 250 subjects			
Other							
Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws		(6,100)	cleft-none identified kindergarten-CFA depicted problems
Box 4. Internal Consistency							
Does the scale consist of effect indicators i.e. is it based on a reflective model? Yes							
Design Requirements							
Was an internal consistency statistic calculated foreach unidimensional scale or subscale separately?	Internal consistency statisticcalculated for each unidimensional scale or subscale	Unclear whether scale or sub scale is unidimensional	Internal consistency statisticNOT calculated for each unidimensional scale or sub scale			(100,113)	
Statistical methods							
For continuous scores: Was Cronbach's alpha oromega calculated?	Cronbach's alpha, or Omega calculated		Only item-total correlations calculated	No Cronbach's alpha and noitem-total correlations calculated	Not Applicable	(100)	
For dichotomous scores: Was Cronbach's alpha or KR-20 calculated?	Cronbach's alpha or KR-20 calculated		Only item-total correlations calculated	No Cronbach's alpha or KR-20 and no item-total correlations calculated	Not Applicable	(6,100,113)	Cleft-0.70 Pre-schoolers-0.71-0.80 Kindergarten-0.75
For IRT-based scores: Was standard error of the theta(SE (θ)) or reliability coefficient of estimated latent trait value (index of (subject or item) separation) calculated?	SE(θ) or reliability coefficient calculated			SE(θ) or reliability coefficient NOT calculated	Not Applicable	(6,100,113)	
Other							

Were there any important flaws in the design or methods of the study?	No other important methodological flaws		Other minor methodological flaws	Other important methodological flaws		(6,100,113)	
Box 5: Cross-cultural validity/Measurement invariance – Not tested							
Box 6. Reliability						(6,113)	
Design Requirements							
Were patients stable in the interim period on the construct to be measured?	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable			Not stated
Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate			Cleft-doubtful Kindergarten-2 week interval
Were the test conditions similar for the measurements? e.g. typeof administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that testconditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar			Not stated
Statistical methods							
For Continuous scores: Was an Intraclass correlation coefficient (ICC) calculated?	ICC calculated and model or formula of the ICC described	ICC calculated but model or formula of the ICC not described or optimal. Pearson or Spearman correlation calculated with evidence provided that no systematic change has occurred	Pearson or Spearman correlation coefficient calculated WITHOUT evidence provided that no systematic change has occurred WITH evidence that systematic change has occurred	No ICC or Pearson or Spearman correlations calculated	Not Applicable		3-ICC=0.83
For dichotomous/nominal/ordinal scores: Was kappa calculated?	Kappa calculated			No kappa calculated	Not Applicable		

For ordinal scores: Was a weighted kappa calculated?	Weighted Kappa Calculated		Unweighted Kappa calculated or not described		Not Applicable		
For ordinal scores: Was the weighting scheme described? E.g. linear, quadratic	Weighting scheme described	Weighting scheme NOT described					
Other							
Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws			
Box 7. Measurement error: absolute measures – Not tested							
Box 8: Criterion Validity – Not tested							
Box 9: Hypothesis testing for construct validity							
9a. Comparison with other outcome measurement instruments (convergent validity)							
Design requirements							
Is it clear what the comparator instrument(s) measure(s)?	Constructs measured by the comparator instrument(s) is clear		Constructs measured by the comparator instrument(s) is not clear			(100)	
Were the measurement properties of the comparatorinstrument(s) sufficient?	Sufficient measurement properties of the comparator instrument(s) in a population similar tothe study population	Sufficient measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information onmeasurement properties of the comparator instrument(s) in any study population	No information on the measurement properties of the comparator instrument(s), OR evidence for insufficient measurement properties of the comparator instrument(s)		(100)	Compared to Preschool anxiety scale (PAS) and generalized anxiety disorder (GAD)
Statistical methods							

Was the statistical method appropriate for the hypotheses to be tested?	Statistical method was appropriate	Assumable that statistical methodwas appropriate	Statistical method applied NOT optimal	Statistical method applied NOT appropriate		(100)	
Other							
Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws		(100)	
9b. Comparison between subgroups (discriminative or known-groups validity) – Not tested							
Box 10: Responsiveness – Not tested							

Appendix 5: COSMIN Risk of Bias Checklist – CHIP CRF/CE

(level of performance highlighted in grey)

Box General requirement for studies that applied Item Response Theory (IRT) Models							
	Very good	**Adequate**	**Doubtful**	**Inadequate**	**Not applicable**	**Reference**	**Notes**
Box 1. PROM development							
1a. PROM design							
General design requirements							
Is a clear description provided of the construct to be measured?	Construct clearly described			Construct not clearly described	Not Applicable	(72)	Domains clearly described
Is the origin of the construct clear: was a theory, conceptual framework or disease model used or clear rationale provided todefine the construct to be measured?	Origin of the construct clear		Origin of the construct not clear			(72)	Adaptation from CHIP AE
Is a clear description provided of the target population for whichthe PROM was developed?	Target population clearly described			Target population not clearly described		(72)	Young children 6-7-years
Is a clear description provided of the context of use	Context of use clearly described		Context of use not clearly described			(72)	
Was the PROM development study performed in a sample representing the target population for which the PROM was developed?	Study performed in a sample representing the target population	Assumable that the study was performed in a sample representing	Doubtful whether the study was performed ina sample	Study not performed in a sample representingthe		(72)	

		the target population, butnot clearly described	representing the target population	target population **(SKIP items 6-12)**			
Concept elicitation (relevance and comprehensiveness)							
Was an appropriate qualitative data collection method used to identify relevant items for a new PROM?	Widely recognized or well justified qualitative methodused, suitable for the construct and study population	Assumable thatthe qualitative method was appropriate and suitable forthe construct and study population, butnot clearly described	Only quantitative (survey) method(s) used or doubtful whether the method was suitable for the construct and study population	Method used not appropriate or not suitable for the construct or study population		(72)	
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewersused	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers nottrained and no experience		Not Applicable	(72)	Not stated – only 'interviewers' used
Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate,but	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable	(72)	

		not clearly described					
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and transcribed verbatim	Assumable that allgroup meetings or interviews were recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings or interviews were recorded and transcribed verbatim or recordings not transcribed verbatim or only notes were made during the group meetings/ interviews	No recording and no notes	Not applicable	(72)	Time of taken to complete noted
Was an appropriate approach used to analyse the data?	A widely recognized or welljustified approachwas used	Assumable that the approach was appropriate, but not clearly described	Not clear what approach was used ordoubtful whether the approach was appropriate	Approach not appropriate		(72)	
Was at least part of the data coded independently?	At least 50% of thedata was coded byat least two researchers independently	11-49% of the data was coded by at least two researchers independently	Doubtful if two researchers were involved in the codingor only 1-10% of the data was coded by at least two researchers independently	Only one researcher was involvedin coding or no coding		(72)	Not described
Was data collection continued until saturation was reached?	Evidence providedthat saturation was reached	Assumable that saturation was reached	Doubtful whether saturation was reached	Evidence suggests that saturation was not reached	Not applicable	(72)	Recruited from various paediatric settings
For quantitative studies (surveys): was the sample sizeappropriate?	≥100	50-99	30-49	<30	Not applicable	(72)	

1b. Cognitive interview study or other pilot test							
Was a cognitive interview study or other pilot test conducted?	YES			**NO (SKIP items 15-35**		(72)	
General design requirements							
Was the cognitive interview study or other pilot test performed ina sample representing the target population?	Study performed in a sample representing the target population	Assumable that the study was performed in a sample representing the target population, butnot clearly described	Doubtful whether the study was performed in a sample representing the target population	Study not performed in a sample representingthe target population			
Comprehensibility							
Were patients asked about the comprehensibility of the PROM?	YES	Not clear (SKIP standards 17-25)	No **(SKIP standards 17-25)**				
Were all items tested in their final form?	All items were tested in their finalform	Assumable that all items were tested in their final form, but not clearly described	Not clear if all itemswere tested in theirfinal form	Items were not tested in their final form or items were not re-tested after substantial adjustments			
Was an appropriate qualitative method used for assessing the comprehensibility of the PROM instructions, items, response options,	Widely recognized orwell justified qualitative method used	Assumable that the method wasappropriate butnot clearly described	Only quantitative (survey) method (s) used or doubtful whether the method was appropriate or not clear if patients were asked about	Method used not appropriate or patients were not asked about the comprehensibility of the items, response options			

and recall period?			the comprehensibility of the items, response options or recall period or patients not asked about the comprehensibility of the PROM instructions	or recall period or patients			
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear				
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewersused	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers not trained and no experience		Not Applicable		
Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate,but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable		
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and	Assumable that allgroup meetings or interviews	Not clear if all group meetings or interviews were recorded and	No recording and no notes	Not applicable		

	transcribed verbatim	were recorded and transcribed verbatim, but not clearly described	transcribed verbatimor recordings not transcribed verbatim or only notes were made during the group meetings/ interviews				
Was an appropriate approach used to analyse the data?	A widely recognized orwell justified approach was used	Assumable that the approach was appropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate			
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only oneresearcher involvedin the analysis				
Were problems regarding the comprehensibility of the PROM instructions, items, response options, and recall period appropriately addressed by adapting the PROM?	No problems found or problems appropriately addressed and PROM was adapted and re-tested if necessary	Assumable that there were no problems or that problems were appropriately addressed, but not clearly described	Not clear if there were problems or doubtful if problems were appropriately addressed	Problems not appropriately addressed or PROM was adapted but items were not re-tested after substantial adjustments	Not applicable		
Comprehensiveness							
Were patients asked about the <u>comprehensiveness</u> of the PROM?	YES		NO or not clear **(SKIP items 27-35)**				

Was the final set of items tested?	The final set of items was tested	Assumable that the final set of items was tested but not clearly described	Not clear if the final set of items was tested or the final set of items was not tested or the set of items were not re-tested after items were removed or added				
Was an appropriate method used for assessing the comprehensiveness of the PROM?	Widely recognized method used	Assumable that method was appropriate but not clearly described or only quantitative (survey) method (s) used	Doubtful whether the method was appropriate or methodused not appropriate				
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear				
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewersused	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers not trained and no experience		Not Applicable		

Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate,but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable		
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and transcribed verbatim	Assumable that allgroup meetings or interviews were recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings or interviews were recorded and transcribed verbatimor recordings not transcribed verbatim or only notes were made during the group meetings/ interviews	No recording and no notes	Not applicable		
Was an appropriate approach used to analyse the data?	A widely recognized orwell justified approach was used	Assumable that the approach was appropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate			
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only oneresearcher involvedin the analysis				
Were problems regarding the comprehensiveness of the PROM	No problems found or problems were	Assumable that there were problems or	Not clear if there were problems or doubtful if the	Problems not appropriately addressed	Not applicable		

appropriately addressed by adapting the PROM?	appropriately addressed and PROM adapted and re-tested if necessary	that problems were appropriately addressed but not clearly described	problems were appropriately addressed or PROM was adapted but items were not re-tested after substantial adjustments				
Box 2 Content validity – Not tested							
Box 3. Structural Validity – Not tested							
Box 4. Internal Consistency							
Does the scale consist of effect indicators i.e. is it based on a reflective model? Yes or No							
Design Requirements							
Was an internal consistency statistic calculated foreach unidimensional scale or subscale separately?	Internal consistency statistic calculated for each unidimensional scale or subscale	Unclear whether scale or sub scale is unidimensional	Internal consistency statisticNOT calculated for each unidimensional scale or sub scale			(72,107)	
Statistical methods							
For continuous scores: Was Cronbach's alpha oromega calculated?	Cronbach's alpha, or Omega calculated		Only item-total correlations calculated	No Cronbach's alpha and noitem-total correlations calculated	Not Applicable	(72,107)	1-0.70-0.82 2->0.65
For dichotomous scores: Was Cronbach's alpha or KR-20 calculated?	Cronbach's alpha or KR-20 calculated		Only item-total correlations calculated	No Cronbach's alpha or KR-20 and no item-total correlations calculated	Not Applicable	(72,107)	
For IRT-based scores: Was standard error of the theta(SE (θ)) or reliability coefficient of estimated	SE(θ) or reliability coefficient calculated			SE(θ) or reliability coefficient NOT calculated	Not Applicable	(72,107)	

latent trait value (index of (subject or item) separation) calculated?							
Other							
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study		(72,107)	2
Box 5: Cross-cultural validity/Measurement invariance							
Design requirements							
Were the samples similar for relevant characteristics except for the group variable?	Evidence provided that samples were similar for relevant characteristics except group variable	Stated (but no evidence provided)that samples were similar for relevant characteristics except group variable	Unclear whether samples were similar for relevant characteristics exceptgroup variable	Samples were NOT similar for relevant characteristics except group variable		(107)	Various health conditions
Statistical Methods							
Was an appropriate approach used to analyse the data?	A widely recognized or well justified approach was used	Assumable that the approach was appropriate, but notclearly described	Not clear what approach was used ordoubtful whether the approach was appropriate	Approach not appropriate	Not Applicable	(107)	Cronbach's t-test ICC Floor/ceiling effects
Was the sample size included in the analysis adequate?	Regression analyses or IRT/Rasch based analyses: 200 subjects per group	150 subjects per group	100 subjects per group	< 100 subjects per group		(107)	Not clear # per health group

	MGCFA*: 7 times the number of items and ≥100	5 times the numberof items and ≥100; OR 5-7 times the number of items but <100	5 times the numberof items but <100	<5 times the number of items			Not clear
Other							
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study		(107)	Rajmil-participants drawn from convenient samples Further research recommended for convergent/criterion validity
Box 6. Reliability							
Design Requirements							
Were patients stable in the interim period on the construct to be measured?	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable		(72,107)	Both Not stated
Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate		(72,107)	Both -One week apart
Were the test conditions similar for the measurements? e.g. typeof administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that testconditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar		(72,107)	
Statistical methods						(72,107)	
For Continuous scores: Was an Intraclass correlation coefficient (ICC) calculated?	ICC calculated and model or formula of the ICC described	ICC calculated but model or formula of the ICC not described or optimal. Pearson or	Pearson or Spearman correlation coefficient calculated WITHOUT evidence provided	No ICC or Pearson or Spearman correlations calculated	Not Applicable	(72,107)	Riley-0.35-0.76 Rajmil-ICC=0.57-0.93

		Spearman correlation calculated with evidence provided that no systematic change has occurred	that no systematic change has occurred WITH evidence that systematic change has occurred				
For dichotomous/nominal/ordinal scores: Was kappa calculated?	Kappa calculated			No kappa calculated	Not Applicable	(72,107)	
For ordinal scores: Was a weighted kappa calculated?	Weighted Kappa Calculated		Unweighted Kappa calculated or not described		Not Applicable	(72,107)	1 and 2-Not calculated
For ordinal scores: Was the weighting scheme described? E.g. linear, quadratic	Weighting scheme described	Weighting scheme NOT described				(72,107)	
Other							
Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws		(72,107)	1-further research in other parts of the country 2-participants drawn from convenient samples Further research recommended for convergent/criterion validity
Box 7. Measurement error: absolute measures – Not tested							
Box 8: Criterion Validity							
Statistical Methods							
For continuous scores: Were correlations, or the area under the receiver operating curve calculated?	Correlations or AUC calculated			Correlations or AUC NOT calculated	Not Applicable	(72)	Correlations calculated

For dichotomous scores: Were sensitivity and specificity determined?	Sensitivity and specificity calculated			Sensitivity and specificity NOT calculated	Not Applicable	(72)	
Other							
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study		(72)	
Box 9: Hypothesis testing for construct validity							
9a. Comparison with other outcome measurement instruments (convergent validity)							
Design requirements							
Is it clear what the comparator instrument(s) measure(s)?	Constructs measured by the comparator instrument(s) is clear		Constructs measured by the comparator instrument(s) is not clear			(72)	1-8 additional OMs
Were the measurement properties of the comparatorinstrument(s) sufficient?	Sufficient measurement properties of the comparator instrument(s) in a population similar tothe study population	Sufficient measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information onmeasurement properties of the comparator instrument(s) in any study population	No information on the measurement properties of the comparator instrument(s), OR evidence for insufficient measurement properties of the comparator instrument(s)		(72)	
Statistical methods							

Was the statistical method appropriate for the hypotheses to betested?	Statistical method was appropriate	Assumable that statistical methodwas appropriate	Statistical method applied NOT optimal	Statistical method applied NOT appropriate		(72)	Pearson's
Other							
Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws		(72)	
9b. Comparison between subgroups (discriminative or known-groups validity)							
Design requirements							
Was there adequate description provided of important characteristics of the subgroups?	Adequate description of the important characteristics of the subgroups	Adequate description of most of the important characteristics of the subgroups	Poor description of the important characteristics of the subgroups			(72)	1-Hospital settings – types of conditions not specified
Statistical methods							
Was the statistical method appropriate for the hypotheses to betested?	Statistical method was appropriate	Assumable that statistical methodwas appropriate	Statistical method applied NOT optimal	Statistical method applied NOT appropriate		(72)	
Other							
Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws		(1)	
Box 10: Responsiveness – Not yet tested							

Appendix 6: COSMIN Risk of Bias Checklist – PROMIS-PGH-7

(level of performance highlighted in grey)

Box General requirement for studies that applied Item Response Theory (IRT) Models							
	Very good	**Adequate**	**Doubtful**	**Inadequate**	**Not applicable**	**Reference**	**Notes**
Box 1. PROM development							
1a. PROM design							
General design requirements							
Is a clear description provided of the construct to be measured?	Construct clearly described			Construct not clearly described	Not Applicable	(73)	
Is the origin of the construct clear: was a theory, conceptual framework or disease model used or clear rationale provided todefine the construct to be measured?	Origin of the construct clear		Origin of the construct not clear			(73)	
Is a clear description provided of the target population for whichthe PROM was developed?	Target population clearly described			Target population not clearly described		(73)	
Is a clear description provided of the context of use	Context of use clearly described		Context of use not clearly described			(73)	
Was the PROM development study performed in a sample representing the target population for which the PROM wasdeveloped?	Study performed in a sample representing the target population	Assumable that the study was performed in a sample representing the target	Doubtful whether the study was performed ina sample representing the target population	Study not performed in a sample representingthe target population **(SKIP items 6-12)**		(73)	

		population, but not clearly described					
Concept elicitation (relevance and comprehensiveness)							
Was an appropriate qualitative data collection method used to identify relevant items for a new PROM?	Widely recognized or well justified qualitative methodused, suitable for the construct and study population	Assumable that the qualitative method was appropriate and suitable for the construct and study population, but not clearly described	Only quantitative (survey) method(s) used or doubtful whether the method was suitable for the construct and study population	Method used not appropriate or not suitable for the construct or study population		(73)	
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewersused	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers nottrained and no experience		Not Applicable	(73)	Interviewers not described
Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate,but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable	(73)	Not clear – but probed using specific questions
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and	Assumable that allgroup meetings or	Not clear if all group meetings or interviews were	No recording and no notes	Not applicable	(73)	Notes taken, interviewees probed

	transcribed verbatim	interviews were recorded and transcribed verbatim, but not clearly described	recorded and transcribed verbatimor recordings not transcribed verbatim or only notes were made during the group meetings/ interviews				Cognitive debriefing
Was an appropriate approach used to analyse the data?	A widely recognized or well justified approach was used	Assumable that the approach was appropriate, but not clearly described	Not clear what approach was used ordoubtful whether the approach was appropriate	Approach not appropriate		(73)	
Was at least part of the data coded independently?	At least 50% of the data was coded by at least two researchers independently	11-49% of the data was coded by at least two researchers independently	Doubtful if two researchers were involved in the codingor only 1-10% of the data was coded by at least two researchers independently	Only one researcher was involvedin coding or no coding		(73)	
Was data collection continued until saturation was reached?	Evidence provided that saturation was reached	Assumable that saturation was reached	Doubtful whether saturation was reached	Evidence suggests that saturation was not reached	Not applicable	(73)	
For quantitative studies (surveys): was the sample sizeappropriate?	≥100	50-99	30-49	<30	Not applicable	(73)	Pilot
1b. <u>Cognitive interview study</u> or other pilot test							
Was a cognitive interview study or other pilot test conducted?	YES			**NO (SKIP items 15-35**		(73)	
General design requirements							

