An Investigation of the Extended Application of the Oxford Knee Score in Research and Clinical Practice



Kristina Knezevic Harris

Wolfson College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2013

Writing this thesis would not have been possible without the help and support of the wonderful people around me. I am truly and deeply indebted to so many people that there is no way to acknowledge them all, or even any of them properly.

First and foremost, I thank my academic supervisors, Professor David Beard and Professor Andrew Price for initially accepting me into their research group and giving me the opportunity to undertake this wonderful journey. Their academic and personal guidance, support, and good nature made this work successful. I am also indebted to my supervisor, Dr Jill Dawson, who has kindly introduced me into the world of Patient Reported Outcome Measures, carefully guided my academic development, and was always ready to give invaluable advice throughout these three years. The personal mentorship and support of my supervisors, their leadership, hard work, and scholarship have set an example I hope to match someday.

I would like to thank Professor Crispin Jenkinson, Professor Chris Wilson, and Professor Andrew Carr for serving as my committee members and for helping to make my thesis defence an enjoyable experience.

I would also like to thank Luke Jones, for his valuable teamwork and help in collecting the data for this thesis. To Sandra Regan, I am thankful for making me feel welcomed on my arrival to Oxford and for being source of kind support during the course of my DPhil. Furthermore, I would like to acknowledge all clinicians and administrative staff at the Nuffield Orthopaedic Centre, and all of our benevolent patients, who have volunteered to participate in the studies.

My day-to-day life at the department was made enjoyable and fun due in a large part to the many friends and colleagues. I very much enjoyed the opportunity to serve as a student representative, which has allowed me to meet and interact with all of the fellow students at the department. Particularly, I would like to thank Annie Ng, Nasim Zargar, Monica Armengol, Paul Monk, Raj Rout, and George Grammatopoulous for being such great colleagues and friends.

A lot of wonderful and talented friends at the Wolfson College have been helping at different occasions and, in one way or another, have influenced this thesis. Particularly, I would like to thank Daniel Corbett, Joe Martin, and Henry Lambert, for their help in proofreading this thesis. Furthermore, I am exceptionally grateful to Antica Culina, for being a continuous source of fun, inspiration, and support.

Lastly, I am deeply indebted to my family for their selfless support throughout my life.

Abstract

The Oxford Knee Score (OKS) is a popular single summary questionnaire developed to measure the effect of knee replacement surgery from the patients' perspective. There has been a recent interest in the use of the OKS in populations of patients and in roles it has not been originally developed for. To date, no evidence has been provided about the measurement properties of the OKS when it is used outside the context or purpose for which it was originally designed.

The general aim of this thesis is to investigate the measurement properties of the OKS when used in extended roles: a) within the population for which the OKS was originally intended and developed for (knee replacement) and, b) when applied on a different population, viz. patients undergoing non-operative treatment for knee osteoarthritis (OA).

Four existing large-scale databases of patients undergoing knee replacement surgery and a database obtained from a prospective study on patients undergoing non-surgical management for knee OA were analyzed.

The results demonstrate that: 1) it is possible to extract separate information on pain and functional disability from the OKS in a meaningful way (in the form of subscales). 2) For the first time, anchor-based Minimal Important Change (MIC) of 9 points and Minimal Important Difference (MID) of 5 points were established for joint replacement surgery. 3) The OKS demonstrated satisfactory evidence reliability, validity, responsiveness, and interpretability, when used in patients who are undergoing non-operative management for their knee OA. 4) Further evidence of validity was demonstrated by fitting the OKS to the Rasch model. 5) Lastly, it was demonstrated that thresholds can be applied on the OKS to distinguish between patients who consider their knee problem to be severe enough to warrant joint replacement surgery versus patients who do not. This supports the potential use of the OKS in decision making aids for secondary care referral.

Overall the thesis provides critical evidence, not previously existing, to support the continued use, and extended use, of the OKS in orthopaedic medicine.

Contents

1 Introduction **1**
1.1 Overview . . . 1
1.2 Overall Scope and Specific Objectives of the Proposed Thesis . . . 4
1.3 Structure of the Thesis . . . 5
1.4 Software licenses and author's role in the conduct of the studies and data preparation . . . 6
1.5 Brief history of health measurement . . . 8
1.5.1 HRQoL measurement . . . 8
1.5.1.1 Measurement of outcomes by health care providers . 10
1.5.2 Psychometrics . . . 11
1.5.2.1 Stipulations for a good health measure . . . 13

2 Literature review **17**
2.1 Introduction . . . 17
2.2 Basic concepts . . . 17
2.2.1 A conceptual model of HRQoL . . . 17
2.2.2 Using patient-reported outcomes to measure HRQoL . . . 19
2.2.2.1 Scales and indexes . . . 20
2.2.2.2 Unidimensional and multidimensional measures . . . 21
2.2.2.3 Standardised and individualised measures . . . 22
2.3 Features that make a good PROM . . . 24
2.3.1 Methods based on CTT . . . 24
2.3.1.1 Validity . . . 25
Content validity . . . 26
Construct validity . . . 27
Criterion validity . . . 28
Cross-cultural validity . . . 29
2.3.1.2 Reliability . . . 29
Internal consistency . . . 30
Test-retest reliability . . . 30
Score precision . . . 32
2.3.1.3 Responsiveness (sensitivity to change) . . . 33
2.3.1.4 Interpretability . . . 35
2.4 Different types of outcome measures used in orthopaedics . . . 36
2.4.1 Generic outcome measures . . . 37

2.4.2 Disease/site specific outcome measures . . . 38
2.4.2.1 Outcome measures for arthroplasty and knee OA . . 39
Knee arthroplasty . . . 39
Knee osteoarthritis . . . 42
2.4.3 Summary . . . 45

3 An Exploratory and Confirmatory Factor Analysis of the OKS: Generation of Subscales Assessing Self-Reported Function and Pain 48
3.1 Introduction . . . 48
3.1.1 Factor analysis . . . 50
3.1.1.1 Introduction . . . 50
3.1.1.2 Exploratory Factor Analysis (EFA) . . . 51
Factor Rotation . . . 55
Interpretation of factor loadings . . . 56
3.1.1.3 Confirmatory Factor Analysis (CFA) . . . 56
3.2 Methods . . . 60
3.2.1 Analyses . . . 64
3.2.1.1 Exploratory Factor Analysis . . . 64
3.2.1.2 Confirmatory Factor Analysis . . . 64
Model 1 . . . 65
Model 2 . . . 65
3.3 Results . . . 65
3.3.1 Exploratory Factor Analysis . . . 65
3.3.2 Confirmatory Factor Analysis . . . 67
3.3.3 Scoring of pain and functional subscales . . . 67
3.4 Discussion . . . 69
3.5 Summary . . . 74

4 Minimal Important Change or Difference for the OKS Following Knee Replacement Surgery 75
4.1 Introduction . . . 75
4.1.1 The importance of context when choosing between estimates of minimal important change/difference . . . 76
4.1.2 Different approaches to estimating meaningful or minimal important change or difference . . . 77
Distribution based methods . . . 77
Anchor based methods . . . 78
4.2 Materials and Methods . . . 79
4.2.1 Subjects/Assessments . . . 79
4.2.2 Statistical Methods . . . 80
4.2.2.1 Anchor based methods for estimating minimal change /difference . . . 80
Minimal Important Change (MIC) . . . 81
Minimal Important Difference (MID) . . . 81
Best cut point . . . 81

4.2.2.2 Distribution based methods for estimating minimal change or difference . . . 82
Effect Sizes (ES) and the 0.5 Standard Deviation (SD) 82
Precision of the change score based on SEM . . . 82
4.3 Results . . . 82
4.3.1 Study population characteristics . . . 82
4.3.2 Assessment of minimal change . . . 83
4.3.2.1 Summary of the range of anchor and distribution based MIC/MID values . . . 85
Distribution based methods . . . 85
Anchor based methods . . . 85
4.4 Discussion . . . 88
4.4.1 Study considerations and limitations . . . 89
4.4.2 Specific considerations related to the choice of anchor or distribution based MIC/MID . . . 90
4.4.3 Study limitations and considerations . . . 93
4.4.4 Implications of the findings and future directions . . . 94
4.5 Summary . . . 95

5 Validation of the OKS for Patients Undergoing Non-Operative Treatment for Knee OA 97
5.1 Introduction . . . 97
5.2 Patients and Methods . . . 98
5.2.1 Study procedures and assessments . . . 99
5.2.2 Statistical methods . . . 101
5.2.2.1 Reliability . . . 102
5.2.2.2 Validity . . . 102
Structural validity . . . 102
Construct validity . . . 103
5.2.2.3 Responsiveness . . . 104
5.2.3 Interpretability . . . 105
5.3 Results . . . 106
5.3.1 Sample characteristics . . . 106
5.3.2 Measurement properties . . . 107
5.3.2.1 Reliability . . . 107
5.3.2.2 Validity . . . 108
Structural validity . . . 108
Construct validity . . . 109
5.3.2.3 Responsiveness . . . 110
5.3.2.4 Interpretability . . . 114
5.4 Discussion . . . 114
5.4.1 Comparative performance of the OKS and its subscales between different populations/contexts . . . 116
5.4.2 Comparative performance of the OKS and its subscales versus the ICOAP and the KOOS-PS in this study . . . 118

5.4.3 Limitations and conclusion 120
5.5 Summary 122

6 Rasch analysis of the OKS **123**
6.1 Introduction 123
6.1.1 One parameter (Rasch) model 124
6.1.2 Polytomous Rasch model 126
6.1.3 Assumptions of IRT and RMT 127
6.2 Methods 128
Overall fit of the data to the Rasch model 129
Targeting of the measure 130
Ordering of thresholds 130
Local dependency 131
Differential Item Functioning (DIF) 131
Internal consistency reliability. 132
Unidimensionality. 132
6.3 Results 133
6.3.1 Sample Characteristics 133
6.3.2 Rasch analysis results 133
6.3.2.1 The OKS 134
6.3.2.2 The OKS-PCS 136
6.3.2.3 The OKS-FCS 138
6.4 Discussion 139
6.5 Summary 141

7 The Use of Thresholds in Assessing the Need for Knee Replacement Surgery **143**
7.1 Introduction 143
7.1.1 The (mis)use of the OKS in allocating access to TKR 144
7.1.2 A focus on patient involvement in treatment-decision making 145
7.1.3 Tackling variation in TKR provision at the GP level 146
7.1.4 Measuring the severity of knee OA symptoms 147
7.2 Methods 149
7.2.1 Patients/assessments 149
7.2.2 Statistical methods 150
7.3 Results 152
7.3.1 The relationship between patients' perception about the severity of their symptom and absolute scores 153
7.3.2 The ability of measures to distinguish between the two groups of patients, and production of threshold values on absolute scores 155
7.4 Discussion 158
7.5 Summary 163

8 Conclusions, Clinical Relevance, and Future Work **165**
8.1 Review of Research Objectives 165
8.2 Conclusions 167
8.3 Clinical Relevance 170
8.4 Recommendations for Future Studies 172

A List of operation codes and rules used in HES dataset recoding **216**

B Chapter 3 supplementary material **219**
B.1 Sample output for one factor model (SWLEOC dataset) 219
B.1.0.1 Correlation Matrix 220
B.1.0.2 Measurement Equations 220
B.1.0.3 Goodness of Fit Statistics 221
B.1.0.4 Fitted Covariance Matrix 222
B.1.0.5 Fitted Residuals 222
B.1.0.6 Standardized Residuals 223
B.1.0.7 The Modification Indices Suggest to Add an Error Covariance 224

C PROMs used in the validation study (non-copyrighted or reproduced with permission) **226**

D K-L grading system **232**

E Chapter 6 Supplementary material **234**
E.1 Category probability curves 234

F Data examination procedure **236**
F.1 Missing data analysis 236
F.1.1 Introduction 236
F.1.2 Methods 237
F.1.3 Results 238
F.1.3.1 Comparative ROC analyses 240
Analysis of cases with incomplete data 240
Complete case analysis 240
Analysis of cases with incomplete data after imputing missing values 241
F.1.4 Discussion 242

List of Figures

2.1 Wilson and Cleary conceptual model of patient outcomes 18
2.2 Hypothetical examples of a reflective and formative model 21

3.1 Hypothesized factor model 58

4.1 ROC curves for the OKS and its subscales (OKS-PCS and OKS-FCS) 87

5.1 A summary of treatments for knee OA as recommended by NICE . . 99
5.2 Cumulative percentages of patients experiencing the change on the OKS from baseline less or equal to the value on the x-axis. Red line marks the minimum detectable change beyond the measurement error of the score (MDC_{90} of 6 points). 111
5.3 Cumulative percentages of patients experiencing the change on the OKS-PCS and ICOAP and the OKS-FCS and KOOS-PS from baseline less or equal to the value on the x-axis. The ICOAP and KOOS-PS have been reversed to enable comparison of the amount of change between the pairs of scores. 112

6.1 Item characteristic curves for four items in a scale, with dichotomous response options and different levels of difficulty. 125
6.2 Threshold map for the summary OKS. 135
6.3 Targeting of the OKS-PCS. 138
6.4 Threshold Probability Curves for item 6 on the OKS-PCS. 138

7.1 ROC plots for the OKS, OKS-PCS, OKS-FCS, ICOAP and KOOS-PS. 156
7.2 Graphical representation of threshold levels for the ICOAP, the OKS-PCS and the OKS 158

E.1 Category probability curves for the OKS 235

List of Tables

2.1 Overview of some most commonly used generic HRQoL measures . . 38

3.1 Patient characteristics and percentage of completed OKS at baseline across the four datasets. 62

3.2 HES/PROMS dataset. Patient diagnosis at baseline and the number (N) and percentages of procedures undertaken. 63

3.3 SWLEOC dataset. Patient diagnosis at baseline and the number (N) and percentage of procedures undertaken. 63

3.4 KAT dataset. Patient diagnosis at baseline and the number (N) and percentage of procedures undertaken. 63

3.5 NOC dataset. Patient diagnosis at baseline and the number (N) and percentage of procedures undertaken. 63

3.6 Factor extraction results. Suggested number of factors using different extraction methods within each dataset. 66

3.7 Item loadings on the first factor, following 1 factor EFA (abbreviated item content next to question number). 67

3.8 Item loadings on the first two factors, following results of 2 factor EFA (abbreviated item content next to question number). 68

3.9 Summary of CFA fit measures for one and two-factor model of the OKS. 68

3.10 Cronbach's alpha (α) and correlations between the OKS subscales. 69

4.1 Methods considered for determining and using minimal important differences in knee replacement surgery. 80

4.2 Mean pre-operative, six month post-operative, change scores and effect sizes for the OKS and its subscales (OKS-PCS and OKS-FCS). . . . 83

4.3 Number and percentages of responses on the global transition item for different response categories, with effect sizes, mean score changes by response category and ANOVA tests for linear trend for the OKS and its subscales (OKS-PCS and OKS-FCS). 84

4.4 Anchor and distribution based MIC/MID values for the OKS and its subscales (OKS-PCS and OKS-FCS). 86

5.1 Baseline scores for the OKS, its subscales (OKS-PCS and OKS-FCS), ICOAP, KOOS-PS, and SF-12 physical and mental summaries (PCS-12 and MCS-12). 107
5.2 Fit indices of one and two-factor model of the OKS. 109
5.3 Fit indices of the KOOS-PS and one and two-factor model of the ICOAP. 109
5.4 Baseline Spearman's correlations between the scores. All correlations were significant at the 0.01 level (2-tailed). The number of cases with complete information that allowed the calculation of the correlation coefficients is in brackets for each correlation. 110
5.5 Significance of change in OKS, its subscales (OKS-PCS and OKS-FCS), ICOAP and KOOS-PS scores at three months (one sample t-test). . 113
5.6 Spearman's correlations between the 3 month changes in the OKS and its subscales (OKS-PCS and OKS-FCS), ICOAP and KOOS-PS. . . 113
5.7 Number (N) and percentage of responses for different response categories with effect sizes (ES), mean score changes by response category and ANOVA tests for linear trend for the OKS and its subscales (OKS-PCS and OKS-FCS). 114
5.8 Number (N) and percentage of responses for different response categories with effect sizes (ES), mean score changes by response category and ANOVA tests for linear trend for the ICOAP and the KOOS-PS. 115
5.9 Anchor based and distribution based MIC/MID values for the OKS, its subscales, ICOAP and KOOS-PS. 115

6.1 Number of observations in the study sample per response category. . 134
6.2 Summary fit statistics for the summary OKS and its subscales (OKS-PCS and OKS-FCS). 136
6.3 Individual item fit of the summary OKS. 137

7.1 Summary of baseline scores for the OKS, its subscales (OKS-PCS and OKS-FCS), the ICOAP and the KOOS-PS. 153
7.2 Significance of differences in age, BMI and in the baseline scores for the OKS, its subscales (OKS-PCS and OKS-FCS), KOOS-PS and ICOAP between the two group of patients, assessed using the Mann-Whitney U test (2-tailed). 154
7.3 ROC analysis: Area under the curve (AUC) with 95% confidence intervals (CIs) for the OKS, its subscales (OKS-PCS and OKS-FCS), ICOAP, KOOS-PS, and the Kellgren-Lawrence (K-L) grading system. 155
7.4 Uncertainty threshold zones for three best performing PROMs (the OKS, its pain subscale and the ICOAP) that take measurement error of the measure at individual level into an account. 157

D.1 K-L criteria of knee OA 233

F.1 Number and percentage of missing items and missing OKS summary scores, with pain and functional component summaries 239
F.2 Number and percentage of individual missing items and summary scores for the ICOAP 239
F.3 Number and percentage of individual missing items and summary scores for the KOOS-PS 240
F.4 AUCs with 95% CIs for each measure using available case analysis. . 241
F.5 AUCs with 95% CIs for each measure 241
F.6 AUCs with 95% CIs for each measure using complete case analysis without K-L grading included. 241
F.7 Percentage of missing values after imputation 242
F.8 AUCs with 95% CIs for each measure using available case analysis following data imputation. 242
F.9 AUCs with 95% CIs for each measure using complete case analysis following data imputation. 242
F.10 Significance of differences between AUCs obtained using different methods of data imputation for the KOOS-PS. 243

Nomenclature

AKS	American Knee Society Score
ANOVA	Analysis of Variance
AUC	Area Under the Curve
BMI	Body Mass Index
CDF	Cumulative Distribution Function
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
CI	Confidence Interval
COSMIN	Consensus Based Standards for the Selection of Health Measurement Instruments
CTT	Classical Test Theory
DWLS	Diagonally Weighted Least Squares
EFA	Exploratory Factor Analysis
ES	Effect Size
FA	Factor Analysis
FDA	Federal Drug Administration
GFI	Goodness of Fit Index
HES	Hospital Episode Statistics
HR-PRO	Health Related Patient Reported Outcome
HRQoL	Health Related Quality of Life
HSS	Hospital for Special Surgery score
ICC	Item Characteristic Curve

ICC(2,1)	Intra-Class Correlation Coefficient
ICD-10	10th revision of the International Classification of Diseases
ICOAP	Measure of Intermittent and Constant Osteoarthritis Pain
IRT	Item Response Theory
ISOQOL	International Society for Quality of Life Research
ISPOR	The International Society for Pharmacoeconomics and Outcomes Research
K-L	Kellgren-Lawrence
KAT	Knee Arthroplasty Trial
KOOS	Knee Injury and Osteoarthritis Outcome Score
KOOS-PS	KOOS-Physical Function Shortform
MCS-12	Mental Component Summary of the SF12
MDC	Minimum Detectable Change
MIC	Minimal Important Change
MID	Minimal Important Difference
MMT	Modern Measurement Theory
MRI	Magnetic Resonance Imaging
NHS	National Health Service
NICE	Receiver Operating CharacteristicNational Institute of Clinical Excellence
NJR	National Joint Registry
NOC	Nuffield Orthopaedic Centre
OA	Osteoarthritis
OKS	Oxford Knee Score
OKS-FCS	Oxford Knee Score-Functional Component Summary Scale
OKS-PCS	Oxford Knee Score-Pain Component Summary Scale
OPCS	Classification of Interventions and Procedures
PA	Horn's Parallel Analysis

PA-MRFA	PA based on Minimum Rank Factor Analysis
PCA	Principal Component Analysis
PCM	Partial Credit Model
PCS-12	Physical Component Summary of the SF12
PCT	Primary Care Trust
PNFI	Parsimonius Normed Fit Index
PROM	Patient Reported Outcome Measure
PSI	Person Separation Index
QALY	Quality Adjusted Life Year
QoL	Quality of Life
RMSEA	Root Mean Square Error of Approximation
RMT	Rasch Measurement Theory
ROC	Receiver Operating Characteristic
RSM	Rating Scale Model
SD	Standard Deviation
SEM	Structural Equation Modelling
SF-36	Short Form 36 item Health Survey
SHA	Strategic Health Authority
SRM	Standardised Response Mean
SRMR	Standardised Root Mean Square Residual
SWLEOC	South West London Elective Operative Centre
TKR	Total Knee Replacement
WHO	World Health Organization
WOMAC	Western Ontario and McMaster Universities Osteoarthritis Index

Chapter 1

Introduction

1.1 Overview

The Oxford Knee Score (OKS) is a Patient Reported Outcome Measure (PROM) that was designed in the mid 1990s to measure the outcome of knee replacement surgery in clinical trials [1]. In 2009, this widely used measure was adopted as a part of the UK National Health Service (NHS) clinical audit programme (the NHS PROMS[1] programme) as an outcome measure for patients undergoing knee replacement[2]. The introduction of PROMs in the NHS represents a fundamental change in focus from clinician-reported to patient-reported outcomes [3] and its routine collection is now part of the government strategy for measuring output and productivity of health services [4].

The increasing popularity of the OKS has resulted in its uptake for different populations and contexts from that for which it was originally developed. Some of

[1]From this point abbreviation "PROMS" will be used in the context of the NHS program.
[2]The OKS is used in conjunction with the EQ-5D, a generic utility measure of health status [2])

the most common extended uses are now highlighted:

(1) One observed extended application of the OKS is its use in the form of pain and function subscales. There have been several publications that have derived and used subscales on the OKS [5, 6, 7], based only on item wording and without any statistical exploration of their measurement properties. The OKS has been designed as a summary score and currently there is no evidence to support its use in the form of subscales.

(2) The use of the OKS in national joint registries and as part of the NHS PROMS programme has allowed assessment and comparison of outcomes following joint replacement surgery and between different types of prosthesis, on a large samples of patients (for example, see [8, 9, 10, 5, 11]). There are two potential concerns related to the interpretation of the OKS from such data. First, researchers are now able to detect (sometimes minute) differences in the outcome between different types/brands of implants and produce claims about their competitive performance. Due to large sample sizes, even the small differences from such datasets will often be statistically significant (e.g. less than 1 point on the OKS). However, it is currently not known if these differences are also detectable by (or important to) the patients and there is a danger of readily interpreting statistically significant differences as clinically relevant. Secondly, benchmarks of what is considered a 'good' surgical outcome that are derived purely on distributional characteristics of the sample[3] (with no clinical anchor) are increasingly considered inappropriate as primary outcome measures of change. Clearly, there is a need to establish the minimal clinically important changes

[3]Such as effect sizes

and differences on the OKS that could guide the interpretation of such data.

(3) There is a growing interest in (and anecdotal evidence of) using the OKS in clinical practice on individual patients who are undergoing non-operative management for their knee osteoarthritis (OA). The OKS was originally developed for the use on patients who are undergoing joint replacement surgery, its measurement properties (reliability, validity, responsiveness, etc.) are not known when applied on this new population and in the extended context. Furthermore, some Primary Care Trusts(PCTs) and Strategic Health Authorities (SHAs) have taken this use even further and are using the OKS as a key component in guiding resource allocation and provision of care for musculoskeletal patients. Here, individual scores are judged against arbitrary thresholds for surgery or secondary care referral such that patients whose scores lies above (better than) a particular OKS threshold may, for example, be considered not severe enough to warrant surgery [12, 13, 14, 15, 16, 17, 18].

The above mentioned (extended) applications assume that the OKS is acceptable for each of these proposed or extended uses, and that the OKS and its subscales are producing valid and meaningful information when applied in these extended contexts and/or different populations of patients. One of the limitations of the PROMs scores, such as the OKS, is that the measurement properties (such as validity, reliability, responsiveness to the change of clinical status in patients, and meaning of the change scores) of a measure are not of the measure alone but are a reflection of the population they have been tested on [19]. Hence, it is unwarranted to assume that the measurement properties of the OKS are invariant across different groups of patients. While the proposed new uses of the OKS are, in principle, potentially valuable, there is an

urgent need to build a body of evidence to support (or refute) the above mentioned applications. This requires greater understanding of its measurement properties in these extended roles in order to help the interpretation of the OKS data for research purposes, in audit, and in clinical practice.

1.2 Overall Scope and Specific Objectives of the Proposed Thesis

This thesis will examine the measurement properties of the OKS when used in extended roles and on different populations than for those which it was originally designed. The broad aims and specific objectives of this thesis are to:

(1) investigate measurement properties for the *extended application* of the OKS when used within the population the score was originally developed for; with the following specific objectives:

1a to explore if it is possible to extract pain and functional component scores from the summary OKS in the form of subscales.

1b to assess the *interpretability* of the OKS and its subscales by determining the Minimal Important Change (MIC) and Minimal Important Difference (MID) that can be used for both research purposes and in clinical practice.

(2) investigate measurement properties of the OKS when applied on the *new population* of patients (patients with knee OA, undergoing non-operative treatment) *and extended uses*, with the following specific objectives:

2a to examine the measurement properties of the OKS and its pain and functional component scores when used in patients (individuals and groups) who are undergoing non-operative management for their knee OA.

2b to explore if it is possible to use the OKS in individual patients to identify which patients might be in need for knee replacement surgery (i.e. for screening purposes).

1.3 Structure of the Thesis

The structure of this thesis is as follows:

Chapter 2 presents a literature review of the taxonomy of measurement properties and the underpinning methodology that was adopted throughout the thesis. The most commonly used patient reported outcome measures used for assessing knee replacement surgery and knee OA are also described.

Chapter 3 reports exploratory and confirmatory factor analyses of the OKS and the generation of pain and functional component scores (subscales).

Chapter 4 describes the generation of a range of minimal important changes and differences on the OKS for joint replacement surgery.

Chapter 5 reports on the results of a prospective cohort study that was undertaken in order to examine the measurement properties of the OKS, when used in patients who are undergoing non-operative treatment for knee OA.

Chapter 6 reports the results of further exploration of the measurement properties of the OKS by fitting the summary scale and its pain and functional subscales to the

Rasch polytomous model.

Chapter 7 explores the relationship between patients' perceptions about the severity of their knee OA symptoms and the absolute OKS score (and other commonly used knee specific measures), with an aim to produce thresholds on the scores that would indicate patients who are suitable candidates for knee replacement surgery.

Chapter 8 summarizes salient results from this thesis, their clinical relevance, and provides recommendations for future work.

1.4 Software licenses and author's role in the conduct of the studies and data preparation

Data analyses were performed using: SPSS software, for which the license was centrally financed by Oxford University; FACTOR a freeware program; LISREL, a licensed software that was funded (equivalent of £ 372.00) by the Department of Public Health at the University of Oxford; R, a freeware program, and a trial version of RUMM2030.

Chapter 3 used four datasets. The HES/PROMS dataset was obtained from the NHS Information Centre (IC) and the licence funding of £ 1,486.00 was secured from the National Institute of Health Research Biomedical Research Unit (NIHR BRU). Kristina K Harris (KKH) has written an application to the NHS IC requesting Hospital Episode Statistics (HES) (282 variables) and PROMS dataset (139 variables), has subsequently merged, recoded and cleaned the datasets using the SPSS. Following

this, a departmental statistician (IR) reproduced the merging, recoding and cleaning protocol with STATA software and the datasets were cross-validated. The Knee Arthroplasty Trial (KAT), Nuffield Orthopaedic Centre (NOC), and the South West London Elective Operative Centre (SWLEOC) datasets were obtained cleaned and recoded, at no charge. The KAT dataset was obtained from the KAT study (Chief Investigator Professor David Murray) which is a NIHR Health Technology Assessment (HTA) funded randomised controlled trial of different knee prostheses. The request for the use of the data was put forward to the study trial committee, who approved the use of the dataset for this purpose. The NOC dataset of 427 consecutive patients who were undergoing knee joint replacement in the Nuffield Orthopaedic Centre in Oxford was obtained from Professor Andrew Price, these patients were under his clinical care. The SWLEOC dataset was obtained from Professor Richard Field, who is the Director of Research at South West London Elective Orthopaedic Centre.

Chapter 4 used the previously merged HES/PROMS dataset.

Chapter 5, 6 and 7 used a dataset that was obtained from a prospective study looking at the natural history of early knee OA, that was set by KKH and Luke D Jones (LDJ). The KKH and LDJ wrote the protocol for their respective parts of the study, prepared ethics application and patient information documentation, managed NHS governance approvals, performed monthly mailing of the questionnaires, oversaw the work of the data entry clerk, performed readings of X-ray and MRI images, and performed database mergers with cleaning and recoding.

All data analyses in this thesis were performed exclusively by the KKH.

1.5 Brief history of health measurement

In the early 1900s, an American surgeon Ernest Codman proposed his 'End Result Idea', which he described as:

> "...the common sense notion that every hospital should follow every patient it treats, long enough to determine whether or not the treatment has been successful, and then to inquire, 'If not, why not?' with a view to preventing similar failures in the future."[20]

This statement is perhaps the first and the closest to the modern understanding of the importance of health measurement.

The purpose of health measurement in clinical medicine is to enable reliable and valid assessment of patients' health status in a domain of interest. Health measurement can involve assessment of physical characteristics (i.e. bone density) or quantification of health in a broader concept, such as measuring the health related quality of life (HRQoL). As presented later in this chapter, HRQoL measures are increasingly used in research and clinical practice. It is of critical importance to use measures that provide valid and reliable evidence, as the data generated from such measurements will ultimately influence the content and delivery of healthcare.

1.5.1 HRQoL measurement

One of the earliest references that impinges upon a definition of HRQoL can be found in Plato's *Crito*, where he describes the conversation between Socrates and his good friend Crito [21]. Facing a death penalty in prison, Socrates told his friend, who was

trying to persuade him to escape into exile, that "...the most important thing is not life, but good life". In this case, Socrates was using an example of a life worth living in a body that is in a 'bad condition' as an analogy for justice.

The advancement of medicine in modern times has dramatically increased the average lifespan, which meant that people with chronic conditions could live much longer (often with an impaired HRQoL). The first formal definition of health to recognize the importance of other factors than just an absence of symptoms was described by the World Health Organization (WHO) in 1948 [22]. In 1948 the WHO defined health as: "...a state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity." This definition of health influenced the development of outcome measurement in research, which eventually evolved into a patient reported format. The same year as the WHO published their definition of health, Karnofsky et al.[23] introduced extended criteria to evaluate the health status of patients undergoing cancer treatment, which ranged from "Dead" to "Normal; no complaints; no evidence of disease" and included the assessment of subjective improvement and their abilities to perform activities of daily life. Later, the importance of patients' perspective in reporting outcomes of treatment was increasingly acknowledged as essential when judging treatment outcomes [24]. This increased acceptance of the importance of patient's perspective in judging outcomes of interventions has been reflected in an exponential increase in the number of papers reporting on the development or evaluation of patient reported questionnaires that assess the quality of life in the past twenty years. There has been an impressive rise in a number of studies focusing on assessment of outcomes in cancer, musculoskeletal disorders and

rheumatology [25].

1.5.1.1 Measurement of outcomes by health care providers

The first known systematic measurement of outcomes started in the UK in 1844, following the introduction of the Lunacy Act. The Act required that psychiatric hospitals record the outcome of treatment and care using three simple categories: relieved, unrelieved and dead. Owing to the advocacy of Florence Nightingale, who argued that published outcome records of hospitals would: "show subscribers how their money was being spent, what amount of good was really done with it, or whether the money was doing mischief rather than good." [26], UK acute hospitals adopted the same measurement system.

Patient reported outcome measures to assess the performance of health providers were first developed in the US, in 1970's[27, 28]. There is a fairly recent shift in focus on (and a growing awareness of) the importance of measuring provider output in the UK in terms of patient-reported HRQoL gained. In 2009 the NHS rolled out its landmark PROMS programme. The PROMS programme uses standardized patient reported questionnaires as primary outcomes in order to systematically measure and monitor the performance of health care providers for four elective surgeries (hip and knee replacement, groin hernia and varicose vein surgery) [29]. This shift of focus from measuring *health care* to measuring *health* also (among other benefits) allows linking the quality of the performance of health care providers to the funding they receive, and may also help patients in choosing the best hospital for their treatment [3]. The PROMS data are now published on a monthly basis and are freely available

to the public [30, 31].

1.5.2 Psychometrics

The first recording of any testing date to about 2000 BC, when the Chinese civil service used oral examinations every three years on the officials to assess if they were fit for work. This later evolved into the use of test batteries that assessed knowledge on topics such as agriculture, law etc. [32, 33]. It is suspected that British missionaries in China encouraged the use of such testing systems in the Western world [34].

Psychometrics, as a science, began in the late 19th century. It was Sir Francis Galton, an anthropometrist with an interest in measuring individual differences between people, that introduced test measurement into mainstream psychology. Psychologist Cattell expanded on Galton's work and coined the term 'mental test' [35]. In 1886 Cattell (a collaborator of Galton) founded the world's first psychometric laboratory within the Cavendish Physics Laboratory at the University of Cambridge.

Quantitative methods in psychometrics evolved greatly in the first half of the 20th century. Pearson's chi-square test [36] and the Student's t-test [37] enabled further development of quantitative data analysis in psychometrics. Spearman [38] and Brown [39] in 1910 introduced the concepts of measurement error and the true score. The definition and estimation of validity and reliability evolved and took various forms during the next several decades [40, 41, 42, 43, 44]. There were numerous attempts to summarize the development in the area to and provide a basic overview in the field [45, 46, 47, 48] which culminated in 1968 with Lord and Novick's [49]

seminal systematic overview of Classical Test Theory (CTT) which (among other contributions) offered an elegant solution for the true score theory and is quoted as a start of a modern era in psychometrics [50].

Lord and Novick's publication also included four chapters by Birnbaum on Item Response Theory (IRT), which was considered an alternative to CTT. Unlike CTT, which is sometimes criticized for its sample dependency and focus on the summary score, IRT is considered to be 'sample free'. In IRT, the focus of analysis is not on the summary score, but on responses to items which are influenced by the individual's level of ability (or, in health measurement, disease severity). Latent trait is the underlying continuum of the ability or disease severity that is being measured by the test. IRT uses various models to describe relationships between the item responses and the position of the individual on the latent trait. Lord and Novick's publication brought together a good deal of work in the field and provided the most comprehensive coverage on the subject to date [51, 52, 53, 54, 55, 56, 57, 58, 59, 60]. This was followed by growing interest about this field in the community and the further development of existing and new IRT models [61, 62, 63, 64, 65, 66, 67, 68].

In parallel with the development of IRT, George Rasch independently developed an IRT model for items with dichotomous response categories (e.g. right/wrong) [69]. This was followed by further extensions of the basic model, which today form the family of Rasch models [70, 71, 72, 73, 74, 75, 76, 77]. Due to differences in the paradigms underpinning Rasch Measurement Theory (RMT) and IRT approach, these are each considered essentially different. These two theories/approaches can be denominated as the Modern Measurement Theory (MMT) [78]. The differences

between the two (and some 'Rasch Family' models) are described in more detail in Chapter 6 of this thesis.

1.5.2.1 Stipulations for a good health measure

Although psychometrics primarily originated and evolved in the field of education and personality testing, tests are being increasingly applied in the health care context [79, 19]. Tests are designed and applied in the healthcare setting on patients for different purposes, such as: to screen patients and/or assign a diagnosis (e.g. screening for alcoholism), provide a description of patients' health status at a single time point or over time (e.g. after an intervention), as prediction tools (e.g. return to work following a specific injury) [80]. Psychometric requirements of health measures depend on the purpose and the context they are applied in. For instance, if a measure is assessing the HRQoL, it is preferred that such measures are reported by the patient, as there is often discordance between patients' own rating of their HRQoL and the rating provided by the clinician or the carer [81]. Such measures are also known as Patient Reported Outcome Measures (PROMs).

Choosing a good PROM for a particular purpose and in particular context (e.g. clinical trial of arthroscopic knee washout for OA), can be a challenging task. There has been a great increase in the number of new PROMs, often with insufficient published evidence of their measurement properties. Also, as the standards on what constitutes acceptable evidence of measurement properties of a measure tend to change over time, it is somewhat challenging for older measures to keep up-to-date with the current criteria without updating this evidence in the form of new prospective studies.

This complicates the choice of an outcome measure for the researchers, who sometimes need to choose between an older, well established measure that might not have evidence on its measurement properties (which is, at that point in time, considered up-to-date), and a new measure that meets some of the new psychometric criteria. Consequently, the lack of consistency in the application of these measures is causing difficulties with comparing the results across studies [82, 83].

Over the years, there have been a number of efforts to define standards that health measures should achieve [84, 85, 86, 87, 19]. The latest (and perhaps the most comprehensive) international Delphi study was performed by Mokknik et. al [88], with an aim to define consensus based standards for the selection of health measurement measures (COSMIN). The study resulted in a broad agreement about important measurement properties and the recommended methodology for their assessment. The COSMIN taxonomy and broadly consists of: reliability, validity, responsiveness, and interpretability.The COSMIN guideline represents the absolute minimum measurement properties that should be assessed in a health status measure, beyond which there was no agreement[4].

While the evolving standards and increasing level of consensus on measurement properties of health related PROMs are extremely valuable for researchers embarking on the creation of a new measure, this currently complicates the choice of an existing measure for researchers or clinicians. When using such guidance to assess the quality of existing measures, (based upon information supplied in journal pub-

[4]For example, consensus was not reached in the choice of methods that should be used to assess minimal important differences or changes following an intervention.

lication) many of them (especially the older ones) are likely to fall short of what is currently considered a minimum standard. One reason for this is a lack of published evidence on measurement properties, due to specialist clinical journals not being keen on using the new 'psychometric language'. Another reason is that at the time of the development of early measures, the development of standards had not occurred. An immediate dismissal of existing, well established, measures on the grounds of the lack of published evidence, cannot necessarily be justified. In order to demonstrate the validity of existing measures, additional evidence of measurement properties would need to be subsequently obtained and published. This could be done in the first instance by contacting the authors or (in the absence of required information) by conducting a new prospective study. While it is reasonably straightforward to adjust a measure by changing item wording or even discarding or adding an item (and then testing the new version of the measure), some faults with measures might be more difficult to fix. For instance, in 2009 the U.S. the Food and Drug Administration (FDA) published a guidance for the industry where it outlines the requirements of the PROMs used as endpoints in clinical trials. The guidance puts emphasis on the measure content (content validity) which is achieved by the use of qualitative methods (patient interviews/focus groups) in the development phase of the measure. This publication also stimulated further publications in this field, with elaborate guidance on the construction of new measures [89, 90, 91] and on the examination of content validity for existing measures [92]. However, achieving the above outlined standards will be unattainable for some of the widely used existing measures as it might require their extensive restructuring, or their discontinued usage.

This chapter presented the overall scope, specific objectives and the structure of this thesis. Furthermore, a brief overview of the history of HRQoL measurement and psychometrics was provided. The next chapter will attempt to outline the basic terminology and requirements of a good outcome measure by reviewing aspects of the methodological quality of measures. The current state of knowledge in the outcome measurement for knee interventions will also be discussed.

Chapter 2

Literature review

2.1 Introduction

PROMs are now an increasingly important component of research, audit and clinical governance. Ultimately, the data generated from their use are responsible for shaping the content and delivery of healthcare and are receiving increasing attention. Consequently, there is a growing demand and requirement on both practising clinicians and researchers for a more comprehensive knowledge of this area.

2.2 Basic concepts

2.2.1 A conceptual model of HRQoL

In 1995 Wilson and Cleary [93] proposed the conceptual model of HRQoL (Figure 2.1). This model describes the relationships between five levels of patient outcomes: biological and psychological variables, symptom status, functional status, general

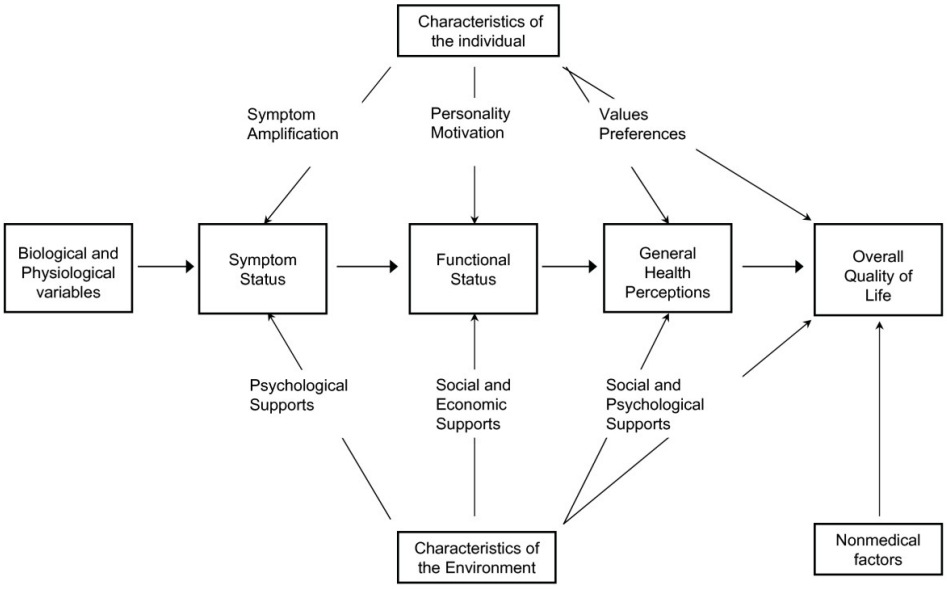


Figure 2.1: Wilson and Cleary conceptual model of patient outcomes

health perceptions and overall quality of life. The model has been extensively used in studies with patients with various health problems such as HIV/AIDS, heart disease and renal problems [94].