Was the cognitive interview study or other pilot test performed ina sample representing the target population?	Study performed in a sample representing the target population	Assumable that the study was performed in a sample representing the target population, but not clearly described	Doubtful whether the study was performed in a sample representing the target population	Study not performed in a sample representingthe target population		(73)	
Comprehensibility							
Were patients asked about the comprehensibility of the PROM?	YES	Not clear (SKIP standards 17-25)	No **(SKIP standards 17-25)**			(73)	Probed by interviewer
Were all items tested in their final form?	All items were tested in their final form	Assumable that all items were tested in their final form, but not clearly described	Not clear if all itemswere tested in theirfinal form	Items were not tested in their final form or itemswere not re- tested after substantial adjustments		(73)	
Was an appropriate qualitative method used for assessing the comprehensibility of the PROM instructions, items, response options, and recall period?	Widely recognized orwell justified qualitative method used	Assumable that the method was appropriate but not clearly described	Only quantitative (survey) method (s) used or doubtful whether the method was appropriate or not clear if patients were asked about the comprehensibility of the items, response options or recall period or patients not asked about the comprehensibility	Method used not appropriate or patients were not asked about the comprehensibility of the items, response options or recall period or patients		(73)	

			of the PROM instructions				
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear			(73)	
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewersused	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers not trained and no experience		Not Applicable	(73)	
Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate,but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable	(73)	
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and transcribed verbatim	Assumable that allgroup meetings or interviews were recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings or interviews were recorded and transcribed verbatimor recordings not transcribed verbatim or only notes were made during the group meetings/ interviews	No recording and no notes	Not applicable	(73)	

Was an appropriate approach used to analyse the data?	A widely recognized orwell justified approach wasused	Assumable that the approach wasappropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate		(73)	
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only one researcher involved in the analysis			(73)	
Were problems regarding the comprehensibility of the PROM instructions, items, response options, and recall period appropriately addressed by adapting the PROM?	No problems found or problems appropriately addressed and PROM was adapted and re-tested if necessary	Assumable that there were no problems or that problems were appropriately addressed, but not clearly described	Not clear if there were problems or doubtful if problems were appropriately addressed	Problems not appropriately addressed or PROM was adapted but items were not re-tested after substantial adjustments	Not applicable	(73)	
Comprehensiveness							
Were patients asked about the comprehensiveness of the PROM?	YES		NO or not clear **(SKIP items 27-35)**			(73)	
Was the final set of items tested?	The final set of items was tested	Assumable that the final set of items was tested but not clearly described	Not clear if the final set of items was tested or the final set of items was not tested or the set of items were not re-tested after items were removed or added			(73)	
Was an appropriate method used for	Widely recognized method used	Assumable that method was	Doubtful whether the method was			(73)	

assessing the comprehensiveness of the PROM?		appropriate but not clearly described or only quantitative (survey) method (s) used	appropriate or methodused not appropriate				
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear			(73)	
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewersused	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers not trained and no experience		Not Applicable	(73)	
Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate,but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable	(73)	
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and transcribed verbatim	Assumable that allgroup meetings or interviews were recorded and transcribed verbatim, but	Not clear if all group meetings or interviews were recorded and transcribed verbatimor recordings not transcribed	No recording and no notes	Not applicable	(73)	

		not clearly described	verbatim or only notes were made during the group meetings/ interviews				
Was an appropriate approach used to analyse the data?	A widely recognized orwell justified approach wasused	Assumable that the approach wasappropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate		(73)	
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only one researcher involved in the analysis			(73)	
Were problems regarding the comprehensiveness of the PROM appropriately addressed by adapting the PROM?	No problems found or problems were appropriately addressed and PROM adapted and re-tested if necessary	Assumable that there were problems or that problems were appropriately addressed but not clearly described	Not clear if there were problems or doubtful if the problems were appropriately addressed or PROM was adapted but items were not re-tested after substantial adjustments	Problems not appropriately addressed	Not applicable	(73)	
Box 2 Content validity							
2a. Asking patients about relevance							
Design requirements							
Was an appropriate method used to ask patients whether eachitem is relevant for their experience with the condition?	Widely recognized orwell justified method used	Only quantitative (survey) method(s)used	Not clear if patients were asked whethereach item is relevantor	Method used not appropriate or patients not asked about the		(73,185)	

		or assumable that the method was appropriate but not clearly described	doubtful whether the method was appropriate	relevance of all items			
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear			(73,185)	
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewersused	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers not trained and no experience		Not Applicable	(73,185)	
Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate,but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable	(73,185)	
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and transcribed verbatim	Assumable that allgroup meetings or interviews were recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings or interviews were recorded and transcribed verbatimor recordings not transcribed verbatim or only	No recording and no notes	Not applicable	(73,185)	Codes used for notes

			notes were made during the group meetings/ interviews				
Analyses							
Was an appropriate approach used to analyse the data?	A widely recognized orwell justified approach wasused	Assumable that the approach wasappropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate		(73,185)	
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only one researcher involved in the analysis			(73,185)	
2b. Asking patients about comprehensiveness							
Was an appropriate method used for assessing the comprehensiveness of the PROM?	Widely recognized or well justified method used	Only quantitative (survey) method(s)used or assumable that the method was appropriate but not clearly described	Doubtful whether the method was appropriate	Method used not appropriate		(73,185)	
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear			(73,185)	

Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewersused	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers nottrained and no experience		Not Applicable	(73,185)	
Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate,but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable	(73,185)	
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and transcribed verbatim	Assumable that allgroup meetings or interviews were recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings or interviews were recorded and transcribed verbatimor recordings not transcribed verbatim or only notes were made during the group meetings/ interviews	No recording and no notes	Not applicable	(73,185)	
Analyses							
Was an appropriate approach used to analyse the data?	A widely recognized orwell justified approach wasused	Assumable that the approach wasappropriate, but not clearly described	Not clear what approach was used or doubtful whether the	Approach not appropriate		(73,185)	

			approach was appropriate				
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only one researcher involved in the analysis			(73,185)	
2c. Asking patients about comprehensibility						(73,185)	
Was an appropriate qualitative method used for assessing the comprehensibility of the PROM instructions, items, response options, and recall period?	Widely recognized orwell justified qualitative method used	Assumable that the method was appropriate but not clearly described	Only quantitative (survey) method (s) used or doubtful whether the method was appropriate or not clear if patients were asked about the comprehensibility of the items, response options or recall period or patients not asked about the comprehensibility of the PROM instructions	Method used not appropriate or patients were not asked about the comprehensibility of the items, response options or recall period or patients		(73,185)	
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear			(73,185)	

Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewersused	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers nottrained and no experience			(73,185)	
Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate,but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable	(73,185)	
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and transcribed verbatim	Assumable that allgroup meetings or interviews were recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings or interviews were recorded and transcribed verbatimor recordings not transcribed verbatim or only notes were made during the group meetings/ interviews	No recording and no notes	Not applicable	(73,185)	
Analyses							
Was an appropriate approach used to analyse the data?	A widely recognized orwell justified approach wasused	Assumable that the approach wasappropriate, but not clearly described	Not clear what approach was used or doubtful whether the	Approach not appropriate		(73,185)	

			approach was appropriate				
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only one researcher involved in the analysis			(73,185)	
2d. Asking professionals about relevance	Not done					(73,185)	
2e. Asking professionals about comprehensiveness	Not done					(73,185)	
Box 3. Structural Validity							
Does the scale consist of effect indicators, i.e. is it based on a reflective model? Yes/No							
Does the study concern unidimensionality or structural validity?							
Statistical methods							
For CTT: Was exploratory or confirmatory factor analysis performed?	Confirmatory factor analysis performed	Exploratory factor analysis performed		No exploratory or confirmatory factor analysis performed	Not Applicable	(73)	
For IRT: Were IRT/Rasch: does the chosen model fit to the research topic	Chosen model fits wellto the research question	Assumable that the chosen model fits wellto the research question	Doubtful if the chosen model fits well to the research question	Chosen model does not fit to the research question	Not Applicable	(73)	
	FA: 7 times the numberof items and ≥100	FA: at least 5 times the number of items and ≥100; OR at least6 times	FA: 5 times the number of items but <100	FA: < 5 times the number of items		(73)	5.2-4.6

		number of items but <100					
	Rasch/1PL models: ≥ 200 subjects	Rasch/1PL models:100-199 subjects	Rasch/1PL models: 50-99 subjects	Rasch/1PL models: <50 subjects			
	2PL parametric IRT models OR Mokken scale analysis: ≥ 1000 subjects	2PL parametric IRT models OR Mokkenscale analysis: 500-999 subjects	2PL parametric IRT models OR Mokken scale analysis: 250-499 subjects	2PL parametric IRT models OR Mokken scale analysis: < 250subjects			
Other							
Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws		(73)	
Box 4. Internal Consistency							
Does the scale consist of effect indicators i.e. is it based on a reflective model? Yes or No							
Design Requirements							
Was an internal consistency statistic calculated foreach unidimensional scale or subscale separately?	Internal consistency statisticcalculated for each unidimensional scale or subscale	Unclear whether scale or sub scale is unidimensional	Internal consistency statisticNOT calculated for each unidimensional scale or sub scale			(73,80)	1=CR0.88 PR0.84 2=CR0.67-0.80 PR0.71-0.80
Statistical methods							
For continuous scores: Was Cronbach's alpha oromega calculated?	Cronbach's alpha, or Omega calculated		Only item-total correlations calculated	No Cronbach's alpha and noitem-total correlations calculated	Not Applicable	(73,80)	
For dichotomous scores: Was Cronbach's alpha or KR-20 calculated?	Cronbach's alpha or KR-20calculated		Only item-total correlations calculated	No Cronbach's alpha or KR-20 and no item-total correlations calculated	Not Applicable		

For IRT-based scores: Was standard error of the theta(SE (θ)) or reliability coefficient of estimated latent trait value (index of (subject or item) separation) calculated?	SE(θ) or reliability coefficient calculated			SE(θ) or reliability coefficient NOT calculated	Not Applicable		
Other							
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study		(73,80)	
Box 5: Cross-cultural validity/Measurement invariance – Not tested							
Box 6. Reliability							
Design Requirements							
Were patients stable in the interim period on the construct to be measured?	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable		(73,80)	
Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate		(73,80)	1=2-weeks 2=not stated
Were the test conditions similar for the measurements? e.g. typeof administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that testconditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar		(73,80)	
Statistical methods							
For Continuous scores: Was an Intraclass correlation coefficient (ICC) calculated?	ICC calculated and model or formula of the ICC described	ICC calculated but model or formula of the ICC not described or optimal. Pearson or Spearman correlation calculated with	Pearson or Spearman correlation coefficient calculated WITHOUT evidence provided that no systematic change has occurred WITH evidence that	No ICC or Pearson or Spearman correlations calculated	Not Applicable	(73,80)	F 2014-ICC=CR0.73 PR0.74 F 2019-ICC=CR0.66-0.81 PR0.71-0/80

		evidence provided that no systematic change has occurred	systematic change has occurred				
For dichotomous/nominal/ordinal scores: Was kappa calculated?	Kappa calculated			No kappa calculated	Not Applicable	(73,80)	
For ordinal scores: Was a weighted kappa calculated?	Weighted Kappa Calculated		Unweighted Kappa calculated or not described		Not Applicable	(73,80)	
For ordinal scores: Was the weighting scheme described? E.g. linear, quadratic	Weighting scheme described	Weighting scheme NOT described				(73,80)	
Other							
Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws		(73,80)	
Box 7. Measurement error: absolute measures – Not tested							
Box 8: Criterion Validity – Not tested							
Box 9: Hypothesis testing for construct validity							
9a. Comparison with other outcome measurement instruments (convergent validity)							
Design requirements							
Is it clear what the comparator instrument(s) measure(s)?	Constructs measured by the comparator instrument(s) is clear		Constructs measured by the comparator instrument(s) is not clear			(80,99,186)	CR vs PR PedsQL, KIDSCREEN
Were the measurement properties of the comparatorinstrument(s) sufficient?	Sufficient measurement properties of the comparator instrument(s) in a population similar	Sufficient measurement properties of the comparator instrument(s) but not sure if	Some information onmeasurement properties of the comparator instrument(s) in	No information on the measurement properties of the comparator instrument(s), OR evidence for		(80,99,186)	

	tothe study population	these apply to the study population	any study population	insufficient measurement properties of the comparator instrument(s)			
Statistical methods							
Was the statistical method appropriate for the hypotheses to be tested?	Statistical method was appropriate	Assumable that statistical methodwas appropriate	Statistical method applied NOT optimal	Statistical method applied NOT appropriate		(80,99,186)	
Other							
Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws		(80,99,186)	
9b. Comparison between subgroups (discriminative or known-groups validity)							
Design requirements							
Was there adequate description provided of important characteristics of the subgroups?	Adequate description of the important characteristics of the subgroups	Adequate description of most of the important characteristics of the subgroups	Poor description of the important characteristics of the subgroups			(99)	chronic, special care
Statistical methods							
Was the statistical method appropriate for the hypotheses to be tested?	Statistical method was appropriate	Assumable that statistical methodwas appropriate	Statistical method applied NOT optimal	Statistical method applied NOT appropriate		(80)	
Other							
Were there any other important flaws in the design or statistical methods of the study?	No other important methodological		Other minor important	Other important methodological flaws		(80)	

	flaws		methodological flaws				
Box 10: Responsiveness – Not tested							

Appendix 7: COSMIN Risk of Bias Checklist – FSIIR

(level of performance highlighted in grey)

Box General requirement for studies that applied Item Response Theory (IRT) Models							
	Very good	**Adequate**	**Doubtful**	**Inadequate**	**Not applicable**	**Reference**	**Notes**
Box 1. PROM development							
1a. PROM design							
General design requirements							
Is a clear description provided of the construct to be measured?	Construct clearly described			Construct not clearly described	Not Applicable	(81,82)	
Is the origin of the construct clear: was a theory, conceptual framework or disease model used or clear rationale provided todefine the construct to be measured?	Origin of the construct clear		Origin of the construct not clear			(81,82)	From adult version
Is a clear description provided of the target population for whichthe PROM was developed?	Target population clearly described			Target population not clearly described		(81,82)	
Is a clear description provided of the context of use	Context of use clearly described		Context of use not clearly described			(81,82)	
Was the PROM development study performed in a sample representing the target population for which the PROM wasdeveloped?	Study performed in a sample representing the target population	Assumable that the study was performed in a sample representing	Doubtful whether the study was performed ina sample	Study not performed in a sample representingthe		(81,82)	

		the target population, but not clearly described	representing the target population	target population **(SKIP items 6-12)**			
Concept elicitation (relevance and comprehensiveness)							
Was an appropriate qualitative data collection method used to identify relevant items for a new PROM?	Widely recognized or well justified qualitative methodused, suitable for the construct and study population	Assumable that the qualitative method was appropriate and suitable for the construct and study population, but not clearly described	Only quantitative (survey) method(s) used or doubtful whether the method was suitable for the construct and study population	Method used not appropriate or not suitable for the construct or study population		(81,82)	
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewersused	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers nottrained and no experience		Not Applicable	(81,82)	(81)-not clear (82)-skilled
Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate,but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable	(81,82)	Interview prepared but no details

Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and transcribed verbatim	Assumable that allgroup meetings or interviews were recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings or interviews were recorded and transcribed verbatimor recordings not transcribed verbatim or only notes were made during the group meetings/ interviews	No recording and no notes	Not applicable	(81,82)	(81,82)-Not clear if notes were taken but interviewer present
Was an appropriate approach used to analyse the data?	A widely recognized or well justified approach was used	Assumable that the approach was appropriate, but not clearly described	Not clear what approach was used ordoubtful whether the approach was appropriate	Approach not appropriate		(81,82)	
Was at least part of the data coded independently?	At least 50% of the data was coded by at least two researchers independently	11-49% of the data was coded by at least two researchers independently	Doubtful if two researchers were involved in the codingor only 1-10% of the data was coded by at least two researchers independently	Only one researcher was involvedin coding or no coding		(81,82)	(81,82)-Not described
Was data collection continued until saturation was reached?	Evidence provided that saturation was reached	Assumable that saturation was reached	Doubtful whether saturation was reached	Evidence suggests that saturation was not reached	Not applicable	(81,82)	
For quantitative studies (surveys): was the sample sizeappropriate?	≥100	50-99	30-49	<30	Not applicable	(81,82)	
1b. <u>Cognitive interview study</u> or other pilot test							
Was a cognitive interview study or	YES			**NO (SKIP**		(81)	

other pilot test conducted?				**items 15-35**			
General design requirements							
Was the cognitive interview study or other pilot test performed ina sample representing the target population?	Study performed in a sample representing the target population	Assumable that the study was performed in a sample representing the target population, but not clearly described	Doubtful whether the study was performed in a sample representing the target population	Study not performed in a sample representingthe target population		(81)	
Comprehensibility							
Were patients asked about the comprehensibility of the PROM?	YES	Not clear (SKIP standards 17-25)	No **(SKIP standards 17-25)**			(81)	
Were all items tested in their final form?	All items were tested in their finalform	Assumable that all items were tested in their final form, but not clearly described	Not clear if all itemswere tested in theirfinal form	Items were not tested in their final form or itemswere not re- tested after substantial adjustments			
Was an appropriate qualitative method used for assessing the comprehensibility of the PROM instructions, items, response options, and recall period?	Widely recognized orwell justified qualitative method used	Assumable that the method was appropriate but not clearly described	Only quantitative (survey) method (s) used or doubtful whether the method was appropriate or not clear if patients were asked about the comprehensibility of the items, response options or recall period or	Method used not appropriate or patients were not asked about the comprehensibility of the items, response options or recall period or patients			

			patients not asked about the comprehensibility of the PROM instructions				
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear				
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewersused	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers not trained and no experience		Not Applicable		
Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate,but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable		
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and transcribed verbatim	Assumable that allgroup meetings or interviews were recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings or interviews were recorded and transcribed verbatimor recordings not transcribed verbatim or only notes were made	No recording and no notes	Not applicable		

			during the group meetings/ interviews				
Was an appropriate approach used to analyse the data?	A widely recognized orwell justified approach wasused	Assumable that the approach wasappropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate			
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only oneresearcher involvedin the analysis				
Were problems regarding the comprehensibility of the PROM instructions, items, response options, and recall period appropriately addressed by adapting the PROM?	No problems found or problems appropriately addressed and PROM was adapted and re-tested if necessary	Assumable that there were no problems or that problems were appropriately addressed, but not clearly described	Not clear if there were problems or doubtful if problems were appropriately addressed	Problems not appropriately addressed or PROM was adapted but items were not re-tested after substantial adjustments	Not applicable		
Comprehensiveness							
Were patients asked about the comprehensiveness of the PROM?	YES		NO or not clear **(SKIP items 27-35)**				
Was the final set of items tested?	The final set of items was tested	Assumable that the final set of items was tested but not clearly described	Not clear if the final set of items was tested or the final set of items was not tested or the set of items were not re-tested				

			after items were removed or added				
Was an appropriate method used for assessing the comprehensiveness of the PROM?	Widely recognized method used	Assumable that method was appropriate but not clearly described or only quantitative (survey) method (s) used	Doubtful whether the method was appropriate or methodused not appropriate				
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear				
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewersused	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers not trained and no experience		Not Applicable		
Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate,but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable		
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and	Assumable that allgroup meetings or interviews were	Not clear if all group meetings or interviews were recorded and	No recording and no notes	Not applicable		

	transcribed verbatim	recorded and transcribed verbatim, but not clearly described	transcribed verbatimor recordings not transcribed verbatim or only notes were made during the group meetings/ interviews				
Was an appropriate approach used to analyse the data?	A widely recognized orwell justified approach wasused	Assumable that the approach wasappropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate			
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only oneresearcher involvedin the analysis				
Were problems regarding the comprehensiveness of the PROM appropriately addressed by adapting the PROM?	No problems found or problems were appropriately addressed and PROM adapted and re-tested if necessary	Assumable that there were problems or that problems were appropriately addressed but not clearly described	Not clear if there were problems or doubtful if the problems were appropriately addressed or PROM was adapted but items were not re-tested after substantial adjustments	Problems not appropriately addressed	Not applicable		
Box 2 Content validity							
2a. Asking patients about relevance							
Design requirements							

Was an appropriate method used to ask patients whether eachitem is relevant for their experience with the condition?	Widely recognized orwell justified method used	Only quantitative (survey) method(s)used or assumable that the method was appropriate but not clearly described	Not clear if patients were asked whethereach item is relevantor doubtful whether the method was appropriate	Method used not appropriate or patients not asked about the relevance of all items		(81)	Does not mention method used to identify problematic items but mentioned that it was noted so that it could be changed appropriately
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear				
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewersused	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers not trained and no experience		Not Applicable	(81)	
Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate,but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable	(81)	
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and	Assumable that allgroup meetings or	Not clear if all group meetings or interviews were	No recording and no notes	Not applicable	(81)	

	transcribed verbatim	interviews were recorded and transcribed verbatim, but not clearly described	recorded and transcribed verbatimor recordings not transcribed verbatim or only notes were made during the group meetings/ interviews				
Analyses							
Was an appropriate approach used to analyse the data?	A widely recognized orwell justified approach wasused	Assumable that the approach wasappropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate		(81)	Briefly mentioned but not detailed
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only oneresearcher involvedin the analysis			(81)	
2b. Asking patients about comprehensiveness	Not done						
2c. Asking patients about comprehensibility	Not done						
2d. Asking professionals about relevance	Not done						
2e. Asking professionals about comprehensiveness	Not done						
Box 3. Structural Validity – Not tested							
Box 4. Internal Consistency							