The arrows demonstrate causal relationships between outcomes, which increase in complexity from the left to the right side of the figure. However, Wilson and Cleary highlighted that this conceptual model only attempts to illustrate the most dominant interactions between the outcomes and the relationships represented by arrows may also be reciprocal. Furthermore, they point out that the absence of arrows should not be taken as a non existence of any relationships between the levels. The model also represents a transition from unidimensional, directly observable, objective, clinician based outcomes on the far left side, to multidimensional, non-observable, subjective,

patient based measures on the right side of the diagram [79].

Ferrans et al. [95] later revised the Wilson and Cleary HRQoL model. The revised model: i) added two more arrows demonstrating that biological and physiological variables are also influenced by characteristics of individuals and the environment, ii) labels on the existing arrows, that represent examples of relationships, were deleted (as these were considered limiting in the characterization of relationships), and iii) the box representing non medical factors was deleted as the authors considered that these can be categorized as either characteristics of individuals or the environment.

2.2.2 Using patient-reported outcomes to measure HRQoL

Using questionnaires to measure HRQoL poses specific conceptual and methodological challenges. Measuring biological and physiological variables (such as blood cholesterol levels) is fairly straightforward, as their values are directly observable (quantifiable). Quantifying HRQoL is more challenging, as this variable/outcome is not directly observable[1]. In psychology, non-observable variables are called *constructs* (or *latent variables*). Constructs can be observed indirectly by measuring observable indicators or *items/responses* on questionnaires (or *response variables*), that are related to the construct. Usually, multiple items are used to measure a construct of interest.

The relationship between the items and the latent variable that is being measured is described in a *conceptual framework* that forms an underlying model of a measure [79, 89]. Building a conceptual framework is usually a first step in the development of

[1] Although, all have some level of measurement error and some involve (subjective) interpretation, e.g. X-rays

a HRQoL measure. This should be initially done using qualitative methods. Ideally, during the development of a new measure, the initial (hypothesized) conceptual framework (that may be based on a literature review and expert opinion) is likely to evolve following patient interviews or focus groups. The final conceptual framework can later be tested (and, if required, adjusted again) using quantitative methods (such as factor analysis or structural equation modelling) [89].

2.2.2.1 Scales and indexes

Depending on the structure of the conceptual framework, outcome measures can be classified as scales or indexes. Figure 2.2 presents a hypothetical example of a scale and an index measure.

The model on the left side of Figure 2.2 outlines the conceptual framework of a scale. *Knee pain* is the construct (latent variable) that is being indirectly measured by the three observable items (response variables). An increase in the construct (pain) will influence all items on the related scale, which are highly correlated. When a change in a construct influences all the items, these items are considered to be *effect indicators*. Conceptual frameworks that are built using items that are effect indicators are also called *reflective models*.

The model on the right outlines the conceptual framework of an index. Three observable response variables are chosen to represent *Job satisfaction*. These three response variables are not necessarily correlated (e.g. promotion is not related to work transport, but salary may be). Also, an increase in the construct (job satisfaction) will not influence all items in a scale. However, an increase in any of the response

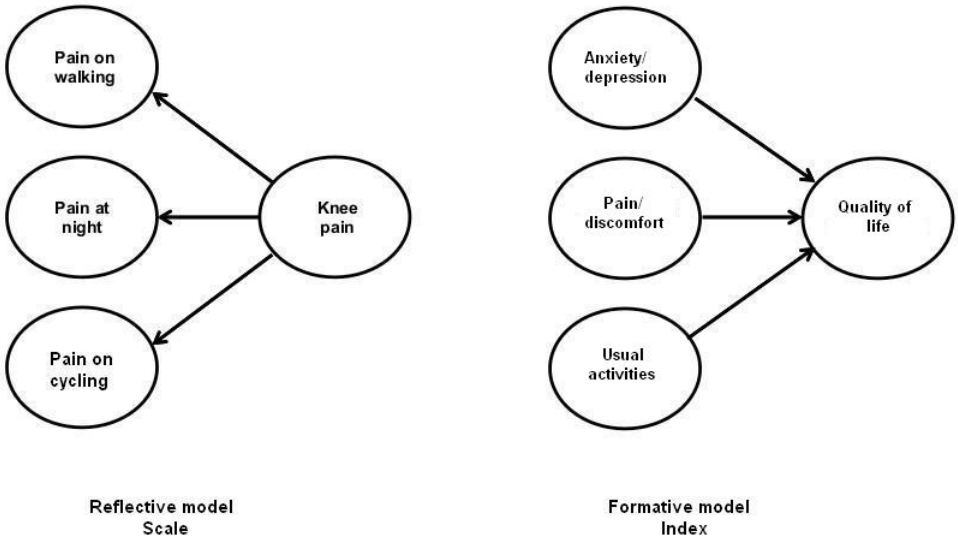


Figure 2.2: Hypothetical examples of a reflective and formative model

variables will influence job satisfaction. This type of a framework is called a *formative model* and items are called *causal indicators* as they 'cause' a construct.

This demonstration is somewhat simplistic. In reality, the distinction between a formative and reflective model is not clear cut as the relationships between the item and a construct can be reciprocal [96, 97, 98].

2.2.2.2 Unidimensional and multidimensional measures

Outcome measures that are built on reflective models are usually *unidimensional* (they assess a single construct). As seen in the example of the scale on the left side of Figure 2.2, such measures have standard items and response categories that are highly correlated and can be summed to a single summary scale. Although, a

high correlation between the items is a prerequisite, it is not a sufficient condition of unidimensionality. Evidence supporting the judgement of unidimensionality should be sought employing further statistical methods [99].

Multidimensional outcome measures have items that are not necessarily correlated. While it is justified to sum multiple items from a unidimensional measure in a single score, there is little justification in summing a multidimensional scale [96]. However, several items in a multidimensional scale representing one construct, can be summed and presented in the form of subscales (profiles), where each subscale (profile) measures one construct [79].

2.2.2.3 Standardised and individualised measures

PROMs are usually devised to reflect the areas of concern that were highlighted by a representative group of patients, during the measure development phase. Final items assigned to an measure usually represent the most commonly voiced areas of concern. However, these *standardised* measures will sometimes fail to fully address all relevant areas of concern of individual patients.

A special type of PROM are the *individualised* measures. Such measures were developed with an aim to overcome the theoretical limitations of standardised measures, by measuring areas of concerns that are nominated by individual patients [100, 101]. By addressing only the areas that are of concern to individual patients, these measures can have very good content validity [102]. Patients are usually asked to nominate several important areas of concerns in their life, that are influenced by their current health problem. Some measures (such as the Patient Generated Index [100]) also asks

patients to assign an importance weighting to each area, that is used to produce a score for each of the items that are ultimately summed in a single index.

However, a number of concerns have been voiced over the validity of individualised measures. One of the concerns is the effect of the 'response shift' on the interpretation of the measure. Response shift is the change over time that can occur in the individual's judgement of their quality of life (QoL) and the relative importance of different aspects of quality of life [103]. Faced with a decline in their quality of life, patients often employ coping mechanisms and adjust their expectations and priorities. For example, a patient might report a reduced QoL due to knee pain during sport and recreation (e.g. regularly playing tennis) and this might represent an important problem. However, once the patient adapts by switching to a different activity (e.g. regular jogging), the same patient might report the lower level of concern in this domain of QoL as it is not that important any more (although the inability to regularly play tennis is still present). This phenomenon can have implications for the responsiveness of the measure (as the measure then does not measure the same concept over time) [104]. Furthermore, completion of these measures requires patients to nominate areas of concern and assign weighting to them. This process can sometimes be confusing for patients and result in low response and completion rate of such questionnaires [105]. For this reason individual measures are often completed with the assistance of trained staff, which makes the use of these measures less practically feasible than the use of standard postal PROMs. As individualised measures are indexes, concerns have been voiced regarding the justification of their scoring in the form of a summary scale. A key problem with using individualised measures is

that their use is of limited value in clinical trials as combining such measures in an aggregate score is challenging (at the least) [106]. Although, individualised measures might be more suitable within a clinical trial context, some of these measures have low test-retest reliability and are presently unsuitable for use on individuals [107].

2.3 Features that make a good PROM

A good PROM will have satisfactory evidence of measurement properties for its intended use. These can be reported either during the measure development phase or in subsequent studies. As outlined in Chapter 1, there are two main measurement theories underlying the construction and evaluation of outcome measures, that are used to define how the scores on the measure represent the construct(s) of interest [97]: the Classical Test Theory (CTT) and the Item Response Theory (IRT).

2.3.1 Methods based on CTT

A number of important works laid the foundations for CTT [108, 40, 109], which culminated in a seminal systematic overview published in 1968 by Lord & Novick [49] that presented the fundamental model of measurement. The basic assertion of CTT is that, in a relation to a particular measure, every person has a true score that would be obtained were it not for the measurement error. Therefore, the observed score on a questionnaire consists of a true score and an error. The following formula [49] illustrates this relationship:

$$X_i = T + e_i$$

Where X is the observed score for item i, T is the true score, and e is the random error for item i. It is assumed that T is a value that would be obtained if the test was administered an infinite number of times and that the random error is normally distributed around the observed score.

Two basic assumptions underpin CTT: a) there is no correlation between the error score and the true score, and b) error components on a measure are uncorrelated [49].

Although these assumptions hold true on the population level, they have been challenged when applied empirically (on a sample) [110, 111, 112]. As CTT calculations rely greatly on means, variances and correlations, they are prone to sampling bias and care should be taken if a test whose measurement properties have been derived from a large sample is used in a study that has a small sample size. Zimmerman [113] used simulations to demonstrate that the results of some CTT calculations vary according to sample size, and as the sample size increases the difference in results between sample and population values decreases.

A description of measurement properties relevant to CTT [86], together with the recommended methods for their evaluation now follows:

2.3.1.1 Validity

The validity of a measure is concerned with whether a measure measures what it purports to [19, 114]. The definition of validity has more recently been defined as: "The degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test" [115]. When assessing the validity of a measure, assessment should be completed in relation to an appropriate

specific population and context (i.e. population of interest with particular impairment). Therefore, the measure can be claimed to be valid only when used with a similar type of population and in a similar context [19].

Several types of validation methods for health measures have been described in the literature that provide evidence on the validity of a measure in a specific context [19, 86].

Content validity A measure has strong content validity if evidence exists that the items adequately cover the construct of interest [19]. For example, questions about intensive sport would be largely inappropriate for a very elderly patient who is attempting to retain mobility by undergoing knee replacement. A PROM is most likely to have good content validity if relevant patients are involved in its development [102, 85], and this is an essential first step when developing new measures.

Since the publication of the FDA guidance that stressed the importance of content validity for PROMs [85], The International Society for Pharmacoeconomics and Outcomes Research (ISPOR) has published detailed protocols of how to establish and report evidence on the content validity of PROMs for medical product evaluation [92, 89, 90]. Generally, the measure development should start with the use of qualitative methods (such as interviews and focus groups) to inform the conceptual model and content of the new measure, followed by the examination of patient understanding of measure content and its revision. At this first stage of measure development, sampling of patients for focus groups/interviews need be carefully considered in order to ensure that the participants are representative of the study population [116, 85].

Cognitive interviews are usually conducted until the saturation occurs (i.e. when interviews stop eliciting new themes/areas of concern). Once the conceptual model of the measure and items have been devised, the newly derived measure can be subjected to further testing that includes quantitative methods.

Construct validity Mokknik et al. [88] defined construct validity as:

> "The degree to which the scores of an Health Related Patient Reported Outcome (HR-PRO) measure are consistent with hypotheses (for instance with regard to internal relationships, relationships to scores of other measures, or differences between relevant groups) based on the assumption that the HR-PRO measure validly measures the construct to be measured."

Construct validity is not assessed by a single statistic that determines the relationship between the test and the condition it is intended to measure. Testing construct validity rather encompasses a number of methods which can be used when assessing a construct that cannot be directly observed (such as pain, daily functioning etc.).

Structural validity Structural validity is one aspect of construct validity; it examines the extent to which the dimensionality of a measure corresponds to the construct that is supposed to be measured [88]. For instance, if a measure is unidimensional, all of its items will measure the same underlying construct [99]. During the development of a measure, this is usually assessed using Exploratory Factor Analysis (EFA) to investigate the number of factors (dimensions/domains) encompassed by the measure (and to identify any ill-loading/redundant items). For existing measures

(where the underlying structure of a measure is already known), Confirmatory Factor Analysis (CFA) is more appropriate[2] [117].

Hypothesis testing Construct validity can also be assessed by testing a series of *a priori* hypotheses about the relationships (or classes of association) of the measure with other (validated) measures [44, 118]. There is no upper limit on hypotheses that can be tested this way, but no single experiment can definitely confirm the construct [19].

Criterion validity Criterion validation assesses the degree of correspondence between a measure and a criterion variable (the gold standard). In order to establish this aspect of validity, an objective, highly reliable standard measure is needed for comparison (which rarely exists). For example, the findings at arthroscopy may be considered the gold standard when evaluating the ability of Magnetic Resonance Imaging (MRI) to detect ligament damage within the knee. When it comes to PROMs, there is often no appropriate gold standard measure, so validity testing must take more indirect approaches. The only exception would be in the case of validating a concise version of an existing measure, where the longer version could represent the 'gold standard'. This approach would imply looking for a strong correlation between the short and the long version of the measure. Two types of temporal designs are common when performing criterion validation of a measure: (1) concurrent validation, which involves gathering both forms of information for comparison at the same time, and (2) predictive validation, when the criterion information becomes known at a later time point

[2]EFA and CFA will be discussed in more detail in Chapter 3

[19]. Predictive validation is usually employed when testing the diagnostic ability of a measure to predict health outcomes, such as in risk prediction models [119].

Cross-cultural validity Cross-cultural validation is required when a measure is to be applied in another language or culture from that for which it was originally developed. The process usually involves forward, then backward translation of the measure, to ensure intended meanings have been retained. After all the translation issues have been resolved, the new measure is usually piloted on a small sample of patients to check comprehension. Following linguistic validation, the measurement properties of the final version of the questionnaire are ideally tested in the same way as occurs following the development of any new questionnaire, ideally using both CTT and MMT methods [79].

2.3.1.2 Reliability

Reliability is an estimation of the consistency and stability of a measure. It includes analysis of the extent to which a measure is free from measurement error. Ideally, a reliable measure, measuring the same characteristic and tested under similar conditions, would produce very similar results on 2 occasions[3]. Reliability (along with responsiveness) is considered an aspect of the validity of the measure. There are three most common classes of reliability estimates.

[3]The results will never be identical, as there is always a degree of random error involved in each application of the measure (as demonstrated in the basic CTT formula in section 2.3.1).

Internal consistency The assessment of internal consistency involves a measurement of inter-correlation of the responses to all items within a measure, arising from a single administration (in the case of scales). The most commonly used method to measure internal consistency uses Cronbach's alpha [43] giving a value (the alpha coefficient) that ranges from zero to one (one indicating perfect reliability). The formula for Cronbach's alpha is:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma_T^2}\right)$$

Where k is the number of items on a measure, σ_i^2 is the variance of the i-th item and σ_T^2 is the variance of all summed items in the score.

Care should be taken when interpreting the results of the alpha, as it tends to increase with the number of items on a scale [120]. Furthermore, high alpha values should not be taken as an indicator of unidimensionality, as even multidimensional scales can have a high inter item correlation [121]. A commonly used rule of thumb when interpreting the alpha coefficient is that it should be $\geq$0.7 to demonstrate internal consistency, and $\leq$0.9. Some authors consider that a score $\geq$0.9 may suggest redundancy of some items [19, 84]. Redundancy of items on a questionnaire also raises potential concerns with its practical use as (for instance) more parsimonious questionnaires tend to have better completion rates [122].

Test-retest reliability Another method of estimating reliability concerns repeatability and this is assessed using the test–retest method. Unlike internal consistency involving a single administration of the measure, the test–retest method requires two

separate administrations of the same measure at a certain time interval, involving the same respondents. In order to avoid the 'practice effect'[4] that can occur in the retest, the interval between two administrations should not be too short in case people remember what they said the first time and try to give the same answers-irrespective of whether it is 'true' or not. At the same time, the period between two administrations should not be too long, because of the possibility of real clinical change having occurred (which needs to be avoided). The most usual retesting period adopted is between 2 and 14 days [19]. The length of the interval depends partly on the condition that people have and how quickly real change is likely to occur.

Test-retest reliability can be expressed in many ways (error standard deviation, coefficient of variation) but the most used and accepted method in recent times for continuous measures is an intra-class correlation coefficient (ICC) [19]. There are six different versions of the ICC [123]. For the purpose of examining the test-retest reliability of PROMS, the $\text{ICC}_{(2,1)}$ is used. This formulation is based on a two-way analysis of variance (ANOVA[5]), where the interest is in assessing the absolute agreement on retesting. The formula for the $\text{ICC}_{(2,1)}$ which focuses on three sources of variance: patients, observers and error is:

$$ICC_{(2,1)} = \frac{\sigma^2_{patients}}{\sigma^2_{patients} + \sigma^2_{observers} + \sigma^2_{error}}$$

If a measure has perfect test-retest reliability (and is free from systematic bias), then

[4]Practice effect occurs when the subjects adjust the responses on retesting due to the learning effect.

[5]ANOVA is a general technique that assesses if the means of several groups are equal, under the assumption of underlying normality (it generalises the t-test on more than two groups).

the observed score will equate to the person's true score. However, the reliability of the measure is never perfect and there is always a degree of error in the observed score.

Score precision One standard deviation (SD) of the distribution of random errors is called the *standard error of measurement* (SEM). The SEM relates directly to the precision of the measure and is expressed as:

$$SEM = \sigma_x \sqrt{1 - R}$$

Where σ_x is the standard deviation of the scores in the sample and R is the test-retest reliability of the score ($\text{ICC}_{(2,1)}$). The SEM can be used to draw confidence intervals (CI) that tell us about the precision of a score at the level of an individual. This is done by multiplying the SEM value with the standard score (z value[6]) that is associated with the desired confidence level (e.g. 1.645 for 90% CI). It is also important to account for the precision of individual scores when measuring change in a construct of interest (e.g. health status) over time. The Minimal Detectable Change (MDC) [124] tells us about the error in the change score that can be expected on repeated testing. At 90% CI this would be:

$$MDC_{90} = 1.645 \times \left(\sqrt{2}\right) \times SEM$$

[6] Z-score indicates how many SDs some observed value is from the population mean. A normal curve with a mean of 0 and SD of 1 will have its 90% of the data within ±1.645 SDs

Where 1.645 is the z-value for 90% CI, $\sqrt{2}$ accounts for the error introduced on two measurement occasions and *SEM* is the standard error of measurement for the measure. As will be discussed later in this chapter (and in more detail in the Chapter 4), the MDC is an important aspect of the interpretability of an measure in identifying real change in health status that is beyond the measurement error.

2.3.1.3 Responsiveness (sensitivity to change)

The ability of a measure to detect meaningful clinical change (where it has occurred) over time is critical for the use and application of an measure [84]. This change might occur following an intervention, or just occur 'naturally' during a period of observation. In order for an instrument to measure change in the construct of interest, more than one measure needs to be made. In the case of interventions, this would normally involve a pre-intervention measure and a post-intervention measure (although there are situations where this may not be possible, such as with conditions that arise through trauma).

There are many statistical methods described for quantifying the responsiveness of a measure, which depend on the type of measure that is being tested and its proposed use [125, 126, 127]. Generally, as with construct validity, responsiveness is assessed by testing *a priori* hypotheses about the relationship of the changes in one measure to the changes in another (validated) measure, or with reference to a change in a gold standard (as with testing criterion validity). Responsiveness can also be tested with reference to a transition item, where the responsiveness is tested only in subjects who had reported that the change has occurred. Although still useful in the evaluation

of responsiveness, methods such as effect sizes (ES) [128] or standardised response mean (SRM) [129] should be seen (and are now widely considered) as indicators of the magnitude of changes measured by the instrument, and not an indicator of the validity of the instrument in measuring change [130].

If a measure does not detect change, this can occur for three reasons:

1. A measure may have compromised validity for the specific population or context in which it is being applied i.e. the content of the items/questions does not tease out particular known or observed clinical changes (that would perhaps be identified by a different outcome measure).

2. The measure is adequate, but it is reflecting a real lack of change in status. Responsiveness should be tested on a sample that has experienced real change in status. A parallel global item that asks directly about change ('transition status') can be given to patients to check whether or not they have experienced relevant change.

3. Any lack of responsiveness may be due to reasons such as the existence of a 'floor' or 'ceiling' effect. Terwee et al. [131] proposed that floor and ceiling effects and interpretability should always be analysed to establish the methodological soundness of a health status measure. For instance, if more than 20% [102] of the respondents are with few/no symptoms, then there is limited scope for demonstrating improvement. On the other hand, if patients cluster at the lower end of the scale, there is no scope for demonstrating deterioration. This may indicate that relevant items at the lower or upper ends of the scale are missing. Floor and ceiling effects can occur when, for instance, administering a measure designed for an older population to a younger,

fitter, sample of respondents who are most likely to score at the higher/better end of the scale.

It is essential to be aware of any issues related to responsiveness before interpreting data.

2.3.1.4 Interpretability

Interpretability is defined as the degree to which one can assign qualitative meaning to a quantitative score [86]. In clinical trials, this issue can concern the question of what is considered to be a 'good', 'bad' or 'indifferent' outcome (as measured by a particular criterion or score) and what is considered to be a clinically relevant change. The minimum amount of change that is discerned as meaningful by patients is particularly important as it affects interpretation of study results (this topic is discussed in some detail in Chapter 4).

Additionally, when reporting results of trials, it is generally useful to present the amount of improvement or deterioration and percentage of patients achieving it. A useful way of visualising the difference between groups is to use the Cumulative Distribution Function (CDF) plot [7], which plots the mean change against the proportion of people who experience change at that or a lower level [85].

Scores obtained at a single point in time, relative to intervention/treatment, can be used (with caveats) to classify outcome in specific situations. In order to provide context a qualitative interpretation, some authors have categorized or graded groups of such 'absolute' scores to represent 'excellent', 'good' health state, and so on. How-

[7] An example of use of a CDF plot can be found in Chapter 5.

ever, 'cut-off points' along the continuum of a score, used to produce such categories, are always arbitrary [19] and do not account of the baseline (e.g. pre-intervention) health status. For instance, if a post-intervention score of 75/100 is generated, in the absence of a context (particularly knowledge of the pre-intervention or baseline score, and hence the amount of change that has occurred), the value may be relatively meaningless.

Finally, it should be noted that while it is extremely important to choose an instrument with the best and most appropriate measurement properties for a specific purpose, even a perfect measure will yield useless data if the process of data collection is poorly conducted. This can lead to missing and potentially biased data [132].

2.4 Different types of outcome measures used in orthopaedics

The focus on outcome measurement in orthopaedics has recently shifted from 'objective' (clinically assessed and reflecting the clinician's perspective) to 'subjective' (patient assessed and reflecting patient's perspective). Traditionally, the success or failure of an orthopaedic intervention was judged exclusively on the basis of empirical assessment of quantitative or technical variables (such as the amount of flexion/extension in an joint, measured in degrees etc.). However, the 'objective' judgement of improvement (or success) of an intervention, such as restored full range of motion does not necessarily reflect improvement in a patient's quality of life (or 'suc-

cess') from their perspective (a 'subjective' outcome) [133]. In more recent times, the importance of assessing outcomes in orthopaedics, that are based on the patient's perspective (PROMs), has been increasingly acknowledged [134, 133]. In the past, it was not uncommon to find indices that combined clinician-based and patient reported outcomes together in a summary score [135, 136]. However, as discussed earlier, indexes (which tend to characterise clinically assessed outcomes) are rarely scalable and (perhaps therefore) rarely have acceptable evidence of validity and reliability (if these were assessed at all).

There are many types of outcome measures used in orthopaedics, but one form of classification contrasts generic ('general health') with disease specific measures.

2.4.1 Generic outcome measures

Generic outcome measures are intended to be suitable for a broad range of patient groups across the general population. They are multidimensional, covering a broad range of aspects of HRQoL dimensions (such as social functioning or general health perception), represented by different subscales (Table 1). Their advantage is that, when used with disease-specific measures, they have the ability to detect the influence of co-morbidities on the specific measure and can also detect unforeseen effects of a health intervention [102]. However, when used as a primary outcome measure for orthopaedic interventions they might not be responsive enough to detect more subtle, clinically important, changes precisely because they also pick up 'noise' from other (irrelevant) conditions.

Table 2.1: Overview of some most commonly used generic HRQoL measures

PROM	Short Form 36 item Health Survey (SF-36)	Nottingham Health Profile (NHP)	Sickness Impact Profile (SIP)
Developer	Ware (1992)	Hunt, McEwen (1980)	Bergner et al. (1981)
Health domains (subscales)	Physical component summary (PCS) and Mental component summary (MCS)	Physical Mobility, Pain, Social Isolation, Emotional Reaction, Energy, Sleep	Physical subscale and Psychosocial subscale
Score	Domain profile (0=worst/100=best	Domain profile (0=best/100=worst)	Overall, category or domain profile (0=best/100=worst)
Number/ type of items	36 Likert type response items	38 dichotomous items	136 dichotomous items
Method of administration	Patient interview/self administered	Self administered	Patient interview/self administered
Time for administration	5-10 minutes	5-10 minutes	20-30minutes

2.4.2 Disease/site specific outcome measures

Disease/site (e.g. joint) specific outcome measures assess patients' health status in relation to a specific disease or health condition, part of the body, or intervention. These measures generally have higher sensitivity than generic measures which allows them to detect more clinically important changes, and thus makes them more clinically relevant when assessing changes in health status over time in relation to a particular intervention. However, they often lack the ability to detect unforeseen effects of a health intervention that may act on other domains other than specific ones measured [102]. As patients are often likely to suffer from more than one condition at a time, including a generic questionnaire together with the condition specific assessment ensures a more comprehensive evaluation of health status [137, 134].

2.4.2.1 Outcome measures for arthroplasty and knee OA

A number of 'structured' systematic reviews of the measurement properties of measures commonly used for assessing outcome of knee arthroplasty and OA are currently available [138, 139, 140, 141]. It should be noted that systematic reviews tend to employ different criteria in evaluating the measurement properties of measures and (as mentioned in Chapter 1) the standards for assessing measurement properties tend to change over time. Furthermore non-published information of measurement properties (due to some journal policies) should not be taken as evidence of absence of the same. While a new, comprehensive evaluation of these measures and their methodological quality is beyond the scope of this literature review, the remainder of this chapter will briefly outline key characteristics of some of the most popular measures.

Knee arthroplasty

Clinician reported measures These measures tend to be indexes (indices) and are generally considered to be methodologically inferior to PROMs in measuring the outcome of interventions in clinical trials, as will be demonstrated by the following example. The best known clinician reported measures for the knee are the Hospital for Special Surgery (HSS) score [136] and its successor, the American Knee Society Score (AKSS) [135]. The AKSS is a knee-specific assessment that consists of one item relating to patient reported pain, three items relating to knee function, and six items relating to the clinical examination of the knee (including assessment of range of motion, laxity, alignment etc.). The score assessed for the pain item is added to

the scores for items relating to the clinical examination to create a 'Knee Score' that ranges from 0 to 100 (best), and scores from the three function items are summed into a 'Functional Score', that also ranges from 0 to 100 (best). There are three major concerns with the use of such measures (and AKSS specifically).

i) The AKSS lacks content validity as its items were devised to measure aspects of a patient's knee and knee function that were considered, by clinicians, to be relevant and important for judging the outcome and success of knee surgery. Furthermore, no 'consensus methodology' was applied to elicitate these items from a sample of clinicians

ii) Index measures (such as the AKSS) have methodological concerns regarding their summary scoring, as their underlying model is formative, with items that are not necessarily correlated.

iii) Empirical evidence has demonstrated that clinician reported measures frequently (such as the AKSS) suffer from low inter-observer / rater and intra-observer/rater reliability [142, 143].

Patient reported measures The evolution and proliferation of PROMs over the years [25] may have led to some confusion when it comes to choosing an appropriate measure for a data collection exercise or a research study. Whether used in research, clinical practice or for audit purposes, PROMs are increasingly required to have adequate evidence of measurement properties following rigorous methodological standards [102, 86, 85, 89, 90, 92]. However, only a few measures for assessing the

outcome of arthroplasty have been shown to meet the criteria applied in systematic reviews of their measurement properties [144, 140, 145]. Presently, the Oxford Knee Score (OKS) is the most widely recommended PROM for this purpose.

The OKS is a 12-item questionnaire asking patients about their pain and functional impairment in relationship to their knee. Originally, Likert-type responses for each item were scored 1 to 5, with a summary score of 12 (best) to 60 (worst) [1]. The scoring system has since changed and the recommendation is to score each question from 0 to 4, with a summary score of 0 (worst) to 48 (best) [146]. The OKS was originally developed using interviews with patients who were undergoing knee arthroplasty, to be used as an outcome measure in clinical trials of total knee replacement (TKR) (at a group level i.e. for making group comparisons). Since then, it has been accepted and widely used within the international orthopaedic research community. The OKS is also used by a number of joint registries for audit purposes [146]. Thus, in 2009, it was adopted by the NHS PROMs programme in England and Wales as a primary outcome measure for knee replacement surgery [29]. As a part of this programme, OKS data are currently collected on patients who are undergoing joint replacement, pre-operatively and at six months post operation. Furthermore, from 2012 the UK National Joint Registry (NJR), which collects information on shoulder, elbow, hip, knee and ankle joint replacements, extended the NHS PROMS programme by collecting additional OKS data at one, three and five years following arthroplasty in patients who participate in the NHS PROMS programme [147]. These data are also linked to Hospital Episode Statistics (HES) data.

Although, the OKS is the only PROM developed specifically for assessing outcomes of arthroplasty, there is on-going work to validate measures for this purpose that were originally developed for other populations of patients. For instance, the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) [148] and the Measure of Intermittent and Constant Osteoarthritis Pain (ICOAP) [149] were originally developed for use with patients with hip and knee OA, but have shown to have acceptable evidence of some measurement properties for assessing patients undergoing knee replacement surgery [150, 151, 152]. However, before the use of these measures for such extended purposes is recommended, a comprehensive scrutiny of their methodological quality is required.

Knee osteoarthritis The measurement of outcome for patients with knee osteoarthritis, rather than arthroplasty, presents some different problems. Whilst many of the issues are similar to assessing outcomes for TKR specifically, there are some aspects peculiar to measuring the progress of a chronic disease, rather than measuring outcomes related to a particular procedure to treat severe disease that occurs on a particular date. Arthroplasty involves a rather dramatic intervention resulting in (on average) a large improvement in measured health status. Thus, the patients' health gain following surgery could easily be detected by using a measure that is somewhat insensitive to detecting subtle changes in measured health status. The assessment of gradual, often slow and subtle changes in function (which may include only temporary improvement), due to disease progression, demands the use of precise and responsive tools capable of detecting such change.

Several PROMs for assessing patients with knee OA are in existence. These are now described.

The WOMAC Index is a widely recommended measure that has been developed, based on expert opinion, for patients with hip/ or knee osteoarthritis [148]. It has been tested and used in a number of validation studies and clinical trials to measure the effect of pharmacological therapies, joint replacement surgery, various physiotherapy interventions and acupuncture [153]. The latest version of the measure WOMAC 3.1 has 24 items, assessing three dimensions (pain, disability and joint stiffness). Response options on each of the items on the WOMAC questionnaire can be Likert-type[8], with 5 response categories raging from 0 (best) to 4 (worst) or in the form of 100 mm Visual Analogue Scale (VAS). The summary score ranges from 0 (best) to 100 (worst). In the context of assessing knee OA (or arthroplasty) concerns exist about the fact that it is not just a knee specific measure, but was developed to assess both hip and knee problems, with items asking about lower limb function in general (which could include many structures, not just hip or knee) [144]. Although the content validity of an older version of the WOMAC questionnaire was subsequently examined on a sample of patients with knee and hip OA [154], in the light of the new FDA guidance (which stresses the importance of patient input in the developmental stages of the PROMs) [85], the content validity of WOMAC can be questioned, as this measure was developed based only on expert input, without the

[8]Likert-type response options in HRQoL measures are often are at an ordinal level; they do imply ordering between the response categories (e.g. worst to best), but do not assume equal distances between response options.

involvement of patients.

The Knee Injury and Osteoarthritis Outcome Score (KOOS) was developed in 1998 using expert panel and patient input. Semi-structured interviews were conducted with 30+ patients with knee OA or injury. Patients were also shown the WOMAC questionnaire, on which the KOOS is based (but is knee specific). The KOOS extends the assessment of WOMAC to 42 questions, which create five domains (pain, symptoms, function in daily living, function in sports and recreation, and knee related quality of life) [155]. Items are scored on a 5 point Likert-type response options ranging from 0 (worst) to 4 (best). Results are not combined in a single score, but presented as five different scores, ranging from 0 (worst) to 100 (best), representing five dimensions of the measure. The KOOS has been recommended as a good choice for long and short term assessment of knee OA, anterior cruciate ligament (ACL) reconstruction and meniscus injury and has good evidence of reliability, validity and responsiveness in this context [144, 156].

ICOAP and KOOS-Physical Function Shortform (KOOS-PS) A set of criteria to be used as an endpoint in clinical trials of disease-modifying osteoarthritis drugs was recently proposed by the OARSI/OMERACT initiative [157]. These consist of three domains (pain, function and structure). With regard to the proposed domains, two new PROMs were developed:

The KOOS-PS [158] consists of seven items with Likert-type response options and was developed, using Rasch analysis, to measure physical function in patients with

various degrees of knee OA from a longer version of the KOOS questionnaire. Each item is scored as the KOOS; from 0 (best) to 4 (worst), with a summary raw score ranging from 0 to 28. The score is converted (via a nomogram) to a true interval score that ranges from 0 (best) to 100 (worst).

The ICOAP [149] is an 11 item questionnaire devised to measure the nature and intensity of pain in people with knee or hip OA [159]. Its items were devised using focus groups with patients who had knee or hip OA. It has two subscales that measure the intermittent and constant pain. Factor analytic results from the developmental study of the questionnaire demonstrated that item loadings on a factor are not consistent with the proposed subscale structure, therefore for most analyses in this thesis scoring system where all items were summed into a single summary score will be used. Likert-type response options on items are scored from 0 (best) to 4 (worst), with a standardized summary score ranging from 0 (best) to 100 (worst). As with the WOMAC, the ICOAP was developed on a sample of patients with hip and knee problems, and might have compromised content validity (and sensitivity) when applied to a single joint.

2.4.3 Summary

- Whether used for research purposes (e.g. in clinical trials), audit, or in clinical practice, PROMs should be chosen carefully after close inspection of evidence of their method of development, validity, reliability, responsiveness, context and purpose.

- Ideally, a PROM will be supported by extensive evidence on measurement properties based both on the CTT and IRT, which may be presented over time in a number of independent research papers.

- There is a noticeable and recent emphasis on using PROMs in patients with knee OA and/or undergoing knee arthroplasty. At the same time, there is a growing awareness about the required standards regarding the quality of measurement properties of these measures. Clinicians and researchers are advised to increase their awareness and knowledge of the current and new tools coming into existence.

- Evolving standards of the acceptable evidence of measurement properties, coupled with selective reporting of such evidence in validation studies, somewhat complicate the choice of the most appropriate outcome measure for researchers and clinicians.

- Whilst the orthopaedic literature is awash with different scoring systems and outcome measures specifically used for assessing the knee, not all measures have evidence of or reach even the minimum psychometric standards for their proposed uses.

As outlined in the section 1.3 of this thesis, the next five chapters will describe studies that were conducted with an aim examine specific measurement properties of the OKS, when used in different populations and/or roles than it was originally designed for.

Chapter 3 that follows will investigate the proposed use of the OKS in a form of pain and functional subscales (an extended role), in a population it has originally been designed for (patients undergoing knee replacement).

Chapter 3

An Exploratory and Confirmatory Factor Analysis of the OKS: Generation of Subscales Assessing Self-Reported Function and Pain

3.1 Introduction

The OKS was devised following extensive interviews with patients undergoing knee replacement surgery and it can be said that it satisfies the basic (qualitative) requirement of face and content validity. The measure, however, lacks quantitative evidence for aspects of its construct validity (structural validity). This chapter will explore quantitatively the factorial structure of the OKS. The number of latent variables

that are encompassed and their relationship with score items will be explored.

The OKS assessment is currently presented as a single summary score. The developmental interviews appeared to suggest that pain and function, from a patient's point of view, were inextricably linked in patients undergoing joint replacement. A similar perception of linked pain and function in patients with knee OA was confirmed in subsequent research [160, 161]. The OKS was therefore designed to be used as a composite measure reflecting patients' perception of pain and functional impairment arising from their knee [1, 162] and in this form, it has demonstrated evidence of validity, reliability and responsiveness [1, 162, 163]. When used as a single scale, the OKS has previously been demonstrated to meet the expectations of the Rasch measurement model and confirmed as a unidimensional measure that appears to measure both pain and function as a higher order construct combining both attributes [164].

There have more recently been several publications [5, 6, 7] analysing pain and function separately by extracting subscales from the OKS data ('Pain' consisting of questionnaire items 1, 4, 5, 8, 9, and 'Function' consisting of items 2, 3, 6, 7, 10, 11, 12). However, these subscales were in each case created on the basis of face validity (presumed item-scale allegiance based on their wording), without any formal statistical exploration of their measurement properties being undertaken. Consequently it is not known if these subscales are producing valid and meaningful information.

The purpose of this chapter is to examine if pain and functioning can be distinguished in the OKS in a meaningful way. This will be done by: (i) conducting Exploratory Factor Analysis (EFA) to explore the factorial structure of the OKS, and (ii) conducting Confirmatory Factor Analysis (CFA) to examine if a two factor

solution is superior to a one factor solution.

3.1.1 Factor analysis

3.1.1.1 Introduction

Factor analysis is a procedure that is widely recommended and used in the construction and validation of PROMs [19, 86, 165]. The main goal of factor analysis is to explain the observed variables (in the case of PROMs, items on a scale) by a smaller number of latent variables (factors) [165, 166].

EFA and CFA are two general techniques for conducting factor analysis and which method is used depends on the purpose of the study. Normally, in the early stages of instrument development EFA is used to identify the underlying structure of the measure or to discard redundant items. If, on the other hand, the underlying structure of the measure is already known and the goal is to check if this structure holds across groups (invariance), CFA is the method of choice. When conducting EFA, there is often no *a priori* knowledge about the relationships between the latent and observed variables, and the purpose of EFA is to identify latent factor solutions that are able to explain the pattern of correlations or covariances between the observed variables. The researcher usually decides on the criteria for determining the number of factors to be retained and on the underlying relationships between the latent variables (factors). Alternatively, CFA can be used statistically to test the fit of an *a priori* hypothesized structure of an instrument. Usually, several competing models that are based on theory and/or empirical research are tested for the goodness of fit. Nevertheless,

when both EFA and CFA are conducted it is important to consider the type of measurement scale represented by the instrument. Measures (such as the OKS) that use Likert-type responses provide a categorical (ordinal) description of an underlying continuous variable. Both EFA and CFA assume normally distributed data using Pearson-product moment correlations and are not robust when items with Likert-type responses are used [167, 168]. In such situations, data analysis should be based on the matrix resulting from polychoric correlations [169], which is also robust to underlying non-normality [170].

3.1.1.2 Exploratory Factor Analysis (EFA)

When constructing a new measure, the researcher usually wants to identify the redundant items on a scale, and/or the latent structure of the measure (that can be used to generate subscales). If the aim of the study is to construct the scale that has the most economic number of items, then the procedure known as Principal Component Analysis (PCA) is used. If the goal is to identify the number of factors that the measure is assessing, then the common Factor Analysis (FA) is used. These are the two most commonly used factor analytic methods [166].

PCA and FA have different approaches, which fundamentally differ in the way they relate to the variance of items (variables). FA assumes that the variance of a variable is composed of the common variance and unique variance. Common variance is the variance that is shared with the common factor (or factors). Unique variance is specific to a single variable and cannot be explained by the common factor. Unique variance consists of the variance is that is specific to the variable and error variance (random

error). In FA common variance is used in the estimation the latent variable (factor), whereas PCA uses both common and unique variance to estimate the *components* of the score. In both analyses, correlation matrices are used to extract factors (or components) that account for the maximum variance in the dataset. A simplified description of such procedure follows.

PCA and FA often start with the correlation matrix of all the items. In PCA, the diagonal elements of the factor correlation matrix are 1 whereas the FA computes and uses only communalities in the diagonal (and does not account for/excludes the unique variance). A series of eigenvectors (a set of weights, applicable to each variable) are computed from the correlation matrix, until convergence occurs (convergence is defined when new a new eigenvector is very similar to the previous one). Then, the elements of the first eigenvector are multiplied by the square root of the eigenvalue[1]. The result is the first component with its factor loadings. After the first component (or factor) is extracted, its item loadings are used to construct the matrix of cross products. This matrix explains the variance that is due to the first factor. The matrix of cross products is then subtracted from the original correlation matrix and the resulting matrix is called the *residual* matrix. The residual matrix is then used to extract the second factor or component (using the same procedure of calculating eigenvectors and eigenvalues), with the influence of the first factor or component removed. Again, the item loadings on the second factor or component are used to compute a second cross product matrix and the procedure can continue until the

[1] An eigenvalue is obtained separately and represents a square root of the sum of squares of the vector which is produced during the calculation of the last eigenvector (that will define convergence).

maximum number of factors or components is extracted (that corresponds to the number of variables). However, often the first one or several factors can explain the majority of variance and the extraction of other factors does not add much to the analysis. A critical step in factor analysis is the decision on the number of factors. There are several methods that can be used to decide on the number of factors or components that can be retained:

Kaiser over 1 rule is the most widely used method and also a default criteria for factor extraction in some widely used statistical programs (such as the SPSS) [171]. This method retains all the factors or components that have eigenvalues over unity. Eigenvalues are calculated as a sum of squared loadings of the item on a factor and, when divided by the number of items, they represent the percentage of variance that is explained by that factor. The commonly recognised problem with this rule is that it tends to produce too many factors [172]. Furthermore, this rule has been criticised to be arbitrary as the components that have eigenvalues just below 1 would automatically be discarded, whereas the factor with an eigenvalue just above 1 would be retained [173].

Scree test is a criterion for factor retention that relies on subjective judgement [174]. A scree plot is built by plotting the eigenvalues on the vertical axis against the factors on the horizontal axis. The plot is usually the highest for the first factor and then decreases. The point on the plot where the 'scree' begins (the slope of the line turns reasonably flat) is taken as the cut point for the number of factors that

are retained. However, the decision of where to put the cut-off point is not always clear-cut and it has demonstrated low reliability of plot interpretations [175].

Velicer's MAP test uses averages of the squared partial correlations from the partial correlation matrices [176]. In the first step, the first principal component is extracted from the matrix of item correlations and the average of squared partial correlations is calculated from the diagonals of the matrix. Then, the first two principal components are extracted from the matrix of squared correlations and the average of squared partial correlations is calculated again. The procedure is repeated for the number of variables minus one times. The point where the average of squared partial correlations is the lowest is taken as the number of factors to retain [177]. Velicer demonstrated that at this point the retained n components will not contain any unique components.

Horn's Parallel Analysis (PA) was proposed as an alternative to the Kaiser over 1 rule [178]. This method produces two scree plots. One scree plot is based on the actual data, and the other is based on a generated random set of uncorrelated variables (using the same number of variables and subjects as in the actual data). These plots are contrasted and the components on the real matrix which have greater eigenvalues than the corresponding components on the simulated matrix are retained.

Zwick and Velicer [179] have demonstrated that MAP and PA have shown to be superior to some widely used methods (such as Kaiser over 1 rule or the Scree test). However, when used on data that has a small number of variables loading on

a component or when these loadings are low, the MAP tends to extract fewer factors and the PA tends to extract more factors. Zwick and Velicer recommended that these two tests should be used in parallel.

PA based on Minimum Rank Factor Analysis (PA-MRFA) is a recently developed method to decide on the number of factors with polytomous ordered items [180]. This procedure is similar to the Horn's PA, with the main difference that it uses the polychoric correlation matrices (instead of Pearson correlations) and that the eigenvalues are calculated on the partial correlation matrices obtained by minimum rank factor analysis (MRFA), both for the observed and randomly generated data. Unlike PCA and common FA, MRFA does not use values from the diagonals. This enables the calculation of the percentage of explained and unexplained common variance (after the set number of factors is accounted for). In the case of PA-MRFA, the maximum number of factors is extracted (number of variables minus one). A simulation study by Timmerman and Lorenzo-Seva [180] has confirmed that both Horn's PA based on Pearson's correlations and PA-MRFA based on polychoric correlation matrices were robust, with PA-MRFA slightly outperforming the PA when the sample size was high.

Factor Rotation The initial factor loading results are often difficult to interpret and factor rotation is used. The simplified method to describe this process is that factors are represented as axes (with their loadings across the continuum) which are placed in the Euclidean space. An item's position would be determined by its loading

on each factor. For instance, if there are two factor axes, the position of an item would be determined by the correlation of the item with each factor (x,y). The items could fall in any of the four quadrants and distinctive factors will usually cluster together. The axes can be rotated in space, so that the positions (loadings) of the items on the factor axes change. However, this procedure does not change the percentage of variance of each factor and any solution is considered mathematically equivalent to the initial solution. There are virtually endless ways factors can be rotated, but the basic distinction is between orthogonal and oblique rotation. If we expect factors to be uncorrelated (which is rarely the case when measuring psychological phenomena [181]), then orthogonal rotation should be used. In this case, the relation of factor axes remains perpendicular during the rotation. If there is an expectation that factors should be correlated, then the oblique rotation is used. Here the factor axes can be rotated independently of each other and the resulting correlation between the factor is a cosine of the angle between the factor axes.

Interpretation of factor loadings Once the factors have been appropriately rotated, factor loadings should be interpreted. Normally, item loadings on a factor of 0.3 and above are usually taken as 'significant' [182].

3.1.1.3 Confirmatory Factor Analysis (CFA)

CFA is a part (or rather, a special case) of a wider collection of statistical techniques that are used to examine the relationship between dependent and independent

variables[2], known as Structural Equation Modelling (SEM) [183]. As CFA is a computationally demanding procedure, its detailed description is outside the scope of this chapter. A simplified description of the technique follows.

In its essence, CFA assesses if the sample data fit the hypothesized factor model. The first step in conducting CFA is the specification of the hypothesized factor model that is grounded in theory (often informed, partially or wholly, by the EFA). The researcher specifies expected relationships between observed and latent variables (factors and items) in the form of a diagram, where latent variables are usually represented as ellipses or circles and observable variables (items) are represented as squares. Figure 3.1 shows an example of a hypothesized factor model with 12 observable variables (items) and one latent variable or factor (g).

The arrows between the factor and the items go in one direction, meaning that the factor predicts the responses on the items. The factor is also an independent variable (as it has no arrows pointing to it) and the items are dependent variables. The arrows are partial regression weights that indicate how much the change in the independent variable (factor) will influence the change in the dependent variable (item). Items also have small arrows pointing to them which represent errors, that are residuals not predicted by the factor.

Following the specification of the model, the hypothesized relationships in the factor model are then translated into a series of equations or matrices, which estimate the population parameters. There are a number of estimation methods that are used under various conditions (e.g. violation of the assumption of normality, large or small

[2]These variables are often continuous or discrete and can be observable or unobservable (latent)

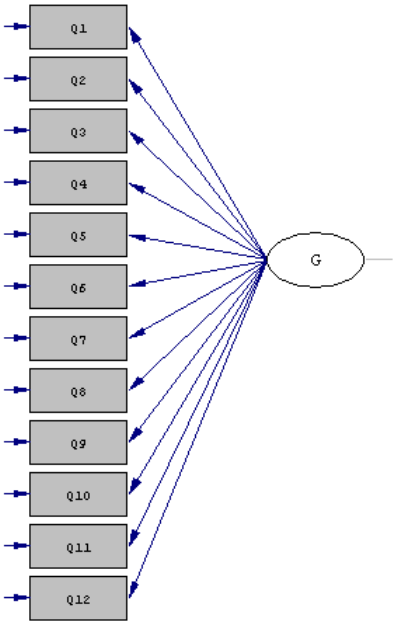


Figure 3.1: Hypothesized factor model

sample sizes etc.) that are offered by different software programs [184].

After the parameters of the model have been estimated (based on the choice of the most appropriate estimation method), it is then used to calculate the estimated population covariance matrix (or implied covariance matrix) which is compared to the covariance matrix based on the sample data. If those two matrices are similar (the difference between them is not statistically significant), the sample data fits the hypothesized factor model. The significance of the difference between the two matrices is estimated by using the chi-square test statistic (χ^2), with the χ^2 value of 0 indicating a perfect fit (no difference between the sample and population covariance matrix). Thus, the indication of a good fit of the model to the data is a non-significant χ^2.

A major concern with using the χ^2 as an exclusive indicator of goodness of fit is

that its calculation involves multiplication with the number of cases in the sample (N-1). Thus, large sample sizes will often yield a significant χ^2 and will lead to model rejection. A plethora of alternative goodness-of-fit indices have been derived to overcome this problem. Generally fit indices can be divided into: comparative fit indices (that assess the estimated model with the perfect fitting or non-fitting model), indices of proportion of variance in the sample covariance matrix accounted for by the population covariance matrix, indices that calculate the degree of parsimony in the model and indices that are based on residuals. These indices are mostly based on the χ^2 values of the matrices, sample sizes and degrees of freedom. They range from 0 (no fit) to 1 (perfect fit) [166]. If a model has a good fit, it will normally demonstrate this based on several types of goodness-of-fit indices.

The most commonly reported alternative fit index is Root Mean Square Error of Approximation (RMSEA) [185]. RMSEA compares the estimated model fit to a perfect model. RMSEA values below 0.05 are considered to indicate close fit, values to 0.08 indicate good fit, but if a model has more than 0.1 its fit is not considered satisfactory. The formula for RMSEA is:

$$RMSEA = \sqrt{\frac{\chi^2_M - df_M}{df_M\,(N-1)}}$$

where the chi-square of the model (χ^2_M) is related to the degrees of freedom of the model (df_M) and the sample size (N). Degrees of freedom are calculated by subtracting the number of parameters in the model to be estimated from the number of unique free parameters that are estimated from the data. A number of unique data

points is estimated using the formula $\frac{p(p+1)}{2}$, where p is the number of observable variables (items). In the case of the hypothesized factor model in Figure 3.1 there would be 78 unique data, 24 parameters to be estimated, resulting with 54 degrees of freedom.

If a model does not fit as expected, its modification can be considered. However, model modification can increase the risk of creating a miss-specified model that better fits the intricateness of the current sample but risks deviations from the true model [186]. Rather than considering model modification, CFA usually tests several *a priori* hypothesized (competing) models for the goodness of fit.

3.2 Methods

Exploratory (EFA) and confirmatory factor analysis (CFA) using PASW Statistics 20, FACTOR [187] and LISREL (structural equation modelling software) were performed by analysing pre-operative data on patients undergoing knee replacement surgery from four independent datasets: (i) the NHS HES/PROMS dataset containing records on patients from April 2009 to December 2011[3], (ii) the South West London Elective Orthopaedic Centre (SWLEOC) dataset containing records on consecutive patients from January 2009 to December 2010, (iii) the Knee Arthroplasty Trial (KAT) dataset on patients who participated in a UK national multi-centre national randomized controlled trial looking at the difference in outcomes between different types of total

[3]HES is the national dataset that contains information on all admissions and outpatients appointments in England. PROMS dataset is produced as a part of a separate NHS programme that collects additional information on all patients undergoing hip and knee replacements).

knee replacement (TKR) from July 1999 to January 2003, and (iv) the Nuffield Orthopaedic Centre dataset containing hospital records on consecutive patients from August 2010 to August 2011.

The datasets were obtained in a form of Excel spreadsheets. All of the databases were obtained after they have been cleaned, which explains high OKS completion rates (Table 3.1). Additionally, detection and correction of out range values was conducted using descriptive statistics (minimum and maximum values) and histograms. Participant demographics and the percentage of completed OKS are presented in Table 3.1. Tables 3.2, 3.3, 3.4, and 3.5 present the clinical characteristics of patients and procedures undertaken within all four datasets. In the HES/PROMS dataset the 'Diagnosis' field in was coded according to the 10th revision of the International Classification of Diseases (ICD-10) codes provided in the HES dataset: Infectious arthropaties=codes starting with M00 to M03; Inflammatory arthropaties=codes starting with M05 to M14; Polyarthrosis=codes starting with M1; Gonarthrosis=codes starting with M17; Other arthrosis=codes starting with M19. The HES dataset contains up to 19 'Diagnosis' fields, which contain the ICD-10 codes that are (sometimes) entered for all known diagnoses and for secondary procedures a patient might have. As the HES data is collected and entered by each health provider, there are some inconsistencies and errors in the coding of diagnosis variables (among others). In this study, a diagnosis was identified as valid, only if a case would have one of the relevant diagnostic ICD-10 codes. The 'Procedure' field was coded according to the relevant Classification of Interventions and Procedures (OPCS) procedure codes (see Appendix A for the OPCS codes used in the coding). Patients who had more than one procedure or diagnosis

Table 3.1: Patient characteristics and percentage of completed OKS at baseline across the four datasets.

Dataset	N	% Gender M/F	Age (range)	Completed OKS (%)
HES/PROMS	156 788	43/56	69 (18-97)	153 742 (98)
SWLEOC	2 405	37/63	71 (26-96)	2 405 (100)
KAT	2 353	44/56	70 (22-93)	2 337 (99)
NOC	427	44/56	69 (37-93)	427 (100)

Note. N=number of cases within each dataset; M=male; F=female.

were classified as 'mixed'. The HES and PROMS datasets were linked in SPSS using the "EPIKEY" field provided in both datasets. After the merger, there was missing or inadequately coded information about diagnosis on the 39% of patients and 34.9% had missing or inadequately coded information on the type of knee replacement undertaken. This is largely due to the fact that the linkage rate between the PROMS and HES dataset was 34.9%. According to the Health and Social Care Information Centre, the PROMS questionnaires did not link to HES for a number of reasons which include: poorly completed patient details on the PROMS questionnaire, un-coded or poorly coded OPCS codes, provider mapping issues, patients dying before, cancelling, or delaying their operation. Full reports on known issues about data quality for HES datasets are available online[188]. Within the HES dataset, 6.3% of patients had missing or inadequately coded diagnosis and no patients had missing information on the type of surgical procedure undertaken.

Table 3.2: HES/PROMS dataset. Patient diagnosis at baseline and the number (N) and percentages of procedures undertaken.

Diagnosis	N (%)	Procedure	N (%)
Infectious arthropaties	14 (0)	Primary TKR	95 140 (60.7%)
Inflammatory polyarthropaties	2 257 (2.4)	Primary UKR	1 163 (0.7)
Polyarthrosis	4 236 (2.7)	Primary hybrid knee	502 (0.3)
Gonarthrosis	83 565 (53.3)	Revision TKR	4 804 (3.1)
Other arthrosis	709 (0.4)	Revision UKR	34 (0%)
Mixed coding	4 873 (3.1)	Revision hybrid knee	30 (0)
		Mixed coding	428 (0.3)
Missing	61 134 (39)	Missing	54 687 (34.9)

Table 3.3: SWLEOC dataset. Patient diagnosis at baseline and the number (N) and percentage of procedures undertaken.

Diagnosis	N (%)	Procedure	N (%)
Primary OA	1 847 (76.8)	Primary TKR	2 246 (93)
Secondary OA	32 (1.3)	Primary UKR	150 (6.2)
Rheumatoid arthritis	54 (2.2)	PFR	9 (0.4)
Other	2 (0.1)		
Missing	470 (19.5)	Missing	0 (0)

Table 3.4: KAT dataset. Patient diagnosis at baseline and the number (N) and percentage of procedures undertaken.

Diagnosis	N (%)	Procedure	N (%)
Knee OA	2 151 (91.4)	Primary TKR	2 337 (99.3)
Rheumatoid arthritis	108 (4.6)		
Both	4 (0.2)		
Missing	90 (3.8)	Missing	16 (0.7)

Table 3.5: NOC dataset. Patient diagnosis at baseline and the number (N) and percentage of procedures undertaken.

Diagnosis	N (%)	Procedure	N (%)
Primary OA	400 (93.7)	Primary TKR	229 (53.6)
Secondary OA	13 (3)	Primary UKR	183 (42.9)
Rheumatoid arthritis	14 (3.3)	UKR with ACL	4 (0.9)
		PFR	6 (1.4)
		Complex primary KR	5 (1.2)
Missing	0	Missing	0

3.2.1 Analyses

3.2.1.1 Exploratory Factor Analysis

The datasets (separately and in combination) were submitted to EFA and CFA. The responses on the OKS are ordinal so the EFA was based on the polychoric correlations matrix [169] which was estimated using a two-step procedure. As the goal was to detect the structure of the measure and its association between factors, Principal Axis Factoring (PAF) was chosen as an extraction method [189, 190]. Several methods of factor retention were explored: Kaiser over 1 rule, Scree test, Velicer's MAP test, Horns PA, and PA-MRFA. Oblique rotation was used (Promax) as it was anticipated that the factors would be correlated with each other [191]. Items were assigned to a factor if their loading on the factor was above 0.3 and the difference between items loading on each factor was more than 0.2.

3.2.1.2 Confirmatory Factor Analysis

Following EFAs, CFAs were conducted. One factor and two factor models of the OKS were tested in LISREL.

The following fit indices were considered satisfactory [166]: Comparative Fit Index (CFI)>0.95; Goodness of Fit Index (GFI)>0.95; Standardised Root Mean Square Residual (SRMR) <0.8 good; <0.5 close fit; RMSEA <0.05 close fit, <0.08 good fit, <0.1 satisfactory fit. If these criteria were not met, the model would be considered as rejected. χ^2 difference tests were not performed, owing to the distributional assumptions of the test and the sample sizes of the datasets. Cronbach's Alpha was used to

test the internal consistency of the subscales.

Model 1 hypothesized that all 12 items characterize the single underlying factor. This model was tested as the one factor model forms a conceptual basis of the OKS. This was further confirmed by its evidence of high internal consistency [192] and on the basis of the number of extracted factors in this study using some of the most commonly suggested methods of factor extraction (Velicer's MAP, PA-MRFA and K-over-1 rule on all datasets and Horn's PA on HES/PROMS and the NOC dataset).

Model 2 tested two first-order correlated factors as suggested by the Scree Test on all four datasets and Horn's PA on KAT and SWLEOC dataset. As the data were ordinal and non-normal, the Diagonally Weighted Least Squares (DWLS) method, based on polychoric correlations and asymptomatic covariances, was used for extraction [170].

3.3 Results

3.3.1 Exploratory Factor Analysis

EFA produced evidence supporting both one-factor and two-factor solutions of the measure, depending on the statistical method employed (Table 3.6).

One and two-factor EFA revealed a generally consistent pattern of factor loadings across all four datasets (Tables 3.7 and 3.8). The overall results of all four merged

Table 3.6: Factor extraction results. Suggested number of factors using different extraction methods within each dataset.

	HES/PROMS	SWLEOC	KAT	NOC
N	156 788	2 405	2 192	427
Velicer's MAP test	1	1	1	1
Horn's PA	1	2	2	1
PA-MRFA	1	1	1	1
Scree Test	2	2	2	2
Kaiser over 1 rule	1	1	1	1

Note. N=number of cases in each dataset.

datasets were the same as in the HES/PROMS dataset. Two-factor EFA revealed that items 2 ("Have you had any trouble with washing and drying yourself [all over] because of your knee?"), 3 ("Have you had any trouble getting in and out of a car or using public transport because of your knee? [whichever you would tend to use]"), 7 ("Could you kneel down and get up again afterwards?"), 11 ("Could you do the household shopping on your own?") and 12 ("Could you walk down one flight of stairs?") loaded saliently on one factor. This factor was labelled 'Function'. Items 1 ("How would you describe the pain you usually have from your knee?"), 8 ("Have you been troubled by pain from your knee in bed at night?") and 6 ("Have you been limping when walking, because of your knee?") loaded on the other factor. This factor was labelled 'Pain'. Items 4 ("For how long have you been able to walk before pain from your knee becomes severe? [with or without a stick]") and 5 ("After a meal [sat at a table], how painful has it been for you to stand up from a chair because of your knee?") were not consistent. These items, together with items 9 ("How much has pain from your knee interfered with your usual work [including housework]?") and 10 ("Have you felt that your knee might suddenly 'give way' or let you down?"),

Table 3.7: Item loadings on the first factor, following 1 factor EFA (abbreviated item content next to question number).

	HES/PROMS	SWLEOC	KAT	NOC
N	156 788	2 405	2 192	427
Q9 (Work)	.818	.768	.821	.801
Q11 (Shopping)	.779	.754	.733	.770
Q12 (Stairs)	.756	.806	.730	.745
Q5 (Standing)	.752	.692	.723	.737
Q3 (Transport)	.722	.717	.671	.717
Q1 (Pain)	.685	.746	.730	.628
Q2 (Washing)	.661	.660	.621	.697
Q10 (Confidence)	.657	.618	.606	.668
Q7 (Kneeling)	.647	.686	.585	.595
Q6 (Limping)	.600	.517	.529	.537
Q4 (Walking)	.594	.652	.590	.589
Q8 (Night Pain)	.585	.564	.529	.570

demonstrated cross-loading within some datasets. These cross loading items were allocated to the factor called 'Pain' (while noting that the content of these items often overlapped with function).

3.3.2 Confirmatory Factor Analysis

CFA indicated that the two-factor model had slightly better fit indices than the one-factor model on all four datasets, but neither the one-factor nor the two-factor models met the criteria for rejection (Table 3.9).

3.3.3 Scoring of pain and functional subscales

Two subscales, based on the 2-factor EFA were derived: the 'OKS Functional Component Score' consisting of items 2, 3, 7, 11, 12 and the 'OKS Pain Component Score'

Table 3.8: Item loadings on the first two factors, following results of 2 factor EFA (abbreviated item content next to question number).

	HES/PROMS		SWLEOC		KAT		NOC	
Number of factors	1	2	1	2	1	2	1	2
Q3 (Transport)	.809	-.046	.726	.025	.837	-.115	.839	-.090
Q12 (Stairs)	.763	.034	.816	-.035	.561	.211	.881	-.036
Q11 (Shopping)	.731	.090	.803	.005	.582	.193	.610	.181
Q2 (Washing)	.717	-.020	.709	.021	.752	-.088	.737	-.049
Q7 (Kneeling)	.578	.102	.591	.030	.591	.030	.543	.178
Q1 (Pain)	-.066	.810	-.061	.758	-.044	.833	-.045	.889
Q8 (Night Pain)	-.043	.674	-.092	.726	.090	.471	.030	.593
Q6 (Limping)	.014	.631	.096	.484	-.119	.688	-.047	.624
Q9 (Work)	.323	.543	.361	.492	.305	.565	.380	.440
Q5 (Standing)	.256	.542	.239	.553	.374	.389	.448	.284
Q4 (Walking)	.224	.405	.404	.216	.010	.620	.365	.328
Q10 (Confidence)	.303	.391	.235	.480	.279	.362	.311	.348
Factor correlations	.786		.772		.778		.760	

Table 3.9: Summary of CFA fit measures for one and two-factor model of the OKS.

Acceptable fit				**>.95**	**>.95**	**<.08**	**<0.1**	**>.05**
	F	χ^2	df	GFI	CFI	SRMR	RMSEA (90% CI)	p[a]
PROMS/HES	1	44 766	54	.99	.98	.04	.073 (.073-.074)	1
	2	26 411	53	1	.99	.03	.057 (.056-.057)	1
SWLEOC	1	774	54	.99	.98	.05	.074 (.070-.079)	0
	2	508	53	.99	.99	.04	.060 (.055-.065)	0
KAT	1	566	54	.99	.98	.04	.066 (.061-.071)	0
	2	364	53	.99	.99	.04	.052 (.047-.057)	.3
NOC	1	183	54	.99	0.98	.05	.075 (.063-.087)	0
	2	149	53	.99	.98	.05	.065 (.053-.078)	.02

Note. P value for test of close fit (RMSEA<.05); F=number of factors; χ^2 =chi-square; df=degrees of freedom; GFI=goodness of fit index; SRMR=standardized root mean square residual; RMSEA=root mean square of approximation; CI=confidence intervals.

Table 3.10: Cronbach's alpha (α) and correlations between the OKS subscales.

	HES/PROMS	SWLEOC	KAT	NOC
N	156 788	2 405	2 192	427
α Total score	.92	.88	.87	.88
α OKS-Pain	.81	.82	.78	.79
α OKS-Function	.82	.82	.78	.82
Subscale correlations	.73	.71	.70	.75

Note. All correlations are significant at the 0.01 level (2-tailed). N=number of cases in each dataset.

and consisting of items 1, 4, 5, 6, 8, 9, 10. Table 3.10 presents α values for the total OKS, subscales and their correlations. When calculating the subscale scores, standardising them to a range from 0 (worst) to 100 (best) by multiplying the sum of the 'OKS Functional Component Score' by 5 and the sum of 'OKS Pain Component Score' by 3.57 is recommended.

3.4 Discussion

Factor analysis is one of a number of techniques that may be used to guide the development of PROMs, although some widely used PROMs have been developed without the use of this technique (such as the EQ-5D [2]). It should be noted that EFA and CFA are complex procedures whose results can depend on the chosen analytic approach (which always requires justification) and care must be taken when generalising beyond a particular study sample [190, 193].

The purpose of this chapter was to examine if pain and functioning can be distinguished in the OKS in a meaningful way. Factor extraction methods indicate that the OKS can be understood as consisting of one or two common factors. In a two factor solution, however, there remains a somewhat unclear distinction between self-

reported pain and function. Findings from other research conducted on this topic [160, 194, 161, 1] suggest that self-reported pain and function are constructs that might have considerable overlap, particularly in the way that patients perceive them, this is because pain influences functional ability [194, 160, 1]. Measures (such as the OKS) that are developed to reflect patients' concerns and perceptions about the influence of the disease on their life would normally reflect this lack of distinction. Nevertheless, although the high correlation between pain and self-reported functional ability in knee OA should be expected (especially where people are affected enough to warrant joint replacement), where people have less severe disease the progress or decay in one factor may not be predictive of the same dynamic in the other due to the complex relationship between pain and function in knee OA [195]. Some distinction between the two constructs can thus be drawn which allows this issue to be further investigated.

The two-factor EFA consistently identified that some items do saliently load on factors named function (items 2, 3,7,11 and 12) and pain (items 1, 6 and 8), both overall and across all four datasets. The results of analyses using the merged datasets were on the whole the same as in the PROMS/HES dataset alone due its large sample size. These eight items clearly drive the related constructs (containing additional items) that could be used as subscales. When considering a more in-depth analysis of pain and functional aspects from the existing OKS data, the cross loading items (4, 5, 9 and 10) that ask about pain in relation to function, can be assigned to the 'OKS pain component' (as here, pain influences functional ability). The cross-loading of these items supports their validity - they were derived to reflect the way in which themes

were expressed by patients in the original interviews and in practice, the results show that they do tap these different (yet overlapping) concepts.

Findings in this chapter support the use of subscale structures that are different from those previously used [5, 6, 7]. These previous structures were created on the basis of subjective judgements regarding item association with the constructs of pain and function, based simply on item wording (face validity). While this approach can be a good starting point, the appropriateness of the structures that result should be supported with more objective evidence of validity, i.e. with evidence from construct validation studies. In this respect it is interesting to note the main difference between these findings and those of previously published subscales. While in former studies items 6 ("Have you been limping when walking, because of your knee?") and 10 ("Have you felt that your knee might suddenly 'give way' or let you down?") have been allocated to the 'OKS functional component', the EFA findings in this study indicate that item 6 is clearly associated with the 'Pain' factor and item 10 with both factors. Thus, while on the face of it, these items might appear to ask about function, they are actually tapping into the construct of pain (albeit in relation to function) and they reflect patient perceptions of, and a complex dynamic between, pain and function in knee OA.

CFA of one and two factor models revealed a high χ^2 with a statistically significant value ($p < 0.05$) (using this criterion, models with p values of less than 0.05 are advised to be rejected). This was expected due to the large sample size, and alternative fit indices were considered [196]. CFI, GFI, SRMR and RMSEA indices demonstrated good fit of the data to both models and, if anything, seem to slightly favour the

two-factor model, but both models are plausible (they are not rejected). The OKS in total, and its subscales, have acceptable internal consistency to be used in research.

In addition to the established use of the OKS as a single summary score, the option to extract and separately analyse information on pain and function from the OKS offers a valuable opportunity in the context of exploring the complex pattern of relationships between pain and function in knee OA. Researchers who are interested in further differentiating assessment of Pain and Function are now given a possibility to extract this information from the existing OKS data in the form of subscales.

The OKS is the only knee specific measure that has been subjected to such a high level of scrutiny in the population of patients undergoing knee replacement surgery. Consistent factor-analytic results in this study provide convincing evidence in favour of the use of the OKS and its subscales across diverse patient groups, but there are some limitations of this study that should be noted. First, the HES/PROMS dataset had 39% of the patients with missing or inadequately coded information on diagnosis and 34.9% of cases with missing or inadequately coded surgical procedure undertaken. Any problem with missing (or inadequately coded) data in this study largely concerns the linkage rates between the HES and the PROMS dataset (and particularly concerns patients' diagnosis and the surgical procedure undertaken). As most of the missing data is due to ID coding between the datasets, it is unlikely that this might introduce a threat to the validity of the data as the dataset represents national data on NHS patients undergoing knee replacement in England and Wales over 1.5 years. This does not rule out the possibility of bias, but does make it likely that the data are fairly representative. Data completion/provision is particularly

good (>98%) for the key OKS items associated with the PROMS dataset-which are crucial for this analysis. Furthermore, once the missingness due to poor linkage between the datasets is partialled out, in the HES dataset there were no cases with missing (or inadequately coded) information on the surgical procedure undertaken and only 6.3% cases with missing or inadequately coded diagnosis. Second, the fact that both models meet the criteria for acceptable fit, does not exclude the possibility that there could be an infinite number of models that could fit the data equally well (as CFA cannot confirm the model, but can only fail to refute it [197]). Third the internal consistency and structural validity of the 'OKS functional component' and 'OKS pain component' was evaluated. Further evaluation of construct validity and responsiveness of the subscales is necessary. Fourth, all of the four datasets contained responses from patients who were already listed for knee replacement surgery and the results could be different when applied to new populations of patients (i.e. post-operatively or in patients prior to listing for knee replacement): as stated earlier, the measurement properties of an measure are not of the measure alone but also reflect the population and context on which it has been tested. Further research should focus on trying to evaluate the generalizability of these findings in extended populations. Fifth, the separation of the responses into separate pain and functioning factors could be argued to be somewhat artificial. While it is possible to split the responses in such way (and indeed the results of these analyses suggest that it can) to a certain extent, this still does not reflect the fact that patients may perceive them as related concepts.

3.5 Summary

- This study used EFA and CFA to provide further evidence of construct validity of the OKS.
- The results demonstrated that the OKS measures a single construct and is primarily meant to be used as a single composite measure, but can also now be used in form of pain and functional subscales.
- The two newly-derived subscales also have high internal consistency and provide the potential for additional analyses in the context of clinical trials or other forms of exploration.
- If used in trials that are specifically targeting either pain, or functional, improvement, these subscales could be used as primary endpoints and to power sample size calculations.

This chapter explored and provided evidence of structural validity of the OKS. The results of the study demonstrated that the OKS can also be used in the form of its functional and pain component summary scales (OKS-FCS and OKS-PCS). The next chapter will explore the interpretability of the OKS and its subscales by estimating values representing minimal clinically important differences and changes and will also look at this in the context of the subscales outlined in this chapter.

Chapter 4

Minimal Important Change or Difference for the OKS Following Knee Replacement Surgery

4.1 Introduction

In addition to having evidence of essential measurement properties (validity, reliability and responsiveness), it is also important that PROMs scores are interpretable, i.e. that qualitative meaning can be assigned to a particular quantitative score (or to a difference/change in the score) [131]. For example, clinical trials are usually powered to detect differences in the measure of interest (at a desired significance level) between two (or more) types of treatment at the end of therapy. However, a statistically significant difference does not necessarily represent one that is clinically relevant. Thus a small, clinically irrelevant difference/change can be statistically significant if

the sample size is large; likewise, a large clinically relevant change can fail to reach statistical significance in a small study. Determining whether statistically significant differences/changes are clinically meaningful or important is essential not just for performing sample size calculations and judging the efficacy of interventions in clinical trials, but also for interpreting score changes in longitudinal cohorts, clinical trials or in individual patients (in clinical practice).

4.1.1 The importance of context when choosing between estimates of minimal important change/difference

The definition of terms and their use in this area can be complicated, as already expounded by King [198] and Terwee et al. [199] It is necessary to make clear that, while the two terms of 'change' and 'difference' are often used interchangeably, a distinction can exist between the concepts [200]. If the interest is in measuring an improvement (or deterioration) in health status over time (in a single patient or in a group/cohort), then a value representing the minimal important *change* (MIC) is generally used. However, if the interest is in comparing outcome in two separate groups of patients (such as a comparison of treatment type in an randomized controlled trial) a value representing the minimal important *difference* (MID) i.e. the difference between the two groups in the stated variable of interest, is often used. These distinctions have also sometimes been referred to as 'within' (relating to MIC) and 'between' (relating to MID) patient minimal important differences [201].

4.1.2 Different approaches to estimating meaningful or minimal important change or difference

Estimating the smallest amount of change or difference in a score on a measure that is meaningful can be undertaken using a number of different methods. Each method has advantages and disadvantages and presently there is no consensus as to which method is best [202]. Generally, methods fall into two main approaches termed 'distribution based' or 'anchor based' [131, 203, 204].

Distribution based methods calculate estimates using the statistical characteristics of scores (such as mean changes, standard deviations etc.) observed in a particular study sample [200] and express the observed change in a standardized metric (such as the effect size (ES) or the standardized response mean (SRM)[1]). Effect sizes offer a non-dimensional, unit-less way of expressing the change in the score of interest. ES is usually calculated by dividing the mean change in scores, in a group/study sample, by the SD of the baseline score and values of 0.2, 0.5 and 0.8 are by convention regarded as a small, medium and large magnitude of change, respectively [205]. ESs are often used as a basis for making sample size calculations. In a review [206] of thirty-eight studies computing minimal important differences of various HRQoL measures across a range of chronic diseases and in different contexts, Norman et al. found that the estimates of minimal important change/difference were generally around 0.5 SD at baseline (equivalent to an ES of 0.5). The authors recommended the use of this value

[1]The ES is obtained by dividing the change in the score with the SD at baseline and SRM by dividing the change in the score by its SD. Both of these statistics quantify the difference between two groups in a dimensionless fashion.

as a starting point in estimating minimal important changes/differences in HRQoL measures for chronic conditions. One commonly used distribution-based measure for this purpose is the SEM, which links the reliability of the measurement measure to the standard deviation of the population (i.e. the SD of the measure's mean scores at one point in time, such as, baseline/pre-intervention) [207, 208, 209]. An extension of this estimate is the MDC, which applies a confidence interval around the SEM [200]. As distribution based methods rely solely on statistical distributions to mark important change/difference, (as opposed to another clinically related variable), they are generally limited in their ability to indicate meaningful clinical change.

Anchor based methods explore how the change or difference in score on the measure is related to a relevant external anchor (for example, a level of response on a global transition item), where the anchor can be rated or set by the patient, clinician or other stakeholder. One example of global transition item set by patients would be a question about the amount of improvement they have experienced as a result of surgery. Anchor based methods are therefore, by definition, more clinically relevant as they relate the score or changes in the score to a clinically meaningful reference. Anchor-based estimates of minimal change (MIC) or difference (MID) can be obtained by direct reference to the anchor question or finding the best cut point on a Receiver Operating Characteristic curve (ROC) analysis [210]. Owing to their clinical relationship, MICs and MIDs can sometimes be referred to as the Minimal Clinically Important Change (MCIC) or the Minimal Clinically Important Difference (MCID) [211] (this nomenclature was not used in this thesis).

Presently, MIC/MID values for the OKS (or its subscales) are not published. As the interpretability of a measure is one of its essential measurement properties, this information is required in order to allow the appropriate interpretation of the changes measured by the OKS in different situations (clinical trials, cohort studies, clinical practice etc.)

In this chapter the OKS data from the NHS PROMS national dataset on patients who underwent joint replacement surgery was used in order to:

1. Present a range of estimates of meaningful or minimal important change/differences for the OKS and its subscales, based on different approaches.

2. Identify the estimates that are most consistent and useful for specific contexts.

4.2 Materials and Methods

An analysis of the NHS PROMS dataset of all hip and knee replacements undertaken from 01/01/2009 to 31/12/2011 in England and Wales was performed. Full PROMs data reports and methodology guides can be found online [212].

4.2.1 Subjects/Assessments

As a part of the NHS PROMS program, patients who were listed for primary joint replacement surgery completed a set of pre-operative questionnaires including the OKS. At six months post-surgery, the measures were repeated and patients also completed a global transition item ("Overall, how are your knee problems now, compared

Table 4.1: Methods considered for determining and using minimal important differences in knee replacement surgery.

	Anchor based methods	Distribution based methods
Minimal important change (MIC)	a) Mean score change in patients reporting to be "a little better" b) Optimum point on the ROC curve discriminating patients who consider themselves "a little better/much better" vs. "not changed/worse"	$\pm MDC_{90}$, ES, and 0.5 SD
Minimal important difference (MID)	Difference in the average change score between the group of patients who consider themselves "a little better" and patients who consider themselves to be "about the same"	

Note. MDC_{90}= minimum detectable change; ES=effect size; SD=standard deviation.

to before your operation?") with five response categories: "much better" (scored 1), "a little better" (2), "about the same" (3), "a little worse" (4), "much worse" (5).

4.2.2 Statistical Methods

Data were analysed using IBM SPSS Statistics (version 20.0; SPSS Inc., Chicago, IL, USA). Change scores followed a normal distribution, which allowed the use of parametric statistics. A range of minimal change values was estimated and examined using both anchor and distribution based methods. Table 4.1 outlines the classification system and methods considered in this study for determining and using minimal important differences in knee and hip replacement surgery.

4.2.2.1 Anchor based methods for estimating minimal change /difference

The validity of the anchor item to establish the minimal change/difference in relationship to the OKS was examined. It is usually recommended that the correlation

coefficients between a global transition item and the measure should be a minimum of r=0.30 [202] before an anchor based MIC (often termed the MCID) can be established, and a minimum of r=0.5 to ensure optimum sensitivity/specificity is achieved for minimal change scores where they are to be applied at an individual level [213]. Directionality of the association between the transition item and the scores was tested using an analysis of variance (ANOVA) with tests for linear trend.

Minimal Important Change (MIC) MIC can be used to assess the meaningfulness of the average change in a single group from baseline (e.g. cohort studies). The MICs for the OKS was calculated as a mean change in the score associated with the response category of "a little better" on the patient reported global transition item.

Minimal Important Difference (MID) MID can be used to compare the difference in average change from baseline between two treatment groups (e.g. in clinical trials, case-control studies). The between-group MIDs for the OKS were calculated as the difference in mean change scores between two categories of response on the patient reported global transition item : "a little better" and "about the same".

Best cut point ROC curves were used to derive a change value in the OKS associated with optimal sensitivity/specificity at which patients who have/have not improved (by their own rating) are most accurately identified. Patients who responded "much better" and "a little better" on the global transition item were categorized as 'improvers' and patients who responded "about the same", "a little worse" or "much worse" were categorized as 'non-improvers'. The value corresponding to the point on

the ROC curve that was closest to the top left corner was (by convention) defined as the optimal cut point [210]. ROC curves were also used to assess the predictive ability of the OKS to detect change. The area under the curve (AUC) represents the diagnostic ability of a measure, with a value of 0.5 denoting performance no better than random chance and a value of 1.0 indicating perfect predictive ability [214].

4.2.2.2 Distribution based methods for estimating minimal change or difference

Effect Sizes (ES) and the 0.5 Standard Deviation (SD) In this analysis, effect sizes for the OKS associated with each level of response on the global transition item were examined.

Precision of the change score based on SEM In this study, the MDC at the 90% level was estimated, using the formula $1.645 \times \left(\sqrt{2}\right) \times SEM$.

4.3 Results

4.3.1 Study population characteristics

At baseline, 156, 799 patients who underwent knee replacement surgery were recruited, out of which 94, 502 (60.3%) had completed both pre-operative and post-operative OKS, with 94, 015 (99.5%) of those completing the global transition item. 94, 040 and 94, 259 patients had completed both the global transition item, and had information on the change OKS-PCS and OKS-FCS, respectively. 55.5% of pa-

Table 4.2: Mean pre-operative, six month post-operative, change scores and effect sizes for the OKS and its subscales (OKS-PCS and OKS-FCS).

	OKS	OKS-PCS (SD)	OKS-FCS (SD)
Scale range	0 (worst)-48	0 (worst) -100	0 (worst) -100
Pre-operative mean (SD)	18.47 (8)	33.93 (17)	44.84 (19)
Post-operative mean (SD)	33.75 (10)	72.54 (23)	67.16 (22)
Mean change (SD)	14.74 (10)[a]	37.65 (23)[a]	20.96 (21)[a]
ES	1.9	2.2	1.1

Note. SD=standard deviation; ES=effect size;[a]All changes were significant at (paired t-test, at p=<0.001).

tients who underwent knee replacement were female. Patients who completed the post-operative OKS had 1.43 (SE 0.042) points higher pre-operative OKS than the patients who did not return the post-operative questionnaires. All differences were statistically significant (independent t-test, $p<0.001$).

4.3.2 Assessment of minimal change

The mean pre-operative and six month post-operative absolute scores, change scores and effect sizes for the OKS and its subscales (with SDs) are shown in Table 4.2. Effect sizes were large (>0.8) and mean changes were statistically significant ($p<0.001$).

The Pearson correlation coefficients for the association between OKS, OKS-PCS, and OKS-FCS change scores, and the global transition item were 0.55, 0.55, and 0.46 respectively. Correlations were significant at the 0.01 level (two-tailed).

Percentages of responses for different response categories, effect sizes and mean score changes by response category are presented in Table 4.3. There is a statistically significant ($p<0.001$) linear trend between greater perceived change measured with the global transition item and an increase in the change scores. Effect sizes in the

Table 4.3: Number and percentages of responses on the global transition item for different response categories, with effect sizes, mean score changes by response category and ANOVA tests for linear trend for the OKS and its subscales (OKS-PCS and OKS-FCS).

		Much better	A little better	About the same	A little worse	Much worse
OKS	N (% of responses)	65 086 (69)	16 901 (18)	5 279 (6)	3 819 (4)	2 930 (3)
	Mean change (SD)	18.64 (8)	9.22 (7)	4.38 (7)	1.53 (7)	-3.18 (7)
	Effect size	2.4	1.2	.6	.2	-.4
	P-value for linear trend	<.001	<.001	<.001	<.001	<.001
OKS-PCS	N (% of responses)	65 104 (69)	16 902 (18)	5 281 (6)	3 823 (4)	2 930 (3)
	Mean change (SD)	46.74 (19)	25.08 (17)	13.03 (17)	6.57 (16)	-4.59 (17)
	Effect size	2.8	1.5	.8	.4	.1
	P-value for linear trend	<.001	<0.001	<.001	<.001	<.001
OKS-FCS	N (% of responses)	65 241 (69)	16 944 (18)	5 293 (6)	3 834 (4)	2 947 (3)
	Mean change (SD)	27.74 (18)	10.96 (16)	3.70 (15)	-1.54 (15)	-9.49 (16)
	Effect size	1.5	.6	.2	-.1	-.5
	P-value for linear trend	<.001	<.001	<.001	<.001	<.001

Note. SD=standard deviation.

group who reported their knee to be "about the same" indicated a medium to large treatment benefit measured by the OKS and the OKS-PCS (0.6-0.8). More post-operative improvement per response category was experienced as measured by the OKS-PCS.

4.3.2.1 Summary of the range of anchor and distribution based MIC/MID values

Distribution based methods The 0.5 SD, SEM, and MDC_{90} values are shown in Table 4.4. The MDC_{90} values are similar to the 0.5 SD values, and the values suggested by the anchor based ('between group') MID are generally half of the anchor based MIC. The MDC_{90} values shown can be interpreted as meaning that in 90% of cases, a patient will have experienced real change (beyond measurement error) if their score has changed by at least 4.15 points on the OKS, 8.91 points on the OKS-PCs and 9.85 points on the OKS-FCS in either direction.