Does the scale consist of effect indicators i.e. is it based on a reflective model? Yes or No							
Design Requirements							
Was an internal consistency statistic calculated foreach unidimensional scale or subscale separately?	Internal consistency statisticcalculated for each unidimensional scale or subscale	Unclear whether scale or sub scale is unidimensional	Internal consistency statisticNOT calculated for each unidimensional scale or sub scale			(81,82)	
Statistical methods							
For continuous scores: Was Cronbach's alpha oromega calculated?	Cronbach's alpha, or Omega calculated		Only item-total correlations calculated	No Cronbach's alpha and noitem-total correlations calculated	Not Applicable	(81,82)	Stein=0.83-0.94 Kromer=not calculated
For dichotomous scores: Was Cronbach's alpha or KR-20 calculated?	Cronbach's alpha or KR-20calculated		Only item-total correlations calculated	No Cronbach's alpha or KR-20 and no item-total correlations calculated	Not Applicable		
For IRT-based scores: Was standard error of the theta(SE (θ)) or reliability coefficient of estimated latent trait value (index of (subject or item) separation) calculated?	SE(θ) or reliability coefficient calculated			SE(θ) or reliability coefficient NOT calculated	Not Applicable		
Other							
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study			
Box 5: Cross-cultural validity/Measurement invariance							
Design requirements							

Were the samples similar for relevant characteristics except for thegroup variable?	Evidence provided that samples were similar for relevant characteristics except group variable	Stated (but no evidence provided)that samples were similar for relevant characteristics except group variable	Unclear whether samples were similar for relevant characteristics exceptgroup variable	Samples were NOT similar forrelevant characteristics except group variable		(82)	Process explained – translation to Spanish and used
Statistical Methods							
Was an appropriate approach used to analyse the data?	A widely recognizedor well justified approach was used	Assumable that the approach was appropriate, but notclearly described	Not clear what approach was used ordoubtful whether the approach was appropriate	Approach not appropriate	Not Applicable	(82)	
Was the sample size included in the analysis adequate?	Regression analysesor IRT/Rasch based analyses: 200 subjects per group	150 subjects per group	100 subjects per group	< 100 subjects per group		(82)	
	MGCFA*: 7 times the number of itemsand ≥100	5 times the numberof items and ≥100; OR 5-7 times the number of items but <100	5 times the numberof items but <100	<5 times the number of items		(82)	
Other							
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study		(82)	
Box 6. Reliability							
Design Requirements							

Were patients stable in the interim period on the construct to be measured?	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable			Test-retest not done
Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate			
Were the test conditions similar for the measurements? e.g. typeof administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that testconditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar			Test-retest not done
Statistical methods							
For Continuous scores: Was an Intraclass correlation coefficient (ICC) calculated?	ICC calculated and model or formula of the ICC described	ICC calculated but model or formula of the ICC not described or optimal. Pearson or Spearman correlation calculated with evidence provided that no systematic change has occurred	Pearson or Spearman correlation coefficient calculated WITHOUT evidence provided that no systematic change has occurred WITH evidence that systematic change has occurred	No ICC or Pearson or Spearman correlations calculated	Not Applicable	(82)	(82)-ICC between English and Spanish versions 0.76-0.88
For dichotomous/nominal/ordinal scores: Was kappa calculated?	Kappa calculated			No kappa calculated	Not Applicable		
For ordinal scores: Was a weighted kappa calculated?	Weighted Kappa Calculated		Unweighted Kappa calculated or not described		Not Applicable		
For ordinal scores: Was the weighting scheme described? E.g. linear, quadratic	Weighting scheme described	Weighting scheme NOT described					
Other							

Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws			
Box 7. Measurement error: absolute measures – Not done							
Box 8: Criterion Validity – Not tested							
Box 9: Hypothesis testing for construct validity							
9b. Comparison between subgroups (discriminative or known-groups validity)	Not tested						
Design requirements							
Was there adequate description provided of important characteristics of the subgroups?	Adequate description of the important characteristics of the subgroups	Adequate description of most of the important characteristics of the subgroups	Poor description of the important characteristics of the subgroups			(81,82)	Both - basic health information included such as chronic, acute
Statistical methods							
Was the statistical method appropriate for the hypotheses to be tested?	Statistical method was appropriate	Assumable that statistical methodwas appropriate	Statistical method applied NOT optimal	Statistical method applied NOT appropriate		(81,82)	
Other							
Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws		(81,82)	
Box 10: Responsiveness – Not tested							

Appendix 8: COSMIN Risk of Bias Checklist – CHU-9D

(level of performance highlighted in grey)

Box General requirement for studies that applied Item Response Theory (IRT) Models							
	Very good	**Adequate**	**Doubtful**	**Inadequate**	**Not applicable**	**Reference**	**Notes**
Box 1. PROM development							
1a. PROM design							
General design requirements							
Is a clear description provided of the construct to be measured?	Construct clearly described			Construct not clearly described	Not Applicable	(74,83,84,87,116,187)	HRQoL
Is the origin of the construct clear: was a theory, conceptual framework or disease model used or clear rationale provided todefine the construct to be measured?	Origin of the construct clear		Origin of the construct not clear			(74,83,84,87,116,187)	More child friendly OMs
Is a clear description provided of the target population for whichthe PROM was developed?	Target population clearly described			Target population not clearly described		(74,83,84,87,116,187)	Paediatric populations with and without medical conditions
Is a clear description provided of the context of use	Context of use clearly described		Context of use not clearly described			(74,83,84,87,116,187)	
Was the PROM development study performed in a sample representing the target population for which the PROM was developed?	Study performed in a sample representing the target population	Assumable that the study was performed in a sample	Doubtful whether the study was performed ina sample	Study not performed in a sample representingthe		(74,83,84,87,116,187)	

		representing the target population, butnot clearly described	representing the target population	target population **(SKIP items 6-12)**			
Concept elicitation (relevance and comprehensiveness)						(74,83,84,87,116,187)	
Was an appropriate qualitative data collection method used to identify relevant items for a new PROM?	Widely recognized or well justified qualitative methodused, suitable for the construct and study population	Assumable thatthe qualitative method was appropriate and suitable forthe construct and study population, butnot clearly described	Only quantitative (survey) method(s) used or doubtful whether the method was suitable for the construct and study population	Method used not appropriate or not suitable for the construct or study population		(74,83,84,87,116,187)	
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewersused	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers were trained or group moderators/ interviewers nottrained and no experience		Not Applicable	(74,83,84,87,116,187)	2-Not stated – only referred to as 'interviewers'
Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate,	Not clear if a topic guide was used or doubtful if topic or interview guide		Not Applicable	(74,83,84,87,116,187)	2-Interviewer guide given

		but not clearly described	wasappropriate or no guide				
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and transcribed verbatim	Assumable that allgroup meetings or interviews were recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings or interviews were recorded and transcribed verbatimor recordings not transcribed verbatim or only notes were made during the group meetings/ interviews	No recording and no notes	Not applicable	(74,83,84,87,116,187)	Notes were taken, time taken
Was an appropriate approach used to analyse the data?	A widely recognized or welljustified approachwas used	Assumable thatthe approach was appropriate, but not clearly described	Not clear what approach was used ordoubtful whether the approach was appropriate	Approach not appropriate		(74,83,84,87,116,187)	
Was at least part of the data coded independently?	At least 50% of the data was coded by at least two researchers independently	11-49% of the data was codedby at least two researchers independently	Doubtful if two researchers were involved in the codingor only 1-10% of the data was coded by at least two researchers independently	Only one researcher was involvedin coding or no coding		(74,83,84,87,116,187)	Not described
Was data collection continued until saturation was reached?	Evidence providedthat saturation was reached	Assumable thatsaturation was reached	Doubtful whether saturation was reached	Evidence suggests that saturation was not reached	Not applicable	(74,83,84,87,116,187)	Children recruited from schools/healthcare settings
For quantitative studies (surveys): was the sample sizeappropriate?	≥100	50-99	30-49	<30	Not applicable	(74,83,84,87,116,187)	

1b. <u>Cognitive interview study</u> or other pilot test						(74,83,84,87,116,187)	
Was a cognitive interview study or other pilot test conducted?	YES			**NO (SKIP items 15-35**		(74,83,84,87,116,187)	
General design requirements							
Was the cognitive interview study or other pilot test performed ina sample representing the target population?	Study performed in a sample representing the target population	Assumable that the study was performed in a sample representing the target population, butnot clearly described	Doubtful whether the study was performed in a sample representing the target population	Study not performed in a sample representingthe target population			
Comprehensibility							
Were patients asked about the <u>comprehensibility</u> of the PROM?	YES	Not clear (SKIP standards 17-25)	No **(SKIP standards 17-25)**				
Were all items tested in their final form?	All items were tested in their finalform	Assumable thatall items were tested in their final form, but not clearly described	Not clear if all itemswere tested in theirfinal form	Items were not tested in their final form or itemswere not re- tested after substantial adjustments			
Was an appropriate qualitative method used for assessing the <u>comprehensibility</u> of the PROM instructions, items, response options,	Widely recognized or well justified qualitative method used	Assumable that the method was appropriate butnot clearly described	Only quantitative (survey) method (s) used or doubtful whether the method was appropriate or not clear if patients were	Method used not appropriate or patients were not asked about the comprehensibility of the items, response options			

and recall period?			asked about the comprehensibility of the items, response options or recall period or patients not asked about the comprehensibility of the PROM instructions	or recall period or patients			
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear				
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewersused	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers not trained and no experience		Not Applicable		
Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate, but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable		
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and	Assumable that allgroup meetings or interviews	Not clear if all group meetings or interviews were recorded	No recording and no notes	Not applicable		

	transcribed verbatim	were recorded and transcribed verbatim, but not clearly described	and transcribed verbatimor recordings not transcribed verbatim or only notes were made during the group meetings/ interviews				
Was an appropriate approach used to analyse the data?	A widely recognized or well justified approach was used	Assumable that the approach was appropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate			
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only oneresearcher involvedin the analysis				
Were problems regarding the comprehensibility of the PROM instructions, items, response options, and recall period appropriately addressed by adapting the PROM?	No problems found or problems appropriately addressed and PROM was adapted and re-tested if necessary	Assumable thatthere were no problems or that problems were appropriately addressed, but not clearly described	Not clear if there were problems or doubtful if problems were appropriately addressed	Problems not appropriately addressed or PROM was adapted but items were not re-tested after substantial adjustments	Not applicable		
Comprehensiveness							
Were patients asked about the comprehensiveness of the PROM?	YES		NO or not clear **(SKIP items 27-35)**				

Was the final set of items tested?	The final set of items was tested	Assumable that the final set of items was tested but not clearly described	Not clear if the final set of items was tested or the final set of items was not tested or the set of items were not re-tested after items were removed or added				
Was an appropriate method used for assessing the <u>comprehensiveness</u> of the PROM?	Widely recognized method used	Assumable that method was appropriate but not clearly described or only quantitative (survey) method (s) used	Doubtful whether the method was appropriate or methodused not appropriate				
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear				
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewersused	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers not trained and no experience		Not Applicable		

Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate, but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable		
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and transcribed verbatim	Assumable that allgroup meetings or interviews were recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings or interviews were recorded and transcribed verbatimor recordings not transcribed verbatim or only notes were made during the group meetings/ interviews	No recording and no notes	Not applicable		
Was an appropriate approach used to analyse the data?	A widely recognized or well justified approach was used	Assumable that the approach was appropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate			
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only oneresearcher involvedin the analysis				

Were problems regarding the comprehensiveness of the PROM appropriately addressed by adapting the PROM?	No problems found or problems were appropriately addressed and PROM adapted and re-tested if necessary	Assumable that there were problems or that problems were appropriately addressed but not clearly described	Not clear if there were problems or doubtful if the problems were appropriately addressed or PROM was adapted but items were not re-tested after substantial adjustments	Problems not appropriately addressed	Not applicable		
Box 2 Content validity – Not tested							
Box 3. Structural Validity							
Does the scale consist of effect indicators, i.e. is it based on a reflective model? Yes							
Does the study concern unidimensionality or structural validity?							
Statistical methods							
For CTT: Was exploratory or confirmatory factor analysis performed?	Confirmatory factor analysis performed	Exploratory factor analysis performed		No exploratory or confirmatory factoranalysis performed	Not Applicable	(116)	Confirmatory factor analysis performed
For IRT: Were IRT/Rasch: does the chosen model fit to the research topic	Chosen model fits wellto the research question	Assumable that the chosen model fits wellto the research question	Doubtful if the chosen model fits well to the research question	Chosen model doesnot fit to the research question	Not Applicable	(116)	
	FA: 7 times the numberof items and ≥100	FA: at least 5 times the number of items and ≥100; OR at	FA: 5 times the number of items but <100	FA: < 5 times the number of items		(116)	FA 4-10, n=100

		least6 times number of items but <100					
	Rasch/1PL models: ≥ 200 subjects	Rasch/1PL models:100-199 subjects	Rasch/1PL models:50-99 subjects	Rasch/1PL models: <50 subjects			
	2PL parametric IRT models OR Mokken scale analysis: ≥ 1000 subjects	2PL parametric IRT models OR Mokkenscale analysis: 500-999 subjects	2PL parametric IRT models OR Mokkenscale analysis: 250- 499 subjects	2PL parametric IRT models OR Mokken scale analysis: < 250 subjects			
Other							
Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws		(116)	
Box 4. Internal Consistency							
Does the scale consist of effect indicators i.e. is it based on a reflective model? Yes or No							
Design Requirements							
Was an internal consistency statistic calculated foreach unidimensional scale or subscale separately?	Internal consistency statistic calculated for each unidimensional scale or subscale	Unclear whether scale or sub scale is unidimensional	Internal consistency statisticNOT calculated for each unidimensional scale or sub scale			(116)	ICC=>0.70
Statistical methods							
For continuous scores: Was Cronbach's alpha oromega calculated?	Cronbach's alpha, or Omega calculated		Only item-total correlations calculated	No Cronbach's alpha and no item-total correlations calculated	Not Applicable	(116)	

For dichotomous scores: Was Cronbach's alpha or KR-20 calculated?	Cronbach's alpha or KR-20 calculated		Only item-total correlations calculated	No Cronbach's alpha or KR-20 and no item-total correlations calculated	Not Applicable	(116)	
For IRT-based scores: Was standard error of the theta(SE (θ)) or reliability coefficient of estimated latent trait value (index of (subject or item) separation) calculated?	SE(θ) or reliability coefficient calculated			SE(θ) or reliability coefficient NOT calculated	Not Applicable	(116)	
Other							
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study		(116)	
Box 5: Cross-cultural validity/Measurement invariance							
Design requirements							
Were the samples similar for relevant characteristics except for thegroup variable?	Evidence provided that samples were similar for relevant characteristics except group variable	Stated (but no evidence provided)that samples were similar for relevant characteristics except group variable	Unclear whether samples were similar for relevant characteristics exceptgroup variable	Samples were NOT similar for relevant characteristics except group variable		(116)	All GenPop
Statistical Methods							
Was an appropriate approach used to analyse the data?	A widely recognizedor well justified approach was used	Assumable that the approach was appropriate, but notclearly described	Not clear what approach was used ordoubtful whether the approach was appropriate	Approach not appropriate	Not Applicable	(116)	Translation process not stated

Was the sample size included in the analysis adequate?	Regression analysesor IRT/Rasch based analyses: 200 subjects per group	150 subjects pergroup	100 subjects per group	< 100 subjects pergroup		(116)	
	MGCFA*: 7 times the number of itemsand ≥100	5 times the numberof items and ≥100; OR 5-7 times the number of items but <100	5 times the numberof items but <100	<5 times the number of items			
Other							
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study		(116)	
Box 6. Reliability							
Design Requirements							
Were patients stable in the interim period on the construct to be measured?	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable		(87,116)	(2)-not stated, sub sample used (6)-stated that some excluded due to ill health/major life event occurred
Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate		(87,116)	not stated for both

Were the test conditions similar for the measurements? e.g. typeof administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar		(87,116)	not stated for both
Statistical methods							
For Continuous scores: Was an Intraclass correlation coefficient (ICC) calculated?	ICC calculated and model or formula of the ICC described	ICC calculated but model or formula of the ICC not described or optimal. Pearson or Spearman correlation calculated with evidence provided that no systematic change has occurred	Pearson or Spearman correlation coefficient calculated WITHOUT evidence provided that no systematic change has occurred WITH evidence that systematic change has occurred	No ICC or Pearson or Spearman correlations calculated	Not Applicable	(87,116)	
For dichotomous/nominal/ordinal scores: Was kappa calculated?	Kappa calculated			No kappa calculated	Not Applicable	(87,116)	(87,116)
For ordinal scores: Was a weighted kappa calculated?	Weighted Kappa Calculated		Unweighted Kappa calculated or not described		Not Applicable	(2)	(2)-weighted and unweighted calculated
For ordinal scores: Was the weighting scheme described? E.g. linear, quadratic	Weighting scheme described	Weighting scheme NOT described				(2)	
Other							
Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws		(87,116)	
Box 7. Measurement error: absolute measures – Not tested							
Box 8: Criterion Validity – Not tested							

Box 9: Hypothesis testing for construct validity							
9a. Comparison with other outcome measurement instruments (convergent validity)							
Design requirements							
Is it clear what the comparator instrument(s) measure(s)?	Constructs measured by the comparator instrument(s) is clear		Constructs measured by the comparator instrument(s) is not clear			(84,87,187)	All OMs described
Were the measurement properties of the comparatorinstrument(s) sufficient?	Sufficient measurement properties of the comparator instrument(s) in a population similar tothe study population	Sufficient measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties of the comparator instrument(s) in any study population	No information on the measurement properties of the comparator instrument(s), OR evidence for insufficient measurement properties of the comparator instrument(s)		(84,87,187)	All measuring HRQoL or similar dimensions
Statistical methods							
Was the statistical method appropriate for the hypotheses to betested?	Statistical method was appropriate	Assumable that statistical methodwas appropriate	Statistical method applied NOT optimal	Statistical methodapplied NOT appropriate		(84,87,187)	
Other							
Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws		(84,87,187)	
9b. Comparison between subgroups (discriminative or known-groups validity)							

Design requirements							
Was there adequate description provided of important characteristics of the subgroups?	Adequate description of the important characteristics of the subgroups	Adequate description of most of the important characteristics of the subgroups	Poor description of the important characteristics of the subgroups			(84,87,187)	Basic descriptions of medical conditions eg. chronic, CP, acute, etc.
Statistical methods							
Was the statistical method appropriate for the hypotheses to betested?	Statistical method was appropriate	Assumable that statistical methodwas appropriate	Statistical method applied NOT optimal	Statistical methodapplied NOT appropriate		(84,87,187)	
Other							
Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws		(84,87,187)	
Box 10: Responsiveness							
10a. Criterion approach (ie. comparison to a gold standard)							
Statistical Methods							
For continuous scores: Were correlations between change scores,or the area under the Receiver Operator Curve (ROC) curve calculated?	Correlations or Areaunder the ROC Curve (AUC) calculated			Correlations or AUC NOT calculated	Not Applicable	(187)	
For dichotomous scales: Were sensitivity and specificity (changed versus not changed) determined?	Sensitivity and specificity calculated			Sensitivity and specificity NOT calculated	Not Applicable	(187)	
Other							
Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws		(187)	None identified

10b. Construct approach (i.e. hypotheses testing; comparison with other outcome measurement instruments)							
Is it clear what the comparator instrument(s) measure(s)?	Constructs measured by the comparator instrument(s) is clear		Constructs measured by the comparator instrument(s) is not clear			(187)	
Were the measurement properties of the comparatorinstrument(s) sufficient?	Sufficient measurement properties of the comparator instrument(s) in a population similar tothe study population	Sufficient measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties of the comparator instrument(s) in any study population	NO information onthe measurement properties of the comparator instrument(s) OR evidence of poor quality of comparator instrument(s)		(187)	SDQ and KIDSCREEN
Statistical Methods							
Was the statistical method appropriate for the hypotheses to betested?	Statistical methodwas appropriate	Assumable that statistical method were appropriate	Statistical method applied NOT optimal	Statistical methodapplied NOT appropriate		(187)	Floor/ceiling effects at baseline and end-of-treatment
Other							
Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws		(187)	
10c. Construct approach: (i.e. hypotheses testing: comparison between subgroups)							
Design requirements							

Was an adequate description provided of important characteristicsof the subgroups?	Adequate description of the important characteristics of the subgroups	Adequate description of most of the important characteristics of the subgroups	Poor or no descriptionof the important characteristics of the subgroups		Mental conditions broken down to more specific conditions	(187)	
Statistical Methods							
Was the statistical method appropriate for the hypotheses to betested?	Statistical methodwas appropriate	Assumable that statistical method was appropriate	Statistical method applied NOT optimal	Statistical methodapplied NOT appropriate		(187)	
Other							
Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws		(187)	
10d. Construct approach: (i.e. hypotheses testing: before and after intervention)							
Design requirements							
Was an adequate description provided of the intervention given?	Adequate description of theintervention		Poor description of the intervention	NO description of the intervention		(187)	All OMs described
Statistical Methods							
Was the statistical method appropriate for the hypotheses to betested?	Statistical methodwas appropriate	Assumable that statistical method was appropriate	Statistical method applied NOT optimal	Statistical methodapplied NOT appropriate		(187)	
Other							

Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws		(187)	

Appendix 9: COSMIN Risk of Bias Checklist – DISABKIDS-TAKE-6

(level of performance highlighted in grey)

Box General requirement for studies that applied Item Response Theory (IRT) Models							
	Very good	**Adequate**	**Doubtful**	**Inadequate**	**Not applicable**	**Reference** (7)	**Notes**
Box 1. PROM development							
1a. PROM design							
General design requirements							
Is a clear description provided of the construct to be measured?	Construct clearly described			Construct not clearly described	Not Applicable		
Is the origin of the construct clear: was a theory, conceptual framework or disease model used or clear rationale provided todefine the construct to be measured?	Origin of the construct clear		Origin of the construct not clear				
Is a clear description provided of the target population for whichthe PROM was developed?	Target populationclearly described			Target population not clearly described			
Is a clear description provided of the context of use	Context of use clearly described		Context of use not clearly described				
Was the PROM development study performed in a sample representing the target population for which the PROM was developed?	Study performed in a sample representing the target population	Assumable that the study was performed in a sample representing	Doubtful whether the study was performed ina sample	Study not performed in a sample representingthe			

		the target population, but not clearly described	representing the target population	target population **(SKIP items 6-12)**			
Concept elicitation (relevance and comprehensiveness)							
Was an appropriate qualitative data collection method used to identify relevant items for a new PROM?	Widely recognized or well justified qualitative methodused, suitable for the construct and study population	Assumable that the qualitative method was appropriate and suitable forthe construct and study population, but not clearly described	Only quantitative (survey) method(s) used or doubtful whether the method was suitable for the construct and study population	Method used not appropriate or not suitable for the construct or study population			Some sort of cog debriefing by identifying ambiguous words and rephrasing and noting it down for future reference
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewersused	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers nottrained and no experience		Not Applicable		Members of the project team, trained for interviews
Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate, but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable		
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were	Assumable that allgroup meetings or	Not clear if all group meetings or interviews were	No recording and no notes	Not applicable		Interviewer guided interviewees,

	recorded and transcribed verbatim	interviews were recorded and transcribed verbatim, but not clearly described	recorded and transcribed verbatimor recordings not transcribed verbatim or only notes were made during the group meetings/ interviews				ambiguous terms were rephrased and noted
Was an appropriate approach used to analyse the data?	A widely recognized or well justified approach was used	Assumable that the approach was appropriate, but not clearly described	Not clear what approach was used ordoubtful whether theapproach was appropriate	Approach not appropriate			IC – Cronbach ICC – Test-retest Convergent val – Pearson r
Was at least part of the data coded independently?	At least 50% of thedata was coded byat least two researchers independently	11-49% of the data was coded by at least two researchers independently	Doubtful if two researchers were involved in the codingor only 1-10% of the data was coded by at least two researchers independently	Only one researcher was involvedin coding or no coding			Not mentioned
Was data collection continued until saturation was reached?	Evidence providedthat saturation was reached	Assumable that saturation was reached	Doubtful whether saturation was reached	Evidence suggests that saturation was not reached	Not applicable		
For quantitative studies (surveys): was the sample sizeappropriate?	≥100	50-99	30-49	<30	Not applicable		
1b. <u>Cognitive interview study</u> or other pilot test							
Was a cognitive interview study or other pilot test conducted?	YES			**NO (SKIP items 15-35**			
General design requirements							

Was the cognitive interview study or other pilot test performed ina sample representing the target population?	Study performed in a sample representing the target population	Assumable that the study was performed in a sample representing the target population, but not clearly described	Doubtful whether the study was performed in a sample representing the target population	Study not performed in a sample representingthe target population			
Comprehensibility							
Were patients asked about the comprehensibility of the PROM?	YES	Not clear (SKIP standards 17-25)	No **(SKIP standards 17-25)**				Interviewers noted down ambiguous words and rephrased words if necessary
Were all items tested in their final form?	All items were tested in their finalform	Assumable that all items were tested in their final form, but not clearly described	Not clear if all itemswere tested in theirfinal form	Items were not tested in their final form or itemswere not re- tested after substantial adjustments			
Was an appropriate qualitative method used for assessing the comprehensibility of the PROM instructions, items, response options, and recall period?	Widely recognized or well justified qualitative method used	Assumable that the method was appropriate but not clearly described	Only quantitative (survey) method (s) used or doubtful whether the method was appropriate or not clear if patients were asked about the comprehensibility of the items, response options	Method used not appropriate or patients were not asked about the comprehensibility of the items, response options or recall period or patients			

			or recall period or patients not asked about the comprehensibility of the PROM instructions				
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear				
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewersused	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers not trained and no experience		Not Applicable		
Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate, but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable		
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and transcribed verbatim	Assumable that allgroup meetings or interviews were recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings or interviews were recorded and transcribed verbatimor recordings not transcribed verbatim or only notes were made	No recording and no notes	Not applicable		

			during the group meetings/ interviews				
Was an appropriate approach used to analyse the data?	A widely recognized or well justified approach was used	Assumable that the approach wasappropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate			
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only one researcher involved in the analysis				
Were problems regarding the comprehensibility of the PROM instructions, items, response options, and recall period appropriately addressed by adapting the PROM?	No problems found or problems appropriately addressed and PROM was adapted and re-tested if necessary	Assumable that there were no problems or that problems were appropriately addressed, but not clearly described	Not clear if there were problems or doubtful if problems were appropriately addressed	Problems not appropriately addressed or PROM was adapted but items were not re-tested after substantial adjustments	Not applicable		Ambiguous words noted and used for rewriting
Comprehensiveness							
Were patients asked about the comprehensiveness of the PROM?	YES		NO or not clear **(SKIP items 27-35)**				
Was the final set of items tested?	The final set of items was tested	Assumable that the final set of items was tested but not clearly described	Not clear if the final set of items was tested or the final set of items was not tested or the set of items were not re-tested after				

			items were removed or added				
Was an appropriate method used for assessing the comprehensiveness of the PROM?	Widely recognized method used	Assumable that method was appropriate but not clearly described or only quantitative (survey) method (s) used	Doubtful whether the method was appropriate or methodused not appropriate				
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear				
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewersused	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers not trained and no experience		Not Applicable		
Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate, but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable		
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and transcribed verbatim	Assumable that allgroup meetings or interviews were recorded and transcribed	Not clear if all group meetings or interviews were recorded and transcribed verbatimor	No recording and no notes	Not applicable		

		verbatim, but not clearly described	recordings not transcribed verbatim or only notes were made during the group meetings/ interviews				
Was an appropriate approach used to analyse the data?	A widely recognized or well justified approach was used	Assumable that the approach wasappropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate			
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only one researcher involved in the analysis				
Were problems regarding the comprehensiveness of the PROM appropriately addressed by adapting the PROM?	No problems found or problems were appropriately addressed and PROM adapted and re-tested if necessary	Assumable that there were problems or that problems were appropriately addressed but not clearly described	Not clear if there were problems or doubtful if the problems were appropriately addressed or PROM was adapted but items were not re-tested after substantial adjustments	Problems not appropriately addressed	Not applicable		
Box 2 Content validity							
2a. Asking patients about relevance							
Design requirements							

Was an appropriate method used to ask patients whether eachitem is relevant for their experience with the condition?	Widely recognized or well justified method used	Only quantitative (survey) method(s)used or assumable that the method was appropriate but not clearly described	Not clear if patients were asked whethereach item is relevantor doubtful whether the method was appropriate	Method used not appropriate or patients not asked about the relevance of all items			
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear				
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewers used	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers not trained and no experience		Not Applicable		
Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate, but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable		
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and transcribed verbatim	Assumable that allgroup meetings or interviews were recorded and transcribed	Not clear if all group meetings or interviews were recorded and transcribed verbatimor	No recording and no notes	Not applicable		

		verbatim, but not clearly described	recordings not transcribed verbatim or only notes were made during the group meetings/ interviews				
Analyses							
Was an appropriate approach used to analyse the data?	A widely recognized or well justified approach was used	Assumable that the approach wasappropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate			
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only one researcher involved in the analysis				
2b. Asking patients about comprehensiveness							
Was an appropriate method used for assessing the comprehensiveness of the PROM?	Widely recognized or well justified method used	Only quantitative (survey) method(s)used or assumable that the method was appropriate but not clearly described	Doubtful whether the method was appropriate	Method used not appropriate		Interviewer noted any ambiguous words	
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear				

Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewersused	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers nottrained and no experience		Not Applicable		
Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate, but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable		
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and transcribed verbatim	Assumable that allgroup meetings or interviews were recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings or interviews were recorded and transcribed verbatimor recordings not transcribed verbatim or only notes were made during the group meetings/ interviews	No recording and no notes	Not applicable		
Analyses							
Was an appropriate approach used to analyse the data?	A widely recognized or well justified approach was used	Assumable that the approach wasappropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate			

Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only one researcher involved in the analysis				
2c. Asking patients about comprehensibility							
Was an appropriate qualitative method used for assessing the comprehensibility of the PROM instructions, items, response options, and recall period?	Widely recognized or well justified qualitative method used	Assumable that the method was appropriate but not clearly described	Only quantitative (survey) method (s) used or doubtful whether the method was appropriate or not clear if patients were asked about the comprehensibility of the items, response options or recall period or patients not asked about the comprehensibility of the PROM instructions	Method used not appropriate or patients were not asked about the comprehensibility of the items, response options or recall period or patients			
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear				
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewersused	Group moderators/ interviewers had limited	Not clear if group moderators/ interviewers				

		experienceor were trained specifically for thestudy	weretrained or group moderators/ interviewers nottrained and no experience				
Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate, but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable		
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and transcribed verbatim	Assumable that allgroup meetings or interviews were recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings or interviews were recorded and transcribed verbatimor recordings not transcribed verbatim or only notes were made during the group meetings/ interviews	No recording and no notes	Not applicable		
Analyses							
Was an appropriate approach used to analyse the data?	A widely recognized or well justified approach was used	Assumable that the approach wasappropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate			
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in	Not clear if two researchers were included in the analysis or only one				

		the analysis, but not clearly described	researcher involved in the analysis				
2d. Asking professionals about relevance							
Was an appropriate method used to ask professionals whethereach item is relevant for the construct of interest?	A widely recognized or well justified approach was used	Only quantitative (survey) method(s) used or assumable that the method wasappropriate but not clearly described	Not clear if professionals were asked whether each item is relevant or doubtful if the method was appropriate	Method used not appropriate or professionals not asked about the relevance of all items			
Were professionals from all relevant disciplines included?	Professionals from allrequired disciplines were included	Assumable that professionals fromall required disciplines were included, but not clearly described	Doubtful whether professionals from all required disciplines were included or relevant professionals were not included				
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear				
Analyses							
Was an appropriate approach used to analyse the data?	A widely recognized or well justified approach was used	Assumable that the approach wasappropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate			

Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only one researcher involved in the analysis				
2e. Asking professionals about comprehensiveness							
Design requirements							
Was an appropriate method used to for assessing comprehensiveness of the PROM?	A widely recognized or well justified approach was used	Only quantitative (survey) method(s) used or assumable that the method wasappropriate but not clearly described	Not clear if professionals were asked whether each item is relevant or doubtful if the method was appropriate	Method used not appropriate or professionals not asked about the relevance of all items			
Were professionals from all relevant disciplines included?	Professionals from allrequired disciplines were included	Assumable that professionals fromall required disciplines were included, but not clearly described	Doubtful whether professionals from all required disciplines were included or relevant professionals were not included				
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear				
Analyses							
Was an appropriate approach used to analyse the data?	A widely recognized or well justified	Assumable that the approach wasappropriate,	Not clear what approach was used or doubtful	Approach not appropriate			

	approach was used	but not clearly described	whether the approach was appropriate				
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only one researcher involved in the analysis				
Box 3. Structural Validity – Not tested							
Box 4. Internal Consistency							
Does the scale consist of effect indicators i.e. is it based on a reflective model? Yes or No							
Design Requirements							
Was an internal consistency statistic calculated foreach unidimensional scale or subscale separately?	Internal consistency statistic calculated for each unidimensional scale or subscale	Unclear whether scale or sub scale is unidimensional	Internal consistency statisticNOT calculated for each unidimensional scale or sub scale				Ranged from 0.64 to 0.71
Statistical methods							
For continuous scores: Was Cronbach's alpha oromega calculated?	Cronbach's alpha, or Omega calculated		Only item-total correlations calculated	No Cronbach's alpha and noitem-total correlations calculated	Not Applicable		
For dichotomous scores: Was Cronbach's alpha or KR-20 calculated?	Cronbach's alpha or KR-20 calculated		Only item-total correlations calculated	No Cronbach's alpha or KR-20 and no item-total correlations calculated	Not Applicable		
For IRT-based scores: Was standard error of the theta(SE (θ)) or reliability coefficient of estimated	SE(θ) or reliability			SE(θ) or reliability coefficient NOT calculated	Not Applicable		

latent trait value (index of (subject or item) separation) calculated?	coefficient calculated						
Other							
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study			None identified but further research is suggested
Box 5: Cross-cultural validity/Measurement invariance							
Design requirements							
Were the samples similar for relevant characteristics except for thegroup variable?	Evidence provided that samples were similar for relevant characteristics except group variable	Stated (but no evidence provided)that samples were similar for relevant characteristics except group variable	Unclear whether samples were similar for relevant characteristics exceptgroup variable	Samples were NOT similar forrelevant characteristics except group variable			Over 30 languages, VMC process described, used internationally
Statistical Methods							
Was an appropriate approach used to analyse the data?	A widely recognizedor well justified approach was used	Assumable that the approach was appropriate, but notclearly described	Not clear what approach was used ordoubtful whether theapproach was appropriate	Approach not appropriate	Not Applicable		
Was the sample size included in the analysis adequate?	Regression analysesor IRT/Rasch based analyses: 200 subjects per group	150 subjects per group	100 subjects per group	< 100 subjects per group			
	MGCFA*: 7 times the	5 times the numberof items and ≥100; OR 5-	5 times the number of items but <100	<5 times the number of items			

	number of itemsand ≥100	7 times the number of items but <100					
Other							
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study			
Box 6. Reliability							
Design Requirements							
Were patients stable in the interim period on the construct to be measured?	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable			
Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate			
Were the test conditions similar for the measurements? e.g. typeof administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that testconditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar			
Statistical methods							
For Continuous scores: Was an Intraclass correlation coefficient (ICC) calculated?	ICC calculated and model or formula of the ICC described	ICC calculated but model or formula of the ICC not described or optimal. Pearson or Spearman correlation calculated with evidence provided that no systematic	Pearson or Spearman correlation coefficient calculated WITHOUT evidence provided that no systematic change has occurred WITH evidence that systematic change has occurred	No ICC or Pearson or Spearman correlations calculated	Not Applicable		

		change has occurred					
For dichotomous/nominal/ordinal scores: Was kappa calculated?	Kappa calculated			No kappa calculated	Not Applicable		
For ordinal scores: Was a weighted kappa calculated?	Weighted Kappa Calculated		Unweighted Kappa calculated or not described		Not Applicable		
For ordinal scores: Was the weighting scheme described? E.g. linear, quadratic	Weighting scheme described	Weighting scheme NOT described					
Other							
Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws			
Box 7. Measurement error: absolute measures – Not tested							
Box 8: Criterion Validity – Not tested							
Box 9: Hypothesis testing for construct validity							
9a. Comparison with other outcome measurement instruments (convergent validity)							
Design requirements							
Is it clear what the comparator instrument(s) measure(s)?	Constructs measured by the comparator instrument(s) is clear		Constructs measured by the comparator instrument(s) is not clear				KINDL, General Health Questionnaire and CHQ
Were the measurement properties of the comparatorinstrument(s) sufficient?	Sufficient measurement properties of the comparator instrument(s) in a population similar tothe study population	Sufficient measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information onmeasurement properties of the comparator instrument(s) in any study population	No information on the measurement properties of the comparator instrument(s), OR evidence for insufficient measurement properties of the			

				comparator instrument(s)			
Statistical methods							
Was the statistical method appropriate for the hypotheses to betested?	Statistical method was appropriate	Assumable that statistical methodwas appropriate	Statistical method applied NOT optimal	Statistical method applied NOT appropriate			Pearson r used
Other							
Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws			
9b. Comparison between subgroups (discriminative or known-groups validity)							
Design requirements							
Was there adequate description provided of important characteristics of the subgroups?	Adequate description of the important characteristics of the subgroups	Adequate description of most of the important characteristics of the subgroups	Poor description of the important characteristics of the subgroups				Types of chronic illnesses described
Statistical methods							
Was the statistical method appropriate for the hypotheses to betested?	Statistical method was appropriate	Assumable that statistical methodwas appropriate	Statistical method applied NOT optimal	Statistical method applied NOT appropriate			Widely used methods were used
Other							
Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws			None identified
Box 10: Responsiveness – Not tested							

Appendix 10: COSMIN Risk of Bias Checklist – TACQoL

(level of performance highlighted in grey)

Box General requirement for studies that applied Item Response Theory (IRT) Models							
	Very good	**Adequate**	**Doubtful**	**Inadequate**	**Not applicable**	**Reference**	**Notes**
Box 1. PROM development							
1a. PROM design						(85)	
General design requirements							
Is a clear description provided of the construct to be measured?	Construct clearly described			Construct not clearly described	Not Applicable		Construct and development described
Is the origin of the construct clear: was a theory, conceptual framework or disease model used or clear rationale provided todefine the construct to be measured?	Origin of the construct clear		Origin of the construct not clear				
Is a clear description provided of the target population for whichthe PROM was developed?	Target population clearly described			Target population not clearly described			
Is a clear description provided of the context of use	Context of use clearly described		Context of use not clearly described				
Was the PROM development study performed in a sample representing the target population for which the PROM wasdeveloped?	Study performed in a sample representing the target population	Assumable that the study was performed in a sample	Doubtful whether the study was performed ina sample	Study not performed in a sample representingthe			

		representing the target population, but not clearly described	representing the target population	target population **(SKIP items 6-12)**			
Concept elicitation (relevance and comprehensiveness)							
Was an appropriate qualitative data collection method used to identify relevant items for a new PROM?	Widely recognized or well justified qualitative methodused, suitable for the construct and study population	Assumable that the qualitative method was appropriate and suitable for the construct and study population, but not clearly described	Only quantitative (survey) method(s) used or doubtful whether the method was suitable for the construct and study population	Method used not appropriate or not suitable for the construct or study population			
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewersused	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers nottrained and no experience		Not Applicable		
Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate,but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable		

Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and transcribed verbatim	Assumable that allgroup meetings or interviews were recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings or interviews were recorded and transcribed verbatimor recordings not transcribed verbatim or only notes were made during the group meetings/ interviews	No recording and no notes	Not applicable		
Was an appropriate approach used to analyse the data?	A widely recognized or well justified approach was used	Assumable that the approach was appropriate, but not clearly described	Not clear what approach was used ordoubtful whether the approach was appropriate	Approach not appropriate			
Was at least part of the data coded independently?	At least 50% of the data was coded by at least two researchers independently	11-49% of the data was coded by at least two researchers independently	Doubtful if two researchers were involved in the codingor only 1-10% of the data was coded by at least two researchers independently	Only one researcher was involvedin coding or no coding			
Was data collection continued until saturation was reached?	Evidence provided that saturation was reached	Assumable that saturation was reached	Doubtful whether saturation was reached	Evidence suggests that saturation was not reached	Not applicable		
For quantitative studies (surveys): was the sample sizeappropriate?	≥100	50-99	30-49	<30	Not applicable		
1b. <u>Cognitive interview study</u> or other pilot test							
Was a cognitive interview study or	YES			**NO (SKIP**			

other pilot test conducted?				**items 15-35**			
General design requirements							
Was the cognitive interview study or other pilot test performed ina sample representing the target population?	Study performed in a sample representing the target population	Assumable that the study was performed in a sample representing the target population, but not clearly described	Doubtful whether the study was performed in a sample representing the target population	Study not performed in a sample representingthe target population			
Comprehensibility							
Were patients asked about the comprehensibility of the PROM?	YES	Not clear (SKIP standards 17-25)	No **(SKIP standards 17-25)**				
Were all items tested in their final form?	All items were tested in their finalform	Assumable that all items were tested in their final form, but not clearly described	Not clear if all itemswere tested in theirfinal form	Items were not tested in their final form or itemswere not re- tested after substantial adjustments			
Was an appropriate qualitative method used for assessing the comprehensibility of the PROM instructions, items, response options, and recall period?	Widely recognized orwell justified qualitative method used	Assumable that the method was appropriate but not clearly described	Only quantitative (survey) method (s) used or doubtful whether the method was appropriate or not clear if patients were asked about the comprehensibility of the items, response options or recall period or	Method used not appropriate or patients were not asked about the comprehensibility of the items, response options or recall period or patients			

			patients not asked about the comprehensibility of the PROM instructions				
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear				
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewersused	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers not trained and no experience		Not Applicable		
Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate,but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable		
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and transcribed verbatim	Assumable that allgroup meetings or interviews were recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings or interviews were recorded and transcribed verbatimor recordings not transcribed verbatim or only notes were made	No recording and no notes	Not applicable		

			during the group meetings/ interviews				
Was an appropriate approach used to analyse the data?	A widely recognized orwell justified approach wasused	Assumable that the approach wasappropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate			
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only oneresearcher involvedin the analysis				
Were problems regarding the comprehensibility of the PROM instructions, items, response options, and recall period appropriately addressed by adapting the PROM?	No problems found or problems appropriately addressed and PROM was adapted and re-tested if necessary	Assumable that there were no problems or that problems were appropriately addressed, but not clearly described	Not clear if there were problems or doubtful if problems were appropriately addressed	Problems not appropriately addressed or PROM was adapted but items were not re-tested after substantial adjustments	Not applicable		
Comprehensiveness							
Were patients asked about the comprehensiveness of the PROM?	YES		NO or not clear **(SKIP items 27-35)**				
Was the final set of items tested?	The final set of items was tested	Assumable that the final set of items was tested but not clearly described	Not clear if the final set of items was tested or the final set of items was not tested or the set of items were not re-tested				

			after items were removed or added				
Was an appropriate method used for assessing the comprehensiveness of the PROM?	Widely recognized method used	Assumable that method was appropriate but not clearly described or only quantitative (survey) method (s) used	Doubtful whether the method was appropriate or methodused not appropriate				
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear				
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewersused	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers not trained and no experience		Not Applicable		
Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate,but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable		
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and	Assumable that allgroup meetings or interviews were	Not clear if all group meetings or interviews were recorded and	No recording and no notes	Not applicable		

	transcribed verbatim	recorded and transcribed verbatim, but not clearly described	transcribed verbatimor recordings not transcribed verbatim or only notes were made during the group meetings/ interviews				
Was an appropriate approach used to analyse the data?	A widely recognized orwell justified approach wasused	Assumable that the approach wasappropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate			
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only oneresearcher involvedin the analysis				
Were problems regarding the comprehensiveness of the PROM appropriately addressed by adapting the PROM?	No problems found or problems were appropriately addressed and PROM adapted and re-tested if necessary	Assumable that there were problems or that problems were appropriately addressed but not clearly described	Not clear if there were problems or doubtful if the problems were appropriately addressed or PROM was adapted but items were not re-tested after substantial adjustments	Problems not appropriately addressed	Not applicable		
Box 2 Content validity – Not tested							
Box 3. Structural Validity – Not tested							
Box 4. Internal Consistency							

Does the scale consist of effect indicators i.e. is it based on a reflective model? Yes or No							
Design Requirements							
Was an internal consistency statistic calculated foreach unidimensional scale or subscale separately?	Internal consistency statisticcalculated for each unidimensional scale or subscale	Unclear whether scale or sub scale is unidimensional	Internal consistency statisticNOT calculated for each unidimensional scale or sub scale			(75,85,188)	Verrips-0.65-0.84 Vogels-0.71-0.89 Li-0.8995
Statistical methods							
For continuous scores: Was Cronbach's alpha oromega calculated?	Cronbach's alpha, or Omega calculated		Only item-total correlations calculated	No Cronbach's alpha and noitem-total correlations calculated	Not Applicable	(75,85,188)	Cronbach's used
For dichotomous scores: Was Cronbach's alpha or KR-20 calculated?	Cronbach's alpha or KR-20calculated		Only item-total correlations calculated	No Cronbach's alpha or KR-20 and no item-total correlations calculated	Not Applicable		
For IRT-based scores: Was standard error of the theta(SE (θ)) or reliability coefficient of estimated latent trait value (index of (subject or item) separation) calculated?	SE(θ) or reliability coefficient calculated			SE(θ) or reliability coefficient NOT calculated	Not Applicable		
Other							
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study		(75,85,188)	
Box 5: Cross-cultural validity/Measurement invariance – Not tested							
Box 6. Reliability							
Design Requirements							

Were patients stable in the interim period on the construct to be measured?	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable		(75)	
Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate		(75)	Not stated
Were the test conditions similar for the measurements? e.g. typeof administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that testconditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar		(75)	
Statistical methods							
For Continuous scores: Was an Intraclass correlation coefficient (ICC) calculated?	ICC calculated and model or formula of the ICC described	ICC calculated but model or formula of the ICC not described or optimal. Pearson or Spearman correlation calculated with evidence provided that no systematic change has occurred	Pearson or Spearman correlation coefficient calculated WITHOUT evidence provided that no systematic change has occurred WITH evidence that systematic change has occurred	No ICC or Pearson or Spearman correlations calculated	Not Applicable	(75)	ICC=0.44-0.61 Spearman's = 0.289-0.790 Pearson's = 0.24-0.60
For dichotomous/nominal/ordinal scores: Was kappa calculated?	Kappa calculated			No kappa calculated	Not Applicable	(75)	
For ordinal scores: Was a weighted kappa calculated?	Weighted Kappa Calculated		Unweighted Kappa calculated or not described		Not Applicable	(75)	
For ordinal scores: Was the weighting scheme described? E.g. linear, quadratic	Weighting scheme described	Weighting scheme NOT described				(75)	
Other							

Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws		(75)	
Box 7: Measurement error – Not tested							
Box 8: Criterion Validity							
Statistical Methods							
For continuous scores: Were correlations, or the area under the receiver operating curve calculated?	Correlations or AUC calculated			Correlations or AUC NOT calculated	Not Applicable		
For dichotomous scores: Were sensitivity and specificity determined?	Sensitivity and specificity calculated			Sensitivity and specificity NOT calculated	Not Applicable		
Other							
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study			
Box 9: Hypothesis testing for construct validity – Not tested							
Box 10: Responsiveness – Not tested							

Appendix 11: COSMIN Risk of Bias Checklist – CHQ

(level of performance highlighted in grey)

Box General requirement for studies that applied Item Response Theory (IRT) Models							
	Very good	**Adequate**	**Doubtful**	**Inadequate**	**Not applicable**	**Reference**	**Notes**
Box 1. PROM development							
1a. PROM design							
General design requirements							
Is a clear description provided of the construct to be measured?	Construct clearly described			Construct not clearly described	Not Applicable	(76,86,119–121,189–192)	
Is the origin of the construct clear: was a theory, conceptual framework or disease model used or clear rationale provided to define the construct to be measured?	Origin of the construct clear		Origin of the construct not clear			(76,86,119–121,189–192)	
Is a clear description provided of the target population for whichthe PROM was developed?	Target population clearly described			Target population not clearly described		(76,86,119–121,189–192)	
Is a clear description provided of the context of use	Context of use clearly described		Context of use not clearly described			(76,86,119–121,189–192)	
Was the PROM development study performed in a sample representing the target population	Study performed in a sample representing the	Assumable that the study was	Doubtful whether the study was	Study not performed in a sample		(76,86,119–121,189–192)	

for which the PROM was developed?	target population	performed in a sample representing the target population, butnot clearly described	performed ina sample representing the target population	representingthe target population **(SKIP items 6-12)**			
Concept elicitation (relevance and comprehensiveness)							
Was an appropriate qualitative data collection method used toidentify relevant items for a new PROM?	Widely recognized or well justified qualitative methodused, suitable for the construct and study population	Assumable thatthe qualitative method was appropriate and suitable forthe construct and study population, butnot clearly described	Only quantitative (survey) method(s) used or doubtful whether the method was suitable for the construct and study population	Method used not appropriate or not suitable for the construct or study population		(76,86,119–121,189–192)	
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewers used	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers nottrained and no experience		Not Applicable	(76,86,119–121,189–192)	(4) only mentioned trained interviewers
Were the group meetings or interviews based on an	Appropriate topic or interview guide	Assumable that the topic or interview	Not clear if a topic guide was used or doubtful		Not Applicable	(76,86,119–121,189–192)	

appropriatetopic or interview guide?		guide was appropriate, but not clearly described	if topic or interview guide wasappropriate or no guide				
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and transcribed verbatim	Assumable that allgroup meetings or interviews were recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings or interviews were recorded and transcribed verbatimor recordings not transcribed verbatim or only notes were made during the group meetings/ interviews	No recording and no notes	Not applicable	(76,86,119–121,189–192)	None mentioned
Was an appropriate approach used to analyse the data?	A widely recognized or welljustified approachwas used	Assumable thatthe approach was appropriate, but not clearly described	Not clear what approach was used ordoubtful whether the approach was appropriate	Approach not appropriate		(76,86,119–121,189–192)	
Was at least part of the data coded independently?	At least 50% of thedata was coded byat least two researchers independently	11-49% of the data was codedby at least two researchers independently	Doubtful if two researchers were involved in the codingor only 1-10% of the data was coded by at least two researchers independently	Only one researcher was involvedin coding or no coding		(76,86,119–121,189–192)	
Was data collection continued until saturation was reached?	Evidence providedthat saturation was reached	Assumable thatsaturation was reached	Doubtful whether saturation was reached	Evidence suggests that saturation was not reached	Not applicable	(76,86,119–121,189–192)	

For quantitative studies (surveys): was the sample sizeappropriate?	≥100	50-99	30-49	<30	Not applicable		
1b. Cognitive interview study or other pilot test							
Was a cognitive interview study or other pilot test conducted?	YES			**NO (SKIP items 15-35**		(76)	
General design requirements							
Was the cognitive interview study or other pilot test performed ina sample representing the target population?	Study performed in a sample representing the target population	Assumable that the study was performed in a sample representing the target population, butnot clearly described	Doubtful whether the study was performed in a sample representing the target population	Study not performed in a sample representingthe target population		(76)	
Comprehensibility							
Were patients asked about the comprehensibility of the PROM?	YES	Not clear (SKIP standards 17-25)	No **(SKIP standards 17-25)**			(76)	
Were all items tested in their final form?	All items were tested in their finalform	Assumable thatall items were tested in their final form, but not clearly described	Not clear if all itemswere tested in theirfinal form	Items were not tested in their final form or itemswere not re- tested after substantial adjustments			
Was an appropriate qualitative method used for assessing the comprehensibility	Widely recognized or well justified qualitative method used	Assumable that the method was appropriate butnot clearly described	Only quantitative (survey) method (s) used or doubtful whether the method was appropriate or	Method used not appropriate or patients were not asked about the comprehensibilit			

of the PROM instructions, items, response options, and recall period?			not clear if patients were asked about the comprehensibility of the items, response options or recall period or patients not asked about the comprehensibility of the PROM instructions	y of the items, response options or recall period or patients			
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear				
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewers used	Group moderators/ interviewers had limited experienceor were trained specifically for thestudy	Not clear if group moderators/ interviewers weretrained or group moderators/ interviewers not trained and no experience		Not Applicable		
Were the group meetings or interviews based on an appropriatetopic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate, but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable		

Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and transcribed verbatim	Assumable that allgroup meetings or interviews were recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings or interviews were recorded and transcribed verbatimor recordings not transcribed verbatim or only notes were made during the group meetings/ interviews	No recording and no notes	Not applicable		
Was an appropriate approach used to analyse the data?	A widely recognized or well justified approach was used	Assumable that the approach was appropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate			
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only oneresearcher involvedin the analysis				
Were problems regarding the comprehensibility of the PROM instructions, items, response options, and recall period appropriately addressed by adapting the PROM?	No problems found or problems appropriately addressed and PROM was adapted and re-tested if necessary	Assumable thatthere were no problems or that problems were appropriately addressed, but	Not clear if there were problems or doubtful if problems were appropriately addressed	Problems not appropriately addressed or PROM was adapted but items were not re-tested after substantial adjustments	Not applicable		

		not clearly described					
Comprehensiveness							
Were patients asked about the comprehensiveness of the PROM?	YES		NO or not clear **(SKIP items 27-35)**				
Was the final set of items tested?	The final set of items was tested	Assumable that the final set of items was tested but not clearly described	Not clear if the final set of items was tested or the final set of items was not tested or the set of items were not re-tested after items were removed or added				
Was an appropriate method used for assessing the comprehensiveness of the PROM?	Widely recognized method used	Assumable that method was appropriate but not clearly described or only quantitative (survey) method (s) used	Doubtful whether the method was appropriate or methodused not appropriate				
Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear				
Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewers used	Group moderators/ interviewers had limited experienceor	Not clear if group moderators/ interviewers weretrained		Not Applicable		

		were trained specifically for thestudy	or group moderators/ interviewers not trained and no experience				
Were the group meetings or interviews based on an appropriatetopic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate, but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide wasappropriate or no guide		Not Applicable		
Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and transcribed verbatim	Assumable that allgroup meetings or interviews were recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings or interviews were recorded and transcribed verbatimor recordings not transcribed verbatim or only notes were made during the group meetings/ interviews	No recording and no notes	Not applicable		
Was an appropriate approach used to analyse the data?	A widely recognized or well justified approach was used	Assumable that the approach was appropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate			
Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that atleast two researchers	Not clear if two researchers were included in the analysis or only				

		were involved in the analysis, but not clearly described	oneresearcher involvedin the analysis				
Were problems regarding the comprehensiveness of the PROM appropriately addressed by adapting the PROM?	No problems found or problems were appropriately addressed and PROM adapted and re-tested if necessary	Assumable that there were problems or that problems were appropriately addressed but not clearly described	Not clear if there were problems or doubtful if the problems were appropriately addressed or PROM was adapted but items were not re-tested after substantial adjustments	Problems not appropriately addressed	Not applicable		
Box 2 Content validity – Not tested							
Box 3. Structural Validity							
Does the scale consist of effect indicators, i.e. is it based on a reflective model? Yes/No							
Does the study concern unidimensionality or structural validity?							
Statistical methods							
For CTT: Was exploratory or confirmatory factor analysis performed?	Confirmatory factor analysis performed	Exploratory factor analysis performed		No exploratory or confirmatory factoranalysis performed	Not Applicable	(190–192)	
For IRT: Were IRT/Rasch: does the chosen model fit to the research topic	Chosen model fits wellto the research question	Assumable that the chosen model fits wellto the research question	Doubtful if the chosen model fitswell to the research question	Chosen model doesnot fit to the research question	Not Applicable	(190–192)	

	FA: 7 times the numberof items and ≥100	FA: at least 5 times the number of items and ≥100; OR at least6 times number of items but <100	FA: 5 times the number of items but <100	FA: < 5 times the number of items		(190–192)	Hepner-stated Drotat and Ferro-not stated
	Rasch/1PL models: ≥ 200 subjects	Rasch/1PL models:100-199 subjects	Rasch/1PL models:50-99 subjects	Rasch/1PL models: <50 subjects			
	2PL parametric IRT models OR Mokken scale analysis: ≥ 1000 subjects	2PL parametric IRT models OR Mokkenscale analysis: 500-999 subjects	2PL parametric IRT models OR Mokkenscale analysis: 250-499 subjects	2PL parametric IRT models OR Mokken scale analysis: < 250 subjects			
Other							
Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws		(190–192)	
Box 4. Internal Consistency							
Does the scale consist of effect indicators i.e. is it based on a reflective model? Yes or No							
Design Requirements							
Was an internal consistency statistic calculated foreach unidimensional scale or subscale separately?	Internal consistency statistic calculated for each unidimensional scale or subscale	Unclear whether scale or sub scale is unidimensional	Internal consistency statisticNOT calculated for each unidimensional scale or sub scale			(76,119,121,189)	Woods=0.62 Landgraf=>0.40 Raat=>0.70 Norby=0.86-0.94 CHQ/HUI=0.65-0.97 one exception=0.52
Statistical methods							

For continuous scores: Was Cronbach's alpha oromega calculated?	Cronbach's alpha, or Omega calculated		Only item-total correlations calculated	No Cronbach's alpha and no item-total correlations calculated	Not Applicable	(76,119,121,189)	
For dichotomous scores: Was Cronbach's alpha or KR-20 calculated?	Cronbach's alpha or KR-20 calculated		Only item-total correlations calculated	No Cronbach's alpha or KR-20 and no item-total correlations calculated	Not Applicable		
For IRT-based scores: Was standard error of the theta(SE (θ)) or reliability coefficient of estimated latent trait value (index of (subject or item) separation) calculated?	SE(θ) or reliability coefficient calculated			SE(θ) or reliability coefficient NOT calculated	Not Applicable		
Other							
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study		(76,119,121,189)	
Box 5: Cross-cultural validity/Measurement invariance							
Design requirements							
Were the samples similar for relevant characteristics except for thegroup variable?	Evidence provided that samples were similar for relevant characteristics except group variable	Stated (but no evidence provided)that samples were similar for relevant characteristics except group variable	Unclear whether samples were similar for relevant characteristics exceptgroup variable	Samples were NOT similar for relevant characteristics except group variable		(76,120,121)	
Statistical Methods							

Was an appropriate approach used to analyse the data?	A widely recognizedor well justified approach was used	Assumable that the approach was appropriate, but notclearly described	Not clear what approach was used ordoubtful whether the approach was appropriate	Approach not appropriate	Not Applicable	(76,120,121)	Translation method described 2-4 languages tested in study
Was the sample size included in the analysis adequate?	Regression analysesor IRT/Rasch based analyses: 200 subjects per group	150 subjects pergroup	100 subjects per group	< 100 subjects pergroup		(76,120,121)	<100 per country
	MGCFA*: 7 times the number of items and ≥100	5 times the numberof items and ≥100; OR 5-7 times the number of items but <100	5 times the numberof items but <100	<5 times the number of items		(76,120,121)	
Other							
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study		(76,120,121)	
Box 6. Reliability							
Design Requirements							
Were patients stable in the interim period on the construct to be measured?	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable		(119,120)	
Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate		(119,120)	Raat=3wks Pistorio=2-16d

							Recommended=2d -2wk
Were the test conditions similar for the measurements? e.g. typeof administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar		(119,120)	
Statistical methods							
For Continuous scores: Was an Intraclass correlation coefficient (ICC) calculated?	ICC calculated and model or formula of the ICC described	ICC calculated but model or formula of the ICC not described or optimal. Pearson or Spearman correlation calculated with evidence provided that no systematic change has occurred	Pearson or Spearman correlation coefficient calculated WITHOUT evidence provided that no systematic change has occurred WITH evidence that systematic change has occurred	No ICC or Pearson or Spearman correlations calculated	Not Applicable	(119,120)	Raat-psychosocial dimension=0.14-0.78, gen behaviour + psych=>0.70, other=0.50-0.70 8=0.4-0.8
For dichotomous/nominal/ordinal scores: Was kappa calculated?	Kappa calculated			No kappa calculated	Not Applicable	(119,120)	
For ordinal scores: Was a weighted kappa calculated?	Weighted Kappa Calculated		Unweighted Kappa calculated or not described		Not Applicable	(119,120)	
For ordinal scores: Was the weighting scheme described? E.g. linear, quadratic	Weighting scheme described	Weighting scheme NOT described				(119,120)	
Other							
Were there any other important flaws in the design or statistical methods of the study?	No other important methodologica l		Other minor important methodological flaws	Other important methodological flaws		(119,120)	

	flaws						
Box 7. Measurement error: absolute measures							
Design requirements							
Were patients stable in the interim period on the construct to be measured?	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable		(190–192)	
Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate		(190–192)	Not stated
Were the test conditions similar for the measurements? e.g. typeof administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar		(190–192)	
Statistical methods							
For continuous scores: Was the Standard Error of Measurement (SEM), Smallest Detectable Change (SDC) or Limits of Agreement(LoA) calculated?	SEM, SDC, or LOA calculated	Possible to calculate LoA from the data presented		SEM calculated based on Cronbach's alpha, or on SD from another population	Not Applicable		
For dichotomous/ nominal/ordinal scores: Was the percentage (positive and negative) agreement calculated?	% positive and negative agreement calculated	% positive agreement calculated		% agreement not calculated	Not Applicable	(190–192)	
Other							
Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor important methodological flaws	Other important methodological flaws		(190–192)	
Box 8: Criterion Validity – Not tested							
Box 9: Hypothesis testing for construct validity							