Anchor based methods The average score change associated with the response "a little better" on the global transition item was 9.22 points for the OKS (ranging from 0 to 48), 25.08 points for the OKS-PCS, and 10.96 points for the OKS-FCS (both subscales ranging from 0 to 100). The average score difference between the response categories "a little better" and "about the same" was ∼5 points for the OKS, ∼12 for the OKS-PCS and ∼7 for the OKS-FCS (Table 4.3). Patients who reported themselves to be "a little worse" showed an improvement of ∼2 points on the OKS, ∼6 points on the OKS-PCS, and ∼-2 points on the OKS-FCS. The average difference

Table 4.4: Anchor and distribution based MIC/MID values for the OKS and its subscales (OKS-PCS and OKS-FCS).

	Distribution-based			Anchor-based					
						ROC method			
	SEM	1/2 SD	MDC_{90}	MID	MIC	AUC (95% CI)	Best cut point	Sensitivity	Specificity
OKS	1.78	3.97	±4.15	4.84	9.22	.905 (.902-.907)	9	.806	.154
OKS-PCS	3.82	8.50	±8.91	12.05	25.08	.905 (.902-.907)	25	.780	.127
OKS-FCS	4.22	9.44	±9.85	7.26	10.96	.847 (.844-.850)	7.5	.809	.286

Note. SEM=standard error of measurement; SD-standard deviation; MDC_{90}=minimum detectable change; MID=minimum important difference; MIC=minimum important change; AUC= area under the curve; CI=confidence interval.

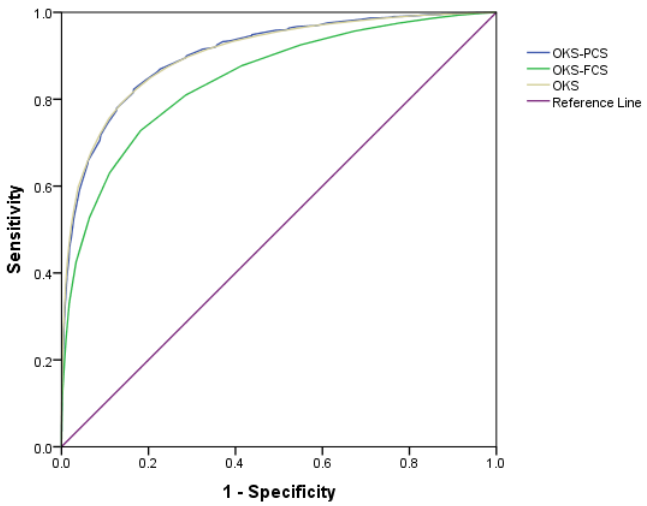


Figure 4.1: ROC curves for the OKS and its subscales (OKS-PCS and OKS-FCS)

between the groups of patients reporting themselves to be "about the same" and "a little worse" was ∼3 points for the OKS, ∼6 points for the OKS-PCS and ∼5 for the OKS-FCS.

Best cut point Table 4.4 summarizes the characteristics of ROC curves for the OKS, OKS-PCs and OKS-FCS. AUCs (with 95% CIs) for all measures show excellent ability to discriminate between patients who improved versus those who deteriorated/stayed the same. The point closest to the top left corner (the best cut point); was 9 points change for the OKS, 25 point for the OKS-PCS, and 7.5 points for the OKS-FCS. In all cases (more for the OKS and the OKS-PCS), this value was close to the MIC.

4.4 Discussion

The purpose of this chapter was to explore the MIC/MIDs for the OKS and its subscales for use in different contexts (such as in clinical trials, cohort studies etc.).

A range of MIC/MID values was produced. The anchor based clinically meaningful MIDs are 5 points for the OKS, 12 points for the OKS-PCs, and 7 points for the OKS-FCS for comparing the difference between groups. MICs of about 9 points for the OKS (out of 48), 25 points for the OKS-PCS (out of 100), and 11 points for the OKS-FCS (out of 100) are useful for assessing the meaningfulness of the change from baseline (e.g. single group design). The MIC values roughly correspond to those generated by the equally clinically meaningful ROC based methods. The MID values generally agreed with the recommended 'rule of thumb' criterion of using 0.5 SD (equivalent to 0.5 of an effect size) [206, 215] often used to set sample sizes in the absence of pilot data. However, without the additional anchor based evidence the use of a distribution based method of 0.5 SD or ES would be less justifiable as, being only statistical distributions, they lack clinical relevance/reference.

Browne et al. [216] produced MID values for the OKS 3.8. However, even though the sample they used had about 500 cases for each score, some of the response option categories had very small sample sizes (e.g. "a little worse" group had n=5), which might explain some discrepancies between the MID values for the OKS.

4.4.1 Study considerations and limitations

Currently, there is a lack of consensus on which method is best for estimating minimal important differences/changes. This lack of consensus is confounded by confusion in terminology and adopted usage [198]. In the recently published paper concerning consensus based standards for the selection of health measurement measures (the COSMIN study), interpretability is highlighted as an important characteristic of any health status measure, but no standards or guidelines were offered for best establishing MIC and MID values [86]. Some guidance, however, has recently been set out by the FDA [85] where the recommendations were that clinical trials should be analysed using *a priori* responder definition. The FDA recommends that the responders to treatment should primarily be defined using an (unspecified) anchor based MID, while values obtained by distribution based methods should only have a 'supportive role'. As there is presently no consensus on the best approach to estimating MICs/MIDs, and methods are constantly evolving, along with the terminology [202, 86, 198] the methods used in this study were chosen on the basis of authors' opinion, albeit grounded in the current relevant literature.

One of the concerns of this study is the low post-operative response rate for both procedures. There were very small, but significant, baseline differences between the groups of patients who did and did not have post-operative OKS (although these values were considerably smaller than the MIDs produced in this study). It has been previously demonstrated that there are baseline differences in patient characteristics and in outcomes between responders and non-responders in patients undergoing

elective hip and knee replacement surgery [217], and that late responders tend to have worse outcome than early responders [218]. However, if there are differences in post-operative outcomes between the responders and non-responders is currently not known [217]. While the problem of missing data and its potential to introduce bias when evaluating the performance of providers (that may have varying response rates), we consider that this was not a source of such concern in our study (which used the overall sample to identify the minimum change detectable by the patient).

4.4.2 Specific considerations related to the choice of anchor or distribution based MIC/MID

With some caveats, the anchor based values can be used for interpretation of data for meaningful changes/differences. The caveats include an appreciation that anchor based methods can generate some idiosyncrasies. For instance, there is a lack of symmetry in the relationship of the OKS change scores with the global transition item. While patients who considered themselves to be "much worse" experienced deterioration of ∼-0.4 ES, and patients who considered themselves to be "much better" experienced very considerable improvement of ∼2.5 ES, the same asymmetry was not observed for the responses closer to "no" or "minimal change". Unexpectedly, patients who considered themselves to be "a little worse" or "about the same" actually experienced small (∼0.2 ES) and medium (∼0.6 ES) levels of improvement respectively. This is even more pronounced for the OKS-PCS, where patients who considered themselves to be much worse actually experienced small benefit (expressed by the ES

statistic).

There are several possible explanations for this lack of symmetry. The lack of congruency between the change measured by the OKS and the change measured by the transition item ratings may suggest a mechanism for how patients perceive improvement and could indicate the presence of recall bias [219]. Patients might have misremembered their hip or knee-related problems as being less severe, prior to surgery, than in fact they were. This is probably more pronounced on the pain subscale, as the pain levels fluctuate and could be more difficult to accurately remember than the (more constant) level of functional impairment. Alternatively, patients might filter their responses on a transition item through their prior expectation for the outcome following surgery, which, by six months, may not be fully realised (and in some cases may have been unrealistic). In this scenario, 'anomalous' transition ratings might represent a proxy measure of disappointment. In the absence of any direct evidence to explain these findings, which would necessitate cognitive interviews, it is only possible to speculate on the explanations behind such discrepancies, which are deserving of further investigation.

Whilst anchor based methods (specifically, the MID) are recommended by the FDA to define responders to treatment in analysis of clinical trials, there might still be merit in using distribution based methods for this purpose. Despite being less clinically linked and more statistically based, the distribution based values were found to be more consistent and not prone to the problems of asymmetry of the anchor methods outlined above. While a distribution based MDC_{90} can say nothing directly about clinical importance, it will provide the lowest value that represents a reliable and real

change in individual patients because the MDC_{90} indicates the measurement error of the change score. If a patient improves or deteriorates beyond the MDC_{90} value of 4 points on the OKS, 9 points on the OKS-PCS or 10 points on the OKS-FCS, it can be stated with a fair amount of certainty that this is not due to random variation in the score [200]. Although, its main value may be in establishing meaningful differences at the level of the individual patient, it could also be used to analyse between group differences i.e. the proportion of patients who have improved/deteriorated beyond measurement error of the measure (responders) can be compared between treatment arms [85].

On a more general note, the fact that that the MIC/MID values are not a fixed attribute of a measure, but are (in common with other attributes of health status measures) dependent on the context, timing of assessment and purpose to which they have been applied should be highlighted [85, 220]. In this study, a patient reported item asking about the perceived change with 5 response categories was used as an anchor. If instead, changes on the OKS had been related to a global item asking about the importance of any change experienced, or about patients' satisfaction with the outcome, different values representing the anchor based MICs/MIDs would very likely have been obtained. This might also have been the case if the same transition item had been asked at 12 months rather than 6 months following surgery.

Likewise, the fact that the anchor can have a varied number of response categories should also be given consideration. It might be argued that had an anchor with 10 responses instead of 5 been used, the MICs/MIDs would likely have been different again.

4.4.3 Study limitations and considerations

This study demonstrated some of the potential issues with anchor and distribution based methods that have been highlighted in the literature. For instance, anchor based methods, when based on the patients' global rating of change, have been criticized for being prone to bias. Various studies have demonstrated that patients' judgements about their past health state can be influenced by their present health state. Response shift and the implicit theory of change are the two most dominant theories that have been proposed to explain this phenomenon [221]. Anchor-based methods have also been criticized because the anchor that is used frequently actually rarely refers to the 'importance' of any change that has been perceived (despite the use of the word 'importance'), where the small amount of change that some individuals perceive might be considered trivial/unimportant by many [199]. As previously reported, distribution based methods have been criticized for only being able to convey whether the observed change is beyond the measurement error of the measure, which says nothing about whether the change is clinically important (or not) [207].

A final point is that the MID was determined based on data derived from a single intervention (viz. a difference between two categories of outcomes - "a little better" and "about the same"). It was concluded in this study that this MID value could then be used to compare outcomes between *different* interventions. Further research is required to determine if this extrapolation is justified.

Each of these issues demonstrate how method dependent the calculation of these values tend to be and there could never be one absolutely correct MIC/MID estimate.

4.4.4 Implications of the findings and future directions

This is the first time that comprehensive estimates of MIC/MID have been reported for the OKS. The findings of this work therefore have potential implications for existing studies and trials that have used different values, particularly to establish sample sizes and set power. The values reported here could be used as a guide for trials and the interpretation of existing studies, but are not necessarily a requirement. Hence trials set up to test for clinical differences of between 3 and 4 points on the OKS (overall score), and identifying such differences, can, in the main, still be considered valid, although the magnitude of the identified effect should be considered. Future trials should however use a clinical difference of at least 5 points to calculate the appropriate sample size. While this will likely result in trials of smaller size, these trials will be easier to interpret. As has been demonstrated in this study, the calculation of MICs and MIDs is far from an exact science and varied opinion on appropriate MIC/MIDs will remain.

The challenges encountered in this study highlight the need for further development and for consensus to be reached on the most appropriate methodology to be used in estimating anchor based MICs/MIDs. As has been demonstrated in this chapter, the calculation of MICs and MIDs is far from an exact science with existing varied opinion on appropriate methods. However, it is likely that methods will evolve further and (hopefully) an international consensus on 'best practice' in calculating MIDs/MICs will be reached. Recommendations presented in this study may therefore change in the future. The challenges encountered in this study highlight

the need for further development and for consensus to be reached, in general, on the most appropriate methodology to be used in estimating anchor based MICs/MIDs.

4.5 Summary

- The choice of MIC or MID, for the interpretation of research studies, including clinical trials, is not straightforward and requires careful consideration.

- For setting sample sizes/assessing group differences; anchor based methods are probably most suitable as they are clinically relevant. The MID value should be used for comparing between group changes (e.g. in clinical trials) and the MIC for single group change from baseline.

- The MDC_{90} can be used when assessing if the health state of individual patients has changed beyond the measurement error of the measure. Therefore, for assessing/interpreting individual scores the MDC_{90} defines the lowest level of difference at which the score retains meaning with a reasonable level of confidence (e.g. anything lower than 4 points for the OKS is within the measurement error of the measure).

- This study derived MIC and MID values using a variety of established methods. Clinically relevant anchor based methods gave consistent values for the MIDs ($\sim$5 points for the OKS, 12 points for the OKS-PCS, and 7 points for the OKS-FCS) and MICs ($\sim$ 9 points for the OKS, 25 points for the OKS-PCS, and 11 points for the OKS-FCS). Distribution based methods showed that the MDC_{90}

for the OKS was 4 points, and 9 and 10 points for the OKS-PCS and OKS-FCS, respectively.

Chapters 3 and 4 assessed specific measurement properties for the OKS when used on the population for which it was originally developed (patients undergoing knee replacement surgery). The next chapter will focus on assessing the measurement properties of the OKS and its pain and functional components when applied to a different population (patients who are undergoing non-operative treatment for knee OA).

Chapter 5

Validation of the OKS for Patients Undergoing Non-Operative Treatment for Knee OA

5.1 Introduction

There appears to be a desire to use the OKS across the spectrum of knee OA disease severity, both in research and clinical practice. As mentioned in the introduction, the OKS is a PROM that was developed for, and validated on, patients who were undergoing knee replacement surgery and it has demonstrated evidence for validity, reliability and responsiveness for assessing outcomes following knee replacement surgery from patients' perspective [162]. More recently, there has been a growing interest in using the OKS for extended purposes, such as in groups of patients who are undergoing non-operative management for earlier stages of knee OA. Indeed, the opportunity to

use PROMs (such as the OKS) in clinical practice (on individual patients) as a means of standardizing clinical assessment, monitoring individuals' self-reported health state and using them as an aid to clinical decision making, is being increasingly highlighted as a great opportunity to achieve maximum benefit from these scores, although it is also considered to be a challenge [3, 222].

These proposed applications go beyond the purpose for which the OKS was originally designed. The measurement properties of a measure reflect the characteristics of the population and the context in which the measure has been developed and tested and these may change when applied in other contexts. End stage OA is usually the indication for knee replacement surgery in the vast majority of cases. The fact that the OKS was developed to be used in the OA context (albeit end stage) is reasonable justification for suggesting its use in people with OA in 'general'. However, it is not known if the OKS remains reliable and valid when used with patients (groups or individuals) who are at earlier stages of their disease management.

The aim of this chapter was to assess the measurement properties of the OKS when used in patients who were undergoing non-operative management for knee OA.

5.2 Patients and Methods

Ethical approval for a prospective cohort study was obtained from a local ethics research committee (11/SC/005). Informed consent was obtained from all participants in the study.

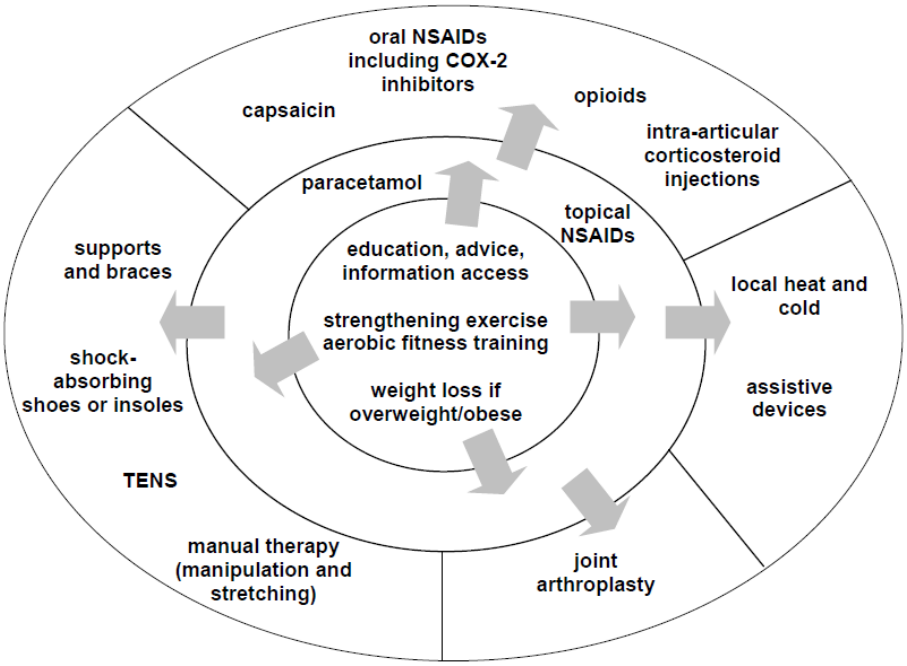


Figure 5.1: A summary of treatments for knee OA as recommended by NICE

5.2.1 Study procedures and assessments

Postal questionnaires were sent to consecutive patients, who had a previously confirmed diagnosis of OA (with varying stages of severity). The patients have been referred to the Nuffield Orthopaedic Centre in Oxford between June 2011 and August 2012, and were not listed for joint replacement surgery. The patients were following an OA management pathway, as recommended by the National Institute of Clinical Excellence (NICE) [223]. As shown in Figure 5.1, the first line of management is the choice of the three core treatments (individually or in combination), followed by treatments in the adjoining sectors. The treatments for patients are tailored individually, taking into account patients' preferences and needs. As such, they represent the standard practice in the NHS.

Patients were sent an invitation letter containing information about the study, consent forms and baseline questionnaires. Patients who consented to participate in the study were asked to complete the OKS [1], the ICOAP [149], the KOOS-PS [158] and the SF12 [224] patient-reported questionnaires[1].

In Chapter 3, evidence was presented, in the context of joint replacement, that supported the original conceptual basis of the OKS and the validity of its composite summary measure, but also offered an option to perform additional analyses using pain and function subscales. OKS Pain Component Score (OKS-PCS) consists of items 2, 3, 7, 11 and 12 and the Functional Component Score (OKS-FCS) consists of items 1, 4, 5, 6, 8, 9 and 10. Raw subscale scores are standardized from 0 (worst) to 100 (best).

Patients were asked to complete the OKS at baseline, 2 and 5 days (for test-retest reliability) and at 3 months. Furthermore, patients were asked to complete the KOOS-PS and ICOAP at baseline and 3 month follow up. The KOOS-PS has 7 items, which consist of 5 Likert-type response categories that are scored from 0 (best) to 4 (worst), with a summary score ranging from 0 (best) to 100 (worst) (following transformation). The ICOAP has 11 items that also consist of 5 Likert-type response categories, scored from 0 (best) to 4 (worst) and with a summary score ranging from 0 (best) to 100 (worst). These scores are developed to measure pain and functional disability related to knee OA and are now a recommended outcome measure by the Osteoarthritis Research Society International (OARSI). Patients also completed the

[1]Non-copyrighted measures (or reproduced with permission) can be found in Appendix C

SF-12, a 12-item general health measure with 8 items that have Likert-type response categories and 4 items with dichotomous (yes/no) response categories. The SF-12 is scored as a Physical Component Summary (PCS) and Mental Component Summary (MCS) ranging from 0 (worst) to 100 (best). Lastly, patients were asked to complete a transition question in regards to the change they experienced from the baseline measurement: "Compared to one week before your clinic visit, please indicate how much your knee problem has changed?" The question had three response options: "1. My knee has got better; 2. My knee has stayed the same; 3. My knee has got worse".

Patient reported outcome data were supplemented with clinical information. The degree of structural changes in the knee was classified using (K-L) grading [225, 226]. An orthopaedic surgeon (LDJ) performed the K-L grading using available knee OA radiographs, In the absence of X-rays, intra-operative documentation from previous knee arthroscopy or available MRI's were assessed to examine the extent of cartilage loss and confirm the diagnosis of osteoarthritis[2].

5.2.2 Statistical methods

The recommended minimum sample sizes for validation studies (based on optimal numbers for correlations) often range from 50 to 100 [79, 182]. For confirmatory factor analysis (CFA) the literature agrees with a minimum sample size of about 100-150 [227] or about 10 subjects per item [228]. These sample sizes are the ones required for data analyses and should be adjusted (i.e. increased) for the risk of loss

[2]The KL grading system (along with the questionnaires used in this study) can be found in the Appendix D

to follow up. In this study the recruitment stopped when the dataset was sufficient to perform the CFA with about 10 subjects per item.

Data analysis was performed using SPSS version 20 and LISREL V 8.80. Baseline and 3 month follow up scores were generally non-normally distributed and change scores approximated to normal (except the ICOAP and the OKS-PCS). Non-parametric statistics were used where appropriate. Psychometric evaluation of the OKS addressed the following:

5.2.2.1 Reliability

Cronbach's alpha was calculated to assess internal consistency of the OKS and its subscales. An intraclass correlation coefficient ($ICC_{2,1}$) [123] was calculated to assess the test-retest reliability of the OKS and its subscales. The precision of individual scores at 90% CI level was calculated by multiplying the SEM by the 2-tailed z value at 90%.

5.2.2.2 Validity

Structural validity was examined by conducting CFA of one and two factor models of the OKS, one facto model of the KOOS-PS, and one and two factor model of the ICOAP using LISREL. The diagonally weighted least squares (DWLS) method was used, based on polychoric correlations and asymptomatic covariance matrix, as the baseline data were non-normal and ordinal [170]. The following indices were considered satisfactory: a non-significant χ^2 ($p>0.05$), SRMR$>$0.08, CFI$\geq$ 0.95, RMSEA$<$0.05 close fit, $<$0.08 good fit, $<$0.1 satisfactory fit, RMSEA p test of close

fit>0.05 [185]. Additionally, χ^2 difference test and Parsimonius Normed Fit Index (PNFI) were used to compare the fit between the two models for the OKS and ICOAP [183]. Chi-square difference tests are calculated looking at the difference of χ^2 of two models along with the difference in their degrees of freedom. The χ^2 difference, with its degrees of freedom is then checked in the χ^2 distribution table. If this value is significant, then the model with more degrees of freedom is favoured.

Construct validity of the OKS and its subscales were examined by testing *a priori* set hypotheses about the expected relationships between the measures at baseline:

(i) the OKS and the physical component summary of the SF12 (PCS-12) are measuring sufficiently similar constructs (PCS-12 measures self-reported physical function and the OKS measures self-reported pain and physical functioning related to the knee), so the correlation between these two measures should be moderate and in the same direction

(ii) the correlation between the OKS and the mental component summary scale of the SF12 (MCS-12) should be weaker than the one between the PCS-12 and OKS as these two constructs are not considered to be related to such an extent

(iii) the OKS and the KOOS-PS are measuring a sufficiently similar construct (the KOOS-PS measures self-reported knee function and the OKS measures self-reported pain and physical functioning related to the knee) that the correlation between these two measures should be strong and negative (as scores go in the opposite direction)

(iv) the OKS and the ICOAP are measuring sufficiently similar constructs (the ICOAP measures self-reported knee pain and the OKS measures self-reported pain and physical functioning related to the knee) that the correlation between these two measures should be strong and negative

(v) the OKS-PCS should be correlated more with the ICOAP than the KOOS-PS and negative, in each case (the OKS-PCS measures self-reported knee pain as does the ICOAP)

(vi) the OKS-FCS should be correlated more with the KOOS-PS than the ICOAP and negative (the OKS-FCS measures self-reported knee function, as does the KOOS-PS). Correlations (r) were classified as: r = 0 to 0.29 as weak; r = 0.30 to 0.69 as moderate; and r $>$ 0.7 as strong.

5.2.2.3 Responsiveness

One sample t-test (2 tailed) was used to assess if the changes at 3 months for the OKS, its subscales (OKS-PCS and OKS-FCS), KOOS-PS and the ICOAP were significantly different from 0. A Cumulative Distribution Function (CDF) plot for the OKS, the OKS-PCS, the OKS-FCS, the ICOAP and the KOOS-PS was constructed to examine the proportion of individual patients who experienced deterioration and improvement beyond the MDC. Paired CDF plots for the (i) OKS-PCS and ICOAP, and (ii) OKS-FCS and KOOS-PS were produced to compare the proportion of change in pain and function between the measures.

As with construct validity, responsiveness was tested by setting *a priori* hypo-

theses about the likely direction and magnitude of changes of the validated comparator measures and the OKS:

(i) the change scores for the OKS should correlate strongly with the change scores in the KOOS-PS and the ICOAP,

(ii) the change scores for the OKS-PCS should correlate more strongly with the change scores in the ICOAP than with the change scores for the KOOS-PS,

(iii) change scores for the OKS-FCS should correlate more strongly with change scores for the KOOS-PS than with the change scores for the ICOAP.

All correlations should be negative. There was a concern about the amount of overall change that can be experienced as a result of such a management pathway (that included a wide range of individually tailored, non-invasive, treatments administered to a heterogeneous sample), so the construct of change was additionally defined using the patient rated item of change. The responses on the item were then used to calculate anchor based values of minimal important change and difference.

5.2.3 Interpretability

Interpretability was assessed by using both anchor and distribution based measures. Average change in the score associated with the group of patients who responded with *better* on the global item of change was taken as the anchor based minimal important change (MIC). The difference in the change score between the groups of patients who responded with *no change* and *better* on the global item of change was taken as the

minimal important difference (MID). Finally, the precision of the individual change score (a change in the score that is beyond measurement error) was estimated by calculating the minimum detectable change (MDC_{90}) [229, 124] formula.

5.3 Results

5.3.1 Sample characteristics

One hundred and thirty-eight patients consented to participate in the study, of whom 67 (50 %) were male, 67 patients were female (information on gender was missing for 4 patients). The mean age of patients was 59 years (SD 11). 69% of the patients had information on Body Mass Index (BMI), out of which 30.5% were classified as obese (BMI>30), 41.1% as overweight (BMI between 25 and 29.9), 28.4% as normal weight (BMI between 18.5 and 24.9) [230]. No one was classified as underweight. All of the patients had a diagnosis of knee osteoarthritis. 2.2% of the patients had K-L grading of 0 (but evidence of cartilage loss on the MRI's), 8% had K-L of 1, 42.8% had K-L of 2, 16.7% had K-L of 3, 3.6% had K-L of 4. In 26.8% of the patients radiological information was missing, out of which in 20% the diagnosis was confirmed on the basis of MRI and 7% of patients did not have X-rays or MRIs accessible (however, these patients had the diagnosis of OA previously confirmed in the primary care setting, different trust, or in private clinics). All patients underwent standard non-operative management of knee OA [231]. The course of management was individually tailored and included exercise, medication, weight loss etc.

Table 5.1: Baseline scores for the OKS, its subscales (OKS-PCS and OKS-FCS), ICOAP, KOOS-PS, and SF-12 physical and mental summaries (PCS-12 and MCS-12).

	N		Mean (SD)	Median	Percentiles	
	Valid	Missing			25	75
OKS	124	14	29.1 (10)	29.5	22	37
OKS-PCS	126	12	57.0 (23)	57	42.8	75
OKS-FCS	130	8	65.5 (22)	65	48.8	85
ICOAP	124	14	37.9 (26)	31.8	16.5	58.5
KOOS-PS	111	27	40.7 (18)	38.6	31.8	51.2
PCS-12	122	16	36.4 (10)	34.9	28.9	44.4
MCS-12	122	16	51 (12)	54.9	42.7	60

104 (75%) out of 138 recruited patients returned the questionnaires at three month follow up. Analysis of patients' characteristics, comparing those who did not respond at three months, versus patients who did complete the questionnaire revealed no significant difference in age, BMI or baseline OKS (t-test, $p<0.05$). A summary of the baseline scores is presented in Table 5.1[3].

5.3.2 Measurement properties

5.3.2.1 Reliability

Cronbach's alpha for the 12-item OKS was 0.94, 0.87 for the OKS-FCS and 0.90 for the OKS-PCS. For the ICOAP Cronbach's alpha was 0.97 and was 0.94 for the KOOS-PS. The alpha value did not change considerably if any of the items were sequentially removed from the total scores.

Test-retest reliability ICCs were 0.93 (95% CI, 0.91-0.95) for the summary OKS, 0.91 (95% CI, 0.88-0.94) for the OKS-PCS and 0.92 (95% CI, 0.90-0.95) for the OKS-

[3]The higher proportion of missing information is perhaps due to the fact that KOOS-PS was calculated from a longer (42 item) KOOS [155] questionnaire.

FCS.

The standard error of measurement (SEM) for the summary OKS was 2.65 and the confidence in individual single score at 90% was ±4.3 OKS points. SEM for OKS-FCS was 5.98 with ±9.87 as 90% CI for individual score, and the SEM for OKS-PCS was 6.66 with ±11 points as 90% CI for individual score (noting that OKS-PCS and OKS-FCS are on a different scale than the OKS). The SEM for ICOAP was 9.9 with ±16 points as 90% CI for individual score. The SEM for ICOAP was calculated using the test-retest reliability reported in the developmental study (0.85) [149]. For KOOS-PS it was not possible to find the test-retest reliability for an English version of the questionnaire, so the value of 0.86 from the validation of the French version of the questionnaire was taken. The SEM for KOOS-PS was 6.8 with ±11 points as 90% CI for individual score.

5.3.2.2 Validity

Structural validity 124 pre-operative OKSs, 124 pre-operative ICOAP and 111 pre-operative KOOS-PS were available for the CFA. Fit indices of one and two factor models for the OKS are presented in Table 5.2. Neither of the one and two factor models were rejected and fit indices of both models were excellent and almost identical. However, the χ^2 statistic favoured the 2 factor model as the reduction in χ^2 in the two factor model was significant (χ^2diff>7.879, with df=1, at the α=0.005 level)[4].

CFA revealed that one-factor KOOS-PS model was rejected by the χ^2 test and its RMSEA was above the highest acceptable threshold of an acceptable fit (0.1) (Table

[4]The significance of the reduction in χ^2 value was obtained from the χ^2 table of proportions

Table 5.2: Fit indices of one and two-factor model of the OKS.

F	1	2
χ^2(p value)	55 (0.45)	48 (0.78)
df	54	53
RMSEA	.01	0
90% CI RMSEA	.0-.06	.0-.04
p test[a]	.89	.98
SRMR	.041	.037
CFI	1	1
PNFI	.8	.79

Note. F=number of factors; χ^2 =chi-square; df=degrees of freedom; RMSEA=root mean square of approximation; CI=confidence intervals; [a]p value for test of close fit (RMSEA<.05); SRMR=standardized root mean square residual; CFI-comparative fit index; PNFI=parsimonious normed fit index.

Table 5.3: Fit indices of the KOOS-PS and one and two-factor model of the ICOAP.

	KOOS-PS	ICOAP (1F)	ICOAP (2F)
χ^2(p value)	42 (<.001)	239 (<.001)	235 (<.001)
df	14	44	43
RMSEA	.14	.19	.19
90% CI RMSEA	.09-.18	.17-.21	.17-.21
p test[a]	0	0	0
SRMR	.047	.064	.057
CFI	.95	.96	.98
PNFI	/	.76	.74

Note. nF=number of factors; χ^2 =chi-square; df=degrees of freedom; RMSEA=root mean square of approximation; CI=confidence intervals; [a]p value for test of close fit (RMSEA<.05); SRMR=standardized root mean square residual; CFI-comparative fit index; PNFI=parsimonious normed fit index.

5.3). SRMR was acceptable and CFI was on the threshold of a good fit. Both one and two factor ICOAP models were rejected by the χ^2 test and both models had RMSEA values far above the lowest threshold of an acceptable fit. However, SRMR and CFI were acceptable for both scores. There was a significant reduction in χ^2 for the 2 factor model of the ICOAP (χ^2diff>, with df=1, at the α=0.05 level).

Construct validity All correlations were generally consistent with *a priori* hypotheses set about the relationships of the OKS with comparator measures. Spearman's

Table 5.4: Baseline Spearman's correlations between the scores. All correlations were significant at the 0.01 level (2-tailed). The number of cases with complete information that allowed the calculation of the correlation coefficients is in brackets for each correlation.

	OKS	OKS-PCS	OKS-FCS
ICOAP	-.883 (115)	-.893 (117)	-.791 (121)
KOOS-PS	-.857 (106)	-.788 (106)	-.865 (110)
PCS-12	.728 (113)	/	/
MCS-12	.381 (113)	/	/

ρ between the baseline OKS, KOOS-PS, ICOAP, MCS-12 and PCS-12 are shown in Table 5.4. The OKS correlated strongly with the KOOS-PS and the ICOAP. The correlation between the PCS-12 and the OKS was slightly higher than expected. As expected, the OKS was most poorly related to the MCS-12. The OKS-PCS correlated more strongly with ICOAP than with KOOS-PS and the OKS-FCS correlated more strongly with the KOOS-PS than with ICOAP. This evidence supports the convergent and divergent validity of the OKS.

5.3.2.3 Responsiveness

Figure 5.2 shows the CDF plot for the OKS. The plot demonstrates that, based on the OKS summary score, approximately 9% of patients in the study experienced deterioration in health state at three months follow up that was beyond the MDC_{90} of 6 points, approximately 26% of patients experienced improvement and 65% of patients did not experience change beyond this value.

Figure 5.3 shows CDF plots of OKS-FCS with KOOS-PS and OKS-PCS with ICOAP. For the OKS-PCS subscale, around 7.5% of patients experienced deterioration, 33% experienced improvement, and 59.5% of patients did not experience change

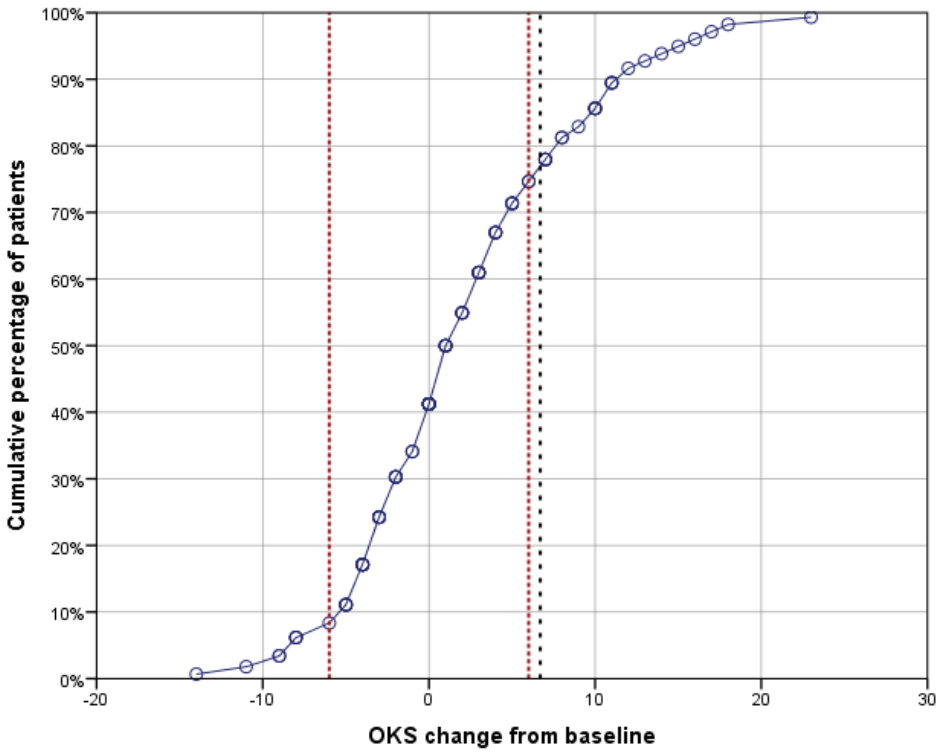


Figure 5.2: Cumulative percentages of patients experiencing the change on the OKS from baseline less or equal to the value on the x-axis. Red line marks the minimum detectable change beyond the measurement error of the score (MDC_{90} of 6 points).

that was beyond the MDC_{90} of 15 points. For the OKS-FCS subscale, around 5% of the patients experienced deterioration, 28% of patients experienced improvement, and 67% of patients did not experience change that was beyond the MDC_{90} of 14 points.

Table 5.5 shows the mean baseline, three month follow-up change scores, p values for the significance of 3 month change and ES for the OKS, OKS-PCS, OKS-FCS, KOOS-PS and ICOAP for the overall cohort. All mean changes were significantly

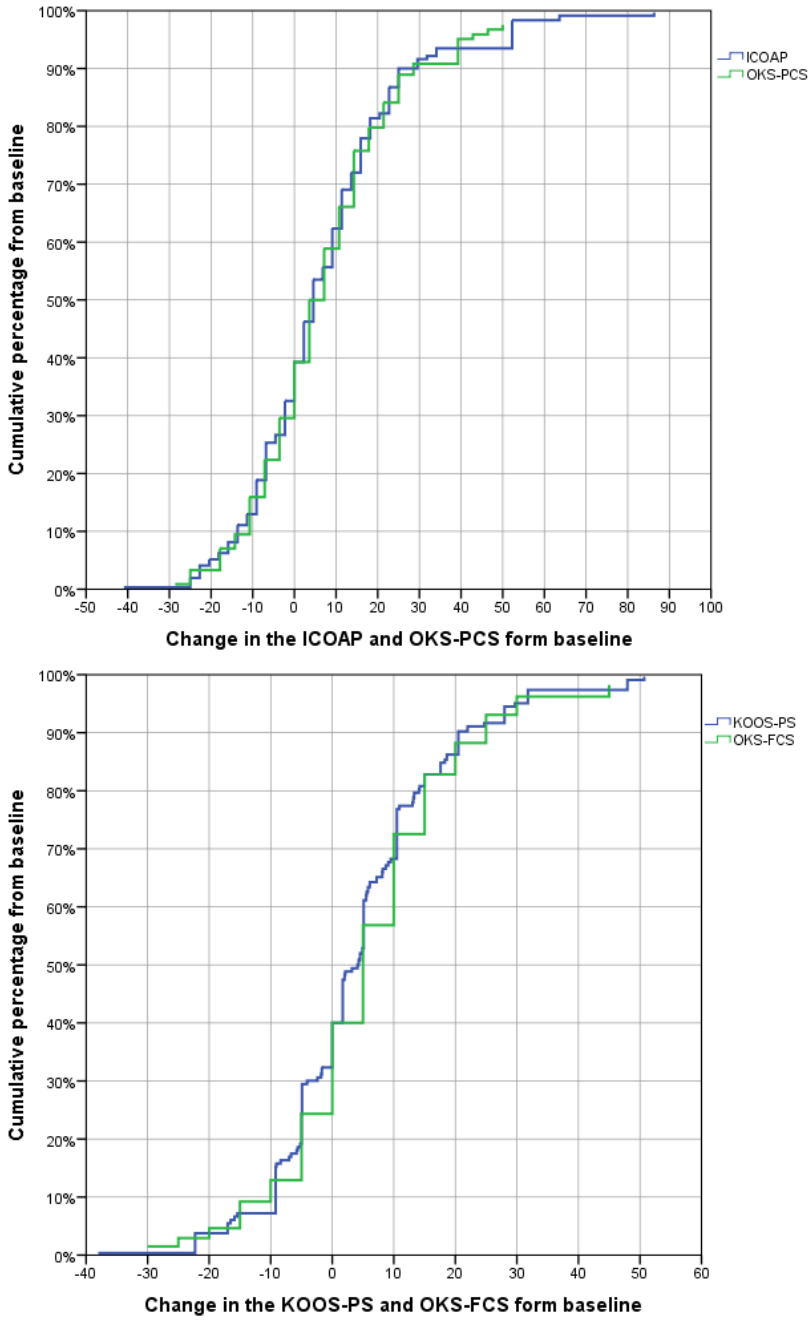


Figure 5.3: Cumulative percentages of patients experiencing the change on the OKS-PCS and ICOAP and the OKS-FCS and KOOS-PS from baseline less or equal to the value on the x-axis. The ICOAP and KOOS-PS have been reversed to enable comparison of the amount of change between the pairs of scores.

Table 5.5: Significance of change in OKS, its subscales (OKS-PCS and OKS-FCS), ICOAP and KOOS-PS scores at three months (one sample t-test).

	N	Baseline (SD)	3 months (SD)	Change (SD)	p-value	ES
OKS	91	30.01 (10)	32.04 (11)	2.03 (7)	<.01	.20
OKS-PCS	94	59.19 (23)	64.78 (24)	5.59 (16)	<.01	.25
OKS-FCS	95	66.00 (22)	68.53 (23)	2.53 (15)	.1	.12
ICOAP[a]	92	37.65 (26)	31.89 (26)	-5.76 (20)	<.01	.22
KOOS-PS[a]	80	40.38 (19)	35.44 (21)	-4.94 (14)	<.01	.27

Note. N=number of complete cases available for calculation of 3 month follow up; SD=standard deviation; ES=effect size;[a]The ICOAP and the KOOS-PS represent severity of the disease in the opposite direction from the OKS and its subscales.

Table 5.6: Spearman's correlations between the 3 month changes in the OKS and its subscales (OKS-PCS and OKS-FCS), ICOAP and KOOS-PS.

	ICOAP	KOOS-PS
OKS	-.665 (84)	-.605 (75)
OKS-PCS	-.654 (87)	-.543 (76)
OKS-FCS	-.585 (88)	-.613 (78)

Note. All correlations are significant at the 0.01 level (2-tailed). The number of cases with complete information that allowed the calculation of the correlation coefficients is in brackets for each correlation.

different from 0 at the 0.01 level (one sample t-test) except the changes measured by OKS-FCS.

As hypothesized, the changes in the OKS at three months were correlated highly with the changes in the KOOS-PS and the ICOAP (0.62 and 0.67 respectively), the changes in the OKS-PCS correlated more with the changes in ICOAP (also assessing knee pain) than KOOS-PS, and the changes in the OKS-FCS correlated more with the changes in KOOS-PS (also assessing knee function) than with the changes in ICOAP (Table 5.5).