9a. Comparison with other outcome measurement instruments (convergent validity)							
Design requirements							
Is it clear what the comparator instrument(s) measure(s)?	Constructs measured by the comparator instrument(s) is clear		Constructs measured by the comparator instrument(s) is not clear			(119,190,192)	Drotar,8-CHAQ Raat-VAS Ferro-family orientated measures
Were the measurement properties of the comparatorinstrument(s) sufficient?	Sufficient measurement properties of the comparator instrument(s) in a population similar tothe study population	Sufficient measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties of the comparator instrument(s) in any study population	No information on the measurement properties of the comparator instrument(s), OR evidence for insufficient measurement properties of the comparator instrument(s)		(119,190,192)	
Statistical methods							
Was the statistical method appropriate for the hypotheses to betested?	Statistical method was appropriate	Assumable that statistical methodwas appropriate	Statistical method applied NOT optimal	Statistical methodapplied NOT appropriate		(119,190,192)	
Other							
Were there any other important flaws in the design or statistical methods of the study?	No other important methodologica l flaws		Other minor important methodological flaws	Other important methodological flaws		(119,190,192)	
9b. Comparison between subgroups (discriminative or known-groups validity)						(119)	
Design requirements							

Was there adequate description provided of important characteristics of the subgroups?	Adequate description of the important characteristics of the subgroups	Adequate description of most of the important characteristics of the subgroups	Poor description of the important characteristics of the subgroups				Raat-Only stated from elementary schools – no indication of presence/absence of medical conditions
Statistical methods							
Was the statistical method appropriate for the hypotheses to betested?	Statistical method was appropriate	Assumable that statistical methodwas appropriate	Statistical method applied NOT optimal	Statistical methodapplied NOT appropriate			
Other							
Were there any other important flaws in the design or statistical methods of the study?	No other important methodologica l flaws		Other minor important methodological flaws	Other important methodological flaws			
Box 10: Responsiveness – Not tested							

Appendix 12: EQ-5D-Y-3L-SC (English Version for South Africa)

NAME OF CHILD :

DATE OF ADMINISTRATION OF FORM :

Health Questionnaire

English version for South Africa

EQ-5D-Y

Describing your health TODAY

Under each heading, please tick the ONE box that best describes your health TODAY

Mobility *(walking about)*

I have no problems walking about ❑

I have some problems walking about ❑

I have a lot of problems walking about ❑

Looking after myself

I have no problems washing or dressing myself ❑

I have some problems washing or dressing myself ❑

I have a lot of problems washing or dressing myself ❑

Doing usual activities *(for example, going to school, hobbies, sports, playing, doing things with family or friends)*

I have no problems doing my usual activities ❑

I have some problems doing my usual activities ❑

I have a lot of problems doing my usual activities ❑

Having pain or discomfort

I have no pain or discomfort ❑

I have some pain or discomfort ❑

I have a lot of pain or discomfort ❑

Feeling worried, sad or unhappy

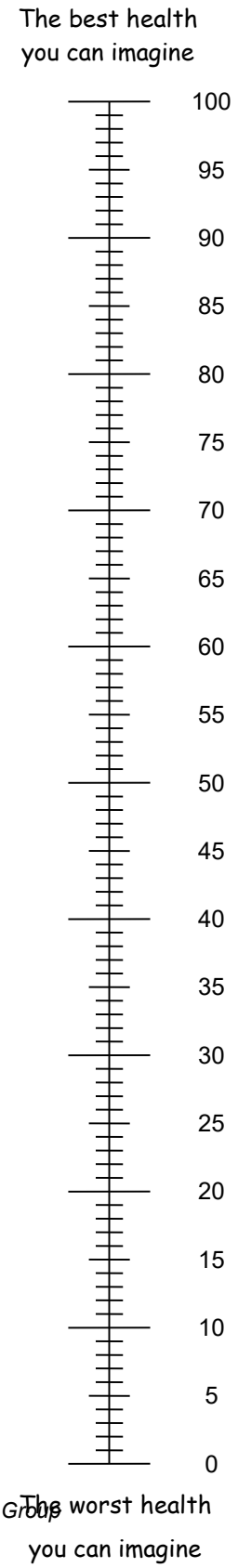
I am not worried, sad or unhappy ❑

I am a bit worried, sad or unhappy ❑

I am very worried, sad or unhappy ❑

How good is your health TODAY

- We would like to know how good or bad your health is TODAY.
- This line is numbered from 0 to 100.
- 100 means the best health you can imagine.
  0 means the worst health you can imagine.
- Please mark with an X on the line to show how good or bad your health is TODAY.



Appendix 13: EQ-5D-Y-3L-IA (English Version for South Africa)

Health Questionnaire

English version for South Africa

VERSION FOR INTERVIEWER ADMINISTRATION

Note to interviewer: although allowance should be made for the interviewer's particular style of speaking, the wording of the questionnaire instructions should be followed as closely as possible. In the case of the EQ-5D-Y descriptive system on page 2 of the questionnaire, the precise wording must be followed.

If the child (respondent) has difficulty choosing a response, or asks for clarification, the interviewer should repeat the question word for word and ask the child to answer in a way that most closely resembles his or her thoughts about his or her health today.

INTRODUCTION

(Note to interviewer: please read the following to the child.)

We are trying to find out what you think about your health. I will explain what to do as I go along, but please stop me if you do not understand something or if things are not clear to you. There are no right or wrong answers. We are interested only in what you think.

First, I am going to read out some questions. Each question has a choice of three answers. Please tell me which answer best describes your health TODAY.

Do not choose more than one answer in each group of questions.

(Note to interviewer: first read all three options for each question. Then ask the child to choose which one applies to him/herself. Repeat the question and options if necessary. Mark the appropriate box under each heading. You may need to remind the child regularly that the timeframe is TODAY.)

EQ-5D DESCRIPTIVE SYSTEM

First, I would like to ask you about WALKING ABOUT (MOBILITY). Would you say that:	
1. You have no problems walking about?	❑
2. You have some problems walking about?	❑
3. You have a lot of problems walking about?	❑

Next, I would like to ask you about LOOKING AFTER YOURSELF. Would you say that:	
1. You have no problems washing or dressing yourself?	❑
2. You have some problems washing or dressing yourself?	❑
3. You have a lot of problems washing or dressing yourself?	❑

Next, I would like to ask you about DOING USUAL ACTIVITIES, for example, going to school, hobbies, sports, playing, doing things with family or friends. Would you say that:	
4. You have no problems doing your usual activities?	❑
5. You have some problems doing your usual activities?	❑
6. You have a lot of problems doing your usual activities?	❑

Next, I would like to ask you about HAVING PAIN OR DISCOMFORT. Would you say that:	
7. You have no pain or discomfort?	❑
8. You have some pain or discomfort?	❑
9. You have a lot of pain or discomfort?	❑

Finally, I would like to ask you about FEELING WORRIED, SAD OR UNHAPPY. Would you say that:	
10. You are not worried, sad or unhappy?	❑
11. You are a bit worried, sad or unhappy?	❑
12. You are very worried, sad or unhappy?	❑

EQ-5D VAS

- **Now, I would like to ask you to say how good or bad your health is TODAY.**

- **I would like you to picture in your mind a vertical line that is numbered from 0 to 100.**

(Note to interviewer: if interviewing face-to-face, please show the child the VAS line.)

- **100 at the top of the line means the <u>best</u> health you can imagine.**

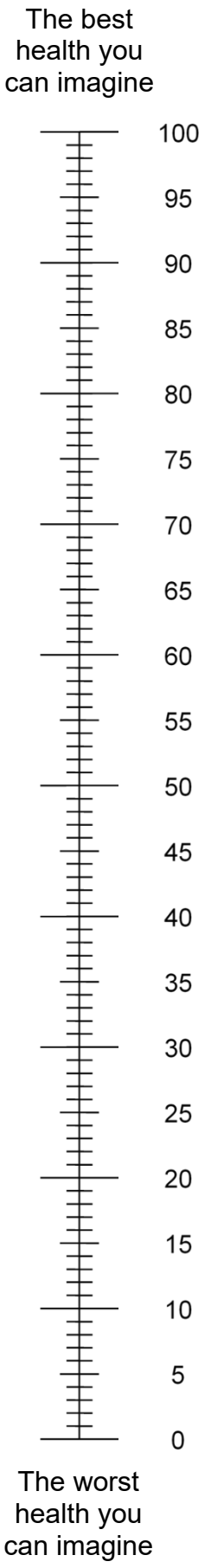
0 at the bottom of the line means the <u>worst</u> health you can imagine.

- **I would now like you to tell me the point on this line where you would put your health TODAY.**

(Note to interviewer: mark the line at the point indicating the child's health today. Now, please write the number you marked on the line in the box below.)

THE CHILD'S HEALTH TODAY = ☐

Thank you for taking the time to answer these questions.



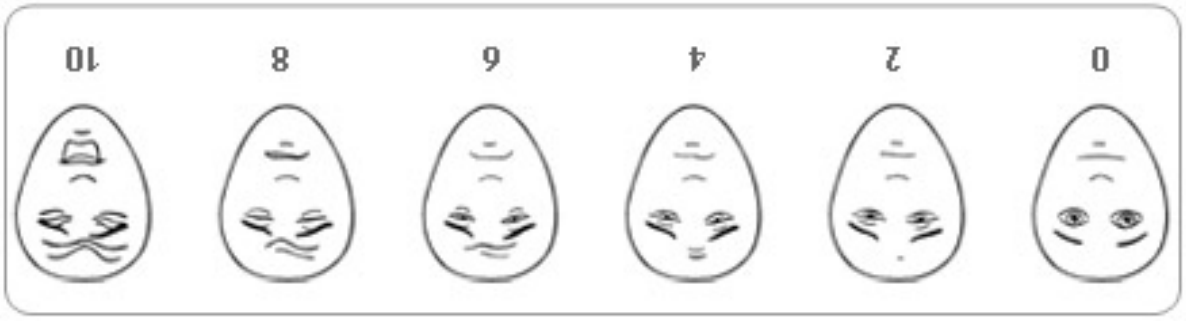
Appendix 14: Faces Pain Scale-R (FPS-R)

In the following instructions, say "hurt" or "pain," whichever seems right for a particular child.
"These faces show how much something can hurt. This face *[point to left-most face]* **shows no pain**.
The faces show more and more pain *[point to each from left to right]* **up to this one** *[point to right-most face]* **- it shows very much pain**.
Point to the face that shows how much you hurt [right now]**."**

*Score the chosen face **0**, **2**, **4**, **6**, **8**, or **10**, counting left to right, so '0' = 'no P/D' and '10' = 'very much P/D.' Do not use words like 'happy' and 'sad'. This scale is intended to measure how children feel inside, not how their face looks.*

(fold along dotted line)

……



Appendix 15: Moods and Feelings Questionnaire (MFQ)

Child Self-Report

MOOD AND FEELINGS QUESTIONNAIRE: Short Version

This form is about how you might have been feeling or acting recently.

For each question, please check (☐)how you have been feeling or acting in the past two weeks.

If a sentence was not true about you, check NOT TRUE.

If a sentence was only sometimes true, check SOMETIMES.

If a sentence was true about you most of the time, check TRUE.

Score the MFQ as follows:

NOT TRUE = 0

SOMETIMES = 1

TRUE = 2

To code, please use a checkmark (X) for each statement	NOT TRUE	SOMETIMES TRUE	TRUE
1. I felt miserable or unhappy.			
2. I didn't enjoy anything at all.			
3. I felt so tired I just sat around and did nothing.			
4. I was very restless.			
5. I felt I was no good anymore.			
6. I cried a lot			
7. I found it hard to think properly or concentrate.			
8. I hated myself.			
9. I was a bad person.			
10. I felt lonely.			
11. I thought nobody really loved me.			
12. I thought I could never be as good as other kids.			
13. I did everything wrong.			

Appendix 16: WeeFIM

WeeFIM® Instrument

LEVELS		
	7 Complete Independence (timely, safely) 6 Modified Independence (device)	NO HELPER
	Modified Dependence 5 Supervision (subject = 100%) 4 Minimal Assistance (subject = 75%+) 3 Moderate Assistance (subject = 50%+) **Complete Dependence** 2 Maximal Assistance (subject =25%+) 1 Total Assistance (subject = less than 25%)	HELPER

	ASSESSMENT	GOAL
Self-Care		
A. Eating	☐	☐
B. Grooming	☐	☐
C. Bathing	☐	☐
D. Dressing - Upper Body	☐	☐
E. Dressing - Lower Body	☐	☐
F. Toileting	☐	☐
G. Bladder	☐	☐
H. Bowel	☐	☐
Self-Care Total	☐	
Mobility		
I. Chair, Wheelchair	☐	☐
J. Toilet	☐	☐
K. Tub, Shower	☐ (W Walk, C Wheelchair, L Crawl, B Combination)	☐
L. Walk, Wheelchair	☐	☐
M. Stairs	☐	☐
Mobility Total	☐	
MOTOR SUBTOTAL RATING	☐	
Cognition		
N. Comprehension	☐ (A Auditory, V Visual, B Both)	☐
O. Expression	☐ (V Vocal, N Nonvocal, B Both)	☐
P. Social Interaction	☐	☐
Q. Problem Solving	☐	☐
R. Memory	☐	☐
Cognition Total	☐	
WEEFIM® TOTAL RATING	☐	

NOTE: Leave no blanks. Enter 1 if patient is not testable due to risk.

Appendix 17: Study specific demographic questionnaire

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD
HEALTH SCIENCES

DHRS
Department of Health and Rehabilitation Sciences
FACULTY OF HEALTH SCIENCES

Divisions of Communication Sciences & Disorders • Disability Stu
Nursing & Midwifery • Occupational Therapy • Physiotherapy

F45 Old Main Building, Groote Schuur Hospital
Observatory, Cape Town, South Africa, 7925
Telephone: +27 (0) 21 406 6401
Website: www.dhrs.uct.ac.za

Demographic questionnaire to be completed by the parent/legal guardian of each child

Dear Parents/Caregivers

Please complete the questionnaire below by filling in your responses in the space provided or crossing the appropriate box. This will enable the researchers to understand your child's health and abilities.

1. Today's date: ……
2. Child's name and surname: ……
3. Child's date of birth: ……
4. Sex of child: MALE ☐ FEMALE ☐
5. What grade is your child in? ……
6. Relationship of guardian to child: ……
7. Has your child been diagnosed with a chronic illness or disability, by a doctor? YES ☐ NO ☐
8. If yes, what type of illness/disability was your child diagnosed with?

……

……

9. Is your child currently using any medication? YES ☐ NO ☐

10. If yes, please list the medication: ……

……

……

11. Is your child currently using a mobility device to get around or wearing orthoses?
If no, please tick the N/A box.
If yes, please tick the correct type of assistive device below.

N/A ☐ Wheelchair ☐ Rollator ☐ Walking frame ☐

Crutches ☐ Orthoses ☐

Other: ……

Thank you for taking the time to complete this form.

Appendix 18: Preference between SC and IA versions

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD
HEALTH SCIENCES

DHRS
Department of Health and Rehabilitation Sciences
FACULTY OF HEALTH SCIENCES

Divisions of Communication Sciences & Disorders • Disability Studies • Nursing & Midwifery • Occupational Therapy • Physiotherapy

F45 Old Main Building, Groote Schuur Hospital
Observatory, Cape Town, South Africa, 7925
Telephone: +27 (0) 21 406 6401
Website: www.dhrs.uct.ac.za

Did you prefer to answer the questions when I read them out to you or when you filled them in yourself?

Preferred IA version ☐ Preferred SC version ☐

Why did you prefer that?

……

……

……

Thank you.

Appendix 19: HREC approval

UNIVERSITY OF CAPE TOWN
Faculty of Health Sciences
Human Research Ethics Committee

G50, G Floor, Old Main Building
Groote Schuur Hospital
Observatory 7925
Telephone [021] 650 1236
Email: hrec-enquiries@uct.ac.za
Website: www.health.uct.ac.za/fhs/research/humanethics/forms

02 February 2021

HREC REF: 369/2020

Ms Des Scott
Health & Rehab Sciences
Division Physiotherapy
F-Floor, Old Main Building
Email: des.scott@uct.ac.za
Student: AMNRAZ002@myuct.ac.za

Dear Ms Des Scott

PROJECT TITLE: *PERFORMANCE OF THE EQ-5D-Y INTERVIEWER ADMINISTERED VERSION IN YOUNG CHILDREN AGED 5-8YEARS (STUDY LINKED TO 128/2020)-MSC CANDIDATE-MS RAZIA AMIEN*

Thank you for submitting your study to the Faculty of Health Sciences Human Research Ethics Committee.

It is a pleasure to inform you that the HREC has **formally approved** the above-mentioned study.

Approval is granted for one year until the 30 August 2021.

This approval is subject to strict adherence to the HREC recommendations regarding research involving human participants during COVID-19, dated 17 March 2020 & 6 July 2020.

Please submit a progress form, using the standardised Annual Report Form if the study continues beyond the approval period. Please submit a Standard Closure form if the study is completed within the approval period.
(Forms can be found on our website: www.health.uct.ac.za/fhs/research/humanethics/forms)

The HREC acknowledge that the student: - Ms Razia Amien will also be involved in this study.

Please quote the HREC REF 369.2020 in all your correspondence.

Please note that for all studies approved by the HREC, the principal investigator **must** obtain appropriate institutional approval, where necessary, before the research may occur.

Please also note that the ongoing ethical conduct of the study remains the responsibility of the principal investigator.

Yours sincerely

PROFESSOR M BLOCKMAN
CHAIRPERSON, FACULTY OF HEALTH SCIENCES HUMAN RESEARCH ETHICS COMMITTEE
Federal Wide Assurance Number: FWA00001637.
Institutional Review Board (IRB) number: IRB00001938
This serves to confirm that the University of Cape Town Human Research Ethics Committee complies to the Ethics Standards for Clinical Research with a new drug in patients, based on the Medical Research Council (MRC-SA), Food and Drug Administration (FDA-USA), International Convention on Harmonisation Good Clinical Practice (ICH GCP), South African Good Clinical Practice Guidelines (DoH 2005), based on the Association of the British Pharmaceutical Industry Guidelines (ABPI), and Declaration of Helsinki (2013) guidelines.
The Human Research Ethics Committee granting this approval is in compliance with the ICH Harmonised Tripartite Guidelines E6: Note for Guidance on Good Clinical Practice (CPMP/ICH/135/95) and FDA Code Federal Regulation Part 50, 56 and 312.

HREC REF NO. 369/2020 SA

Form FHS007: Amendment – study staff

HREC office use only (FWA00001637; IRB00001938)			
☑ Approved			
This serves as notification that all changes to the study staff and documentation described below are approved.			
Chairperson of the HREC signature/ Designee		Date	23/1/21

Note:
Please note that incomplete amendment submissions will not be reviewed.

Please email this form and supporting documents (if applicable) in a combined pdf-file to hrec-enquiries@uct.ac.za.

Please clarify your plan for research-related activities during COVID-19 lockdown.

Principal Investigator to complete the following:

1. Protocol Information

Date (when submitting this form)	20/01/2021
HREC REF Number	HREC 369/2020
Protocol title	Performance of the EQ-5D-Y interviewer administered version in young children aged 5-8 years (study linked to 128/2020)
Protocol number (if applicable)	
Principal Investigator	Ms Des Scott
Department / Office Internal Mail Address	Department of Health and Rehab Sciences Division of Physiotherapy F-Floor, Old Main Building Groote Schuur Hospital Des.scott@uct.ac.za

1.1 Does this protocol receive US Federal funding?	☐ Yes	☒ No

2.1 Staff changes (tick ✓)

Are new personnel being added to this research?	☒ Yes	☐ No
Are current personnel being removed from this research?	☐ Yes	☒ No

Is the principal investigator for this research being changed? If yes, please attach revised conflict of interest and PI declaration statements. (Refer: sections 7 and 8.3 in the New Protocol Application Form - FHS013)	☐ Yes	☒ No
Do the consent and assent forms need modification to reflect these staff changes? If yes, please attach copies of the revised forms, with all changes highlighted or tracked and listed in the documents for approval.	☒ Yes	☐ No

2.2 Amended study staff details

Title, first name, surname	Department/Division	E-mail	Role of new staff member
Ms Razia Amien	Division of Physiotherapy	AMNRAZ001@myuct.ac.za	Student Investigator

3. List of documentation for approval

Please list below all staff documentation such as CVs, declarations, GCP certificates and revised consent forms which need approval. This information must correspond to all 'yes' answers in 2.1 above. This form will be signed and returned to the PI as notification of approval. Please add extra pages if necessary.

-CV Ms Razia Amien
-HPCSA Registration Ms Razia Amien
-Ethics Training Certificate Ms Razia Amien
-Revised Consent form with tracked changes
-Revised Assent form with tracked changes

4. Signature

My signature certifies that I will maintain the anonymity and/ or confidentiality of information collected in this research. If at any time I want to share or re-use the information for purposes other than those disclosed in the original approval, I will seek further approval from the HREC.