Table 5.7: Number (N) and percentage of responses for different response categories with effect sizes (ES), mean score changes by response category and ANOVA tests for linear trend for the OKS and its subscales (OKS-PCS and OKS-FCS).

		Better	Same	Worse
OKS	N (% of responses)	26 (33)	25 (31)	29 (36)
	Mean change (SD)	6.80 (6)	0.6 (7)	-1.72 (5)
	ES	.7	.1	-.2
	P-value for linear trend	<.001	<.001	<.001
OKS-PCS	N (% of responses)	27 (33)	27 (33)	29 (35)
	Mean change (SD)	16.30 (17)	2.65 (14)	-3.32 (16)
	ES	.7	.1	-.2
	P-value for linear trend	<.001	<.001	<.001
OKS-FCS	N (% of responses)	28 (33)	26 (31)	30 (36)
	Mean change (SD)	10.54 (12)	1.35 (16)	-4.83(13)
	ES	.5	.1	-.2
	P-value for linear trend	<.001	<.001	<.001

5.3.2.4 Interpretability

Tables 5.7 and 5.8 present the percentage of responses for different response categories, effect sizes and mean score changes by response category. Independent sample t-tests were conducted for the equality of means between the mean scores for groups of patients who responded *better* and *same* on the transition item. Only the OKS, OKS-PCS, and OKS-FCS had significant differences between the means (2 tailed, $p<0.05$) of groups who responded that they were *better/same*. Table 5.9 presents the summary of interpretability indices.

5.4 Discussion

The aim of this chapter was to examine the measurement properties of the OKS and its postulated subscales, when applied to patients who are undergoing non-operative

Table 5.8: Number (N) and percentage of responses for different response categories with effect sizes (ES), mean score changes by response category and ANOVA tests for linear trend for the ICOAP and the KOOS-PS.

		Better	Same	Worse
ICOAP	N (% of responses)	28 (34)	26 (32)	28 (34)
	Mean change (SD)	-13.56 (23)	-6.29 (17)	4.14 (16)
	ES	-.5	-.3	.2
	P-value for linear trend	.003	.003	.003
KOOS-PS	N (% of responses)	21 (30)	26 (37)	23 (33)
	Mean change (SD)	-12.10 (15)	-4.42 (13)	1.04 (12)
	Effect size	-.7	-.2	.1
	P-value for linear trend	.006	.006	.006

Table 5.9: Anchor based and distribution based MIC/MID values for the OKS, its subscales, ICOAP and KOOS-PS.

	Distribution based	Anchor based	
	MDC_{90}	MID	MIC
OKS	±6.17	6.2	6.8
OKS-PCS	±15.50	13.65	16.30
OKS-FCS	±13.91	9.19	10.54
ICOAP	±23.03	7.27	13.56
KOOS-PS	±15.82	7.68	12.10

Note. MDC_{90}=minimum detectable change; MID=minimum important difference; MIC=minimum important change.

treatment for their knee OA (i.e. in the population it has not been designed for). The OKS, and its OKS-PCS and OKS-FCS subscales were validated against the KOOS-PS, the ICOAP (measures validated for use in patients with knee OA) and the SF-12 (a generic health measure) by testing logical *a priori* hypotheses regarding the construct validity and responsiveness of the OKS and its subscales in comparison to these (validated) measures. The OKS and its subscales were all found to have acceptable measurement properties for their use in patients who are undergoing non-operative management of knee OA. CFA confirmed the fit of the existing one and two-factor models in this population of patients. Furthermore, the test-retest reliability of the OKS and its subscales demonstrated that they can be used both at group and individual levels (clinical practice) [232].

5.4.1 Comparative performance of the OKS and its subscales between different populations/contexts

This is the first study examining in detail the measurement properties of the English version of the OKS in patients at earlier stages of their disease management. The average age of patients in this sample was 59, about 10 years less than in the original validation sample of patient who underwent knee joint replacement. More than 70% of the patients in this sample with X-rays, had earlier stages of knee OA (K-L up to 2); the developmental study of the OKS [1] involved patients who had, mostly advanced knee OA. In addition to the difference in the disease severity based on radiographic evidence, this group of patients was considerably different in terms of self

reported pain and functional disability from the patients undergoing knee replacement surgery[5].

Cronbach's alpha obtained from this study was similar to the alpha on the post-operative sub-sample from the OKS developmental study (0.93) [1]. Test-retest reliability was almost identical for both populations (0.92 for patients undergoing joint replacement surgery and 0.93 in this sample). This chapter and Chapter 3, provided evidence of the structural validity of the OKS and its subscales in both populations of patients[6].

Chapter 4 provided MIC/MID values for the OKS, that can be used in patients undergoing knee replacement surgery. This study provided MIC/MID values for patients who are undergoing a range of individually tailored non-operative treatments that form a 'standard practice' of knee OA management in the NHS (as recommended by NICE). The MIC/MID values are somewhat similar to the values provided in Chapter 4, except the anchor based MIC for the pain subscale. The MIC for the pain subscale is almost twice as large in the population of patients that are undergoing knee replacement surgery, than it is in the patients who are undergoing non surgical management. This could be due to longer recall period in patient undergoing knee replacement surgery (6 months, versus 3 months in this study). Indeed, the lack of symmetry between the patient rated change and the change in the OKS is less pronounced in this group of patients. The existence of extensive surgical intervention (coupled with long rehabilitation) might have influenced patients' memory of pre-

[5]See reported patient characteristics in chapter 3

[6]Even though, as mentioned in Chapter 3, further evidence is needed in regards to construct validity of subscales in the population of patients undergoing knee replacement surgery

operative pain, difference in patient expectations between the two populations, the difference in the wording of the global transition item, and/or the small sample size in this study which was used to generate this value could all have contributed to these differences in anchor-based benchmarks of change in the pain subscale.

5.4.2 Comparative performance of the OKS and its subscales versus the ICOAP and the KOOS-PS in this study

Even though the ICOAP and the KOOS-PS are currently widely used as outcome measures for knee OA, the OKS performed better in this study on several accounts.

The 11-item ICOAP had a Cronbach's alpha of 0.97 (compared to the alpha of the OKS-PCS of 0.9) and the alpha was 0.94 for the KOOS-PS (compared to the alpha of 0.87 for the OKS-FCS). A high alpha value can mean that some of the items on a scale are redundant and this seems to be more of a concern for the ICOAP and KOOS-PS than for the OKS subscales. Furthermore, the precision of the score on the individual level was better for the OKS and its subscales than for the KOOS-PS and ICOAP.

There was evidence to support both one and two factor models of the OKS, but there was no acceptable evidence of structural validity the KOOS-PS or the ICOAP. The KOOS-PS and the ICOAP models were rejected by the χ^2 test and the RMSEAs were unacceptably high for both scales. Other fit indices demonstrated a more acceptable fit (although not as good as on the OKS). For instance, CFI for the ICOAP demonstrated a very good fit. It should be noted that CFI tends to be high

when the correlations between the items in a scale are high (which is particularly the case with the ICOAP) [233]. However, the exploration of the sources of poor fit of these measures is beyond the scope of this thesis and future studies should investigate this problem further (perhaps also using exploratory factor analysis).

CDF plots demonstrated the pairs of scores are overall measuring similar amounts and proportion of change. Furthermore, the results in this study were consistent with the *a priori* hypotheses about the general responsiveness of the scores.

There are some concerns about the interpretability of the scores. It seems that ICOAP and KOOS-PS performed less good than the OKS in this account. First, due to the fact that the ICOAP has low precision of the score at individual level (MDC_{90} is double that of the MIC). This makes it less suitable to interpret change score at the individual level. Second, although one third of the patients reported being better following 3 months of non-operative management for knee OA, neither the ICOAP or the KOOS-PS had statistically significant differences in the change score between the groups of patients who reported themselves to be *better* or *the same* (unlike the OKS and its subscales). This might be due to the fact that the sample sizes in this study were small (about 30 in each group), but could also indicate problems with the sensitivity of the scores to change. Third, while there was some lack of symmetry between the mean change in the OKS score and its subscales in relation to the patient rated item of change, this lack of symmetry seems to be more pronounced for the KOOS-PS and ICOAP.

5.4.3 Limitations and conclusion

There are some limitations in this study that should be noted.

Even though the reliability, construct validity and responsiveness of the OKS and its subscales have been proven to be satisfactory when used in patients undergoing non-operative management for their knee OA, there might be a need to explore further its content validity [92]. The items for the OKS have been originally devised using a representative sample of patients with end stage disease, who are undergoing knee replacement surgery. It may be argued that the measure in its current form does not fully represent the concerns of this extended population of patients who are undergoing non-operative management for knee OA. If a measure is used in a different context or with different types of patients than the context for which interviews occurred in its design/development, then the content validity may be suspect (in relation to the new/different usage) [168]. On the other hand it may be assumed to be appropriate if the context is considered to be "similar enough" [19]. Another argument is that it is unrealistic to have a new/different measure (and the study conducted to design and test one) for every possible sub category of patient or type of treatment within all diseases or conditions. In such cases a researcher should make a judgement about the best available/closest measure, but at least check that the measurement properties still otherwise are maintained. Ideally, a further examination of the content validity of the OKS would necessitate a new study (based on qualitative interviews), that would be used to inform the modification of the instrument (and the subsequent validation of the new version of the instrument). This is beyond the

scope of this thesis (and with questionable value). However, this issue will be partially addressed in the next chapter, where the validity of the OKS and its subscales will be explored further. Rasch analysis will be used to assess (among other factors) if the OKS is well targeted (i.e. is it appropriate to use in patients with this level of functioning and pain). Furthermore, it may be argued that the sample size used to perform the analysis of the interpretability of the scores was somewhat low and may have produced idiosyncratic results.

In conclusion, findings in relation to objectives in this study suggest that the OKS and its subscales can be used to assess the self-reported health state in patients with all stages of knee OA who are undergoing either non-operative management or joint replacement surgery. The OKS subscales can be used to specifically target the improvement or deterioration in pain or function, whether in research (as an endpoint or for sample size calculations) or in clinical practice. Anchor based MIC of $\sim$7 for the OKS, $\sim$16 for the OKS-PCS, and $\sim$11 for the OKS-FCS can be used in cohort studies to assess if the change in the OKS (from baseline) is clinically relevant. Anchor based MID of $\sim$6 for the OKS, $\sim$14 for the OKS-PCS, and $\sim$9 for the OKS-FCS can be used in clinical trials to assess if the difference in change between two arms of treatment is clinically relevant. Finally, changes in individual patient scores beyond the MDC_{90} ($\sim$6 points for the OKS, $\sim$15 points for the OKS-PCS, and $\sim$14 points for the OKS-FCS) can be used as a benchmark of improvement or deterioration that is beyond the measurement error of the score.

5.5 Summary

- The OKS and its subscales demonstrated excellent psychometric properties for application in patients undergoing non-operative management for knee OA (i.e. in a different population than it was designed for).
- This evidence supports the proposed use of the OKS across the spectrum of the OA disease progression both in research and clinical practice.
- This study also identified concerns about the structural validity and interpretability of the ICOAP and KOOS-PS and the need for further exploration of psychometric properties of these measures.

This chapter demonstrated that, when tested following CTT principles the OKS has acceptable evidence supporting the validity, reliability and responsiveness for patients who are at earlier stages of their OA disease progression (as well as in patients who are in the stage immediately prior to joint replacement). In the next chapter, Rasch analysis will be used to further explore validity (specifically, internal validity) of the OKS and its pain and functional component scores.

Chapter 6

Rasch analysis of the OKS

6.1 Introduction

The majority of measures today are continuing to be developed using classical test theory (CTT), however, a number of problems with CTT have been historically put forward [234, 19, 235]. The most commonly highlighted problems are:

(i) Sample dependency: CTT-based properties of a measure are dependent on the characteristics of the developmental sample, and application of a measure to an extended population would ideally require re-examination of its measurement properties

(ii) Item equivalence assumption: ordinal response options and items are usually summed as if they were linear when this, in reality, is not necessarily the case.

To resolve these difficulties, item response theory (IRT) approach can be applied during the scale construction [49, 69].[1]

IRT was originally developed to be used with items that have dichotomous (e.g. yes/no) response options; it has more recently been adapted and applied to items with larger number of ordinal response options. A great majority of IRT models assess a single underlying construct. In IRT, measured constructs are called 'latent traits' (denoted as θ). IRT uses modelling to describe a relationship between an individual person's performance on the test and the underlying latent trait. Both the difficulty (or in health status, severity) of an item on a scale and the performance of each individual can be plotted along the latent trait continuum (θ). The association between the two is described for any given item within a scale with an item characteristic curve (ICC[2]).

6.1.1 One parameter (Rasch) model

The ICCs are based on the simplest, one parameter logistic model:

$$P_i(\theta) = \frac{e^{\theta - b_i}}{1 + e^{\theta - b_i}}$$

Where $P_i(\theta)$ represents a proportion of people, with a given level of trait θ, who answer positively (endorse) the item i, and b_i representing the difficulty of the item i.

Figure 6.1 demonstrates ICCs for a scale with four dichotomous items, where item

[1] Although, the superiority of IRT over the CTT might be theoretical rather than empirical as several studies have found little or no empirical difference between the two frameworks [236, 237, 238].

[2] This is not to be confused with Intraclass Correlation Coefficient ($ICC_{2,1}$), a statistic used for assessing the test-retest reliability, that was used in the other chapters of this thesis.

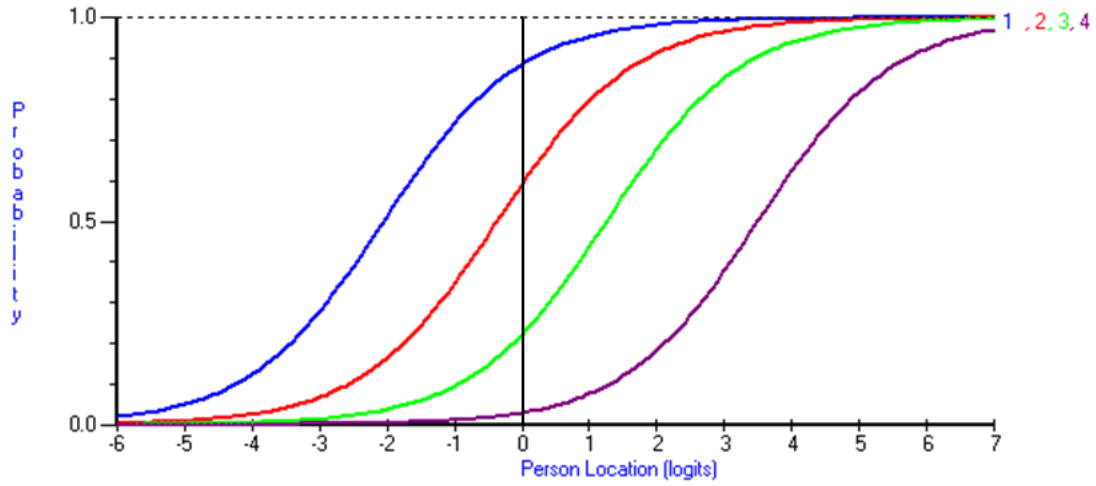


Figure 6.1: Item characteristic curves for four items in a scale, with dichotomous response options and different levels of difficulty.

1 is the easiest and item 4 the most difficult. Latent trait (θ) is plotted on the x-axis and the probability of an individual answering positively on that item is plotted on the y-axis.

If we take an example of a person who has $\theta = 0$, they will have about 90% probability to answer positively on item 1, about 60 % probability to answer positively item 2, about 20% to answer positively item 3, and less than 5% probability to answer positively item 4. If another person is located to the left on the x axis, they would have lesser probability of answering positively on these items (persons with lower severity) and persons on the right side would be more likely to answer positively on the items.

The one parameter model is also called the Rasch model [69] and it forms a part of a family of Rasch models or Rasch Measurement Theory (RMT) (which fundamentally differ from the IRT). In the Rasch model, the ICCs differ from each other only on the difficulty parameter (b). In IRT, there are also derivations of two and three para-

meter models which allow ICCs to account for the difficulties in item discrimination, respondent guessing, and also non-parametric models [234].

One of the fundamental differences between IRT and RMT is that IRT aims to find a model that best explains the given data, wherein RMT the data are fitted to the existing Rasch model. The IRT approach is useful in using the model to explain the data. However, RMT is a recommended method to validate measures by testing the hypothesis of unidimensionality [239]. An excellent overview of the differences in paradigms between the IRT and RMT has been given in a paper by Andrich [78]

As outlined in the introduction, the aim of this chapter is to analyse OKS data using a Rasch measurement model, therefore the remainder of this section will focus on a review of the RMT.

6.1.2 Polytomous Rasch model

If an item has more than two response categories (e.g. ordinal) which indicate increasing order of difficulty/severity, then the extended (polytomous) Rasch model is used. There are two parametrizations of the polytomous Rasch model: the Rating Scale Model (RSM) [240], which assumes equal distances between the response category thresholds or the Partial Credit Model (PCM) [72] which allows the response category threshold distances to vary. A threshold represents a point (boundary) in the single continuum logit scale between the two adjacent response options where the likelihood of response to either of the two responses is 50%. If an item has Likert-type response options, equal distances between thresholds cannot be straightforwardly as-

sumed, and the PCM is often used.

The PCM is essentially a dichotomous Rasch model, but where the model is applied to each pair of successive categories of responses to an item. For example, if an item has five response options (A,B,C,D and E) these would be treated as four dichotomous items within a scale. The five response options on an item would be dichotomised and four thresholds created (e.g. threshold 1=A vs. B, C, D, E, threshold 2=A, B vs. C, D, E, threshold 3=A, B, C vs. D, E etc.). The formula for the Rasch polytomous model is:

$$P_i(\theta) = \frac{e^{\theta - b_{ik}}}{1 + e^{\theta - b_{ik}}}$$

where the newly added subscript k corresponds to the thresholds between the item response options. The ICC graph for one item with five response options would look exactly like the one in Figure 6.1, except that the curves would now represent threshold response curves.

6.1.3 Assumptions of IRT and RMT

Two main assumptions underpin IRT (and RMT): scale unidimensionality and local independence. If a measure is unidimensional, all of its items tap into the same latent trait [99], and once the data are fitted to the Rasch unidimensional model, there should be no further association between items. Furthermore, if a response on one item determines the response to another item (for instance one item asks about ability to walk and the other about ability to run), then there could be some local

(statistical) dependency between the two items. If local dependency is present, this could (among influencing other parameter estimates) result in inflated reliability of the score [241]. Applying the Rasch model in health status measurement might not be always appropriate as some measures will be multidimensional, with items that are clinically relevant but not necessarily satisfying the assumption of the Rasch model [242].

6.2 Methods

For this analysis, data from the validation study described in Chapter 5 was used (obtained from the sample of patients with knee OA, having non surgical management for their knee OA). Baseline OKS, OKS-PCS, and OKS-FCS item values, together with variables containing information on age, gender, and BMI were saved from the original dataset as a FIXED-TEXT format Data file (.dat) using SPSS Statistics V.20 and imported into RUMM 2030 software [243].

The likelihood ratio test was used to determine which version of the polytomous Rasch model should be applied; the (restricted) RSM [71] assumes that the distances between item response thresholds are equal across items, while the (unrestricted) PCM [72] relaxes this assumption. Likelihood ratio test compares two nested models (where one model is a simpler, special case of the more complex model) and assesses how many times more likely the data will fit under one model than the other. If the likelihood ratio test demonstrates that the difference between the two is $p<0.05$, then the Partial Credit Model should be used.

If the scale is well targeted (that is, the items appropriately encompass the level of the ability of the sample), the sample size of 108 would allow item calibrations to be stable within ±0.5 logit, with 99% CIs (for dichotomous items) [244]. For polytomous items, the recommendation is to have at least 10 observations per response category [245].

The following analyses were performed:

Overall fit of the data to the Rasch model RUMM2030 offers three overall fit statistics:

Item-person interactions assess if persons respond to items in an expected manner (following the Guttman pattern[3]) [246]. The degree of differences between the data expected by the model and the observed data are compared. Ideally, the residual means of items or persons will have a mean of 0 and an SD of ±1, which represents a perfect fit.

Item-trait interaction examines the invariance of the hierarchical ordering of the items across a trait, and is expressed as a chi-square statistic(χ^2) [247]. A non-significant statistic is indicative of a good fit.

Additionally, individual items and person statistics are examined. These are also expressed in terms of fit statistics and residuals. Individual item residual fit statistics should not be above or below ±2.5. Values above 2.5 indicate that the item might be measuring a different construct and values below -2.5 indicate that the item might be

[3]Guttman pattern represents ordering of items in a unidimensional scale that follow hierarchical ordering according to their difficulty.

redundant. Also, if the (χ^2) p values are significant after a Bonferroni adjustment[4] [248], it is highly likely that the item is not fitting to the model. The individual (χ^2) values are summed into the item-trait interaction statistic. Individual person fit should also be examined in order to identify any extreme scoring individuals who might influence the overall fit of the items.

Targeting of the measure In Rasch analysis the mean difficulty of all items on the scale is centred on zero logits. Items and persons can be plotted along the continuum according to their difficulty/performance. If the overall scale is well targeted, the mean location (expressed in logits) of persons should be around zero. A positive mean value indicates a ceiling effect (people are generally performing on the higher end of the scale, which might indicate that the measure is too easy) and a negative mean value indicates the presence of the floor effect. Individual targeting of the items should also be examined.

Ordering of thresholds In the case of polytomous items, examination of the category structure is required in order to estimate if the response categories on each item follow the level of trait that is being measured. Threshold categories on an item should be hierarchically ordered to justify their scoring system [249]. If the response categories on an item are confusing (e.g. 'sometime', 'rarely') and might seem interchangeable, the responders might have consistent difficulties in discrimin-

[4]Bonferroni adjustment is a method used in multiple significance testing in order to preserve the error rate to a desired level. For instance, if four statistical tests are being performed and we wish to set the critical level of significance at 5%, then we can conclude than statistical significance was achieved if p was less than 0.0125 (or 0.05/4)

ating between these two options. If disordered thresholds are identified, immediate re-scoring of items by collapsing categories can be used as a remedial action to improve the overall fit to the model. Alternatively, the response categories might need rewording, although this would require a qualitative exploration (e.g. using patient interviews) into the source of the problem, with subsequent item rewording.

Local dependency Examining local dependency can highlight the existence of redundant items in the scale. The association between items should only be through the latent trait that is being measured [49]. If this basic postulate is violated, the results of any RMT analyses will be compromised [250]. Ideally, a response to any of the items on the scale will be independent of the response to any other item on the scale. For instance, two questions asking about different levels of walking ability could be considered to be locally dependent as the answer on one question will be related to the answer to the other question. Presence of local dependency can artificially inflate the reliability of the overall scale. This problem can be solved by combining the two items into one. To identify local dependency, the residual correlation matrix is examined. Correlations between any two item residuals exceeding 0.3 indicate their local dependency.

Differential Item Functioning (DIF) DIF was originally called item bias [251]. It occurs if two subgroups within the same sample, with the same level of the underlying attribute, respond in a different manner to the same item. DIF can be uniform (where one group scores consistently lower or higher) or non-uniform (where the dif-

ference in responses varies across different levels of the latent continuum). Uniform DIF can be remedied by creating separate file groups for each DIF group, and calibrating the item accordingly. Items that have a non-uniform DIF are more difficult to remedy and usually need to be removed from the scale, especially if its removal improves the overall fit of the model. In this study DIF was examined for age and gender using the analysis of variance (ANOVA) test for each item. If the item had a Bonferroni adjusted significance of $<5\%$, the item was examined graphically.

Internal consistency reliability. The Person Separation Index (PSI) [77] is used to estimate how well the scale performs in distinguishing between responders with different levels of a given trait. The PSI is interpreted as the Cronbach's α [43] and is calculated in a similar manner (α uses the raw score, and the PSI uses person estimation in logits). PSI values of minimum 0.7 for group use and 0.85 for individual use are recommended as a rule of thumb [252].

Unidimensionality. After the issues of fit, threshold disordering and DIF are resolved, a test of unidimensionality is conducted. Smith's test was used to test for unidimensionality [253]. Principal Component Analysis (PCA) of the residuals was performed, to identify the items that load positively and negatively on the first factor. Two subsets of items were created (items that loaded positively and items that loaded negatively on the first factor) and compared using independent t-tests, to examine the difference in estimates (for each person). If more than 5% of tests were beyond the range of ± 1.96 (or if the lower confidence interval of the binomial test for the

proportions was above 5%) this was considered indicative of the multidimensionality of the scale [254].

6.3 Results

6.3.1 Sample Characteristics

121 patients had full OKS baseline data out of which 3 (2.48%) patients had extreme scores which were removed before the analysis. The mean age was 59.6 years, and 47.1% patients were men. 86 (71%) patients had information on BMI, out of which 29.1% were classified as normal (BMI between 18.5 and 25), 28.1% were classified as overweight (BMI between 25 and 30) and 31.4% were classified as obese (BMI over 30). The mean OKS summary score was 29.38 (SD 10.2). The mean OKS-PCS was 16.10 (SD 6.4) and the mean OKS-FCS was 13.23 (4.3). Table 6.1 outlines the number of baseline observations per response category. All items except items 6 and 8 had fewer than 10 observations in at least one response category. Items 2, 4 and 11 had less than 10 responses per category, in two categories.

6.3.2 Rasch analysis results

The likelihood ratio test for the OKS and its subscales was significant, so the PCM was used for all analyses.

Table 6.1: Number of observations in the study sample per response category.

	Category response freq.				
	0	1	2	3	4
Q1 (Pain)	2	13	35	50	18
Q2 (Washing)	54	34	24	5	1
Q3 (Transport)	18	36	47	17	0
Q4 (Walking)	45	36	26	6	5
Q5 (Standing)	17	38	39	24	0
Q6 (Limping)	14	41	23	16	24
Q7 (Kneeling)	5	25	43	31	14
Q8 (Night Pain)	22	20	34	25	17
Q9 (Work)	15	36	38	22	7
Q10 (Confidence)	29	47	22	18	2
Q11 (Shopping)	55	27	23	7	6
Q12 (Stairs)	33	35	33	16	1

6.3.2.1 The OKS

The OKS generally had good fit (Table 6.2). Closer examination of individual item fit statistics revealed that all items had probabilities below the Bonferroni adjusted value of 0.004 and residuals within the ±2.5 limit (Table 6.3). Only item 8 ("Have you been troubled by pain from your knee in bed at night?") had a higher fit residual of 3.31. Removal of this item worsened the overall fit of the scale: the overall fit statistic was χ^2 (22) =43.16, p=0.005. The OKS was reasonably well targeted (mean 1.010 ±1.615). Items 4 ("For how long have you been able to walk before pain from your knee becomes severe? [with or without a stick]") and 11 ("Could you do the household shopping on your own?") had disordered thresholds (Figure 6.1). The disordered thresholds were between response categories "5 to 15 minutes" / "around house only" and "around house only/not at all-pain severe when walking" for item 4. For item 11 the disordered thresholds were between the response categories "with moderate

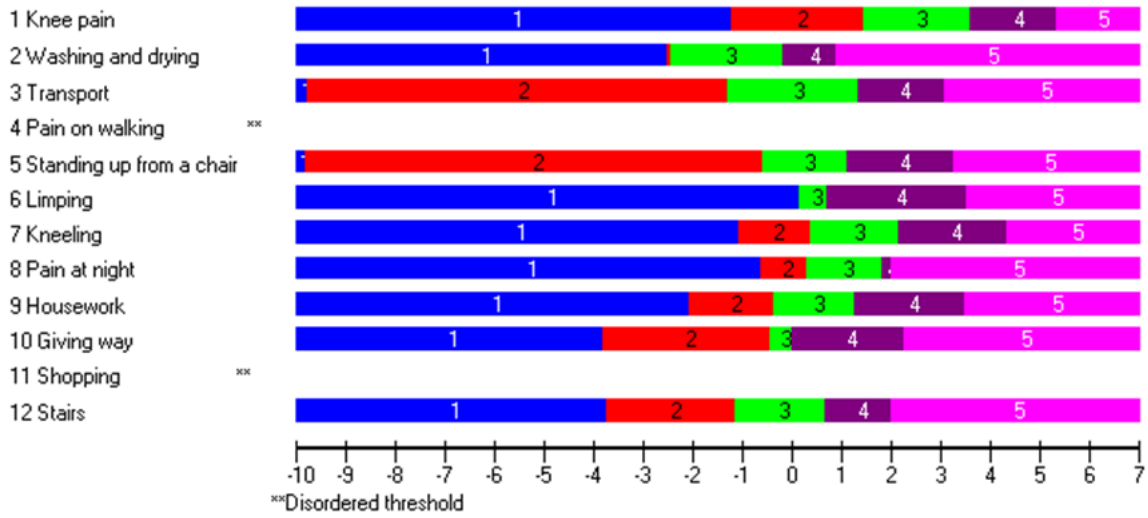


Figure 6.2: Threshold map for the summary OKS.

difficulty" / "with extreme difficulty" and "with extreme difficulty" / "no, impossible"[5]. Due to large standard errors, the differences between disordered threshold estimates were not statistically significant (<0.05).

Collapsing of categories for items 4 and 11 lowered the maximum possible score to 46 and did not improve the overall fit (χ^2 (24) =36.49, p=0.05), so this intervention was not considered worthwhile. Examination of the residual correlation matrix revealed that there was no residual correlation between items that exceeded 0.3. There was no DIF for age, gender or BMI. The PSI was excellent (Table 6.2), allowing the use of the OKS in a clinical setting. The lower bound of binomial confidence interval overlapped 5%, which supported the unidimensionality of the test.

[5]Item characteristic curves can be found in the Appendix E

Table 6.2: Summary fit statistics for the summary OKS and its subscales (OKS-PCS and OKS-FCS).

Scale	OKS	OKS-PCS	OKS-FCS	Ideal values
Number of of items	12	7	5	
Item Residual (SD)	.161 (1.6)	.254 (1.4)	-.022 (1.1)	0 (<1.5)
Person Residual (SD)	-.144 (1.1)	-.200 (1.0)	-.310 (.9)	0 (<1.5)
Chi-square p-value	(p=.06)	(p=.51)	(p=.18)	>.05
PSI	.93	.89	.84	>.85
Coefficient α	.93	.90	.87	
t-test %(CI)	11 (5-17)	5 (1-9)	3 (0-7)	<5

Note. SD=standard deviation; PSI=person separation index; α= Cronbach's alpha; t-test %= percentage of t-tests between the subsets of items that load positively and negatively on the first factor for each patient (Smith's test) that were beyond the range of ±1.96.

6.3.2.2 The OKS-PCS

The OKS-PCS had excellent overall fit (Table 6.2). Individual item fit revealed that (as in the OKS summary scale) item 8 had fit residual of 2.76, which is somewhat out of the ±2.5 range. The measure was very well targeted, with a mean of 0.674 and SD of 1.715 (Figure 6.2). Item 6 ("Have you been limping when walking, because of your knee?") had overlapping thresholds (Figure 6.3). Collapsing of second ("Sometimes, or just at first") and third ("Often, not just at first") response categories resulted in ordered thresholds for this item and improved the overall fit (χ^2 (14) =11.88, p=0.62). Examination of the residual correlation matrix of the original version of the subscale revealed that there were no correlations between items higher than 0.3. In both original and modified version of the subscale, there was no DIF for age, gender or BMI. PSI was above 0.85, confirming the suitability of the subscale's use in clinical practice, as well as for research purposes. There were less than 5% of cases with significant t-test outside the ±1.96 range, confirming the unidimensionality of the

Table 6.3: Individual item fit of the summary OKS.

Item	*During the past 4 weeks.....*	Fit statistics		
		Location	Chi-square	Residual
1	How would you describe the pain you usually have from your knee?	2.301	.309	-1.207
2	Have you had any trouble with washing and drying yourself (all over) because of your knee?	-1.053	.399	.715
3	Have you had any trouble getting in and out of a car or using public transport because of your knee (whichever you would tend to use)	-1.655	.392	-.609
4	For how long have you been able to walk before pain from your knee becomes severe? (with or without a stick)	-.477	.46	.27
5	After a meal (sat at a table), how painful has it been for you to stand up from a chair because of your knee?	-1.505	.032	-1.34
6	Have you been limping when walking, because of your knee?	1.131	.856	.428
7	Could you kneel down and get up again afterwards?	1.466	.03	2.204
8	Have you been troubled by pain from your knee in bed at night?	.878	.034	3.307
9	How much has pain from your knee interfered with your usual work (including housework)?	.583	.395	-1.253
10	Have you felt that your knee might suddenly 'give way' or let you down?	-.498	.561	1.977
11	Could you do the household shopping on your own?	-.599	.306	-1.319
12	Could you walk down one flight of stairs?	-.552	.459	-1.237
	Ideal values		>.0042	within ±2.5

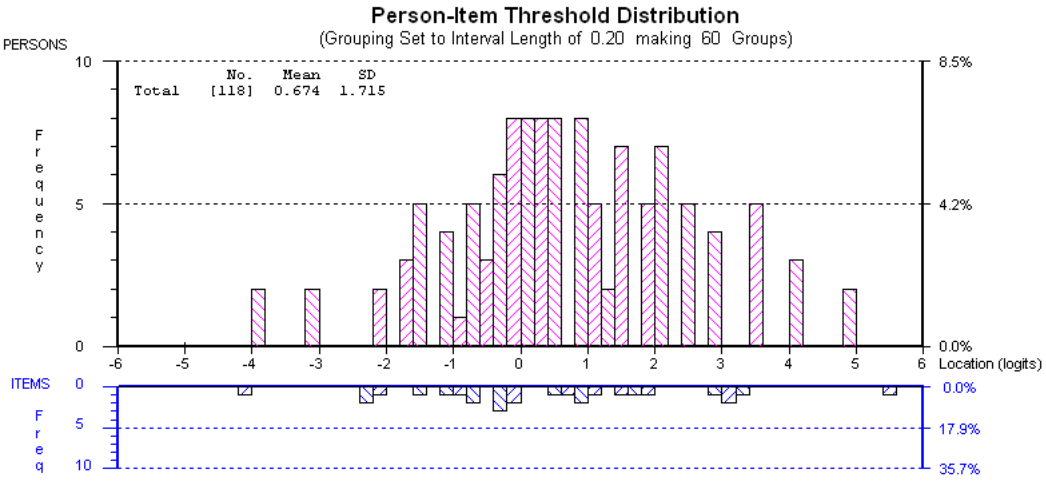


Figure 6.3: Targeting of the OKS-PCS.

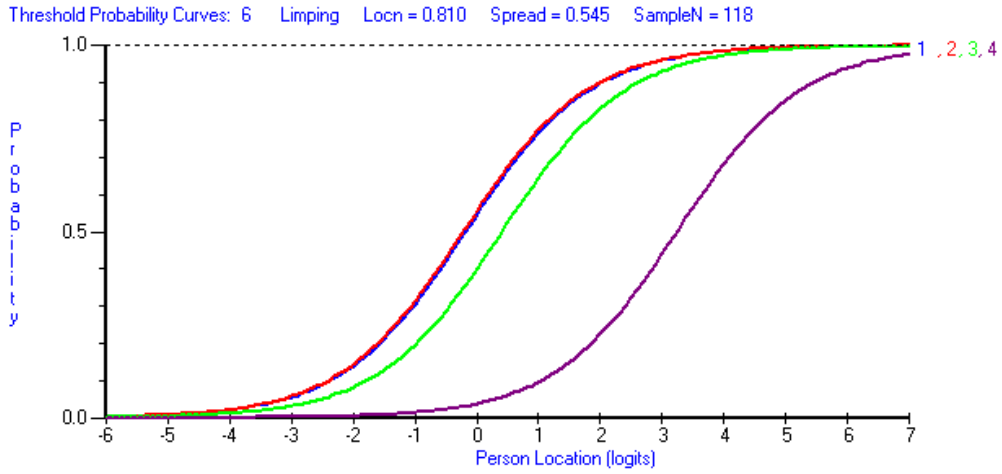


Figure 6.4: Threshold Probability Curves for item 6 on the OKS-PCS.

subscale.

6.3.2.3 The OKS-FCS

The OKS-FCS had a good overall fit (Table 6.2). All items had an excellent fit, with no item fit residuals outside the ±2.5 range. As in the summary scale, item 11 had disordered thresholds. Collapsing of the first two categories resolved the disordering of thresholds but did not improve the overall fit ($\chi^2(10)$ =14.19, p=0.17). Mean

location of persons was 1.551 with an SD of 1.809, suggested that the performance of the sample for this subscale was located on a higher level than the average of the scale. The residual correlation matrix demonstrated that there was no residual correlation between items above 0.3. There was no evidence of DIF for age, gender or BMI. PSI was 0.84, just below the 0.85 threshold for clinical use. A t-test demonstrated that the test had good evidence of unidimensionality.

6.4 Discussion

The RMT (and IRT) approach overcomes some of the major weaknesses of the CTT, and offers a complementary analysis of the scale validity. RMT can be used to provide further evidence of internal construct validity of a measure (specifically unidimensionality), to test the invariance of items, the appropriateness of item response category ordering, and whether or not if the scale is free from bias when measuring different subgroups in a sample (differential item functioning) [252].

The results of this study demonstrate that the OKS and its subscales generally fit the expectations of the PCM. The OKS and its subscales are unidimensional, have a good fit and are free of DIF. The PSI and the reliability of the score are sufficiently high for their use in clinical practice.

The OKS summary scale and the OKS-PCS demonstrated good targeting, but the OKS-FCS subscale indicated the presence of a moderate skew (ceiling effect), which is caused by having a greater number of patients with high functional ability in this sample. However, the PSI was high, which indicates the excellent ability of the scale

to discriminate between patients.

Some threshold disordering was found for both the summary scale and its subscales. In the OKS summary scale, items 4 and 11 had disordered thresholds. Item 11 also exhibited disordered thresholds in the OKS-FCS, and in the OKS-PCS item 4 was not disordered-but now item 6 had disordered thresholds.

Although some threshold disordering was confirmed in earlier studies, the findings in this study lacked sufficient power to justify collapsing of the categories, with subsequent item re-scoring. Previous work [164] has demonstrated threshold disordering on the summary OKS for items 4, 6, 8 and 11. It is possible that patients have difficulties in discriminating between finer increments of response options for these items. However, in order to produce stable a threshold a minimum of 10 responses per category is required [245]. In this study items 4 and 11 did not have more than 10 responses in 2 (adjacent) categories of disordered thresholds. This produced unstable threshold estimates, with large standard errors, that were not significantly different from each other, so re-scoring of the items on the basis of this evidence was not undertaken. The fact that re-scoring did not improve the overall fit of the measures (OKS and OKS-FCS), supports this decision.

Further research (with larger sample sizes) is required before any intervention on the OKS (such as re-scoring) should be considered. The results of Rasch analysis merely point to the problem, and not the source of the problem [78]; so if the threshold disordering turns out to be a consistent feature of the OKS (and its subscales), any remedial actions would have to include extensive examination of the source of the problem. Feedback from patients about distinguishing between successive response

options, with a potential subsequent modification (either by creating new response categories or collapsing the existing ones) would have to be explored.

6.5 Summary

- Rasch analysis (in theory) overcomes some of the problems of sample dependency and item equivalence, problems that have been often highlighted with CTT.

- In this study, the OKS and its subscales demonstrated that they generally met the expectations of the Rasch model when used in patients who are at earlier stages of their disease progression, as opposed to those patients who are severe enough to require TKR. The measures have good patient separation, and are free of DIF for age, gender and BMI. Some threshold disordering was noted on items 4, 6 and 11. However, due to a low frequency of responses per category, these estimates were unstable and no remedial action (such as collapsing categories and re-scoring items) was adopted.

- Further work (with larger sample sizes) is required in order to explore if the threshold disordering of these items is statistically significant.

Chapters 5 and 6 have demonstrated evidence for the validity of the OKS and its subscales when applied in patients who are at earlier stages of their knee OA progression (both on an individual and group level). The next chapter will take this use further and explore if the OKS (and other commonly used indicators of disease severity) can

be used to benchmark patients' perceptions about their disease severity in relation to their perceived need (i.e. thresholds) for joint replacement surgery.

Chapter 7

The Use of Thresholds in Assessing the Need for Knee Replacement Surgery

7.1 Introduction

With a need to deliver up to £20 billion efficiency savings in the NHS over the next few years, the provision of high-cost procedures (such as joint replacement surgery) is under increased scrutiny.

A recent NHS review revealed that there is an unwarranted amount of variation in expenditure on total knee joint replacement across the country which highlighted: (i) professional opinion (bias) regarding need for surgery in some surgeons, and (ii) differences in access to health services (geographic/socio-economic factors), as possible sources of this variation [255]. Deprived areas seem to be associated with lower

surgical rates for TKR, perhaps because of greater pre-operative symptom severity is associated with certain manual occupations (e.g. ones involving lifelong kneeling) [256].

Currently, there are no uniform guidelines in the NHS for who should have (or be prioritised for) knee replacement surgery and decisions are made on a case-by case basis [257]. However, there has been a recent interest in supporting this decision making process by using priority scoring tools that would guide referral to secondary care and facilitate equitable access to joint replacement surgery across the country [258].

7.1.1 The (mis)use of the OKS in allocating access to TKR

In the UK there have been some examples of the use of score thresholds on the OKS for the purpose of standardising access to secondary care and rationing access to TKR. Currently, some UK Primary Care Trusts (PCTs) [12, 14, 13, 15, 16, 17, 18] agree to fund knee replacements only in patients who (among other clinical factors) score less than a pre-specified value on the OKS. Such cut-off scores or thresholds could be useful, but are only meaningful if they are referenced from appropriate and agreed data.

Presently, there is no evidence to support the use of the OKS for the purpose of regulating access to TKR. There is some evidence that pre-operative OKS scores can predict the outcome of joint replacement surgery to some extent (in terms of pain and functional disability) [259]. However, no current universal reference points and agreed

pre-operative thresholds that could be used in predicting post-operative outcomes are in existence [260], nor is there evidence of the validation [261, 119, 262] or of the likely wider impact of such a system on patients and the NHS [263]. Therefore, any threshold systems currently used on the OKS, to guide access to surgery are arbitrary. Furthermore, the use of blanket decision-making thresholds to govern access to surgery works directly against the 2012 Health and Social Care Act that charges the NHS Commission Board with the duty of promoting patient involvement and patient choice in treatment decision-making.

7.1.2 A focus on patient involvement in treatment-decision making

It has been increasingly acknowledged that patient involvement is an essential part of treatment-decision making in orthopaedic surgery [264, 265]. When it comes to the decision to have a TKR (and many other treatment interventions), it is the individual patient, with their personal expectations and preferences in relation to their disease, that should be pivotal in the decision making process [266, 267]. Ideally, the patient who has knee OA (and has by now, usually exhausted conservative management options, received appropriate information about their disease, and considered the risks and benefits of surgery) should make the final decision of whether or not to proceed to surgery.

Research has shown that when patients are given appropriate information about their disease and treatment options, they are less likely to choose the more intensive

treatment [268]. Even though more research is needed to fully evaluate the effect of shared decision making on outcomes in TKR, the patient-centred approach is likely to continue to grow as a preferred healthcare model in the future. This is to ensure that all patients receive "care they need (no less), and care they want (and no more)" [269]. Evidently, designing blanket decision-making criteria to govern access to TKR, while respecting patient individuality, is a challenging objective.

7.1.3 Tackling variation in TKR provision at the GP level

One method to deal with unwarranted variation in surgery provision may involve acting at the level of GP referral. This is because some authors have suggested that the variability in surgery provision can also be caused by differences in GP referral rates [3]. There is also a general support within the GP community to use referral guidelines that incorporate patient preferences [258].

In 2004, the NIHR commissioned the Realistic Effective Facilitation of Elective Referral (REFER) project to improve the referral process to secondary care of patients who might benefit from elective surgical procedures [258]. The study concluded that referral to secondary care in patients with knee OA should be guided both by the severity of the symptoms (knee pain/function/activities of daily living) and patients' preferences about their referral to secondary care. The referral guidelines for knee OA were:

- "Patients should be referred if they have severe knee symptoms and have a strong preference in favour of referral.

- Patients should not be referred if they have mild knee symptoms and have a strong preference against referral or no referral preference either way.
- For all other patient groups, defined according to symptom severity and referral preference, there was no consensus with regard to referral.
- Age, co-morbidity, and body mass should not affect the appropriateness of referral."

However, these referral guidelines are somewhat broad and unclear and and would benefit from further development and validation, before they could be implemented in any form. A standardised measurement of patient's preferences and symptom severity should form the basis of the guideline. One way to do this is to incorporate existing (validated) measures that measure the severity of knee OA symptoms into the referral guideline and complement it with information about patient preferences regarding their treatment.

7.1.4 Measuring the severity of knee OA symptoms

One of the measures that could be used for this purpose (and already is used in a similar fashion) is the OKS. The OKS is a PROM that assesses patients' perceptions of pain and functional impairment arising from their knee [162]. Although it has been developed for patients who are undergoing joint replacement surgery, the work in this thesis (specifically, Chapters 5 and 6) demonstrated that it can also be used in patients (individuals and groups) who are in earlier stages of OA disease progression.

The proposed use of the OKS in this role has some potential challenges.

- The OKS is a standardized measure and not individualized, so will only approximate the concerns of individuals.

- The OKS is originally designed to be preferably used as a change score, where patients act as their own control. Whether 'absolute' scores (or level of score) have some intrinsic meaning (i.e. severe symptoms/mild symptoms), that is in any sense 'stable' or generalizable, has yet to been examined.

- Although the OKS has been used for similar purposes, there might be other knee OA measures that could well be more suitable for this purpose, such as the ICOAP or the KOOS-PS.

The purpose of this study was to examine the relationship between patients' perceived severity of their knee problem (with their absolute scores on the OKS, its pain and function component scores, as well as with other most commonly used PROMs for assessing knee OA: the ICOAP and the KOOS-PS) in order to attempt to give intrinsic meaning to pre-operative absolutes scores.

The aims of this study were to:

(i) explore the relationship between patients' perceptions about the severity of their symptoms (in relationship to their perceived need of knee replacement) and their absolute scores on the OKS, ICOAP and KOOS-PS,

(ii) examine the ability of these measures to distinguish between the group of patients who consider their knee symptoms to be severe enough to warrant knee replacement surgery and the group of patients who consider their symptoms to not be that severe to warrant surgery, and (if satisfactory) to produce threshold values

on the scores that can be used to distinguish between the two groups of patients.

7.2 Methods

7.2.1 Patients/assessments

Consecutive patients who attended the out-patient knee clinics at the Nuffield Orthopaedic Centre from June 2011 to August 2012, who had not been listed for knee replacement surgery, were invited to participate in a study to examine the validity of the OKS for use with patients who are undergoing non-operative treatment for knee OA (details reported in Chapter 5). Patients who consented to participate were sent a questionnaire consisting of the following measures: the OKS, the ICOAP, and the KOOS-PS, together with items that rated: 1) patients' perceived severity of knee OA symptoms and need for, and their perceived need for surgery, along with 2) and their attitude towards surgery.

The item that assessed patients' perceived severity of knee OA asked: "Which of the following statements do you most agree with?: (1) my knee is so bad, that I am desperate to have it replaced, (2) my knee is so bad, that I am beginning to seriously think about having it replaced, (3) while my knee is a problem, I don't think it is bad enough to need replacing, (4) while my knee is currently a problem, I think it may yet improve, (5) while my knee is currently a problem, I am expecting it to get back to normal". According to their responses, patients were divided into two groups: (i) patients who perceived their symptoms to be so severe that they warranted surgery

(responses 1 and 2); group 2: patients with symptoms not severe enough to warrant surgery (responses 3, 4 and 5).

The second item asked: "If your knee was causing you serious problems and a surgeon advised you to have it replaced, would you agree to have knee replacement surgery?" Patients who answered that they would not consider surgery, even in the case of serious problems, were excluded from the analyses as their responses about the need for knee replacement surgery might be influenced by other factors (such as personal attitudes towards having surgery), unrelated to disease severity.

Patient reported outcomes data were supplemented with demographic and clinical data that included information about age, gender, BMI, and radiological grading of knee OA (K-L grading [225][1]).

7.2.2 Statistical methods

Data analysis was performed using R 2.13.2 (The R Foundation for Statistical Computing) with a pROC package [270].

There was less than 5% of missing data for the OKS and ICOAP, 14.5% missing data for the KOOS-PS and 22.6% missing data on radiological grading. Available case analysis (pairwise deletion), without data imputation was used. Details of the sensitivity analysis can be found in the Appendix F.

To compare the differences between the two subgroups, non-parametric statistics were used, as all data were non-normally distributed. Mann-Whitney U tests were used to compare the differences in the OKS, ICOAP, KOOS-PS, age, and BMI

[1] K-L grading scale can be found in Appendix D

between the two subgroups.

ROC analysis [210] was used to assess the ability of the summary OKS, OKS-PCS, OKS-FCS, ICOAP, and KOOS-PS to distinguish between the two groups of patients. Also, ROC analysis on the K-L grading system was performed to assess if the severity of observed structural changes could be used for this purpose. ROC curves provide corresponding sensitivities and specificities for every possible threshold on the measure of choice. Graphical representations of ROC curves demonstrate visually the diagnostic quality of the test to predict the two groups (a perfect test will have a ROC curve that crosses through the top left corner of the graph, which represents 100% sensitivity and 100% specificity). The Area Under the Curve (AUC) was calculated for each measure to assess their predictive ability. AUC's range from 0 to 1, with any value below 0.5 considered worse than random guessing. AUC's >0.7 and <0.8 were considered acceptable and >0.8 were considered excellent [271].

The difference in AUCs between best performing measures was compared using the DeLong method [272] for the measures in the same direction and non-stratified bootstrap method [273] (n=10, 000) for the measures in the opposite direction. The null hypothesis was that the true difference between AUC's was equal to zero.

Threshold values for each measure were extracted based on the desired specificity of 100%. Specificity was defined as the percentage of patients who (from their rating) were correctly identified as not considering their knee symptoms to be severe enough that they needed surgery. Maximum specificity was chosen in order to allow identification of patients who could safely be categorised as not being in need of joint replacement.

If the measure is applied at the level of an individual patient, the measurement error needs to be taken into account [19]. Here, 90% confidence intervals were calculated around the individual scores using the SEM formula: $SD_{baseline}\sqrt{1 - ICC_{2,1}}$ [232].

7.3 Results

Out of 138 patients who consented to participate in the study, 128 responded to the item about the perceived severity of their symptoms at baseline. Out of this number, 124 confirmed that they would have surgery if their health state was bad enough and 4 patients said that they would never have such a procedure. These four patients were excluded from further analyses as it was considered that their responses about the need for surgery could be influenced by other factors, besides the severity of symptoms or impact of the disease on their lives.

Out of 124 patients in the final cohort, 59 (47.6%) were female and 65 (52.4%) were male. The mean age was 59 years (SD 10). Patients in this study were on average 10 years younger than the original OKS validation sample of patients who underwent joint replacement [1]. In 3 patients, information on age was missing. 90 patients had information on their BMI, of whom 32.2% were classified as obese (BMI$\geq$30), 42.2% as overweight (BMI $\geq$ 25, $\leq$29.99), and 25.6% as normal (BMI $\geq$ 18.5 $\leq$ 24.99) [230]. 2.4% of the patients had KL knee radiograph grading of 0, 8.9% had K-L of 1, 41.4% had K-L of 2, 18.5% had K-L of 3, 3.2% had K-L of 4, and in 22.6% of the patients this information was missing. There was no difference in the OKS between the group of

Table 7.1: Summary of baseline scores for the OKS, its subscales (OKS-PCS and OKS-FCS), the ICOAP and the KOOS-PS.

		OKS	OKS-PCS	OKS-FCS	KOOS-PS	ICOAP
Possible range		0*-48	0*-100	0*-100	0-100*	0-100*
N	Valid	115	117	121	106	120
	Missing	9	7	3	18	4
Median		29	57.14	65	38.6	35.23
Percentiles	25	23.00	42.86	47.50	31.30	18.18
	75	37.00	75.00	82.50	48.50	59.09

Note. *=most severe score

patients who had X-rays and the group of patients who did not (independent samples Mann-Whitney U test, sig= 0.108) A summary of baseline PROMs values is presented in Table 7.1.

7.3.1 The relationship between patients' perception about the severity of their symptom and absolute scores

Table 7.2 shows differences in and any statistical significance between the subgroups, using the Mann-Whitney U test (Group 1: Patients who perceived their symptoms to be so severe that they warranted surgery; group 2: Patients with symptoms not severe enough to warrant surgery). All differences were significant, except the difference in respondents' age (p=0.137) or BMI (p=0.691).

Table 7.2: Significance of differences in age, BMI and in the baseline scores for the OKS, its subscales (OKS-PCS and OKS-FCS), KOOS-PS and ICOAP between the two group of patients, assessed using the Mann-Whitney U test (2-tailed).

	Groups	N (%)	Mean Rank	Sum of Ranks	M-W U	sig.
	1	36 (31)	34	1 241		
OKS	2	79 (69)	69	5 429	575	.000
	Total	115				
	1	37 (32)	32	1 185		
OKS-PCS	2	80 (68)	71	5 718	482	.000
	Total	117				
	1	37 (31)	43	1 608		
OKS-FCS	2	84 (69)	69	5 773	905	.001
	Total	121				
	1	30 (28)	70	2 096		
KOOS-PS	2	76 (72)	47	3 575	649	.000
	Total	106				
	1	36 (30)	85	3 077		
ICOAP	2	84 (70)	50	4 183	613	.000
	Total	120				
	1	28 (31)	52	1 445		
BMI	2	62 (69)	43	2 651	698	.137
	Total	90				
	1	37 (31)	59	2 187		
Age	2	84 (69)	62	5 195	1 484	.691
	Total	121				

Note. Group 1: Patients who perceived their symptoms to be so severe that they warranted surgery; group 2: Patients with symptoms not severe enough to warrant surgery; N=number of patients in each group; M-W U=Mann-Whitney U test; sig.=level of significance; BMI=body mass index.

Table 7.3: ROC analysis: Area under the curve (AUC) with 95% confidence intervals (CIs) for the OKS, its subscales (OKS-PCS and OKS-FCS), ICOAP, KOOS-PS, and the Kellgren-Lawrence (K-L) grading system.

	K-L grade	OKS	OKS-PCS	OKS-FCS	ICOAP	KOOS-PS
AUC	.65	.80	.84	.71	.80	.72
(95%CI)	(.54-.76)	(.72-.88)	(.76-.91)	(.62-.80)	(.71-.88)	(.61-.82)

7.3.2 The ability of measures to distinguish between the two groups of patients, and production of threshold values on absolute scores

Figure 7.1 shows the plotted ROC curves and Table 7.3 the corresponding AUCs (and 95% CIs) of the measures. Their discriminatory ability, according to AUC value was classified as excellent (>0.8) for OKS-PCS, and acceptable (>0.7 and <0.8) for all other measures, except for the K-L grading system which performed poorly. The null hypothesis that the true difference in AUCs for the OKS and the ICOAP was equal to zero was not rejected (bootstrap method, D=0.068, p=0.95). The null hypothesis was rejected for the difference in AUCs for the OKS and OKS-PCS (DeLong, z=2.352, p=0.02) and the null hypothesis that the true difference in AUCs for the OKS-PCS and the ICOAP was equal to zero was not rejected (bootstrap method, D=1.51, p=0.13) at the 0.05 level.

Optimal cut-off scores with 100% specificity were calculated from ROC curves. For the OKS this was 40 (out of 48) with 26% sensitivity, for the OKS-PCS it was 82 (out of 100), with 25% sensitivity. For the OKS-FCS it was 95 (out of 100) with corresponding 14% sensitivity, for the KOOS-PS the optimum cut-off score was 85 (out of 100) with 12% sensitivity, and the best cut-off score for the ICOAP was 93

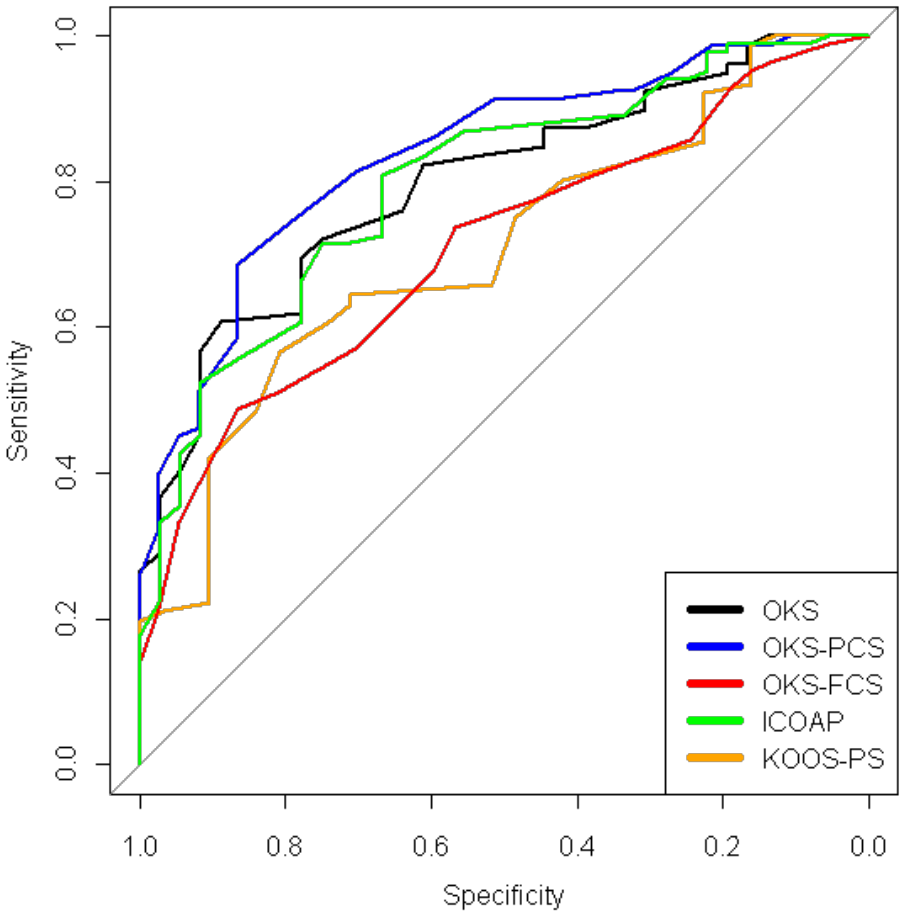


Figure 7.1: ROC plots for the OKS, OKS-PCS, OKS-FCS, ICOAP and KOOS-PS.

Table 7.4: Uncertainty threshold zones for three best performing PROMs (the OKS, its pain subscale and the ICOAP) that take measurement error of the measure at individual level into an account.

	Scaling	Measurement error	Optimum threshold	Uncertainty zone
OKS	0 to 48 (best)	±4	40	36-44
OKS-PCS	0 to 100 (best)	±11	79	68-90
ICOAP	0 to 100 (worst)	±16	9	0-25

(out of 100), with 17% sensitivity.

When applying these measures at the individual level, some methodological issues should be kept in mind, such as the amount of measurement error [232, 19]. Table 7.4 demonstrates the derived thresholds and their relationships to the measurement error of the individual scores for the three best performing measures (OKS, OKS-PCS, and ICOAP). Measurement error values for the individual OKS summary score and OKS-PCS were calculated using the same dataset (as described in Chapter 5). A reported ICC of 0.85 from the ICOAP developmental paper [149] was used and the measurement error was calculated by using the baseline SD from this study (SD was not reported in the developmental study).

When the measurement error was related to the threshold values, the choice of the optimum measure was narrowed. When the measurement error is considered, the ICOAP cannot be used to discriminate between individual patients with certainty as one of the 90% CIs marking the uncertainty zone encompasses 0. For the OKS-PCS, the upper value is 90 and for the OKS it is 44. This is also graphically represented in the Figure 7.2.

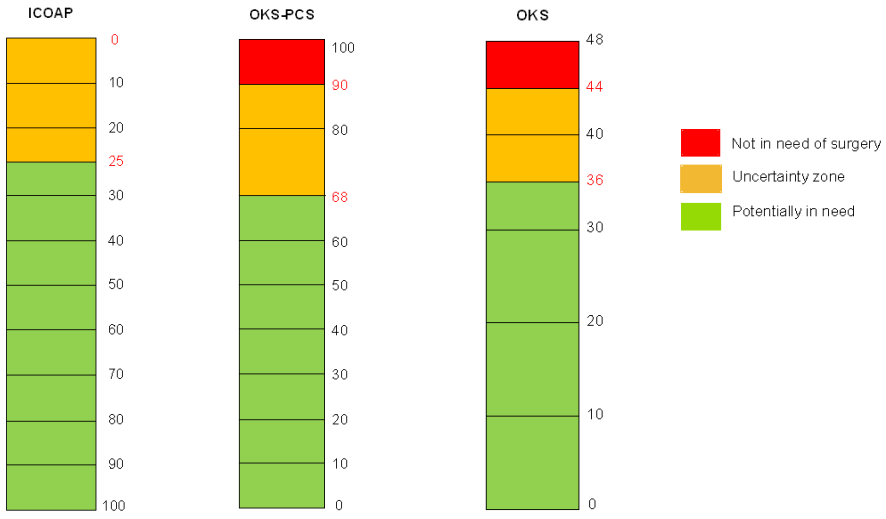


Figure 7.2: Graphical representation of threshold levels for the ICOAP, the OKS-PCS and the OKS

7.4 Discussion

The purpose of this study was to examine the relationship of the OKS, its pain and function component scores, the ICOAP, and the KOOS-PS with patients' perceptions about their severity of symptoms in relationship to the need for joint replacement surgery, and to produce evidence towards the intrinsic meaning of absolute scores. This information could ultimately be used to create clinically meaningful thresholds that can be used as a part of referral guidelines for secondary care in patients with knee OA.

Scores on the OKS, ICOAP and KOOS-PS were related to patients' perceptions about their severity of symptoms. The diagnostic ability for each of the measures to distinguish between patients who considered their symptoms to be severe enough that

they would consider joint replacement surgery or not were acceptable or excellent, with measures that predominantly address pain performing somewhat better (the OKS, OKS-PCS and the ICOAP). The diagnostic ability of the K-L grading was poor, which is understandable as it is generally accepted that the relationship of pain to radiological changes is not that strong (although the evidence for this is limited [274]).

Optimum score thresholds of 40 points for the OKS, 79 points for the OKS-PCS and 9 points for the ICOAP to distinguish between the group of patients not were generated (one group considered themselves in need for joint replacement surgery and the other group considered that their symptoms were not that severe to warrant surgery). However, when the application of thresholds at the individual level was considered, the OKS and the OKS-PCS both demonstrated superiority in terms of measurement precision over the ICOAP. Due to its ease of scoring, the OKS might prove to be a more practical choice for this purpose. For the OKS, the measurement error at the individual level is ± 4 points (as described in Chapter 5). Therefore, patients who score between 36 and 44 on the OKS cannot be easily classified as belonging to one group or the other as this range of scores forms an 'uncertainty zone' where individual cases should be examined in more detail. Having in mind the low sensitivity of the chosen cut-off score, the use of a definite upper threshold of 44 is a better choice as it does not have the risk of missing patients who might be experiencing severe symptoms. This threshold (44) can be used in discriminating between patients who do not currently consider their knee problem to be bad enough to warrant further consideration for surgery versus patients who might currently consider themselves to

be in need of surgery.

This is the only study that has attempted to link patients' perceived severity of symptoms in relation to their preferences about treatment with the OKS, ICOAP or the KOOS-PS. A similar study on thresholds for surgery for the ICOAP and the KOOS-PS was performed in 2011 by the OARSI/OMERACT Task Force. In that study the surgeons' decision (made during outpatient clinical appointments) of whether or not a patient should have a joint replacement was used as a gold standard [275] (surgeons did not have access to the PROMs). The AUCs of the ICOAP and the KOOS-PS were in the region of 0.6, which was insufficient to distinguish between the groups who were/were not considered eligible for surgery. The authors concluded that perhaps patient-specific factors, other than the severity of symptoms (as measured by the ICOAP and/or the KOOS-PS) might play a more important role in surgeons' decisions about patients need for joint replacement. However, the study did not record if all of the patients identified as candidates for surgery by the surgeon actually decided to proceed with the procedure.

Other ways of using PROMs measures to guide access to surgery or secondary care include a limited number of published studies that assess the use of pre-operative OKS thresholds on the post-operative outcome. Judge at al. [260] failed to find acceptable diagnostic ability (and a pre-operative threshold) of the OKS that would predict patient's post-operative satisfaction with the surgery. This lack of a relationship is not very surprising as the post-operative satisfaction is mostly influenced by patients' pre-operative expectations [276]. Dakin et al. [277] assessed if a pre-operative threshold can be set on the OKS to predict patients who will not experience improvement

beyond the NICE cost effectiveness threshold of £20 000-£30 000 Quality Adjusted Life Year gained (QALY[2]) [223]. The authors concluded that knee replacements could be cost effective (£20 000-£30 000/ QALY gained) in some patients who score pre-operatively on the OKS to up to 43.

The pre-operative threshold of 44 on the OKS obtained in this study could be used in guiding access to secondary care; where patients with knee OA who score below 44 are granted access to secondary care. This threshold is also supported by the fact that patients with a pre-operative score of 44 and above cannot experience improvement following surgery beyond the minimum detectable change (MDC_{90}) of 4 points (as described in Chapter 4). Also, this threshold fortuitously agrees with the cost-effectiveness thresholds for joint replacement surgery, set by NICE, at the level of <£20 000–£30 000 QALY [277]. Therefore, even if 'high scoring' patients have indications for, and make an informed decision to undergo, knee replacement surgery, they are likely to experience substantial benefit from the procedure. Evidently, the results of this study (and Dakin et al.) provide further evidence against 'rationing' thresholds currently in use by some PCTs, ranging anywhere from 18 to 32 [12, 14, 13, 15, 16, 17, 18].

The preceding guidelines/findings should be interpreted with caution. Currently the model of what constitutes an acceptable criteria for secondary care referral or candidacy for knee replacement surgery has not yet been operationalized. This study only provides further evidence into the construct validity of the OKS, which can be

[2]QALY is a measure of disease burden that relates both to quality and quantity of life. It is often used in judging the effectiveness of health interventions. For instance, one QALY represents one year of perfect health gained (1 Year of Life x 1 Utility Value=1 QALY) or 2 years at a quality of life of 0.5 (2 Years of Life x 0.5 Utility Value=1 QALY).

used to help in understanding and building the appropriate referral model to be used in decision making aids, that can then be tested (and possibly applied) in the real world. Thresholds provided in this study do not have such normative authority and should not be taken, as they are, as a guide for any clinical decision making as this may not be only misleading, but also unethical.

A conceptual limitation is that the OKS (and other measures used in this study) are standardised and not individualised. The OKS was developed to reflect the concerns of a representative group of patients. When applied on individual patients, the OKS might not encompass the multiplicity of individual patient concerns. An alternative to this problem would theoretically be to use individualised measures, where patients generate their own items and assign weights to them (according to their perceived importance). However, as outlined in Chapter 2, these types of measures have various methodological concerns (such as concerns with reliability and interpretability), especially when in the context of its use on individual level [278]. Also, by using standardised measures it is very difficult to benchmark different individual scores against each other or an external criterion at a single point in time.

Another concern is that the test-retest reliability of the global item of change was not assessed. Recently published International Society for Quality of Life Research (ISOQOL) recommendations for the minimum standard for PROMs state that test-retest reliability of a single item measure should be reported as a minimum[3][87].

A major limitation in this study is that patient recruitment and data collection

[3]This was based on the online survey administered to ISOQOL members where 60% of the respondents considered that the reliability of the single measure should be assessed by the test-retest method (as a minimum standard). 43% of the respondents considered this to be a minimum standard for multi-item unidimensional measures.

were conducted in a single specialised knee clinic, after the patients had already been deemed severe enough to be referred to secondary care. If the data had been collected on a more heterogeneous sample of patients with knee OA, such as in GP practices, the results might have been different. Also, it would be useful if the findings of this study were cross-validated on a larger sample, ideally in multiple centres. Furthermore, additional evidence on the relationship of the pre-operative OKS to post-operative benefit relevant to patients, would be useful to support the potential use of this threshold in referral aids.

7.5 Summary

- There is a need to tackle unwarranted variation in TKR provision across the NHS in England and Wales.
- It has been suggested that the best method to tackle unwarranted variation is to standardize access to secondary care at the GP level, by taking into account symptom severity and patient preferences.
- Symptom severity could be measured using the existing HRQoL measures for knee OA, that would need to be validated for this purpose.
- The results of this study demonstrate that the OKS, ICOAP and KOOS-PS (measures of pain and function in relationship to knee OA) are all related to patients' perceptions about their knee OA symptom severity in relation to the need for joint replacement.

- Based on the diagnostic ability, measurement precision and the ease of use, the OKS outperformed its subscales, the ICOAP, and the KOOS-PS.

- A pre-operative threshold of 44 on the OKS can be used to distinguish between patients who feel that their knee OA is not so bad that they would consider themselves as in need for consideration of TKR and patients who might feel that their symptoms are bad enough that they would consider TKR.

- Further research is needed to confirm the validity of these thresholds at a primary care level, and to explore the relationship of a pre-operative OKS on a post-operative outcome. These findings could ultimately be used in the development of decision making aids for secondary care referral.

Chapter 8

Conclusions, Clinical Relevance, and Future Work

8.1 Review of Research Objectives

PROMs are becoming increasingly important in measuring health status and outcomes of health care interventions from the patient's perspective in research, audit and clinical practice. With this increased endorsement of PROMs, existing measures are occasionally used in new and extended roles, but sometimes this occurs without essential understanding of their validated utilities and limitations.

One example of using established measures in extended roles is the recent uptake of the OKS. The OKS is now being routinely used as a primary outcome measure for joint replacement surgery across the NHS to monitor provider performance [279]. With its increased popularity, further extended uses of the OKS in research and within the clinical community have been noted. There is evidence of the OKS being

used in patients with earlier stages of knee OA (both in a research context [280, 281, 282, 283, 284] and there is anecdotal evidence of its use in clinical practice). Some publications have analysed the OKS by splitting it into pain and functional subscales [5, 6, 7]. The OKS is even being used as a screening tool to guide access to secondary care or to knee arthroplasty [12, 14, 13, 15, 16, 17, 18]. However, the OKS (or any other PROM) should only be used for purposes for which it has been validated, and until recently there was no evidence to justify its use in these extended roles.

The scope of this thesis was to explore the measurement properties of the OKS when used in some of these extended roles. This broadly included investigating: (i) measurement properties of the OKS when used in extended roles on the population it has been developed for (patients undergoing knee replacement surgery), (ii) the measurement properties of the OKS when applied to a new population of patients (patients with knee OA, undergoing non-operative treatment) and when used in extended roles with this population of patients.

Specific research objectives of this thesis were:

1. To explore if it is possible to define/identify pain and functional component subscales from the summary OKS.

2. To determine minimal clinically important change (MIC) and minimal clinically important difference (MID) for the OKS when used in patients who are undergoing knee replacement surgery, that can be used for both research purposes and in clinical practice.

3. To examine the measurement properties of the OKS summary score and its pain

and functional component subscales when used in patients who are undergoing non-operative management for their knee OA, using both principles of classical test theory (CTT) and item response theory (IRT)

4. To explore if it is possible to use the OKS in individual patients to identify which patients might be in need of knee replacement surgery

8.2 Conclusions

A review of literature (Chapter 2) set out the framework of measurement properties that were used to investigate the validity of the OKS in extended roles. Furthermore, this review highlighted the limited systematic approach in published evidence on the validity of knee specific measures.

A factor analytic study (Chapter 3) was used to explore and confirm the structure of the OKS (construct validity). This was done on four independent, large scale, datasets. Factor extraction methods suggested that, depending on the method employed, both one and two factor solutions are meaningful. Overall, and in each data set, some cross loading occurred and item loadings were consistent across two factors. On confirmatory factor analysis, both one and two factor models had acceptable fit indices. This allowed the creation of the 'OKS pain component score' (OKS-PCS) and the 'OKS functional component score' (OKS-FCS). In summary, factor analysis confirmed the original conceptual basis of the OKS but offered an option to perform additional analyses using pain and functional subscales.

The objective of Chapter 4 was to present a range of estimates of meaningful or

minimal important changes/differences for the OKS based on different approaches, and to identify the estimates that are most consistent and useful for application in specific contexts. The results of the study demonstrated that the anchor based methods produced MID values of 5 points for the OKS (and 12 points for the OKS-PCS and 7 points for the OKS-FCS), and MIC of ∼9 points for the OKS (and 25 points for the OKS-PCS and 11 points for the OKS-FCS). Distribution based methods showed that the MDC_{90} for the OKS was 4 points (and 9 and 10 points for the OKS-PCS and OKS-FCS respectively). The results of this study suggest that the MID of 5 points on the OKS (and 12 points for the OKS-PCS and 7 points for the OKS-FCS) can be used to indicate meaningful differences between patient groups and the MIC of 9 points on the OKS (and 25 points for the OKS-PCS and 11 points for the OKS-FCS) to indicate meaningful change from baseline in a single group design undergoing joint replacement. The MDC_{90} of 4 points for the OKS (and 9 for the OKS-PCS and 10 points for the OKS-FCS) can be used to indicate change beyond measurement error on individual patients. Other values may be appropriate in special circumstances.

The aim of the prospective study in Chapter 5 was to examine the measurement properties of the OKS when used with patients who are undergoing non-operative treatment for knee osteoarthritis (in contrast to the original developmental cohort of TKR patients). Results of this study demonstrated that the OKS and its pain and functional component subscales had acceptable internal consistency and test-retest reliability when used for groups of patients with knee OA (research/audit) and for individuals (clinical practice). The results in this study were consistent with *a priori* set hypotheses concerning the relationship of the OKS with other validated measures

(KOOS-PS, ICOAP, and SF12), which provided evidence of construct validity and comparative responsiveness of the measure. CFA confirmed the structural validity of the OKS. The conclusion of this study was that the OKS and its pain and functional component subscales, have excellent measurement properties when used with patients undergoing non-operative management for their knee OA. This evidence provided support for the use of the OKS across the spectrum of OA disease severity, both in research and clinical practice. A major limitation of the study was that the OKS was originally developed using interviews with patients who were undergoing knee replacement surgery and the content validity of the measure might be suspect in this new population of patients. However, even though the KOOS and ICOAP were developed specifically for the population of patients with knee OA, their measurement properties were poor (when compared to the performance of the OKS) in this study.

Analysis of validity of the OKS was further extended in Chapter 6, where the OKS and its subscales were subjected to Rasch analysis. The study demonstrated that the OKS and its subscales generally meet the expectations of the Rasch model, and that its items are performing well across a spectrum of patients that differ in age, gender, and BMI (the measure is of DIF). Some threshold disordering was noted on items 4, 6 and 11. This could imply that patients do not perceive the response options on these items to be in an increasing order of difficulty. Re-scoring of items was not considered as an appropriate outcome as the study had concerns of insufficient sample size. Furthermore, collapsing the categories of disordered thresholds did not improve the fit of the model. It was concluded that there is a need for further research to examine if the threshold disordering encountered in this study is statistically significant.

Having examined and confirmed the suitability of the use of the OKS in patients undergoing non-operative management for knee OA (Chapters 5 and 6), an examination of the potential use of the OKS (and other commonly used knee rating systems) in decision making aids (specifically, to inform secondary care referral) was performed. The aim of the study was to examine the validity of absolute scores, by assessing their diagnostic ability in predicting the patients who considered their symptoms to be so severe that they would consider having joint replacement surgery. The OKS outperformed all measures and a clinically relevant threshold of 44 was argued to be the most appropriate option. This threshold corresponded to the health economic threshold of £20 000-£30 000 per QALY, set by NICE and would allow patients to experience improvement that is beyond the measurement error of the OKS ($MDC_{90}$of 4 points). The study concluded that any patient that scores <44 on the OKS summary score should be granted access to secondary care, if they have such preferences. The results of this study can ultimately be used to inform development of referral guidelines.

8.3 Clinical Relevance

The work presented in this thesis contributed to further understanding of the measurement properties of the OKS, when used in some extended roles beyond those it was originally designed for. The findings in this thesis can be used to inform some extended uses of the OKS in patients with knee OA. This information is likely to be useful to clinicians and stakeholders:

- In addition to the existing use of the OKS in the familiar form of a summary score, it is now also possible to analyse pain and function separately. The OKS was originally developed to be used as a single summary score encompassing patients' perceptions of pain and functional disability related to their knee. It is now possible to use items 1, 4, 5, 6, 8, 9 and 10 to create a subscale assessing pain (OKS-PCS) and items 2, 3, 7, 11, and 12 to create a subscale assessing self-reported function. The subscales can be used to analyse the status and improvement or deterioration in these domains, for research purposes (such as in clinical trials of the analysis of joint registries) or in clinical practice.

- Clinicians can now use the OKS and its subscales in clinical practice to analyse if the change an individual patient has experienced as a result of knee replacement surgery is clinically relevant. A change in score in any direction on the OKS beyond the MDC_{90} (4 points for the OKS, and 9 and 10 points for the OKS-PCS and OKS-FCS respectively) represents a change that is beyond the instrument's measurement error, and an improvement in the OKS beyond 9 points (and 25 points for the OKS-PCS and 11 points for the OKS-FCS) is likely to represent a minimum change in health state that is detectable by patients.

- Validation studies in Chapters 5 and 6 provided evidence that support the use of the OKS and its pain and functional subscales in patients who are at earlier stages of their knee OA progression (than in those who are about to have knee replacement surgery) and are undergoing non-operative treatment. When used with these patients, the OKS has demonstrated evidence of validity, reliability,

and responsiveness in measuring the health state of individuals. The measure can, for instance, be used in clinical practice to monitor disease progression in individual patients undergoing non-operative management for their knee OA, or for hospital audit where the information from groups of patients is analysed to assess the effectiveness of current patient management pathways in terms of health gain/deterioration.

- Lastly, there is some (limited) support for the use of the OKS as a health status measure in primary care to guide referral to secondary care. The study in Chapter 7 demonstrated that the information obtained from the OKS can be linked to patients preferences regarding their treatment. The results demonstrated that no patient who scores less than 44 should be denied access to secondary care (if they have such preferences about their treatment), as ultimately they are likely to experience substantial benefit from such procedures. However, stakeholders should guard against the use of thresholds on the OKS to govern access to knee replacement surgery, as further research is required concerning the predictive validity and prospective impact of such referral guidelines.

8.4 Recommendations for Future Studies

Although this thesis examined in some detail the measurement properties of the OKS when used in extended roles, more work is needed to support its proposed extended usage, and to increase understanding about the potential limits of its use in orthopaedics. Throughout the thesis, some common methodological problems in

the field have been encountered that merit further work in this area. It is hoped that the research and results presented in this thesis will stimulate further interest and research studies on the use of the OKS (and PROMs in general) in orthopaedics.

The review of literature identified a difficulty in obtaining up-to-date information on measurement properties of PROMs; which also complicates the choice for clinicians and researchers when it comes to choosing the best outcome measure for an intended purpose. There has been a substantial growth in the number of new outcome measures in the past 20 years and this is expected to continue to rise in the future [25]. Also, the evidence for instruments' measurement properties usually gets updated over time. For a researcher or a clinician that wants to choose the best outcome measure, keeping track of all the relevant literature sources can be challenging. Furthermore, shifting sands in the criteria of what is currently considered evidence of good measurement properties, coupled with the selective tolerance of 'psychometric language' by some journals (so that not all relevant information is published), make the task of choosing the best measure even more difficult. The consequence of this lack of consensus about the minimum psychometric standards for a measure can be that researchers or clinicians choose a poor outcome measure (for e.g. clinical trials) and that the results obtained from that outcome measure are then used as a base for (poor) decision making (ultimately influencing the life of patients). This has not just scientific, but also ethical implications. A central database of up-to-date evidence of measurement properties (based on consensus criteria) would be of significant help in facilitating high quality research.

Chapter 3 provided evidence about the structural validity of the OKS. The results

of this study allowed generation of pain and functional subscales on the OKS that can be used for patients who are undergoing knee replacement surgery. Further research is needed to explore the construct validity and responsiveness of these subscales.

There is a need to achieve a consensus about the taxonomy of minimal important differences/changes in orthopaedics. In Chapter 4, this taxonomy was created based on the author's literature review and personal knowledge. Published articles reporting on minimal changes/differences use similar approach and, as a result, there is a great heterogeneity (and often confusion) within the community about the appropriate methodology. Unfortunately, COSMIN guidance [86] fell short on giving prescriptive instructions on how to do this. The latest FDA guidance [85] offers good, but somewhat narrow guidance (just for clinical trials) on how to assess minimum important differences. An international Delphi study of experts in the field (such as the COSMIN study) would prove to be valuable in creating a consensus criteria and set standards in the measurement of minimal clinically important changes/differences for various settings and study designs.

The study in Chapter 5 explored the measurement properties of the OKS when used with patients who are undergoing non-operative treatment for their knee OA and found good evidence of its reliability, validity and responsiveness. However, the ICOAP and the KOOS-PS (established measures for patients with knee OA), which were used as comparator instrument for construct validation, did not perform as well, and there were concerns about their structural validity. Further exploration of measurement properties of the ICOAP and KOOS-PS is needed. Consequently, additional evidence supporting the construct validity of the OKS, when applied in

this population of patients it would be useful in confirming the findings in this study.

The study in Chapter 6 explored how the OKS and its subscales fit the expectations of the Partial Credit Rasch model and found possible threshold disordering on similar items as it was previously reported [164]. However, the results are not definite as in this study the frequency per response category was low [245], which resulted in unstable thresholds with large standard errors. Unfortunately, the previous study did not report the frequency of responses per category. If one wishes to perfect the OKS, further research, with sufficient power, is needed to confirm or refute these findings. If there is evidence of consistent threshold disordering, that would also likely necessitate a qualitative exploration to identify the source of this problem. However, the merits of doing this are unclear. The OKS has been in use for many years and is an established measure, translated into numerous languages. To date, its measurement properties have been shown to be 'good enough' (and the results of this thesis reinforce this). However, there can never be a 'perfect' measure and the decision of whether or not to accept some 'flaws' should be based on its potential to be a meaningful and useful measure [285].

Another important line of research that was underlined in this thesis is the increasingly important problem of standardizing referral and access to surgery. The results of the study in Chapter 7 demonstrated some support towards the use of absolute scores on the OKS for measuring patient perceived knee OA symptom severity in relation to patient's need for knee joint replacement. This information could be used to identify likely candidates for joint replacement surgery and ultimately incorporated into the referral guidelines for secondary care. However, even though the

study provided some evidence towards the use of referral decision aids, developing such guidance is a complex task and there is more work that is needed before the use of such systems can be advocated. The sample used in the study (patients who were seen in orthopaedic clinics) was not representative of the setting (patients presenting with knee pain at the GP clinic) of the proposed use of the measure, therefore the results obtained in this study might be misleading. A study conducted on patients in a primary care setting would provide stronger evidence about the validity of absolute OKS scores in this context. Furthermore, the development and validation of referral guidelines calls for extensive research that involves evaluation of not just its validity (including the ability to predict patients who are likely to benefit from the referral) and reliability, but also of its practical feasibility, influence on health outcomes and economic impact over time.

Bibliography

[1] J. Dawson, R. Fitzpatrick, D. Murray, and A. Carr. Questionnaire on the perceptions of patients about total knee replacement. *Journal of Bone and Joint Surgery. British Volume*, 80(1):63–9, 1998.

[2] R. Brooks. EuroQol: the current state of play. *Health policy*, 37(1):53–72, 1996.

[3] N.J. Devlin, J. Appleby, and M. Buxton. Getting the most out of PROMs: putting health outcomes at the hart of NHS decision making. Kings Fund, 2010.

[4] D. Chauhan and J. Sussex, editors. *NHS outcomes, performance and productivity: report of the Office of Health Economics Commission.* Office of Health Economics, 2008.

[5] P.N. Baker, J.H. Van der Meulen, J. Lewsey, and P.J. Gregg. The role of pain and function in determining patient satisfaction after total knee replacement: data from the National Joint Registry for England and Wales. *Journal of Bone and Joint Surgery-British Volume*, 89(7):893, 2007.