Signature of PI		Date	20/01/2021

Appendix 20: Ministerial permission to conduct non-therapeutic research with minors (Form A)

DEPARTMENT OF HEALTH APPLICATION FOR MINISTERIAL CONSENT FOR NON-THERAPEUTIC RESEARCH WITH MINORS

1. **INSTRUCTIONS**

1. This application form must be completed for all protocols that are classified as "non-therapeutic" and involve the participation of minors. *Non therapeutic research is defined in the regulations relating to research on human participants as "research that does not hold out the prospect of direct benefit but holds out the prospect of generalizable knowledge". Minors are defined as persons under the age of 18 by section 17 of the Children's Act (No. 38 of 2005).*

2. This application form should be submitted with a copy of the protocol and supporting documents.

3. This application should be submitted to the Minister of Health or the delegated authority in terms of section 92(a) of the Act.

4. This application form should describe how 'non-therapeutic' research protocols with minors meet the conditions set out in section 71 (3)(b) of the Act (described below).

5. All sections of the form must be completed in full.

6. Ministerial Consent may be granted for non-therapeutic health research with minors when certain conditions set out in section 71 (3)(b) of the Act are met and these conditions are:

    1. The research objectives cannot be achieved except by the enrolment of minors;
    2. The research is likely to lead to an improved scientific understanding of conditions, or disorders affecting children;

3. Any consent given to the research must be in line with public policy; and
4. The research does not pose a significant risk to minors, and if there is some risk, the benefit of the research outweighs the risk.

INVESTIGATORS' DETAILS

Name of principal investigator	Mrs Desiree Scott
Title of research protocol	Performance of the EQ-5D-Y Interviewer Administered Version in young children aged 5-7-years
Institutional affiliation	Faculty of Health Sciences, Division of Physiotherapy, University of Cape Town
Postal Address	F45 Old Main Building, Groote Schuur Hospital, Observatory 7925
Physical Address	F45 Old Main Building, Groote Schuur Hospital, Observatory 7925
Email Address	des.scott@uct.ac.za
Phone	0214066401/7679 or 083 949 8333
Fax	0214066401/7679
Date of Application	23 June 2020
Signature of Applicant	

2. **APPLICATION**

    1. **Condition 1: The research objectives cannot be achieved except by the participation of minors**

Describe the scientific justification for the enrolment of minors. Explain why this research must be done with minors as participants:

The EQ-5D-Y-3L has recently been developed with an expanded number of response options to measure Health Related Quality of Life in Children and Adolescents aged 8-15-years. It is a self-complete measure with child friendly language and layout. As children's opinions are important in healthcare decision-making it is imperative that they be given the opportunity to self-report on such measures. Thus, this research aims to establish whether the measure is valid and reliable for use in children aged 8-15 years in South Africa.

    2. **Condition 2: The research is likely lead to an improved scientific understanding of certain conditions, diseases or disorders affecting minors**

Describe how the research might, or aims to, advance knowledge affecting the health and welfare of minors as a class. Note that 'condition' is defined in the Regulations as 'physical and psycho- social characteristics understood to affect health' allowing that this research does not only involve children with an illness.

This research aims to show that the EQ-5D-Y-3L is a valid and reliable measure of HRQoL in English children in South Africa. If found to be valid and reliable this measure could be used in the future for population health surveys in children, to assist in healthcare decision-making on an individual basis or for population health with economic evaluation, it could further assist in tracking disease trajectory across the lifespan in diseases which were previously classified as 'childhood conditions.'

3. **Condition 3: Any consent given to the research is in line with public policy**

Consent given by authorised persons must be in line with public policy considerations. Describe how consent to the research will be in line with public policy or would be acceptable, for example, show how the research poses acceptable risks and promotes the rights of minors:

Permission will be obtained from the University of Cape Town Human Research Ethics Committee, Department of Education and the corresponding hospital management and ethics committees before commencement. Consent will be obtained from the parent/legal guardian to allow the individual to participate in the research, and assent will be obtained from the child. The purpose, procedure, risks and benefits, compensation and voluntary participation will be explained in the consent and assent forms. The school, parent/legal guardian and the child or adolescent participating in the study may discontinue with the study at any point, even after signing the consent and assent forms. The research will cause no harm to the school, the parents/legal guardians, or the children and adolescents participating in the study.

4. **Condition 4: The research does not pose a significant risk to minors; and if there is some risk, the benefit of the research outweighs the risk.**

Describe how the potential risks from the research procedures and/or intervention to minor participants will be minimized and describe any possible benefits from the research to society in the form of knowledge:

There will be no harm caused to the minor participants. Data will be collected by means of a survey only. Participation in the study will not affect the medical treatment that the child is receiving or any future treatment or management. This research will provide insight into the effectiveness of this generic measure, the EQ-5D-Y-3L, in assessing the health-related quality of life in children and adolescents.

Appendix 21: Permission from the Western Cape Educational Department to conduct research at schools

Western Cape Government
Education

Directorate: Research

Audrey.wyngaard@westerncape.gov.za
tel: +27 021 467 9272
Fax: 0865902282
Private Bag x9114, Cape Town, 8000
wced.wcape.gov.za

REFERENCE: 20200901-7797
ENQUIRIES: Dr A T Wyngaard

Ms Razia Amien
23 Penlyn Avenue
Penlyn Estate
7780

Dear Ms Razia Amien

RESEARCH PROPOSAL: PERFORMANCE OF THE EQ-5D-Y INTERVIEWER ADMISTRATED VERSION IN YOUNG CHILDREN AGED 5 – 8 YEARS

Your application to conduct the above-mentioned research in schools in the Western Cape has been approved subject to the following conditions:

1. Principals, educators and learners are under no obligation to assist you in your investigation.
2. Principals, educators, learners and schools should not be identifiable in any way from the results of the investigation.
3. You make all the arrangements concerning your investigation.
4. Educators' programmes are not to be interrupted.
5. The Study is to be conducted from **25 January 2021 till 30 September 2021.**
6. No research can be conducted during the fourth term as schools are preparing and finalizing syllabi for examinations (October to December).
7. Should you wish to extend the period of your survey, please contact Dr A.T Wyngaard at the contact numbers above quoting the reference number?
8. A photocopy of this letter is submitted to the principal where the intended research is to be conducted.
9. Your research will be limited to the list of schools as forwarded to the Western Cape Education Department.
10. A brief summary of the content, findings and recommendations is provided to the Director: Research Services.
11. The Department receives a copy of the completed report/dissertation/thesis addressed to:

    The Director: Research Services
    Western Cape Education Department
    Private Bag X9114
    CAPE TOWN
    8000

We wish you success in your research.

Kind regards.
Signed: Dr Audrey T Wyngaard
Directorate: Research
DATE: 04 September 2020

Appendix 22: Letter to schools

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD
HEALTH SCIENCES

DHRS
Department of Health and Rehabilitation Sciences
FACULTY OF HEALTH SCIENCES

Divisions of Communication Sciences & Disorders • Disability Studies • Nursing & Midwifery • Occupational Therapy • Physiotherapy

F45 Old Main Building, Groote Schuur Hospital
Observatory, Cape Town, South Africa, 7925
Telephone: +27 (0) 21 406 6401
Website: www.dhrs.uct.ac.za

To establish whether children aged 5-7-years years can reliably report on their Health-Related Quality of Life when compared to older children.

Date: 12th September 2020

PERMISSION TO PERFORM RESEARCH AT YOUR SCHOOL

INFORMATION SHEET

Investigators: Ms Razia Amien; Dr Janine Verstraete and Mrs Des Scott

University of Cape Town

Department of Health and Rehabilitation Sciences

Division of Physiotherapy

Title of study: Performance of the EQ-5D-Y-3L Interviewer Administered Version in young children aged 5-7-years

We are physiotherapists from the University of Cape Town, and we are conducting a research project investigating whether younger children (between five-seven years old) can report on their Health Related Quality of Life (which is the impact that an individual's health has on their overall quality of life) and comparing this to older children (between eight-10 years). We would like your permission to conduct this research project at your facility. The full details of the study are outlined below, and if you require any additional information you may contact us at any time. Our contact details are found at the end of the form.

Why are we doing this study?

The research project aims to determine the reliability and validity of the Interviewer Administrated questionnaire (called the EQ-5D-Y-3L IA), when administered to the younger age-group compared to the older group, in children from the general population and children at special needs schools with a functional disability. In order to do this, we will be comparing the results from that Interviewer

Administered questionnaire to the self-complete questionnaire in the older children. We will also be comparing the results of the Interviewer Administered questionnaire to other questionnaires (called the Faces Pain Scale-Revised and the Mood & Feelings Questionnaire). These short questionnaires ask children to report on any problems they might have with walking around, washing and dressing themself, doing daily activities like going to school and playing games with friends, having pain and whether the child is worried or sad about anything.

How will we go about conducting this study?

Once we have received your permission to recruit children from your school for the study, we will send informed consent forms home with learners, explaining the research study and what will be required of their child, to the legal guardians of the children selected from your school, to sign if they consent to their child taking part in the study. We will recruit children aged five-10 years from your school. Children with intellectual disabilities or medically diagnosed as unable to hear with assistive technology will be excluded from the study due to data collection being done by means of an interview. Children with physical disabilities will be included and will be assisted with writing or assigned a scribe if needed. We will arrange a meeting with you and the teachers at a time which is suitable for you, to discuss when we may conduct the interviews with the children. Once a time has been arranged and parents have returned informed consent forms, we will collect small groups of children from the classrooms and take them to another quiet room in the school. The study will be explained to the children and they will be asked for assent before commencing with the interviews. Any child who does not wish to participate will be taken back to the classroom, with no negative results. The children would only be out of the classroom (at a suitable time that the class teacher and the researchers agree upon) for about 30 minutes to complete three/four questionnaires with the researcher. The children will also be asked general questions regarding whether they understood the questionnaires or not. The WeeFIM, an observational functional outcome measure, will be completed by a trained researcher to assess the child's functional ability.

In order to determine reliability, the Interviewer Questionnaire only, will need to be re-administered again 2 days later. This should only take 10 minutes to do. After the interviews, the child's and the school's involvement in the study is over.

Are there any risks in taking part in the study?

There are no risks in taking part in the study, to either your school or the children. The researchers will interview the children and get them to answer the questionnaires which will be read out to them.

They will be allowed to stop participating at any time and will not be forced to complete the questionnaires if they do not wish to.

Are there any benefits to participating in the study?

While there are no direct benefits to the children for participating in the study, the research will provide important information on whether younger children and children whose reading skills are limited, are able to self-report on their own Health Related Quality of Life, if an interviewer reads out the questions to them. This would mean that another person would not have to report on the child's behalf and the reporting would reflect the child's views more accurately. This information will be made available to the school in a report once the study is completed.

Each of the schools who participate in the study will be compensated for their time and receive a suitable hamper of educational materials to the value of R2000.

How will confidentiality and school details be handled?

The permission form signed by the parents will be returned in sealed envelopes and kept secure until the researchers collect them. When the children complete the questionnaires, their names will not be used, but they will be given a code instead. Only researchers will have access to the information on the questionnaires from the study. The completed questionnaires will be kept locked cupboard in a secure office. No names of the child or the school will be used when analysing the data or writing up the report.

Ethical approval has been granted by the Human Research Ethics Committee at UCT.
Permission has been granted by the Western Cape Education Department to conduct this study at various schools in the Western Cape.

Permission for the school to be involved in the study:

CONSENT

I have read and understand the provided information	Yes	No
I have had the opportunity to ask questions	Yes	No
I understand that the participation of the school and the children is voluntary	Yes	No

I voluntarily agree to allow this school to take part in this study	Yes	No

Signature of School Representative: _____

Date: _____

Witness: _____

Date: _____

How to contact us:

If you have any questions, or if you would like any further information with regards to our study, please do not hesitate to contact us at the address below. If you have any ethical concerns or questions about your and/or the children's rights as they participate in the study, please contact Professor Blockman at the Human Research Ethics Committee.

Dr Janine Verstraete	Mrs. Des Scott	Razia Amien	Prof Marc Blockman
Division of Physiotherapy Department of Health and Rehabilitation Sciences University of Cape Town Groote Schuur Hospital Anzio Road Observatory Cell: 082 840 9293 E-mail: Janine.verstraete@uct.ac.za	Division of Physiotherapy Department of Health and Rehabilitation Sciences University of Cape Town Groote Schuur Hospital Anzio Road Observatory, 7925 Cell: 083 949 8333 E-mail: des.scott@uct.ac.za	Division of Physiotherapy Department of Health and Rehabilitation Sciences University of Cape Town Groote Schuur Hospital Anzio Road Observatory, 7925 Cell: 076 106 2681 E-mail: Amnraz001@myuct.ac.za	Health Sciences Human Research Ethics Committee University of Cape Town Groote Schuur Hospital Anzio Road Observatory, 7925 Tel number: 021 406 6338

Appendix 23: Permission from healthcare facilities

DR T KERBELKER
Acting Manager: Medical Services
Red Cross War Memorial Children's Hospital
Email: Tamara.Kerbelker@westerncape.gov.za
Tel: +27 21 658 5383 Fax: +27 21 658 5006/5166

19 January 2021

Dr J Verstraete
Health and Rehabilitation Sciences
Division of Physiotherapy

Dear Dr Verstraete,

RESEARCH: RXH: RCC 254 / WC_202011_047

PROJECT TITLE: Performance of the EQ-5D-Y Interviewer Administered Version in young children aged 5-8 years

It is a pleasure to inform you that the hospital Research Review Committee has approved your application to conduct above-mentioned study in the wards and outpatient clinics at Red Cross War Memorial Children's Hospital.

Kindly note that this approval is subject to strict adherence to the HREC recommendations regarding research involving participants during COVID-19, dated 17 March 2020 (UCT HREC notice attached).

Yours sincerely,

DR T KERBELKER
ACTING MANAGER: MEDICAL SERVICES

Appendix 24: Informed consent (Parent/Legal Guardian)

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD
HEALTH SCIENCES

DHRS
Department of Health and Rehabilitation Sciences
FACULTY OF HEALTH SCIENCES

Divisions of Communication Sciences & Disorders • Disability Studies • Nursing & Midwifery • Occupational Therapy • Physiotherapy

F45 Old Main Building, Groote Schuur Hospital
Observatory, Cape Town, South Africa, 7925
Telephone: +27 (0) 21 406 6401
Website: www.dhrs.uct.ac.za

Date

INFORMED CONSENT FOR CHILDREN TO TAKE PART A RESEARCH STUDY

INFORMATION SHEET

Investigators: Dr Janine Verstraete, Mrs Des Scott and Ms Razia Amien

University of Cape Town

Department of Health and Rehabilitation Sciences

Division of Physiotherapy

Title of study: Performance of the EQ-5D-Y-3L Interviewer Administered Version in young children aged 5-7-years

Dear Parent/Legal Guardian

We are physiotherapists from the University of Cape Town, and we are conducting a research project looking at whether younger children (between five-seven years old) can report on their Health-Related Quality of Life (which is the impact that an individual's health has on their overall quality of life) and comparing this to older children (between eight-10 years).

We would like your permission to include your child in this research project. The full details of the study are outlined below, and if you require any additional information you may contact us at any time. Our contact details are found at the end of the form.

Why are we doing this study?

The research project aims to look at how young children report on their health when a researcher reads out the questionnaire to them (EQ-5D-Y-3L Interviewer Administered) and to compare this with how older children report on their health when they complete the questionnaire themselves (EQ-5D-Y-3L Self Complete). In order to do this, we will be asking your child between five–seven years to report

on their health with the researcher reading the questions and recording their answers. Children between eight-10 years will first complete the forms on their own and then have the researcher reading out the questions, to find out which version they preferred. We will also be asking your child to complete two other short questionnaires (called the Faces Pain Scale-Revised and the Mood & Feelings Questionnaire). The questionnaires ask children to report on any problems they might have with walking around, washing and dressing themself, doing daily activities like going to school and playing games with friends, having pain and whether the child is worried or sad about anything.

How will we go about conducting this study?

It is your choice to allow your child to take part in this study and you do not have to agree if you don't want to. If you do not agree, nothing bad will happen to you or your child. We will not ask them to complete any questionnaires.

We will be thankful if you do agree to your child taking part in the study and will ask you to sign consent at the end of this letter. We will also be asking you to agree to fill in a form with some of your child's details (name; date of birth; sex; grade; and whether they have any chronic illness or disability – if so, we will be asking for some further details about the illness or disability). Again, it is your choice to fill this in or not.

Your child will also be asked if they want to participate in the study, before they are asked to complete the questionnaires. If all is agreed upon, your child will be asked to fill in the three or four questionnaires.

If you child is attending an outpatient clinic, the interview will take place in a quiet consultation room. Your child will not lose their place in the queue. If the interview is not completed when they are called in for the doctor's consultation the researcher will stop, and you will be given the option to complete the study after the consultation or withdraw from the study.

If your child is at school, a suitable time for completing the questionnaires, during school, will be arranged with the schoolteacher.

If your child is at a school, we will be asking them to complete the Interviewer Administered form only, 2 days later.

Are there any risks to my child taking part in the study?

There are no risks in taking part in the study. If you are worried about anything to do with the research, please let us know and we will answer any questions you may have.

The medical treatment that your child is getting if they are in hospital or attending an outpatient clinic will not be changed. If your child is at a school, the time away from the teacher will be limited to 30 minutes and the way your child is viewed at school will not be changed by taking part in the study.

Are there any benefits to participating in the study?

While there are no direct benefits to the children for participating in the study, the research will provide important information on whether younger children and children whose reading skills are limited, are able to self-report on their own Health Related Quality of Life, if an interviewer reads out the questions to them. This would mean that another person would not have to report on the child's behalf and the reporting would reflect the child's views more accurately.

How will confidentiality and hospital details be handled?

All the information will be kept confidential. The permission form you sign, the information about your child and your child's completed questionnaires will be kept in locked cupboard in a secure office. When the child completes the questionnaires, their names will not be used, but they will be given a code instead. Only the researchers will have access to the information on the questionnaires from the study. No names of the child or the hospital or school will be used when analysing the data or writing up the report.

Permission for my child to be involved in the study:

CONSENT

I have read and understand the provided information	Yes	No
I have had the opportunity to ask questions	Yes	No
I understand that the participation of my child is voluntary	Yes	No
I voluntarily agree to allow my child to take part in this study	Yes	No

Signature of parent / legal guardian: _____

Date: _____

Witness: _____

Date: _____

How to contact us: If you have any questions, or if you would like any further information with regards to our study, please do not hesitate to contact us at the address below. If you have any ethical concerns or questions about your and/or the children's rights as they participate in the study, please contact Professor Blockman at the Human Research Ethics Committee.

Dr Janine Verstraete	Mrs. Des Scott	Ms Razia Amien (student)	Prof Marc Blockman
Division of Physiotherapy Department of Health and Rehabilitation Sciences University of Cape Town Groote Schuur Hospital Anzio Road Observatory Cell: 082 840 9293 E-mail: Janine.verstraete@uct.ac.za	Division of Physiotherapy Department of Health and Rehabilitation Sciences University of Cape Town Groote Schuur Hospital Anzio Road Observatory, 7925 Cell: 083 949 8333 Email: des.scott@uct.ac.za	Division of Physiotherapy Department of Health and Rehabilitation Sciences University of Cape Town Groote Schuur Hospital Anzio Road Observatory, 7925 Cell: 0761062681 E-mail: amnraz001@myuct.ac.za	Health Sciences Human Research Ethics Committee University of Cape Town Groote Schuur Hospital Anzio Road Observatory, 7925 Tel number: 021 406 6338

Appendix 25: Assent form for child

UNIVERSITY OF CAPE TOWN

Faculty of Health Sciences

Department of Health and Rehabilitation Sciences

Division of Physiotherapy

F45 Old Main Building, Groote Schuur Hospital

Observatory, Cape Town, W Cape, 7925

Tel: +27 (0) 21 406 6401/ 6428/ 6628/ 6534

INFORMED ASSENT FOR CHILDREN TAKING PART IN THE STUDY

INFORMATION SHEET

Investigators: Dr Janine Verstraete, Mrs Des Scott and Ms Razia Amien

University of Cape Town

Department of Health and Rehabilitation Sciences

Division of Physiotherapy

Project name: Performance of the EQ-5D-Y-3L Interviewer Administered Version in young children aged 5-7-years

About the Project

We are physiotherapists at the University of Cape Town. We are doing a research project where we would like to find out about your health. We are looking at the answers from three to four different short forms which asks about your health. This will help us to know how children feel about their health. There are no right or wrong answers, it is just about how you feel.

Your parent has said that it is okay for you to be part of our project if you would like to be. It is your choice if you want to take part or not and nothing bad will happen to you if you don't want to be part of the project.

If you do want to take part, we will ask you to answer three or four short forms about your health. They are called the EQ-5D-Y-3L (interviewer administered and self-complete); the Faces Pain Scale-Revised and the Mood & Feelings Questionnaire. If you cannot read, we will read the questions to you and write your answers on the forms. If you can read, we will ask you to answer some of the questions yourself and then we will read some questions to you. Afterwards we will ask you a few questions about how you filled in the forms.

The questions will ask you: how you walk around, wash and dress yourself, play with friends, play sport, how you are coping at school, whether you have any pain or not and whether you are feeling worried, sad our unhappy. It should take 30 minutes to answer all of the questions. If you don't understand something, we will be there to help you.