[6] W.G. Harcourt, S.H. White, and P. Jones. Specificity of the Oxford knee status

questionnaire. The effect of disease of the hip or lumbar spine on patients' perception of knee disability. *The Journal of bone and joint surgery. British volume*, 83(3):345, 2001.

[7] C. E. Scott, C. R. Howie, D. MacDonald, and L. C. Biant. Predicting dissatisfaction following total knee replacement: a prospective study of 1217 patients. *J Bone Joint Surg Br*, 92(9):1253–8, 2010.

[8] J. Chard, M. Kuczawski, N. Black, and J. Van der Meulen. Outcomes of elective surgery undertaken in independent sector treatment centres and NHS providers in England: audit of patient outcomes in surgery. *BMJ: British Medical Journal*, 343, 2011.

[9] P.N. Baker, D.J. Deehan, D. Lees, S. Jameson, P.J. Avery, P.J. Gregg, and M.R. Reed. The effect of surgical factors on early patient-reported outcome measures (PROMS) following total knee replacement. *Journal of Bone & Joint Surgery, British Volume*, 94(8):1058–1066, 2012.

[10] P.N. Baker, T. Petheram, S.S. Jameson, P.J. Avery, M.R. Reed, P.J. Gregg, and D.J. Deehan. Comparison of patient-reported outcome measures following total and unicondylar knee replacement. *Journal of Bone & Joint Surgery, British Volume*, 94(7):919–927, 2012.

[11] V. Wylde, A.W. Blom, S.L. Whitehouse, A.H. Taylor, G.T. Pattison, and G.C. Bannister. Patient-reported outcomes after total hip and knee replacement: comparison of mid-term results. *Journal of Arthroplasty*, 24(2):210–216, 2009.

[12] NHS-Derbyshire. Derbyshire county prior approval: Procedures of limited clinical value summary.

[13] NHS-Manchester. Topics for Inclusion or Change in the PCT Non-commissioned Procedures List, 2010.

[14] NHS-Gloucestershire. Guidelines for management of orthopaedic conditions, 2010.

[15] Berkshire West & Buckinghamshire (MOBBB) NHS-Milton Keynes O, Berkshire East. Priorities committee. Minutes of meeting held on Wednesday, 15 December 2010.

[16] NHS-Oxfordshire. Interim treatment threshold statement: Knee arthroplasty, 2010.

[17] NHS-Warwickshire. Report to the PCT board 10th November 2010.

[18] NHS-Worcestershire. Commissioning policy for musculoskeletal surgical interventions, 2010.

[19] D.L. Streiner and G.R. Norman. *Health measurement scales: a practical guide to their development and use.* Oxford University Press, 2008.

[20] E.A. Codman. *The shoulder: rupture of the supraspinatus tendon and other lesions in or about the subacromial bursa.* T.Todd Company, 1934.

[21] G.M.A. Grube and others. *Five Dialogues.* Hackett Publishing, 2002.

[22] World Health Organization. Constitution of the World Health Organization, 1948.

[23] D.A. Karnofsky, W.H. Abelmann, L.F. Craver, and J.H. Burchenal. The use of the nitrogen mustards in the palliative treatment of carcinoma. with particular reference to bronchogenic carcinoma. *Cancer*, 1(4):634–656, 1948.

[24] M.L. Slevin, H. Plant, D. Lynch, J. Drinkwater, and WM Gregory. Who should measure quality of life, the doctor or the patient? *British Journal of Cancer*, 57(1):109, 1988.

[25] A. Garratt, L. Schmidt, A. Mackintosh, and R. Fitzpatrick. Quality of life measurement: bibliographic study of patient assessed health outcome measures. *Bmj*, 324(7351):1417, 2002.

[26] F. Nightingale. *Notes on hospitals*. Longman, Green, Longman, Roberts, and Green, 1863.

[27] B.S. Gilson, J.S. Gilson, M. Bergner, R.A. Bobbit, S. Kressel, W.E. Pollard, and M. Vesselago. The sickness impact profile. development of an outcome measure of health care. *American Journal of Public Health*, 65(12):1304–1310, 1975.

[28] R.H. Brook, J.E. Ware, A. Davies-Avery, A.L. Stewart, C.A. Donald, W.H. Rogers, K.N. Williams, and S.A. Johnston. Overview of adult health status measures fielded in rand's health insurance study. *Medical Care*, 17(7):i–131, 1979.

[29] Department of Health. Guidance on the routine collection of Patient Reported Outcome Measures (PROMs), 2008.

[30] The Health and Social Care Information Centre. Provisional monthly patient reported outcome measures (PROMs) in england April 2009 - March 2010: Pre - and postoperative data: Experimental statistics, 2011.

[31] The NHS Information Centre. Publications calendar, 2012.

[32] P.H. DuBois. *A test-dominated society: China, 1115B. C.-1905A. D.* Invitational Conference on Testing Problems, Educational Testing Services, Princeton, New Jersey, 1964.

[33] P.H. DuBois. *A history of psychological testing.* Allyn and Bacon Boston, 1970.

[34] J.S. Wiggins. *Personality and prediction: Principles of personality assessment.* Addison-Wesley Reading, MA, 1973.

[35] J.M.K. Cattell. V.–Mental tests and measurements. *Mind*, 15:373–381, 1890.

[36] K. Pearson. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.

[37] Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

[38] C. Spearman. Correlation calculated from faulty data. *British Journal of Psychology, 1904-1920*, 3(3):271–295, 1910.

[39] W. Brown. Some experimental results in the correlation of mental abilities1. *British Journal of Psychology, 1904-1920*, 3(3):296–322, 1910.

[40] G.F. Kuder and M.W. Richardson. The theory of the estimation of test reliability. *Psychometrika*, 2(3):151–160, 1937.

[41] L. Guttman. A basis for analyzing test-retest reliability. *Psychometrika*, 10(4):255–282, 1945.

[42] L.J. Cronbach. Test 'reliability': Its meaning and determination. *Psychometrika*, 12(1):1–16, 1947.

[43] L. J. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334, 1951.

[44] L.J. Cronbach and P.E. Meehl. Construct validity in psychological tests. *Psychological bulletin*, 52(4):281, 1955.

[45] L.L. Thurstone. *The reliability and validity of tests: Derivation and interpretation of fundamental formulae concerned with reliability and validity of tests and illustrative problems.* Edwards Brothers, 1931.

[46] H. Gulliksen. *Theory of mental tests.* John Wiley & Sons Inc, 1950.

[47] R.C. Tryon. Reliability and behavior domain validity: Reformulation and historical critique. *Psychological Bulletin*, 54(3):229, 1957.

[48] R.B. Cattell. Validity and reliability: A proposed more basic set of concepts. *Journal of Educational Psychology*, 55(1):1, 1964.

[49] F.M. Lord, M.R. Novick, and A. Birnbaum. *Statistical theories of mental test scores*. Addison-Wesley, 1968.

[50] C. Lewis. Test theory and psychometrika: The past twenty-five years. *Psychometrika*, 51(1):11–22, 1986.

[51] L.L. Thurstone. A method of scaling psychological and educational tests. *Journal of Educational Psychology*, 16(7):433, 1925.

[52] L.L. Thurstone. A law of comparative judgment. *Psychological Review; Psychological Review*, 34(4):273, 1927.

[53] J. Guilford. *Psychometric Methods*. McGraw-Hill, 1936.

[54] D.N. Lawley. On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh. Section A. Mathematical and Physical Sciences*, 61(03):273–287, 1943.

[55] L.R. Tucker. Maximum validity of a test with equivalent items. *Psychometrika*, 11(1):1–13, 1946.

[56] S.A. Stouffer, L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Star, and J.A. Clausen. *Measurement and prediction.* Princeton University Press, 1950.

[57] F. Lord. *A theory of test scores (Psychometric Monograph No.7)*. Iowa City, IA: Psychometric Society, 1952.

[58] F.M. Lord. An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika*, 18(1):57–76, 1953.

[59] A. Birnbaum. Efficient design and use of tests of a mental ability for various decision-making problems (Series Report 58-16, Project No. 7755-23). *USAF School of Aviation Medicine, Randolph Air Force Base, Texas*, 1957.

[60] A. Birnbaum. On the estimation of mental ability (Series Report 15, Project No. 7775-23). *USAF School of Aviation Medicine, Randolph Air Force Base, Texas*, 1958.

[61] R.D. Bock and M. Aitkin. Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, 46(4):443–459, 1981.

[62] R. Darrell Bock and M. Lieberman. Fitting a response model forn dichotomously scored items. *Psychometrika*, 35(2):179–197, 1970.

[63] R. Darrell Bock. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37(1):29–51, 1972.

[64] R.D. Bock, W.J. Linden, and RK Hambleton. The nominal categories model. *Handbook of modern item response theory*, pages 33–49, 1997.

[65] F. Samejima. *Estimation of latent ability using a response pattern of graded scores (Psycometric Monograph No.7).* Richmond, VA: Psychometric Society, 1969.

[66] F. Samejima. In *Handbook of modern item response theory*, chapter Graded response model, pages 85–100. Springer, 1997.

[67] D. Thissen and L. Steinberg. A taxonomy of item response models. *Psychometrika*, 51(4):567–577, 1986.

[68] J.H. Albert. Bayesian estimation of normal ogive item response curves using gibbs sampling. *Journal of Educational and Behavioral Statistics*, 17(3):251–269, 1992.

[69] G. Rasch. *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests, Volume 1.* Danmarks paedagogiske Institut, 1960.

[70] G. Rasch. Objektivitet i samfundsvidenskaberne: Et metodeproblem. *Nationaløkonomisk tidsskrift*, 110:161–196, 1972.

[71] D. Andrich. A rating formulation for ordered response categories. *Psychometrika*, 43(4):561–573, 1978.

[72] G.N. Masters. A Rasch model for partial credit scoring. *Psychometrika*, 47(2):149–174, 1982.

[73] E.B. Andersen. Conditional inference for multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology*, 26(1):31–44, 1973.

[74] E.B. Andersen. Sufficient statistics and latent trait models. *Psychometrika*, 42(1):69–81, 1977.

[75] G.H. Fischer. A measurement model for the effect of mass-media. *Acta Psychologica*, 36(3):207–220, 1972.

[76] G.H. Fischer. *Mathematical models for social psychology*, chapter Some probabilistic models for the description of attitudinal and behavioral changes under the influence of mass communication, pages 102–151. Wiley, 1977.

[77] D. Andrich. An extension of the rasch model for ratings providing both location and dispersion parameters. *Psychometrika*, 47(1):105–113, 1982.

[78] D. Andrich. Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care*, 42(1):I–7, 2004.

[79] H.C.W. De Vet, C.B. Terwee, L.B. Mokkink, and D.L. Knol. *Measurement in Medicine: a Practical Guide*. Cambridge University Press, 2011.

[80] S.M. Turner, S.T. DeMers, H.R. Fox, and G. Reed. APA's guidelines for test user qualifications: An executive summary. *American Psychologist*, 56(12):1099, 2001.

[81] M.A.G. Sprangers and N.K. Aaronson. The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease: a review. *Journal of clinical epidemiology*, 45(7):743–760, 1992.

[82] C. Sanders, M. Egger, J. Donovan, D. Tallon, and S. Frankel. Reporting on quality of life in randomised controlled trials: bibliographic study. *BMJ*, 317(7167):1191–1194, 1998.

[83] J.P.T. Higgins and S. Green, editors. *Cochrane handbook for systematic reviews of interventions*, volume 5.1.0. The Cochrane Collaboration, 2011.

[84] R. Fitzpatrick, C. Davey, M.J. Buxton, and D.R. Jones. Evaluating patient-based outcome measures for use in clinical trials. *Health Technology Assessment*, 2(14):1–74, 1998.

[85] US Department of Health, Human Services Food, and Drug Administration. *Guidance for Industry: Patient-Reported Outcome Measures: Use in medical product development to support labeling claims.* 2009.

[86] L.B. Mokkink, C.B. Terwee, D.L. Patrick, J. Alonso, P.W. Stratford, D.L. Knol, L.M. Bouter, and H.C.W. de Vet. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology*, 63(7):737–745, 2010.

[87] B.B. Reeve, K.W. Wyrwich, A.W. Wu, G. Velikova, C.B. Terwee, C.F. Snyder, C. Schwartz, D.A. Revicki, C.M. Moinpour, L.D. McLeod, et al. ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. *Quality of Life Research*, pages 1–17, 2013.

[88] L.B. Mokkink, C.B. Terwee, D.L. Patrick, J. Alonso, P.W. Stratford, D.L. Knol, L.M. Bouter, and H.C.W. de Vet. 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*, 19(4):539–549, 2010.

[89] D.L. Patrick, L.B. Burke, C.J. Gwaltney, N.K. Leidy, M.L. Martin, E. Molsen, and L. Ring. Content validity-Establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO good research practices task force report: part I-Eliciting concepts for a new PRO instrument. *Value in Health*, 14(8):967–977, 2011.

[90] D.L. Patrick, L.B. Burke, C.J. Gwaltney, N.K. Leidy, M.L. Martin, E. Molsen, and L. Ring. Content Validity-Establishing and Reporting the Evidence in Newly Developed Patient-Reported Outcomes (PRO) Instruments for Medical Product Evaluation: ISPOR PRO Good Research Practices Task Force Report: Part II-Assessing Respondent Understanding. *Value in Health*, 14(8):978–988, 2011.

[91] S. Magasi, G. Ryan, D. Revicki, W. Lenderking, R.D. Hays, M. Brod, C. Snyder, M. Boers, and D. Cella. Content validity of patient-reported outcome measures: Perspectives from a PROMIS meeting. *Quality of Life Research*, 21(5):739–746, 2012.

[92] M. Rothman, L. Burke, P. Erickson, N.K. Leidy, D.L. Patrick, and C.D. Petrie. Use of Existing Patient-Reported Outcome (PRO) Instruments and Their Modification: The ISPOR Good Research Practices for Evaluating and Documenting

Content Validity for the Use of Existing Instruments and Their Modification PRO Task Force Report. *Value in Health*, 12(8):1075–1083, 2009.

[93] I.B. Wilson and P.D. Cleary. Linking clinical variables with health-related quality of life. *JAMA: the journal of the American Medical Association*, 273(1):59–65, 1995.

[94] T. Bakas, S.M. McLennon, J.S. Carpenter, J.M. Buelow, J.L. Otte, K.M. Hanna, M.L. Ellett, K.A. Hadler, and J.L. Welch. Systematic review of health-related quality of life models. *Health and Quality of Life Outcomes*, 10(1):134, 2012.

[95] C.E. Ferrans, J.J. Zerwic, J.E. Wilbur, and J.L. Larson. Conceptual model of health-related quality of life. *Journal of Nursing Scholarship*, 37(4):336–342, 2005.

[96] P.M. Fayers, D.J. Hand, K. Bjordal, and M. Groenvold. Causal indicators in quality of life research. *Quality of life research*, 6(5):393–406, 1997.

[97] J.R. Edwards and R.P. Bagozzi. On the nature and direction of relationships between constructs and measures. *Psychological methods*, 5(2):155, 2000.

[98] P.M. Fayers and D.J. Hand. Causal variables, indicator variables and measurement scales: an example from quality of life. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 165(2):233–253, 2002.

[99] J. Hattie. Methodology review: assessing unidimensionality of tests and items. *Applied Psychological Measurement*, 9(2):139–164, 1985.

[100] D.A. Ruta, A.M. , M. Leng, I.T. Russell, and L.M. MacDonald. A new approach to the measurement of quality of life: the Patient-Generated Index. *Medical care*, pages 1109–1126, 1994.

[101] P. Tugwell, C. Bombardier, WW Buchanan, CH Goldsmith, E. Grace, B. Hanna, et al. The MACTAR Patient Preference Disability Questionnaire–an individualized functional priority approach for assessing improvement in physical disability in clinical trials in rheumatoid arthritis. *The Journal of rheumatology*, 14(3):446, 1987.

[102] R. Fitzpatrick, C. Davey, M.J. Buxton, and D.R. Jones. Evaluating patient-based outcome measures for use in clinical trials. *Health Technology Assessment*, 2(14):1–74, 1998.

[103] M.A.G. Sprangers and C.E. Schwartz. Integrating response shift into health-related quality of life research: a theoretical model. *Social science & medicine*, 48(11):1507–1515, 1999.

[104] W. Miles, F. Roberta, W. Lucy, L. Carolyn, and M.M. Marion. Validation of an individualised quality of life measure in older day hospital patients. *Health and Quality of Life Outcomes*, 6, 2008.

[105] C. Macduff and E. Russell. The problem of measuring change in individual health-related quality of life by postal questionnaire: use of the patient-generated index in a disabled population. *Quality of Life Research*, 7(8):761–769, 1998.

[106] K.K. Patel, D.L. Veenstra, and D.L. Patrick. A review of selected patient-generated outcome measures and their application in clinical trials. *Value in Health*, 6(5):595–603, 2003.

[107] D.A. Ruta, A.M. Garratt, and I.T. Russell. Patient centred assessment of quality of life for patients with four common conditions. *Quality in Health Care*, 8(1):22–29, 1999.

[108] C. Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101, 1904.

[109] L. Guttman. A basis for analyzing test-retest reliability. *Psychometrika*, 10(4):255–282, 1945.

[110] L. Guttman. Reliability formulas that do not assume experimental independence. *Psychometrika*, 18(3):225–239, 1953.

[111] W.W. Rozeboom. *Foundations of the theory of prediction*. Dorsey Press Homewood, Illinois, 1966.

[112] D.W. Zimmerman and R.H. Williams. The theory of test validity and correlated errors of measurement. *Journal of Mathematical Psychology*, 16(2):135–152, 1977.

[113] D.W. Zimmerman. Sampling variability and axioms of classical test theory. *Journal of Educational and Behavioral Statistics*, 36(5):586–615, 2011.

[114] L.B. Mokkink, C.B. Terwee, D.L. Patrick, J. Alonso, P.W. Stratford, D.L. Knol,

L.M. Bouter, and H.C.W. De Vet. The COSMIN checklist manual. *Amsterdam: VU University Medical Center*, 2009.

[115] AERA, APA, and NCME. *Standards for educational and psychological testing*. American Educational Research Association, 1999.

[116] N. Mays and C. Pope. Rigour and qualitative research. *BMJ: British Medical Journal*, 311(6997):109, 1995.

[117] D. Child. *The essentials of factor analysis*. Continuum Intl Pub Group, 2006.

[118] M.E. Strauss and G.T. Smith. Construct validity: advances in theory and methodology. *Annual review of clinical psychology*, 5:1, 2009.

[119] D.G. Altman, Y. Vergouwe, P. Royston, and K.G.M. Moons. Prognosis and prognostic research: validating a prognostic model. *Bmj*, 338, 2009.

[120] S.B. Green, R.W. Lissitz, and S.A. Mulaik. Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement*, 37(4):827–838, 1977.

[121] J.M. Cortina. What is coefficient alpha? an examination of theory and applications. *Journal of Applied Psychology*, 78(1):98–104, 1993.

[122] Jill Dawson, Ray Fitzpatrick, David Murray, and Andrew Carr. Comparison of measures to assess outcomes in total hip replacement surgery. *Quality in Health Care*, 5(2):81–88, 1996.

[123] P.E. Shrout and J.L. Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420, 1979.

[124] H.C. De Vet, C.B. Terwee, R.W. Ostelo, H. Beckerman, D.L. Knol, and L.M. Bouter. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. *Health and Quality of Life Outcomes*, 4(1):54, 2006.

[125] J. G. Wright and N. L. Young. A comparison of different indices of responsiveness. *Journal of Clinical Epidemiology*, 50(3):239–46, 1997. Wright, J G Young, N L Comparative Study Research Support, Non-U.S. Gov't England Journal of clinical epidemiology J Clin Epidemiol. 1997 Mar;50(3):239-46.

[126] D.E. Beaton. Understanding the relevance of measured change through studies of responsiveness. *Spine*, 25(24):3192–3199, 2000.

[127] C.B. Terwee, F.W. Dekker, W.M. Wiersinga, M.F. Prummel, and P.M.M. Bossuyt. On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. *Quality of Life Research*, 12(4):349–362, 2003.

[128] J. Cohen. *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum, 1988.

[129] C.A. McHorney and A.R. Tarlov. Individual-patient monitoring in clinical practice: are available health status surveys adequate? *Quality of Life Research*, 4(4):293–307, 1995.

[130] C.B. Terwee, F.W. Dekker, W.M. Wiersinga, M.F. Prummel, and P.M.M. Bossuyt. On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. *Quality of Life Research*, 12(4):349–362, 2003.

[131] C. B. Terwee, S. D. Bot, M. R. de Boer, D. A. van der Windt, D. L. Knol, J. Dekker, L. M. Bouter, and H. C. de Vet. Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology*, 60(1):34–42, 2007.

[132] J. Dawson, H. Doll, R. Fitzpatrick, C. Jenkinson, and A.J. Carr. The routine use of patient reported outcome measures in healthcare settings. *BMJ*, 340, 2010.

[133] D. Jackowski and G. Guyatt. A guide to health measurement. *Clinical Orthopaedics and Related Research*, 413:80–89, 2003.

[134] P.B. Pynsent, J.C.T. Fairbank, and A.J. Carr. *Outcome Measures in Orthopaedics and Orthopaedic Trauma*. Arnold, London, second edition edition, 2004.

[135] J.N. Insall, L.D. Dorr, R.D. Scott, and W.N. Scott. Rationale of the knee society clinical rating system. *Clinical Orthopaedics and Related Research*, 248(1):13–14, 1989.

[136] C.S. Ranawat, J. Insall, and J. Shine. Duo-condylar knee arthroplasty: hospital for special surgery design. *Clinical Orthopaedics and Related Research*, 120:76–82, 1976.

[137] P.A. Dieppe. Recommended methodology for assessing the progression of osteoarthritis of the hip and knee joints. *Osteoarthritis and Cartilage*, 3(2):73–77, 1995.

[138] A.M. Garratt, J.K. Moffett, and A.J. Farrin. Responsiveness of generic and specific measures of health outcome in low back pain. *Spine*, 26(1):71–77, 2001.

[139] D. Wang, M.H. Jones, M.M. Khair, and A. Miniaci. Patient-reported outcome measures for the knee. *Journal of Knee Surgery*, 23(3):137, 2010.

[140] MJ Alviar, J. Olver, C. Brand, J. Tropea, T. Hale, M. Pirpiris, and F. Khan. Do patient-reported outcome measures in hip and knee arthroplasty rehabilitation have robust measurement attributes? A systematic review. *Journal of rehabilitation medicine: official journal of the UEMS European Board of Physical and Rehabilitation Medicine*, 43(7):572, 2011.

[141] T. E. Howe, L.J. Dawson, G. Syme, L. Duncan, and J. Reid. Evaluation of outcome measures for use in clinical practice for adults with musculoskeletal conditions of the knee: A systematic review. *Manual Therapy*, 17(2):100–118, 2012.

[142] Y. Sun, T. Sturmer, KP Gunther, and H. Brenner. Reliability and validity of clinical outcome measurements of osteoarthritis of the hip and knee-a review of the literature. *Clinical rheumatology*, 16(2):185–198, 1997.

[143] R.Y.L. Liow, K. Walker, M.A. Wajid, G. Bedi, and C.M.E. Lennox. The reli-

ability of the American Knee Society Score. *Acta Orthopaedica*, 71(6):603–608, 2000.

[144] A. M. Garratt, S. Brealey, and W. J. Gillespie. Patient-assessed health instruments for the knee: A structured review. *Rheumatology*, 43(11):1414–23, 2004.

[145] S. Smith, S. Cano, D. Lamping, S. Staniszewska, J. Browne, J. Lewsey, J. van der Meulen, J. Cairns, and N. Black. Patient-Reported Outcome Measures (PROMs) for routine use in treatment centres: recommendations based on a review of the scientific evidence. *Final Report to the Department of Health. London: Health Services Research Unit, London School of Hygiene and Tropical Medicine*, 2005.

[146] D.W. Murray, R. Fitzpatrick, K. Rogers, H. Pandit, D.J. Beard, A.J. Carr, and J. Dawson. The use of the Oxford hip and knee scores. *Journal of Bone and Joint Surgery-British Volume*, 89(8):1010, 2007.

[147] The NJR Centre. 8th Annual Report National Joint Registry for England and Wales. Technical report, 2011.

[148] N. Bellamy, W.W. Buchanan, C.H. Goldsmith, J. Campbell, and LW Stitt. Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. *The Journal of Rheumatology*, 15(12):1833, 1988.

[149] GA Hawker, AM Davis, MR French, J. Cibere, JM Jordan, L. March, M. Suarez-Almazor, JN Katz, and P. Dieppe. Development and preliminary psychometric testing of a new OA pain measure-an OARSI/OMERACT initiative. *Osteoarthritis and Cartilage*, 16(4):409–414, 2008.

[150] AM Davis, LS Lohmander, R Wong, V Venkataramanan, and GA Hawker. Evaluating the responsiveness of the icoap following hip or knee replacement. *Osteoarthritis and Cartilage*, 18(8):1043–1045, 2010.

[151] G Hawker, C Melfi, J Paul, R Green, C Bombardier, et al. Comparison of a generic (sf-36) and a disease specific (womac)(western ontario and mcmaster universities osteoarthritis index) instrument in the measurement of outcomes after knee replacement surgery. *The Journal of rheumatology*, 22(6):1193, 1995.

[152] A Escobar, JM Quintana, A Bilbao, I Arostegui, I Lafuente, and I Vidaurreta. Responsiveness and clinically important differences for the womac and sf-36 after total knee replacement. *Osteoarthritis and cartilage*, 15(3):273–280, 2007.

[153] S. McConnell, P. Kolopack, and A.M. Davis. The Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC): a review of its utility and measurement properties. *Arthritis Care & Research*, 45(5):453–461, 2001.

[154] N. Bellamy and WW Buchanan. A preliminary evaluation of the dimensionality and clinical importance of pain and disability in osteoarthritis of the hip and knee. *Clinical rheumatology*, 5(2):231–241, 1986.

[155] E. M. Roos, H. P. Roos, L. S. Lohmander, C. Ekdahl, and B. D. Beynnon.

Knee Injury and Osteoarthritis Outcome Score (KOOS)-development of a self-administered outcome measure. *Journal of Orthopaedic and Sports Physical Therapy*, 28(2):88–96, 1998.

[156] E. Roos. KOOS 2012 User's Guide 1.1, December 2012.

[157] M. Dougados, G. Hawker, S. Lohmander, A. M. Davis, P. Dieppe, J. F. Maillefert, and L. Gossec. OARSI/OMERACT criteria of being considered a candidate for total joint replacement in knee/hip osteoarthritis as an endpoint in clinical trials evaluating potential disease modifying osteoarthritic drugs. *Journal of Rheumatology*, 36(9):2097–9, 2009.

[158] A. V. Perruccio, L. Stefan Lohmander, M. Canizares, A. Tennant, G. A. Hawker, P. G. Conaghan, E. M. Roos, J. M. Jordan, J. F. Maillefert, M. Dougados, and A. M. Davis. The development of a short measure of physical function for knee OA KOOS-Physical Function Shortform (KOOS-PS) - an OARSI/OMERACT initiative. *Osteoarthritis and Cartilage*, 16(5):542–50, 2008.

[159] GA Hawker, L. Stewart, MR French, J. Cibere, JM Jordan, L. March, M. Suarez-Almazor, and R. Gooberman-Hill. Understanding the pain experience in hip and knee osteoarthritis–an OARSI/OMERACT initiative. *Osteoarthritis and Cartilage*, 16(4):415–422, 2008.

[160] R. Gooberman-Hill, G. Woolhead, F. MacKichan, S. Ayis, S. Williams, and P. Dieppe. Assessing chronic joint pain: lessons from a focus group study. *Arthritis Care & Research*, 57(4):666–671, 2007.

[161] P.H. Heuts, J.W. Vlaeyen, J. Roelofs, R.A. de Bie, K. Aretz, C. van Weel, and O.C.P. van Schayck. Pain-related fear and daily functioning in patients with osteoarthritis. *Pain*, 110(1-2):228–235, 2004.

[162] J. Dawson, M. Fitzpatrick, D. Churchman, A. Verjee-Lorenz, and D. Claysonm. *User Manual for the Oxford Knee Score (OKS).* Isis Outcomes, 2010.

[163] D. W. Murray, R. Fitzpatrick, K. Rogers, H. Pandit, D. J. Beard, A. J. Carr, and J. Dawson. The use of the Oxford hip and knee scores. *Journal of Bone and Joint Surgery. British Volume*, 89-B(8):1010–1014, 2007.

[164] P.G. Conaghan, M. Emerton, and A. Tennant. Internal construct validity of the Oxford Knee Scale: evidence from Rasch measurement. *Arthritis Care and Research*, 57(8):1363–1367, 2007.

[165] F.J. Floyd and K.F. Widaman. Factor analysis in the development and refinement of clinical assessment instruments. *Psychological assessment*, 7(3):286, 1995.

[166] B.G. Tabachnick, L.S. Fidell, and S.J. Osterlind. *Using multivariate statistics.* Allyn and Bacon Boston, 2001.

[167] K.A. Bollen. *Structural equations with latent variables.* Johne Wiley and Sons Inc.

[168] J. C. Nunnally and I. H. Bernstein. *Psychometric theory.* McGraw-Hill, New York, 1994.

[169] U. Olsson. Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44(4):443–460, 1979.

[170] D.B. Flora and P.J. Curran. An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological methods*, 9(4):466, 2004.

[171] H.F. Kaiser. The application of electronic computers to factor analysis. *Educational and psychological measurement*, 1960.

[172] W.R. Zwick and W.F. Velicer. Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, 99(3):432, 1986.

[173] R.D. Ledesma and P. Valero-Mora. Determining the number of factors to retain in EFA: an easy-to-use computer program for carrying out parallel analysis. *Practical Assessment, Research & Evaluation*, 12(2):1–11, 2007.

[174] R.B. Cattell. The scree test for the number of factors. *Multivariate behavioral research*, 1(2):245–276, 1966.

[175] D.L. Streinier. Factors affecting reliability of interpretations of scree plots. *Psychological reports*, 83(2):687–694, 1998.

[176] W.F. Velicer. Determining the number of components from the matrix of partial correlations. *Psychometrika*, 41(3):321–327, 1976.

[177] B.P. O'Connor. SPSS and SAS programs for determining the number of com-

ponents using parallel analysis and Velicer's MAP test. *Behavior Research Methods*, 32(3):396–402, 2000.

[178] J.L. Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185, 1965.

[179] W.R. Zwick and W.F. Velicer. Comparison of five rules for determining the number of components to retain. *Psychological bulletin*, 99(3):432, 1986.

[180] M.E. Timmerman and U. Lorenzo-Seva. Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological methods*, 16(2):209, 2011.

[181] R.B. Cattell. *The scientific use of factor analysis in behavioral and life sciences*. Plenum Press New York, 1978.

[182] P. Kline. *An easy guide to factor analysis*. Routledge, 1993.

[183] R.E. Schumacker and R.G. Lomax. *A beginner's guide to structural equation modeling: Second Edition*. Lawrence Erlbaum Associates, 2004.

[184] B. Byrne. *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Lawrence Erlbaum Associates, 1998.

[185] M.W. Browne and R. Cudeck. Alternative ways of assessing model fit. *Testing structural equation models*, 154:136–162, 1993.

[186] G. Maruyama. *Basics of structural equation modeling*. Sage Publications, Inc, 1997.

[187] U. Lorenzo-Seva and P.J. Ferrando. Factor: A computer program to fit the exploratory factor analysis model. *Behavior research methods*, 38(1):88–91, 2006.

[188] The Health and Social Care Information Centre. Annual publication data quality notes library. HESonline web site, November, 2011.

[189] G.R. Norman and D.L. Streiner. *Biostatistics: the bare essentials*. Pmph USA Ltd, 2008.

[190] F.J. Floyd and K.F. Widaman. Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, 7(3):286, 1995.

[191] L.R. Fabrigar, D.T. Wegener, R.C. MacCallum, and E.J. Strahan. Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods*, 4(3):272, 1999.

[192] J. Dawson, R. Fitzpatrick, A. Carr, and D. Murray. Questionnaire on the perceptions of patients about total hip replacement. *Journal of Bone and Joint Surgery. British Volume*, 78(2):185–90, 1996.

[193] A.B. Costello and J.W. Osborne. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research and Evaluation*, 10, 2005.

[194] K. Boersma and S.J. Linton. How does persistent pain develop? An analysis

of the relationship between psychological variables, pain and function across stages of chronicity. *Behaviour Research and therapy*, 43(11):1495–1507, 2005.

[195] J. Dekker, G.M. van Dijk, and C. Veenhof. Risk factors for functional decline in osteoarthritis of the hip or knee. *Current opinion in rheumatology*, 21(5):520, 2009.

[196] R.B. Kline. *Principles and practice of structural equation modeling*. The Guilford Press, 2010.

[197] N. Cliff. Some cautions concerning the application of causal modeling methods. *Multivariate behavioral research*, 18(1):115–126, 1983.

[198] M.T. King. A point of minimal important difference (MID): a critique of terminology and methods. *Expert review of pharmacoeconomics & outcomes research*, 11(2):171–184, 2011.

[199] C.B. Terwee, L.D. Roorda, J. Dekker, S.M. Bierma-Zeinstra, G. Peat, K.P. Jordan, P. Croft, and H.C.W. de Vet. Mind the MIC: large variation among populations and methods. *Journal of Clinical Epidemiology*, 63(5):524–534, 2010.

[200] H.C. De Vet, C.B. Terwee, R.W. Ostelo, H. Beckerman, D.L. Knol, and L.M. Bouter. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. *Health and Quality of Life Outcomes*, 4(1):54, 2006.

[201] D. E. Beaton, C. Bombardier, J. N. Katz, and J. G. Wright. A taxonomy for responsiveness. *Journal of Clinical Epidemiology*, 54(12):1204–1217, 2001.

[202] D. Revicki, R.D. Hays, D. Cella, and J. Sloan. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *Journal of Clinical Epidemiology*, 61(2):102–109, 2008.

[203] A.G. Copay, B.R. Subach, S.D. Glassman, D.W. Polly, and T.C. Schuler. Understanding the minimum clinically important difference: a review of concepts and methods. *The Spine Journal*, 7(5):541–546, 2007.

[204] R.W.J.G. Ostelo, R.A. Deyo, P. Stratford, G. Waddell, P. Croft, M. Von Korff, L.M. Bouter, and H.C. de Vet. Interpreting change scores for pain and functional status in low back pain: towards international consensus regarding minimal important change. *Spine*, 33(1):90, 2008.

[205] J. Cohen. *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum, 1988.

[206] G.R. Norman, J.A. Sloan, and K.W. Wyrwich. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Medical care*, 41(5):582, 2003.

[207] R.D. Crosby, R.L. Kolotkin, and G. R. Williams. Defining clinically meaningful change in health-related quality of life. *Journal of Clinical Epidemiology*, 56(5):395–407, 2003.

[208] K.W. Wyrwich, W.M. Tierney, and F.D. Wolinsky. Further evidence supporting an sem-based criterion for identifying meaningful intra-individual changes in health-related quality of life. *Journal of clinical epidemiology*, 52(9):861–873, 1999.

[209] K.W. Wyrwich, N.A. Nienaber, W.M. Tierney, and F.D. Wolinsky. Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life. *Medical care*, 37(5):469–478, 1999.

[210] R.A. Deyo and R.M. Centor. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. *Journal of chronic diseases*, 39(11):897–906, 1986.

[211] R.W. Ostelo and H.C. de Vet. Clinically important outcomes in low back pain. *Best practice & research. Clinical rheumatology*, 19(4):593, 2005.

[212] NHS Information Centre. The processing cycle and HES data quality. HES Online website, 2012.

[213] H.C.W. de Vet, B. Terluin, D.L. Knol, L.D. Roorda, L.B. Mokkink, R.W.J.G. Ostelo, E.J.M. Hendriks, L.M. Bouter, and C.B. Terwee. Three ways to quantify uncertainty in individually applied 'minimally important change' values. *Journal of Clinical Epidemiology*, 63(1):37–45, 2010.

[214] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982.

[215] J. Sloan, T. Symonds, D. Vargas-Chanes, and B. Fridley. Practical guidelines for assessing the clinical significance of health-related quality of life changes within clinical trials. *Drug Inf J*, 37(1):23–31, 2003.

[216] J.P. Browne, J.H. van der Meulen, J.D. Lewsey, D.L. Lamping, and N. Black. Mathematical coupling may account for the association between baseline severity and minimally important difference values. *Journal of clinical epidemiology*, 63(8):865–874, 2010.

[217] A. Hutchings, J. Neuburger, K.G. Frie, N. Black, and J. van der Meulen. Factors associated with non-response in routine use of patient reported outcome measures after elective surgery in england. *Health and Quality of Life Outcomes*, 10(1):34, 2012.

[218] A. Hutchings, K.G. Frie, J. Neuburger, J. van der Meulen, and N. Black. Late response to patient-reported outcome questionnaires after surgery was associated with worse outcome. *Journal of Clinical Epidemiology*, 66(2):218–25, 2012.

[219] M. Ross. Relation of implicit theories to the construction of personal histories. *Psychological Review*, 96(2):341, 1989.

[220] D.E. Beaton, M. Boers, and G.A. Wells. Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. *Current opinion in rheumatology*, 14(2):109, 2002.

[221] G. Norman. Hi! How are you? Response shift, implicit theories and differing epistemologies. *Quality of life Research*, 12(3):239–249, 2003.

[222] J. Valderas and J. Alonso. Patient reported outcome measures: a model-based classification system for research and clinical practice. *Quality of Life Research*, 17(9):1125–1135, 2008.

[223] National Institute for Health and Clinical Excellence (NICE). *Social Value Judgements: Principles for the development of NICE guidance. 2nd ed.* London: NICE, 2008.

[224] J. Ware, M. Kosinski, and S. D. Keller. A 12-item short-form health survey: construction of scales and preliminary tests of reliability and validity. *Medical Care*, 34(3):220–33, 1996. Ware, J Jr Kosinski, M Keller, S D United states Medical care Med Care. 1996 Mar;34(3):220-33.

[225] J.H. Kellgren and J.S. Lawrence. Radiological assessment of osteo-arthrosis. *Annals of the Rheumatic Diseases*, 16(4):494–502, 1957.

[226] J.H. Kellgren. *The epidemiology of chronic rheumatism: atlas of standard radiographs of arthritis.* Blackwell Scientific, Oxford, 1963.

[227] L. Ding, W.F. Velicer, and L.L. Harlow. Effects of estimation methods, number of indicators per factor, and improper solutions on structural equation modeling fit indices. *Structural Equation Modeling: A Multidisciplinary Journal*, 2(2):119–143, 1995.

[228] P.M. Bentler and C.P. Chou. Practical issues in structural modeling. *Sociological Methods & Research*, 16(1):78–117, 1987.

[229] H. Beckerman, M.E. Roebroeck, G.J. Lankhorst, J.G. Becher, P.D. Bezemer, and A.L.M. Verbeek. Smallest real difference, a link between reproducibility and responsiveness. *Quality of Life Research*, 10(7):571–578, 2001.

[230] WHO Consultation. Obesity: preventing and managing the global epidemic. *World Health Organization Technical Report Series*, (894), 2000.

[231] National Institute for Health and Clinical Excellence (NICE). Osteoarthritis. National clinical guideline for care and management in adults, 2008.

[232] R.A. Charter and L.S. Feldt. Confidence intervals for true scores: Is there a correct approach? *Journal of Psychoeducational Assessment*, 19(4):350–364, 2001.

[233] J. Wang and W. Xiaogian. *Structural Equation Modeling: Applications Using MPlus (Wiley Series in Probability and Statistics)*. Wiley and Sons, 2012.

[234] S.E. Embretson and S.P. Reise. *Item response theory for psychologists*, volume 4. Lawrence Erlbaum, 2000.

[235] F.M. Lord. The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*, 1953.

[236] X. Fan. Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and psychological measurement*, 58(3):357–381, 1998.

[237] C. Stage. *Classical test theory or item response theory: The Swedish experience*, volume 42. Centro de Estudios Publico, 2003.

[238] P. Macdonald and S.V. Paunonen. A monte carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement*, 62(6):921–943, 2002.

[239] R.W. Massof. Understanding rasch and item response theory models: applications to the estimation and validation of interval latent trait measures from responses to rating scale questionnaires. *Ophthalmic epidemiology*, 18(1):1–19, 2011.

[240] D. Andrich. A rating formulation for ordered response categories. *Psychometrika*, 43(4):561–573, 1978.

[241] I. Marais and D. Andrich. Formalizing dimension and response violations of local independence in the unidimensional rasch model. *J Appl Meas*, 9(3):200–15, 2008.

[242] H.G. Dickson and F. Köhler. The multi-dimensionality of the FIM motor items precludes an interval scaling using Rasch analysis. *Scandinavian journal of rehabilitation medicine*, 28(3):159, 1996.

[243] D. Andrich, B. Sheridan, and G. Luo. Rumm 2030. *Western Australia: RUMM Laboratory Pty Ltd*, 2009.

[244] J.M. Linacre. Sample size and item calibration stability. *Rasch Measurement Transactions*, 7(4):328, 1994.

[245] J.M. Linacre. Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, 3(1):85–106, 2002.

[246] B.D. Wright and M.H. Stone. *Best Test Design (Rasch Measurement Series).* Univ of Chicago Social Research, 1979.

[247] B.D. Wright and N. Panchapakesan. A procedure for sample-free item analysis. *Educational and Psychological measurement*, 29(1):23–48, 1969.

[248] J.M. Bland and D.G. Altman. Multiple significance tests: the Bonferroni method. *BMJ (Clinical Research Ed.)*, 310(6973):170, 1995.

[249] D. Andrich. An elaboration of Guttman scaling with Rasch models for measurement. *Social Methodology*, 15:33–80, 1985.

[250] B. Purya. Local dependency and Rasch measures. *Rasch Measurement Transactions*, 21(3):1105–1106, 2007.

[251] F.M. Lord. *Applications of item response theory to practical testing problems.* Lawrence Erlbaum Associates, 1980.

[252] A. Tennant and P.G. Conaghan. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Care and Research*, 57(8):1358–1362, 2007.

[253] E.V. Smith Jr. Understanding rasch measurement: Detecting and evaluating

the impact of multidimenstionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*, 2002.