If you take part, we might ask you to answer one of the question papers (EQ-5D-Y-3L-IA) again two days later.

Risks and Benefits

There are no dangers to taking part in this project. You will only be answering questions about how you feel and there are no right or wrong answers. Please let us know if at any time you do not want to answer the questions. You will be able to stop at any time, even if you have already started. Nothing bad will happened to you if you don't want to finish. Nothing bad will happen to you if you don't want to be part of our project. It will not change the way that you are treated at school or at the hospital.

You will not get any money for taking part in this project. We hope that this project will help us to know how young children answer questions about their health.

Your answers

Only the people doing the study will see the papers where you fill in your answers. No one else will see your answers. We will keep your answers in a locked cupboard in a safe office. When we put your answers on the computer, your name will not be there so that no one else will know you took part in the project. The computers will also have a password, which only the people doing the study will know. When we write a report of our project, we will not be using any names so other people won't know that you helped us by answering questions.

It is your choice to take part in the project

You can choose if you want to take part in our project or not. Nothing bad will happen if you don't want to take part in the project. If you sign this paper, it tells us that you would like to be in this project. You can stop being in the project at any time, even if you have signed this paper already.

If you have any questions about the study, you may contact the researchers:

RESEARCHERS:

Janine Verstraete at:	Des Scott at:	Razia Amien at:
Cell: 082 840 9293	Cell: 083 949 8333	Cell: 076 106 2681
E-mail: Janine.verstraete@uct.ac.za	E-mail: Des.scott@uct.ac.za	E-mail: amnraz001@myuct.ac.za

The UCT FHS Human Research Ethics Committee can be contacted on 021 406 6338 in case participants have any questions regarding your rights and welfare as research subjects on the study.

I AGREE THAT

Please circle yes or no

I have read, and I understand the information	Yes	No
I have asked all of my questions	Yes	No
I know that I can choose to take part or not take part	Yes	No
I know that I can stop at any time and I do not need to say why	Yes	No
I would like to take part in this project	Yes	No

Your signature _____ Date _____

Researcher's signature _____Date _____

Witness signature _____Date _____

Appendix 26: Telephonic consent form (Parent/Legal guardian)

UNIVERSITY OF CAPE TOWN

Faculty of Health Sciences

Department of Health and Rehabilitation Sciences

Division of Physiotherapy

F45 Old Main Building, Groote Schuur Hospital

Observatory, Cape Town, W Cape, 7925

Tel: +27 (0) 21 406 6401/ 6428/ 6628/ 6534

Title of study: Performance of the EQ-5D-Y-3L Interviewer Administered Version in young children aged 5-7-years

Good Day Mr/Mrs _____

My name is _____ and I am a physiotherapist from the University of Cape Town. I am part of a group doing a research project looking at how children aged 5-10 years feel about their health. We would like to invite (your child's name _____) to participate in this study. If you agree for your child to take part, we will also ask your child if they would like to participate in the study, before asking them to answer three or four short forms about their Health-Related Quality of Life (which is the impact that an individual's health has on their overall quality of life). We will not be doing anything to your child, we will only be asking them to fill in their answers to the questions.

(Child's name _____) will only be invited to take part in the study if you have given us permission. We will explain the study to (child's name _____) including that it is their choice to take part and they can refuse to take part or stop the study at any point without anything bad happening to them. Their decision to take part in the study will not change the way that they are treated at school or at the hospital and no one will know what their answers are to the questions. If they agree to participate, they will be asked to provide informed assent.

If your child cannot read, we will read the questions out to them and write their answers down for them. The questionnaires that we will ask (child's name _____) to fill in are called: EQ-5D-Y-3L Interviewer Administrated and Self-Complete, the Faces Pain Scale-Revised and the Mood & Feelings Questionnaire. The questions that they will be asked will include how they walk, wash and dress themselves, about their participation at school and sport, their friends, their pain and whether they are worried, sad or unhappy and their general health. This should not take longer than 30

minutes. If (Child's name _____) is attending a school, they will be asked to complete one questionnaire (EQ-5D-Y-3L Interviewer Administrated) again two days later.

It is your choice for (child's name _____) to take part in this study and you do not have to agree if you don't want to. If you do not agree, nothing bad will happen to you or your child. If you agree now but decide later that you don't want to do it anymore, you can let us know and all of the information that you have given us will be taken out of the study. We will be thankful if you agree to your child taking part in the study by filling in the questionnaires. The medical treatment that your child is getting if they are in hospital, or the way they are viewed at school will not be changed and will carry on as normal.

There are no risks in taking part in this study.
While there are no direct benefits to you or the children for participating in the study, the research will provide important information on whether younger children and children whose reading skills are limited, are able to self-report on their own Health Related Quality of Life, if an interviewer reads out the questions to them. This would mean that another person would not have to report on the child's behalf and the reporting would reflect the child's views more accurately.

All the information that your child gives us will be confidential. Only the people doing the study will see (Child's name _____) answers to the questions. We will keep the answers in a locked cupboard in a safe and secure office. The computers used to store the information will have a password, which only the people doing the study will know. Your child's name will not be used when we analyse the information or write a report about the study findings.

If you have any questions or concerns about the study, you may contact the researchers:

RESEARCHERS:

Janine Verstraete at:	Des Scott at:	Razia Amien at:
Cell: 082 840 9293	Cell: 083 949 8333	Cell: 076 106 2681
E-mail: Janine.verstraete@uct.ac.za	E-mail: Des.scott@uct.ac.za	E-mail: amnraz001@myuct.ac.za

The UCT FHS Human Research Ethics Committee can be contacted on 021 406 6338 in case participants have any questions regarding their rights and welfare as research subjects on the study.

Consent Form

Declaration	Yes	No
Do you understand the information that has been provided?		
Do you understand that your consent is required for your child to take part?		
Do you understand that it is voluntary to participate and that you can refuse to consent without any consequences to yourself or your child?		
Do you understand that refusing to give consent will not affect the current or future health care or school treatment of your child?		
Do you understand that neither you nor your child will be identified should this research study be published?		
Do you consent to your child taking part in this research of your own free will?		

Caregivers name _____

Child's Name _____

Researcher signature _____

Date _____

Appendix 27: Data Management Plan

"Performance of the EQ-5D-Y Interviewer Administered version in young children aged 5-8 years"

A Data Management Plan created using DMPRoadmap

Creator: Razia Amien

Affiliation: University of Cape Town (UCT-Generic)

Template: University of Cape Town (UCT-Generic)

Project abstract:

The EQ-5D-Y-3L IA has the potential to lower the age of completion of the instrument allowing younger children the opportunity to voice their feelings and make their needs known. It will allow health professionals to make collaborative decisions regarding treatment plans with insight from caregivers and children. As a result, treatment plans will be more patient-centered and more effective (8,9). This will further negate the need to use proxy instruments in this age group and avoid transition between proxy and self-reported health in longitudinal studies or between age groups. Studies suggested that proxy responses may not accurately capture the health states of children therefore emphasizing the importance of this study to ensure self-report in children (8,11,37,38).

Last modified: 01-04-2021

"Performance of the EQ-5D-Y Interviewer Administered version in young children aged 5-8 years"

Data summary

Briefly introduce the types of data the research will create. Why did you decide to use these data types?

Both quantitative and qualitative data will be collected using various questionnaires which will be self-completed and/or interviewer administered depending on the age of the participant.

Data collection

Give details on the proposed methodologies that will be used to create the data. Advise how the project team selected will be suitable for the data/digital aspects of the work, including details of how the institution's data support teams may need to support the project

Data collection will be done by means of face-to-face interviews using various questionnaires. Interviews will take place at various mainstream and special needs schools and healthcare facilities in the Western Cape.

Short-term storage

How will the data be stored in the short term?

The information obtained will be entered into a password protected Excel spread sheet under the code allocated to each participant. Spreadsheets will be password protected and saved to the student and supervisor's password protected computers and/or encrypted and secure UCT cloud storage. The hard copies of the questionnaires will be stored a locked cupboard in the secure office of one of the supervisors, at UCT. All forms will be stored in a locked cupboard at the respective schools until the researchers collect them. Only the researcher and supervisors will have access to the raw data. No identifying information of participants will be recorded for data analysis or dissemination of results.

Short-term storage

How the data will be stored in the long term?

The hard copies of the questionnaires will be stored a locked cupboard in the secure office of one of the supervisors, at UCT and will be destroyed 10 years after publication. Only the researcher and supervisors will have access to the raw data. No identifying information of participants will be recorded for data analysis or dissemination of results.

Data sharing

How the data will be shared and the value it will have to others

All institutions involved in data collection will have access to the final thesis.

How the data will enhance the area and how it could be used in the future?

The research may prove useful in lowering the age for self-report of HRQoL, to younger than eight years, negating the need for proxy reporting. The IA version may also be useful in older children with lower literacy levels or who are unable to self-complete a questionnaire.

Releasing the data – advise when you will be releasing and justify if not releasing in line with AHRC guidelines of a minimum of three years. If the data will have value to different audiences, how these groups will be informed?

Estimated completion is February 2022.

Will the data need to be updated? Include future plans for updating if this is the case.

N/A

Will the data be open or will you charge for it? Justify if charging to access the data

The data will be openly accessible via UCT.

Financial requirements of sharing – include full justification in the JoR

N/A

Ethical and legal considerations

Any legal and ethical considerations of collecting the data

Ethical considerations

- Autonomy: informed consent and assent will be obtained before data collection begins
- Beneficence/non-maleficence: There are no known risks associated with the research. Participants will not be reimbursed as they will not incur costs however, a donation to the various institutions will be given by means of an educational hamper.
- Justice: No child will be excluded based on ethnic group, gender preference, religion or any other reason.

Legal and ethical considerations around releasing and storing the data – anonymity of any participants, following promises made to participants

The confidentiality of each participant will be maintained by allocating a code to each questionnaire for data capturing and analysis. Names need to be collected on raw data to identify individuals for re-test. The questionnaires will be kept in a locked cupboard in a secure office. The electronic files will be password protected on a secure computer. No participants, healthcare facilities or schools will be identified in the analysis or write-up of the research.

The data collection will not affect the schooling or medical treatment or the way in which the child is perceived at the school or health facility. The participants will not be reimbursed as they will not incur any costs to participate in the research. The participating schools will receive an educational hamper of materials to the value of R2000, regardless of the number of children who have participated. There are no known risks and therefore no insurance will be required for research-related injuries.

Should any child become emotionally distressed during the research, the interview will be terminated, and the child will be returned to a familiar adult. In the absence of a caregiver in an in-patient setting, a healthcare professional will be called for further assistance. If any signs of neglect or abuse are noted, referral to the necessary authority will be made in line with legal requirements.

Appendix 28: Cognitive debriefing template

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD
HEALTH SCIENCES

DHRS
Department of Health and Rehabilitation Sciences
FACULTY OF HEALTH SCIENCES

Divisions of Communication Sciences & Disorders • Disability Studies • Nursing & Midwifery • Occupational Therapy • Physiotherapy

F45 Old Main Building, Groote Schuur Hospital
Observatory, Cape Town, South Africa, 7925
Telephone: +27 (0) 21 406 6401
Website: www.dhrs.uct.ac.za

Interviewee questionnaire to be completed in children aged 5-7-years and 8-10-years after they have completed EQ-5D-Y-3L-IA

The interviewer (student researcher) should view the child's responses on the EQ-5D-Y-3L IA version.

1. For each of the dimension scores ask the child why they scored the corresponding number of problems and record their reason:
   ASK: Why did you say that you had xxx problems with (dimension)?

DIMENSION	LEVEL OF REPORTING	Why did you say that you had xxx problems with (dimension)?
Mobility	(1) No problems	
	(2) Some problems	
	(3) A lot of problems	
Looking After Myself	(1) No problems	
	(2) Some problems	

	(3) A lot of problems	
Usual Activities	(1) No problems	
	(2) Some problems	
	(3) A lot of problems	
Pain/Discomfort	(1) No problems	
	(2) Some problems	
	(3) A lot of problems	
Worried, Sad or Unhappy	(1) No problems	
	(2) Some problems	
	(3) A lot of problems	

2. If there are any apparent inconsistencies in reporting, this should also be probed.
   e.g. If child reports Level 3 (a lot of problems) on Mobility, but level 1 (no problems) on Usual Activities.
   OR
   Level 1 (no problems) on Mobility but is apparently confined to bed.

Description of Inconsistency noted by interviewer:

Probe: Why did you say that you had xx problems on (dimension) but xx problems on (dimensions)?

Child's answer:

3. Did you understand all of the questions that I asked you?

YES ☐ NO ☐

4. If No, ask: "Which ones did you not understand?"

For each one that was poorly understood, ask: "Why was it difficult?"

Question that was difficult	Reason

Appendix 29: Interviewer questionnaire

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD
HEALTH SCIENCES

DHRS
Department of Health and Rehabilitation Sciences
FACULTY OF HEALTH SCIENCES

Divisions of Communication Sciences & Disorders • Disability Studies • Nursing & Midwifery • Occupational Therapy • Physiotherapy

F45 Old Main Building, Groote Schuur Hospital
Observatory, Cape Town, South Africa, 7925
Telephone: +27 (0) 21 406 6401
Website: www.dhrs.uct.ac.za

Interviewer Questionnaire

Please comment for each child:

1. Do you feel the child understood the questions posed to them on the EQ-5D-Y-3L-IA?

YES ☐ NO ☐

If no, why do you feel that the child did not understand the questions?

……

……

……

2. Did you notice any uncertainty or indecisiveness when the child was reporting?

YES ☐ NO ☐

If yes, which questions was there uncertainty or indecisiveness with?

……

……

……

Appendix 30: Letter to healthcare facilities and schools outlining research findings

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD
HEALTH SCIENCES

DHRS
Department of Health and Rehabilitation Sciences
FACULTY OF HEALTH SCIENCES

Divisions of Communication Sciences & Disorders • Disability Studies • Nursing & Midwifery • Occupational Therapy • Physiotherapy

F45 Old Main Building, Groote Schuur Hospital
Observatory, Cape Town, South Africa, 7925
Telephone: +27 (0) 21 406 6401
Website: www.dhrs.uct.ac.za

"Performance of the EQ-5D-Y Interviewer Administered version in young children"

To Whom It May Concern:

Thank you for allowing your facility to take part in our research project. Your assistance with recruitment, communication with parents/legal guardians and allowing us to use your space for interviews was greatly appreciated!

We have compiled an information sheet outlining the results from our project and how it may help your facility going forward.

Our project focused on **HEALTH-RELATED QUALITY OF LIFE** in young children. The term health-related quality of life refers to the **subjective** measure of one's **physical** and **psychosocial** wellbeing. A group called EuroQoL developed a questionnaire, the **EQ-5D-Y-3L**, which assesses health-related quality of life in young children by looking at five dimensions: Mobility (walking about), Self-care (washing and dressing), doing Usual Activities (playing, hobbies, sport, school and doing things with family/friends), having Pain/Discomfort and feeling Worried, Sad or Unhappy. This questionnaire is recommended for self-completion in children from the age of 8-years. However, if we consider South Africa, our literacy levels are not on par with our international counterparts therefore, self-complete may not be as easy for South African 8-year-olds. This may not only be important to children from the general population but also children with health conditions who require lengthy hospital stays or endless appointments which may affect their school attendance and ability to progress their literacy skills.

Our project aimed to assess a newly developed questionnaire, the **EQ-5D-Y-3L Interviewer Administered Version**, assessing health-related quality of life in young children aged 5-10-years who understand the concept of health but may struggle to read and self-complete a questionnaire due to their age or literacy skills which may or may not be related to their socioeconomic status.

Being that this questionnaire is **INTERVIEWER-BASED**, it allows children to self-report on their health without having to self-complete or rely on someone else reporting on their health for them (known as proxy-report). Previous research has found that proxy-report, in some cases, may not be as reliable as self-report especially regarding psychosocial dimensions.

If the EQ-5D-Y-3L Interviewer Administered Version proved to be **successful** in this age-group, it has the potential to give children the opportunity to be heard and feel empowered when it comes to their health and potential treatment.

We compared the EQ-5D-Y-3L Interviewer Administered Version to the Self-complete version in children aged 8-10-years and asked them for their preference and the reason for their preference. We found that 60% of children preferred the Interviewer Administered Version with a large amount of 8- and 9-year-olds associating their preference with their developing literacy skills.

We then compared the dimensions on the EQ-5D-Y-3L Interviewer Administered Version to other outcome measures which assess similar constructs therefore determining if the dimensions accurately recording what it was meant to. We found no concerning differences between outcome measures therefore showing that the EQ-5D-Y-3L Interviewer Administered Version was able to record dimensions accurately.

To conclude, we found the EQ-5D-Y-3L Interviewer Administered Version to work just as well as the self-complete version in children aged 8-10-years therefore can be used interchangeably.

We also found that the Interviewer Administered Version can be reliably used in younger children from the age of 5-years, however, one needs to pay attention to what the child is able to do or expected to do according to their developmental age to ensure activities are appropriate. For example, it may not be appropriate to associate independent dressing with tying shoe laces as they might have not learnt how to complete such an advanced dressing task yet.

Our **recommendation** for older children would be to individually assess each child's literacy skills to determine which version is most appropriate. We would also recommend the routine use of the Interviewer Administered Version in younger children to ensure the retrieval of information is done

as subjectively as possible. These types of questionnaires also provides a means of monitoring children's progress.

We hope this information is helpful to you.

If you have any questions, or if you would like any further information with regards to our study, please do not hesitate to contact us at the addresses below.

Dr Janine Verstraete (supervisor)	Mrs Des Scott (supervisor)	Ms Razia Amien (student)
Division of Physiotherapy Department of Health and Rehabilitation Sciences University of Cape Town Groote Schuur Hospital Anzio Road Observatory Cell: 082 840 9293 E-mail: Janine.verstraete@uct.ac.za	Division of Physiotherapy Department of Health and Rehabilitation Sciences University of Cape Town Groote Schuur Hospital Anzio Road Observatory, 7925 Cell: 083 949 8333 E-mail: des.scott@uct.ac.za	Division of Physiotherapy Department of Health and Rehabilitation Sciences University of Cape Town Groote Schuur Hospital Anzio Road Observatory, 7925 Cell: 076 106 2681 E-mail: Amnraz001@myuct.ac.za

Appendix 31: Letter to participants outlining research findings

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD
HEALTH SCIENCES

DHRS
Department of Health and Rehabilitation Sciences
FACULTY OF HEALTH SCIENCES

Divisions of Communication Sciences & Disorders • Disability Studies • Nursing & Midwifery • Occupational Therapy • Physiotherapy

F45 Old Main Building, Groote Schuur Hospital
Observatory, Cape Town, South Africa, 7925
Telephone: +27 (0) 21 406 6401
Website: www.dhrs.uct.ac.za

"Performance of the EQ-5D-Y Interviewer Administered version in young children"

Dear Participants

Thank you so much for taking part in our research project. Your answers were so interesting and very helpful to us.

We have made a little summary of what your answers meant to us.

Firstly, our project was all about health in young children and we wanted to find out about your physical, mental and social health. We wanted to be able to get this information directly from you instead of asking your parents/legal guardians/caregivers because they don't always know exactly how you are feeling.

We know that as young children, you may understand what health is, and can tell us all about your health but if we were to give you a questionnaire to read and complete, you might not be able to because you are still too young, learning how to read or might struggle a bit with reading.

That's why we were testing a new questionnaire called the EQ-5D-Y-3L Interviewer Administered version! This questionnaire does not require you to read anything but only listen and answer therefore allowing you to report on your health yourself.

We tested this questionnaire in older children (8-10-years) along with a version that they had to complete by themselves. Most of the children, especially the 8- and 9-year-olds liked the interviewer version because they were still learning to read and therefore struggled to read and complete the questionnaire on their own. We also tested this new questionnaire in younger children (5-7-year-olds) and found that they were able to also report on their health if the questions were read out to them.

This is all great news because, it means that this new questionnaire was able to accurately record your health without you having to read or complete it by yourself!

We hope this helped you understand the reason behind our project.

If you have any questions about our project, you are welcome to ask your parent/caregiver to contact us and we will try and answer your questions as best as we can.

Dr Janine Verstraete (supervisor)	Mrs Des Scott (supervisor)	Ms Razia Amien (student)
Division of Physiotherapy Department of Health and Rehabilitation Sciences University of Cape Town Groote Schuur Hospital Anzio Road Observatory Cell: 082 840 9293 e-mail: Janine.verstraete@uct.ac.za	Division of Physiotherapy Department of Health and Rehabilitation Sciences University of Cape Town Groote Schuur Hospital Anzio Road Observatory, 7925 Cell: 083 949 8333 Email: des.scott@uct.ac.za	Division of Physiotherapy Department of Health and Rehabilitation Sciences University of Cape Town Groote Schuur Hospital Anzio Road Observatory, 7925 Cell: 076 106 2681 Email: Amnraz001@myuct.ac.za