[254] A. Tennant and JF Pallant. Unidimensionality matters!(A tale of two Smiths?). *Rasch Measurement Transactions*, 20(1):1048–51, 2006.

[255] Right Care. The NHS Atlas of Variation in Healthcare: Reducing unwarranted variation to increase value and improve quality. Technical report, November 2011.

[256] A. Judge, N.J. Welton, J. Sandhu, and Y. Ben-Shlomo. Equity in access to total joint replacement of the hip and knee in england: cross sectional study. *BMJ: British Medical Journal*, 341, 2010.

[257] P. Dieppe, K. Lim, and S. Lohmander. Who should have knee joint replacement surgery for osteoarthritis? *International Journal of Rheumatic Diseases*, 14(2):175–180, 2011.

[258] A. Clarke, N. Musila, N. Le Maistre, N. Black, and J. van der Meulen. The REFER Project: Realistic Effective Facilitation of Elective Referral for elective surgical assessment. Technical report, National Co-ordinating Centre for the National Institute for Health Research Service Delivery and Organisation Programme (NCCSDO), 2010.

[259] A. Judge, N.K. Arden, C. Cooper, M.K. Javaid, A.J. Carr, R.E. Field, and P.A. Dieppe. Predictors of outcomes of total knee replacement surgery. *Rheumatology*, 2012.

[260] A. Judge, NK Arden, A. Price, S. Glyn-Jones, D. Beard, AJ Carr, J. Dawson, R. Fitzpatrick, and RE Field. Assessing patients for joint replacement. Can pre-operative Oxford hip and knee scores be used to predict patient satisfaction following joint replacement surgery and to guide patient selection? *Journal of Bone & Joint Surgery, British Volume*, 93(12):1660–1664, 2011.

[261] P. Royston, K.G.M. Moons, D.G. Altman, and Y. Vergouwe. Prognosis and prognostic research: developing a prognostic model. *Bmj*, 338, 2009.

[262] K.G.M. Moons, P. Royston, Y. Vergouwe, D.E. Grobbee, and D.G. Altman. Prognosis and prognostic research: what, why, and how? *Bmj*, 338, 2009.

[263] K.G.M. Moons, D.G. Altman, Y. Vergouwe, and P. Royston. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *Bmj*, 338, 2009.

[264] J. Slover, J. Shue, and K. Koenig. Shared decision-making in orthopaedic surgery. *Clinical Orthopaedics and Related Research®*, pages 1–8, 2012.

[265] General Medical Council. Good medical practice. GMC London, 2013.

[266] G. Elwyn, D. Frosch, R. Thomson, N. Joseph-Williams, A. Lloyd, P. Kinnersley, E. Cording, D. Tomson, C. Dodd, and S. Rollnick. Shared decision making: A model for clinical practice. *Journal of General Internal Medicine*, 27(10):1361–1367, 2012.

[267] R.M. Epstein and E. Peters. Beyond information. *JAMA: The Journal of the American Medical Association*, 302(2):195–197, 2009.

[268] D. Stacey, C.L. Bennett, M.J. Barry, N.F. Col, K.B. Eden, M. Holmes-Rovner, H. Llewellyn-Thomas, A. Lyddiatt, F. Légaré, and R. Thomson. Decision aids for people facing health treatment or screening decisions. *Cochrane Database Syst Rev*, 10, 2011.

[269] C. Elwyn G. Mulley, A. Trimble. Patients' preferences matter. Technical report, The King's Fund, 2012.

[270] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.C. Sanchez, and M. Muller. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12(1):77, 2011.

[271] J.A. Swets. Measuring the accuracy of diagnostic systems. *Science*, 240(4857):1285–1293, 1988.

[272] E.R. DeLong, D.M. DeLong, and D.L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a non-parametric approach. *Biometrics*, pages 837–845, 1988.

[273] J.A. Hanley and B.J. McNeil. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148(3):839, 1983.

[274] M.B. Kinds, P.M.J. Welsing, E.P. Vignon, J.W.J. Bijlsma, M.A. Viergever, A.C.A. Marijnissen, and F. Lafeber. A systematic review of the association between radiographic and clinical osteoarthritis of hip and knee. *Osteoarthritis and Cartilage*, 19(7):768–778, 2011.

[275] L. Gossec, S. Paternotte, JF Maillefert, C. Combescure, PG Conaghan, AM Davis, K.P. Gunther, G. Hawker, and M. Hochberg. The role of pain and functional impairment in the decision to recommend total joint replacement in hip and knee osteoarthritis: an international cross-sectional study of 1909 patients. Report of the OARSI-OMERACT Task Force on total joint replacement. *Osteoarthritis and Cartilage*, 19(2):147–154, 2011.

[276] P.C. Noble, M.A. Conditt, K.F. Cook, and K.B. Mathis. The John Insall Award: Patient expectations affect satisfaction with total knee arthroplasty. *Clinical orthopaedics and related research*, 452:35–43, 2006.

[277] H. Dakin, A. Gray, R. Fitzpatrick, G. Maclennan, and D. Murray. Rationing of total knee replacement: a cost-effectiveness analysis on a large trial data set. *BMJ Open*, 2(1), 2012.

[278] R. Fitzpatrick. Assessment of quality of life as an outcome: finding measurements that reflect individuals' priorities. *Quality in Health Care: QHC*, 8(1):1, 1999.

[279] Department of Health. Guidance on the routine collection of patient reported outcome measures (PROMs), 2008.

[280] S. Clarke, V. Lock, J. Duddy, M. Sharif, J.H. Newman, and J.R. Kirwan. Intra-articular hylan GF 20 (Synvisc®) in the management of patellofemoral osteoarthritis of the knee (POAK). *The Knee*, 12(1):57–62, 2005.

[281] A.A. Alturfan, E. Uslu, E.E. Alturfan, G. Hatemi, I. Fresko, and E. Kokoglu.

Increased serum sialic acid levels in primary osteoarthritis and inactive rheumatoid arthritis. *The Tohoku journal of experimental medicine*, 213(3):241–248, 2007.

[282] A. McGinley and B. Lucas. Telenursing: A pilot of telephone review after intra-articular knee injection. *Journal of Orthopaedic Nursing*, 10(3):144–150, 2006.

[283] T.R. Stanton, C.W.C. Lin, R.J.E.M. Smeets, D. Taylor, R. Law, and G.L. Moseley. Spatially defined disruption of motor imagery performance in people with osteoarthritis. *Rheumatology*, 51(8):1455–1464, 2012.

[284] A.T. Bremner-Smith, P. Ewings, and A.E. Weale. Knee scores in a normal elderly population. *The Knee*, 11(4):279–282, 2004.

[285] J.M. Linacre. The Rasch model cannot be Disproved. *Rasch Measurement Transactions*, 10(3):512–514, 1996.

[286] R.J.A. Little and D.B. Rubin. *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. Wiley-Interscience, 2002.

[287] Jean-Benoit Hardouin, Ronan Conroy, and Veronique Sebille. Imputation by the mean score should be avoided when validating a patient reported outcomes questionnaire by a Rasch model in presence of informative missing data. *BMC Medical Research Methodology*, 11(1):105, 2011.

[288] G. Hawker. Measure of Intermittent and Constant Osteoarthritis Pain: ICOAP. User's Guide. Version 6, July 2010.

Appendix A

List of operation codes and rules used in HES dataset recoding

KNEE SURGERY QUESTIONNAIRE				
Procedure	**Clinical codes (OPCS 4.5)**			**ICD-10 Diagnosis codes #**
	Definitions	**Combinations & sites ***	**Exclusions****	
Unilateral Knee Replacement				
Total knee replacements (primary)	W40.1, W40.8, W40.9, W41.1, W41.8, W41.9, W42.1, W42.8, W42.9	N/A	Z94.1 ("bilateral operation") or where a pair of unilateral hip replacements take place in the same spell accompanied by both Z94.2 ("Right-sided") & Z94.3 ("Left-sided")	N/A
Unicondylar / Unicompartmental knee operations (primary)	W52.1, W52.8, W52.9, W53.1, W53.8, W53.9, W54.1, W54.8, W54.9	Z76.5 ("lower end of femur NEC"), Z77.1 ("condyle of tibia"), Z77.4 ("upper end of tibia NEC"), Z84.4 ("patellofemoral joint"), Z84.5 ("tibiofemoral joint") or Z84.6 ("knee joint")	Z94.1 ("bilateral operation") or where a pair of unilateral hip replacements take place in the same spell accompanied by both Z94.2 ("Right-sided") & Z94.3 ("Left-sided")	N/A
Hybrid prosthetic knee replacements (primary)	O18.1, O18.8, O18.9	N/A	Z94.1 ("bilateral operation") or where a pair of unilateral hip replacements take place in the same spell accompanied by both Z94.2 ("Right-sided") & Z94.3 ("Left-sided")	N/A
Total knee replacements (revision)	W40.0, W40.2, W40.3, W40.4, W41.0, W41.2, W41.3, W41.4, W42.0, W42.2, W42.3, W42.4, W42.5, W42.6	N/A	Z94.1 ("bilateral operation") or where a pair of unilateral hip replacements take place in the same spell accompanied by both Z94.2 ("Right-sided") & Z94.3 ("Left-sided")	N/A

Copyright © 2010, The Health and Social Care Information Centre. All Rights Reserved.

Unicondylar / Unicompartmental knee operations (revision)	W52.0, W52.2, W52.3, W53.0, W53.2, W53.3, W54.0, W54.2, W54.3, W54.4	Z76.5 ("lower end of femur NEC"), Z77.1 ("condyle of tibia"), Z77.4 ("upper end of tibia NEC"), Z84.4 ("patellofemoral joint"), Z84.5 ("tibiofemoral joint") or Z84.6 ("knee joint")	Z94.1 ("bilateral operation") or where a pair of unilateral hip replacements take place in the same spell accompanied by both Z94.2 ("Right-sided") & Z94.3 ("Left-sided")	N/A
Hybrid prosthetic knee replacements (revisions)	O18.0, O18.2, O18.3, O18.4	N/A	Z94.1 ("bilateral operation") or where a pair of unilateral hip replacements take place in the same spell accompanied by both Z94.2 ("Right-sided") & Z94.3 ("Left-sided")	N/A

Reference Information

****: Specified "Combination" & "Site" codes must appear in combination with the relevant definition codes for the procedure to count.***

*****: Activity should be excluded if it would be coded with these codes.***

#: These ICD-10 Diagnoses would need to be present for the activity to count.

Copyright © 2010, The Health and Social Care Information Centre. All Rights Reserved.

Appendix B

Chapter 3 supplementary material

B.1 Sample output for one factor model (SWLEOC dataset)

The following lines were read from file C:\DTB\swl1.SPJ:

Latent Variables Q
Relationships
B1 = Q
B2 = Q
B3 = Q
B4 = Q
B5 = Q
B6 = Q
B7 = Q
B8 = Q
B9 = Q
B10 = Q
B11 = Q
B12 = Q
Path Diagram
Print Residuals
Method of Estimation: Diagonally Weighted Least Squares
End of Problem
Sample Size = 2405

B.1.0.1 Correlation Matrix

B1 B2 B3 B4 B5 B6 B7 B8 B9 B10 B11 B12

——— ——— ——— ——— ——— ———

B1 1.00

B2 0.38 1.00

B3 0.43 0.67 1.00

B4 0.39 0.41 0.41 1.00

B5 0.57 0.49 0.52 0.39 1.00

B6 0.37 0.32 0.33 0.38 0.39 1.00

B7 0.34 0.38 0.46 0.31 0.48 0.29 1.00

B8 0.48 0.36 0.37 0.28 0.48 0.32 0.29 1.00

B9 0.52 0.53 0.52 0.50 0.58 0.48 0.46 0.48 1.00

B10 0.41 0.46 0.43 0.37 0.48 0.44 0.34 0.46 0.56 1.00

B11 0.42 0.56 0.55 0.49 0.50 0.40 0.54 0.38 0.62 0.50 1.00

B12 0.38 0.54 0.56 0.46 0.54 0.33 0.52 0.37 0.57 0.51 0.65 1.00

Number of Iterations = 5

LISREL Estimates (Robust Diagonally Weighted Least Squares)

B.1.0.2 Measurement Equations

B1 = 0.63*Q, Errorvar.= 0.60 ,R^2= 0.40

(0.017) (0.046)

37.77 12.90

B2 = 0.71*Q, Errorvar.= 0.49 , R^2 = 0.51

(0.014) (0.045)

51.49 10.97

B3 = 0.73*Q, Errorvar.= 0.47 , R^2= 0.53

(0.013) (0.044)

57.24 10.67

B4 = 0.59*Q, Errorvar.= 0.65 , R^2= 0.35

(0.018) (0.045)

33.20 14.54

B5 = 0.73*Q, Errorvar.= 0.46 , R^2 = 0.54

(0.013) (0.046)

54.30 10.05

B6 = 0.54*Q, Errorvar.= 0.71 , R^2 = 0.29

(0.018) (0.046)

29.37 15.44

B7 = 0.60*Q, Errorvar.= 0.64 , R^2 = 0.36

(0.017) (0.045)

34.97 14.13

B8 = 0.57*Q, Errorvar.= 0.67 , R^2 = 0.33

(0.018) (0.047)

32.76 14.38
B9 = 0.80*Q, Errorvar.= 0.37 , R^2 = 0.63
(0.011) (0.045)
69.33 8.20
B10 = 0.67*Q, Errorvar.= 0.56 , R^2 = 0.44
(0.015) (0.046)
44.59 12.23
B11 = 0.77*Q, Errorvar.= 0.40 , R^2 = 0.60
(0.011) (0.043)
68.16 9.24
B12 = 0.75*Q, Errorvar.= 0.43 , R^2= 0.57
(0.013) (0.045)
56.59 9.65

B.1.0.3 Goodness of Fit Statistics

Degrees of Freedom = 54
Normal Theory Weighted Least Squares Chi-Square = 1507.14 (P = 0.0)
Satorra-Bentler Scaled Chi-Square = 774.24 (P = 0.0)
Chi-Square Corrected for Non-Normality = 496.82 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 720.24
90 Percent Confidence Interval for NCP = (633.99 ; 813.92)
Minimum Fit Function Value = 0.19
Population Discrepancy Function Value (F0) = 0.30
90 Percent Confidence Interval for F0 = (0.26 ; 0.34)
Root Mean Square Error of Approximation (RMSEA) = 0.074
90 Percent Confidence Interval for RMSEA = (0.070 ; 0.079)
P-Value for Test of Close Fit (RMSEA < 0.05) = 0.00
Expected Cross-Validation Index (ECVI) = 0.34
90 Percent Confidence Interval for ECVI = (0.31 ; 0.38)
ECVI for Saturated Model = 0.065
ECVI for Independence Model = 13.92
Chi-Square for Independence Model with 66 Degrees of Freedom = 33435.71
Independence AIC = 33459.71
Model AIC = 822.24
Saturated AIC = 156.00
Independence CAIC = 33541.13
Model CAIC = 985.09
Saturated CAIC = 685.25
Normed Fit Index (NFI) = 0.98
Non-Normed Fit Index (NNFI) = 0.97
Parsimony Normed Fit Index (PNFI) = 0.80
Comparative Fit Index (CFI) = 0.98

Incremental Fit Index (IFI) = 0.98
Relative Fit Index (RFI) = 0.97
Critical N (CN) = 252.72
Root Mean Square Residual (RMR) = 0.048
Standardized RMR = 0.048
Goodness of Fit Index (GFI) = 0.99
Adjusted Goodness of Fit Index (AGFI) = 0.99
Parsimony Goodness of Fit Index (PGFI) = 0.69

B.1.0.4 Fitted Covariance Matrix

	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12
B1	1.00											
B2	0.45	1.00										
B3	0.46	0.52	1.00									
B4	0.37	0.42	0.43	1.00								
B5	0.46	0.52	0.54	0.43	1.00							
B6	0.34	0.39	0.39	0.32	0.40	1.00						
B7	0.38	0.43	0.44	0.35	0.44	0.33	1.00					
B8	0.36	0.41	0.42	0.34	0.42	0.31	0.35	1.00				
B9	0.50	0.57	0.58	0.47	0.58	0.43	0.48	0.46	1.00			
B10	0.42	0.48	0.49	0.39	0.49	0.36	0.40	0.38	0.53	1.00		
B11	0.49	0.55	0.57	0.46	0.57	0.42	0.47	0.44	0.62	0.52	1.00	
B12	0.48	0.54	0.55	0.44	0.55	0.41	0.45	0.43	0.60	0.50	0.58	1.00

B.1.0.5 Fitted Residuals

	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12
B1	0.00											
B2	-0.07	0.00										
B3	-0.03	0.15	0.00									
B4	0.02	-0.01	-0.02	0.00								
B5	0.10	-0.04	-0.01	-0.04	0.00							
B6	0.03	-0.07	-0.06	0.07	-0.01	0.00						
B7	-0.04	-0.05	0.02	-0.05	0.04	-0.04	0.00					
B8	0.12	-0.05	-0.05	-0.06	0.06	0.01	-0.05	0.00				
B9	0.02	-0.04	-0.06	0.03	0.00	0.05	-0.02	0.02	0.00			
B10	-0.01	-0.02	-0.06	-0.02	-0.01	0.08	-0.06	0.08	0.03	0.00		
B11	-0.07	0.01	-0.02	0.04	-0.07	-0.02	0.07	-0.07	0.00	-0.02	0.00	
B12	-0.10	0.00	0.01	0.01	-0.02	-0.07	0.07	-0.07	-0.03	0.00	0.07	0.00

Smallest Fitted Residual = -0.10
Median Fitted Residual = 0.00
Largest Fitted Residual = 0.15

- 8|6
- 6|322998621
- 4|9974337651
- 2|96648320
- 0|976651188710000000000000
0|24774468
2|13913
4|024
6|577718
8|1
10|59
12|
14|6

B.1.0.6 Standardized Residuals

B1 B2 B3 B4 B5 B6 B7 B8 B9 B10 B11 B12
——— ——— ——— ——— ——— ———

B1 - -
B2 -3.31 - -
B3 -1.59 11.19 - -
B4 0.94 -0.43 -1.19 - -
B5 6.18 -1.97 -0.61 -2.27 - -
B6 1.33 -3.12 -3.04 3.32 -0.32 - -
B7 -1.87 -2.81 1.03 -2.27 2.31 -1.79 - -
B8 6.35 -2.30 -2.59 -2.85 3.70 0.69 -2.52 - -
B9 0.84 -1.99 -3.21 1.91 -0.02 3.00 -1.02 1.30 - -
B10 -0.58 -0.87 -3.23 -1.21 -0.45 4.22 -3.02 4.61 1.96 - -
B11 -3.71 0.51 -1.01 2.39 -4.09 -0.97 4.37 -3.59 0.11 -1.00 - -
B12 -4.40 -0.07 0.45 0.84 -1.04 -3.39 3.93 -3.38 -1.63 0.23 5.09 - -

Smallest Standardized Residual = -4.40
Median Standardized Residual = -0.05
Largest Standardized Residual = 11.19

- 4|41
- 2|7644322100986533300
- 0|9866220000096644310000000000000
0|125578890339
2|0340379

4|2461
6|23
8|
10|2

Residual for B2 and B1 -3.31
Residual for B6 and B2 -3.12
Residual for B6 and B3 -3.04
Residual for B7 and B2 -2.81
Residual for B8 and B3 -2.59
Residual for B8 and B4 -2.85
Residual for B9 and B3 -3.21
Residual for B10 and B3 -3.23
Residual for B10 and B7 -3.02
Residual for B11 and B1 -3.71
Residual for B11 and B5 -4.09
Residual for B11 and B8 -3.59
Residual for B12 and B1 -4.40
Residual for B12 and B6 -3.39
Residual for B12 and B8 -3.38

Residual for B3 and B2 11.19
Residual for B5 and B1 6.18
Residual for B6 and B4 3.32
Residual for B8 and B1 6.35
Residual for B8 and B5 3.70
Residual for B9 and B6 3.00
Residual for B10 and B6 4.22
Residual for B10 and B8 4.61
Residual for B11 and B7 4.37
Residual for B12 and B7 3.93
Residual for B12 and B11 5.09

B.1.0.7 The Modification Indices Suggest to Add an Error Covariance

Between and/ Decrease in Chi-Square /New Estimate

B2 B1 16.0 -0.12
B3 B2 1938.9 3.27
B5 B1 114.0 0.40
B5 B4 9.9 -0.10
B6 B2 15.3 -0.12
B6 B3 15.4 -0.12
B6 B4 19.7 0.14
B7 B2 13.8 -0.11

B7 B4 8.9 -0.09
B7 B5 11.9 0.11
B8 B1 91.8 0.32
B8 B2 9.4 -0.10
B8 B3 11.0 -0.10
B8 B4 13.5 -0.11
B8 B5 32.8 0.19
B8 B7 9.9 -0.09
B9 B3 15.6 -0.11
B9 B4 9.9 0.11
B9 B6 23.0 0.17
B10 B3 17.7 -0.12
B10 B6 38.0 0.20
B10 B7 15.1 -0.12
B10 B8 50.1 0.23
B10 B9 9.3 0.10
B11 B1 22.9 -0.15
B11 B4 17.6 0.16
B11 B5 28.9 -0.15
B11 B7 71.8 0.34
B11 B8 23.1 -0.15
B12 B1 29.0 -0.17
B12 B6 18.9 -0.14
B12 B7 46.2 0.25
B12 B8 20.8 -0.14
B12 B11 90.9 0.35

Appendix C

PROMs used in the validation study (non-copyrighted or reproduced with permission)

PROBLEMS WITH YOUR KNEE

Tick (✓) one box for every question.

Q1. During the past 4 weeks...

How would you describe the pain you usually have from your knee?

None	Very mild	Mild	Moderate	Severe
☐	☐	☐	☐	☐

Q2. During the past 4 weeks...

Have you had any trouble with washing and drying yourself (all over) because of your knee?

No trouble at all	Very little trouble	Moderate trouble	Extreme difficulty	Impossible to do
☐	☐	☐	☐	☐

Q3. During the past 4 weeks...

Have you had any trouble getting in and out of a car or using public transport because of your knee? (whichever you would tend to use)

No trouble at all	Very little trouble	Moderate trouble	Extreme difficulty	Impossible to do
☐	☐	☐	☐	☐

Q4. During the past 4 weeks...

For how long have you been able to walk before pain from your knee becomes severe? (with or without a stick)

No pain/More than 30 minutes	16 to 30 minutes	5 to 15 minutes	Around the house only	Not at all/pain severe when walking
☐	☐	☐	☐	☐

Q5. During the past 4 weeks...

After a meal (sat at a table), how painful has it been for you to stand up from a chair because of your knee?

Not at all painful	Slightly painful	Moderately painful	Very painful	Unbearable
☐	☐	☐	☐	☐

Q6. During the past 4 weeks...

Have you been limping when walking, because of your knee?

Rarely/ never	Sometimes, or just at first	Often, not just at first	Most of the time	All of the time
☐	☐	☐	☐	☐

© Isis Innovation Limited, 1998. All rights reserved.

Q7. During the past 4 weeks...

Could you kneel down and get up again afterwards?

Yes easily	With little difficulty	With moderate difficulty	With extreme difficulty	No, impossible
☐	☐	☐	☐	☐

Q8. During the past 4 weeks...

Have you been troubled by pain from your knee in bed at night?

No nights	Only 1 or 2 nights	Some nights	Most nights	Every nights
☐	☐	☐	☐	☐

Q9. During the past 4 weeks...

How much has pain from your knee interfered with your usual work (including housework)?

Not at all	A little bit	Moderately	Greatly	Totally
☐	☐	☐	☐	☐

Q10. During the past 4 weeks...

Have you felt that your knee might suddenly 'give way' or let you down?

Rarely/ never	Sometimes, or just at first	Often, not just at first	Most of the time	All of the time
☐	☐	☐	☐	☐

Q11. During the past 4 weeks...

Could you do the household shopping on your own?

Yes easily	With little difficulty	With moderate difficulty	With extreme difficulty	No, impossible
☐	☐	☐	☐	☐

Q12. During the past 4 weeks...

Could you walk down one flight of stairs?

Yes easily	With little difficulty	With moderate difficulty	With extreme difficulty	No, impossible
☐	☐	☐	☐	☐

Finally, please check back that you have answered each question.

Thank you very much.

© Isis Innovation Limited, 1998. All rights reserved.

A Measure of Intermittent and Constant Osteoarthritis Pain, ICOAP: KNEE Version

People have told us that they experience different kinds of pain (including aching or discomfort) in their knee. To get a better sense of the different types of knee pain you may experience, we would like to ask you about any "constant pain" (pain you have all the time) separately from any pain that you may experience less often, that is, "pain that comes and goes". The following questions will ask you about the pain that you have experienced in your knee in the PAST WEEK. Please answer ALL questions.

A) CONSTANT PAIN

For each of the following questions, please select the response that best describes, on average, your *constant knee pain* in the PAST WEEK.

1. In the past week, how intense has your *constant knee pain* been?

□0	□1	□2	□3	□4
Not at all/ No constant knee pain	Mildly	Moderately	Severely	Extremely

2. In the past week, how much has your *constant knee pain* affected your sleep?

□0	□1	□2	□3	□4
Not at all/ No constant knee pain	Mildly	Moderately	Severely	Extremely

3. In the past week, how much has your *constant knee pain* affected your overall quality of life?

□0	□1	□2	□3	□4
Not at all/ No constant knee pain	Mildly	Moderately	Severely	Extremely

4. In the past week, how frustrated or annoyed have you been by your *constant knee pain?*

□0	□1	□2	□3	□4
Not at all/ No constant knee pain	Mildly	Moderately	Severely	Extremely

5. In the past week, how upset or worried have you been by your *constant knee pain*?

□0	□1	□2	□3	□4
Not at all/ No constant knee pain	Mildly	Moderately	Severely	Extremely

B) PAIN THAT COMES AND GOES

For each of the following questions, please select the response that best describes your *knee pain that comes and goes*, *on average,* in the PAST WEEK.

6. In the past week, how intense has your most severe *knee pain that comes and goes* been?

□0	□1	□2	□3	□4
Not at all/ No knee pain that comes and goes	Mildly	Moderately	Severely	Extremely

7. In the past week, how frequently has this *knee pain that comes and goes* occurred?

□0	□1	□2	□3	□4
Never/ No knee pain that comes and goes	Rarely	Sometimes	Often	Very Often

8. In the past week, how much has your *knee pain that comes and goes* affected your sleep?

□0	□1	□2	□3	□4
Not at all/ No knee pain that comes and goes	Mildly	Moderately	Severely	Extremely

9. In the past week, how much has your *knee pain that comes and goes* affected your overall quality of life?

□0	□1	□2	□3	□4
Not at all/ No knee pain that comes and goes	Mildly	Moderately	Severely	Extremely

10. In the past week, how frustrated or annoyed have you been by your *knee pain that comes and goes*?

□0	□1	□2	□3	□4
Not at all/ No knee pain that comes and goes	Mildly	Moderately	Severely	Extremely

11. In the past week, how upset or worried have you been by your *knee pain that comes and goes*?

□0	□1	□2	□3	□4
Not at all/ No knee pain that comes and goes	Mildly	Moderately	Severely	Extremely

THANK YOU!

KOOS-Physical Function Shortform (KOOS-PS)

Today's date:_____/_____/_____ Date of birth: _____/_____/_____

Name: _____

INSTRUCTIONS: This survey asks for your view about your knee. This information will help us keep track of how well you are able to perform different activities.
Answer every question by ticking the appropriate box, only one box for each question. If you are unsure about how to answer a question, please give the best answer you can so that you answer all the questions.

The following questions concern your level of function in performing usual daily activities and higher level activities. For each of the following activities, please indicate the degree of difficulty you have experienced in the **last week** due to your knee problem.

1. Rising from bed

None	Mild	Moderate	Severe	Extreme
❑	❑	❑	❑	❑

2. Putting on socks/stockings

None	Mild	Moderate	Severe	Extreme
❑	❑	❑	❑	❑

3. Rising from sitting

None	Mild	Moderate	Severe	Extreme
❑	❑	❑	❑	❑

4. Bending to floor

None	Mild	Moderate	Severe	Extreme
❑	❑	❑	❑	❑

5. Twisting/pivoting on your injured knee

None	Mild	Moderate	Severe	Extreme
❑	❑	❑	❑	❑

6. Kneeling

None	Mild	Moderate	Severe	Extreme
❑	❑	❑	❑	❑

7. Squatting

None	Mild	Moderate	Severe	Extreme
❑	❑	❑	❑	❑

Appendix D

K-L grading system

Grade 0	None
Grade 1	Doubtful narrowing of joint space and possible osteophytic lipping
Grade 2	Definite osteophytes and possible narrowing of joint space
Grade 3	Moderate multiple osteophytes, definite narrowing of joint space and some sclerosis and possible deformity of bone ends
Grade 4	Large osteophytes, marked narrowing of joint space, severe sclerosis and definite deformity of bone ends

Table D.1: K-L criteria of knee OA

Appendix E

Chapter 6 Supplementary material

E.1 Category probability curves

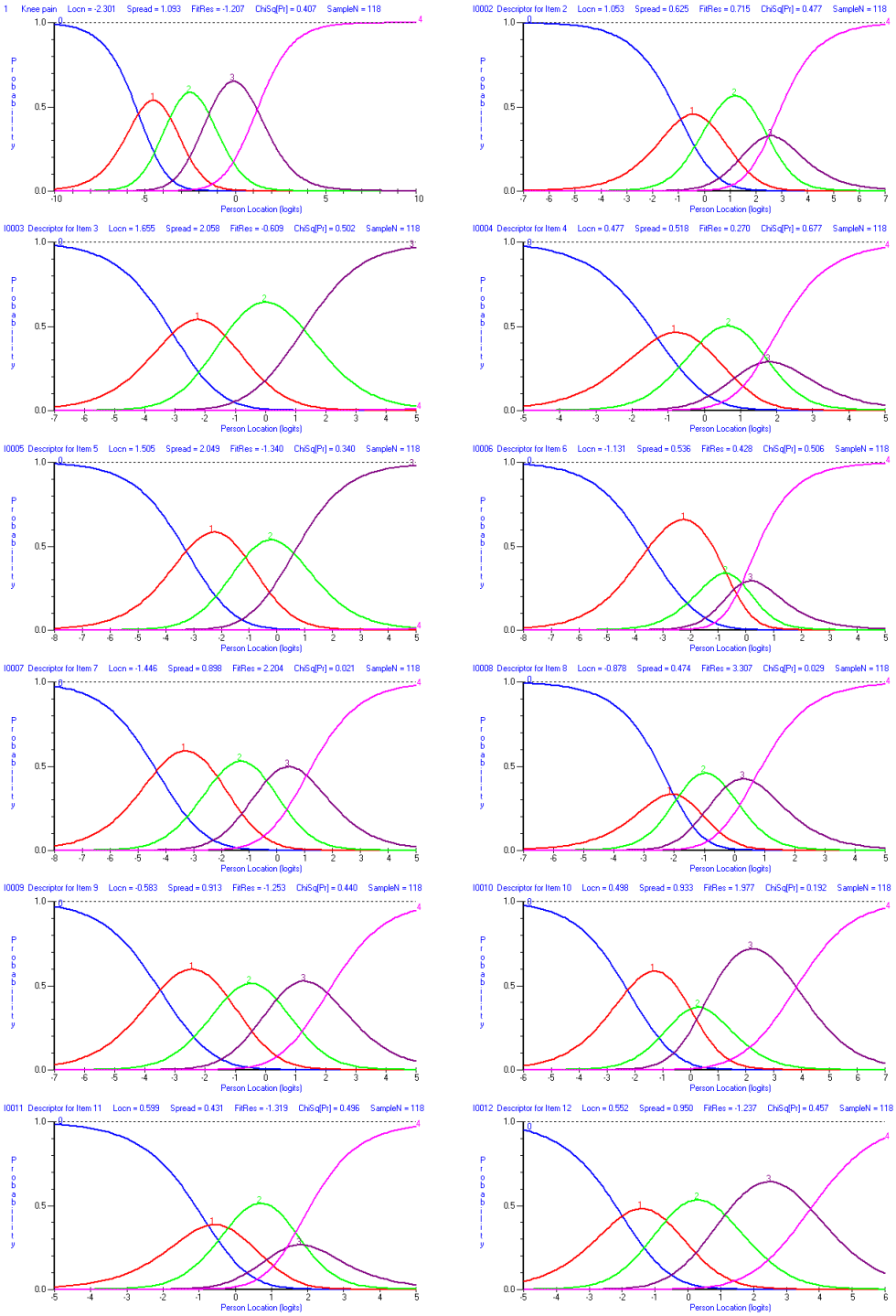


Figure E.1: Category probability curves for the OKS

Appendix F

Data examination procedure

Before any data analysis was performed (as described in Chapters 5, 6, and 7), the integrity of the data entry and coding was checked against the original paper files. 35 (25%) of completed questionnaires were randomly extracted and compared to the coded dataset. No discrepancies were identified.

F.1 Missing data analysis

F.1.1 Introduction

Missing data is a common problem in research that can cause biases in study results [286]. There are several classifications of missing data.

Missing completely at random (MCAR). If the data is MCAR, any missing data can be considered as a random sub-sample that has no ability to bias the results of the analysis.

Missing at random (MAR). A similar situation to MCAR is when the data are MAR; in this case, the data will be missing in relationship to another variable. When controlled for the missing variable, the data can be considered missing at random and the results will more likely be free of bias. Even though, MAR intuitively sounds as missingness that can be ignored, it is likely to produce biased estimates (especially in complete case analysis).

Missing not at random (MNAR). MNAR is a term that is used when there if the data are missing dues some systematic differences that cannot be controlled for (as in MAR).

Different methods of data analysis should be employed, depending if the data is MCAR, MAR or MNAR.

For validation studies in Chapters 5 and 6, analysis of all cases with incomplete data (pairwise deletion) was performed, as imputation is not generally recommended for such analyses [287]. For the ROC analysis, sensitivity analysis of most commonly used methods of dealing with missing data was performed [19].

F.1.2 Methods

In this study AUCs were calculated and compared the results of several different methods of handling missing data:

(i) Analysis of cases with incomplete data. This type of approach uses all available data. This type of analysis is generally statistically more powerful that the complete case analysis. If the data are not missing at random, the results

obtained using this method may be biased.

ii) Complete case analysis. This type of analysis considers only subjects who have data on all the variables of interest. This approach is increasingly problematic as there number of variables per subject increases. Furthermore, if the data are not missing completely at random the results obtained using this method could be biased.

(iii) Imputation of missing values with complete case analysis. Imputation of PROMs data was performed according to the each questionnaire guidelines. For the OKS and ICOAP, this was be a rounded average of item responses. Imputation was be done only if one item was missing in each subscale [288, 162]. The KOOS-PS does not allow imputation and this questionnaire was excluded from the overall analysis.

(iv) Imputation of missing values and analysis of incomplete data. Following the imputation of PROMs data according to each questionnaire guidelines, all available data was be analysed.

DeLong's test [272] was performed to formally assess the significance of the differences between the AUCs obtained from different methods of handling missing data.

F.1.3 Results

Out of 124 patients in the final cohort 98 (79%) had the complete baseline data on the OKS, KOOS-PS and ICOAP. Data analyses were performed for all three cases.

Table F.1: Number and percentage of missing items and missing OKS summary scores, with pain and functional component summaries

Univariate Statistics			
	N	Missing	
		Count	Percent
OKS1	122	2	1.6
OKS2	123	1	.8
OKS3	123	1	.8
OKS4	120	4	3.2
OKS5	123	1	.8
OKS6	123	1	.8
OKS7	123	1	.8
OKS8	124	0	.0
OKS9	123	1	.8
OKS10	122	2	1.6
OKS11	123	1	.8
OKS12	123	1	.8
Sum	115	9	7.3
OKS-PCS	117	7	5.6
OKS-FCS	121	3	2.4

Table F.2: Number and percentage of individual missing items and summary scores for the ICOAP

	N	Missing	
		Count	Percent
ICOAP1	123	1	.8
ICOAP2	122	2	1.6
ICOAP3	122	2	1.6
ICOAP4	122	2	1.6
ICOAP5	121	3	2.4
ICOAP6	124	0	.0
ICOAP7	124	0	.0
ICOAP8	123	1	.8
ICOAP9	124	0	.0
ICOAP10	124	0	.0
ICOAP11	123	1	.8
Sum	120	4	3.2

Table F.3: Number and percentage of individual missing items and summary scores for the KOOS-PS

	N	Missing	
		Count	Percent
KOOS-PS 1	122	2	1.6
KOOS-PS 2	123	1	.8
KOOS-PS 3	123	1	.8
KOOS-PS 4	124	0	.0
KOOS-PS 5	122	2	1.6
KOOS-PS 6	114	10	8.1
KOOS-PS 7	116	8	6.5
sum	106	18	14.5

KOOS-PS had most missing data, particularly items 6 (Kneeling) and 7 (Squatting). Mann Whitney U-test was used to compare the difference between subgroups of patients who have not responded to the 'Kneeling' question in the KOOS-PS. There was no significant difference between the groups in age (p=0.215), BMI (p=0.780), OKS FCS (p=0.119), OKS PCS (p=0.072) or ICOAP (p=0.160). There was no difference between groups who did not respond to the item 'Squatting' in age (p=0.507), BMI (p=0.752), OKS FCS (p=0.176), OKS PCS (p=0.296) or ICOAP (p=0.149). Neither was there difference between two subgroups of patients who did not have the summary KOOS-PS score in age (p=0.486), BMI (p=0.449), OKS FCS (p=0.093), OKS PCS (p=0.097) or ICOAP (p=0.069).

F.1.3.1 Comparative ROC analyses

Analysis of cases with incomplete data

Complete case analysis Out of 76 available cases, 53 considered themselves not to be in need for surgery and 23 considered to be in need for surgery.

Table F.4: AUCs with 95% CIs for each measure using available case analysis.

	K-L grading	OKS	OKS-PCS	OKS-FCS	ICOAP	KOOS-PS
AUC (95% CI)	.652 (.540-.764)	.798 (.716-.880)	.837 (.763-.912)	.709 (.615-.802)	.797 (.714-.880)	.715 (.611-.820)
Positive/Negative	31/65	36/79	37/80	37/84	36/84	30/76
Missing	28	9	7	3	4	18

Table F.5: AUCs with 95% CIs for each measure

	K-L grading	OKS	OKS-PCS	OKS-FCS	ICOAP	KOOS-PS
AUC (95% CI)	.657 (.528-.785)	.749 (.637-.861)	.776 (.667-.887)	.684 (.563-.806)	.739 (.619-.858)	.677 (.547-.807)

When K-L grading was excluded from the comparative analysis, the results were more similar to the previous analysis of cases with incomplete data:

Out of 98 available cases, 70 considered themselves not to be in need for surgery and 28 considered to be in need for surgery.

Analysis of cases with incomplete data after imputing missing values For the OKS and ICOAP one imputation per subscale was allowed. Missing value was substituted by the mean of the responses within the subscale. KOOS-PS does not allow any imputation [158] so this measure was excluded from the analysis. Out of 124 available cases, 82 considered themselves not to be in need for surgery, 35 considered

Table F.6: AUCs with 95% CIs for each measure using complete case analysis without K-L grading included.

	OKS	OKS-PCS	OKS-FCS	ICOAP	KOOS-PS
AUC (95% CI)	.799 (.710-.887)	.829 (.745-.914)	.729 (.629-.830)	.791 (.697-.885)	.745 (.642-.848)

Table F.7: Percentage of missing values after imputation

	N	Missing	
		Count	Percent
ICOAP	122	2	1.6
OKS PCS	120	4	3.2
OKS FCS	122	2	1.6
OKS sum	119	5	4.0

Table F.8: AUCs with 95% CIs for each measure using available case analysis following data imputation.

	OKS	OKS-PCS	OKS-FCS	ICOAP
AUC (95% CI)	.802 (.722-.882)	.842 (.770-.914)	.709 (.616-.803)	.792 (.708-.876)
Positive/negative	37/82	38/82	37/85	36/86
Missing	5	4	2	2

to be in need for surgery and outcome data was missing across 7 cases.

Analysis of complete cases after imputing missing values Out of 117 available cases, 82 considered themselves not to be in need for surgery and 35 considered to be in need for surgery.

The KOOS-PS exhibited most variation between different methods. However, in this study these differences were not statistically significant.

F.1.4 Discussion

The purpose of this study was to perform sensitivity analysis of several different methods of handling missing data. There were no substantial differences between

Table F.9: AUCs with 95% CIs for each measure using complete case analysis following data imputation.

	OKS	OKS-PCS	OKS-FCS	ICOAP
AUC (95% CI)	0.811 (0.733-0.890)	0.846 (0.773-0.920)	0.740 (0.650-0.831)	0.801 (0.718-0.884)

Table F.10: Significance of differences between AUCs obtained using different methods of data imputation for the KOOS-PS.

KOOS-PS IC			
KOOS-PS CC	D=-0.390 (p=0.70)		
KOOS-PS CCX	D=0.446 (p=0.66)	D=0.799 (p=0.43)	
	KOOS-PS IC	KOOS-PS CC	KOOS-PS CCX

Note. IC: analysis of incomplete cases, CC: complete case analysis (with K-L grading), CCX: complete case analysis (without K-L grading).

methods. The largest variation in AUCs between methods (0.068) was in the case of KOOS-PS; this occurred between two variations of complete case analysis that depend on the availability of X-rays. However, it is highly unlikely that there might be a causal relationship between the two, as the availability of X-rays depended on the patient's clinical pathway (some patients who had confirmed diagnosis by MRI were not included in the analysis in Chapter 7). In both groups that had or had not information on the KOOS-PS about 77% of patients had X-rays.

The final choice of the method dealing with missing data analysis that was adopted for the main study (Chapter 7) was dependent on several factors. As the KOOS-PS does not allow data imputation, only approaches that do not rely on imputation were considered. The 'non-imputed available case analysis' was chosen to allow comparative evaluation of AUCs from PROMs and AUCs from the radiological grading. This method was preferred to 'complete case analysis' which seemed to be influenced by the large number of missing radiological information.

No similar studies dealing with the issue of missing data in comparative analysis of diagnostic properties of multiple scoring systems were found. The final approach of dealing with missing data for the main analysis was influenced by the availability of data specific to this study (specifically the inability to impute KOOS-PS and the

large amount of missing radiological information).