2014-09-23

Psychometric Evaluation of Joint-Specific Patient-Reported Outcome Measures Before and After Total Knee Replacement: A Dissertation

Barbara L. Gandek
University of Massachusetts Medical School

PSYCHOMETRIC EVALUATION OF JOINT-SPECIFIC PATIENT-REPORTED OUTCOME MEASURES BEFORE AND AFTER TOTAL KNEE REPLACEMENT

A Dissertation Presented

By

BARBARA LYNNE GANDEK

Submitted to the Faculty of the
University of Massachusetts Graduate School of Biomedical Sciences, Worcester
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

SEPTEMBER 23, 2014

CLINICAL AND POPULATION HEALTH RESEARCH

PSYCHOMETRIC EVALUATION OF JOINT-SPECIFIC PATIENT-REPORTED OUTCOME MEASURES BEFORE AND AFTER TOTAL KNEE REPLACEMENT

A Dissertation Presented
By

BARBARA LYNNE GANDEK

The signatures of the Dissertation Defense Committee signify completion and approval as to style and content of the Dissertation

John E. Ware, Jr. Ph.D., Thesis Advisor

Jeroan J. Allison, M.D., M.S., Member of Committee

Milena Anatchkova, Ph.D., Member of Committee

Courtland Lewis, M.D., Member of Committee

The signature of the Chair of the Committee signifies that the written dissertation meets the requirements of the Dissertation Committee

Patricia D. Franklin, M.D., M.P.H., M.B.A., Chair of Committee

The signature of the Dean of the Graduate School of Biomedical Sciences signifies that the student has met all graduation requirements of the school.

Anthony Carruthers, Ph.D.
Dean of the Graduate School of Biomedical Sciences

Clinical and Population Health Research

September 23, 2014

ACKNOWLEDGEMENTS

I would like to express my gratitude to my mentor, John E. Ware, PhD, for his support and guidance throughout my doctoral program, his challenging questions, and multiple critical readings of every chapter in this dissertation. His vision of what patient-reported outcomes measurement can be has elevated this research far beyond my original intent. It has been a privilege to complete this dissertation under his direction.

I also am very grateful to my dissertation committee of Patricia Franklin, MD, MPH, MBA (chair), Jeroan Allison, MD, MS, Milena Anatchkova, PhD, and Courtland Lewis, MD for their time, their advice, and their thoughtful comments on drafts of this dissertation. I learned a great deal from them and became a better researcher and writer in the process.

I have benefitted so much from the dedication of the faculty of the Clinical & Population Health Research program. In particular, I would like to thank Terry Field, DSc, Stephenie Lemon, PhD, Kate Lapane, PhD, and Carole Upshur, EdD for everything they did to help me navigate through the doctoral process.

So many people have created the wonderful resource that is FORCE-TJR, from those who designed the registry, to those who contacted the surgeons and patients and collected the data, to those who constructed the databases. I especially would like to thank Wenyun Yang, MS and Hua Zheng, PhD for providing me with the data and computer access that I needed for this dissertation, and Christine Goddard, Celeste Lemay, RN, MPH, Janet Sevigny, Betsy Costello and Yubettys Baez for their help.

My colleagues in the Outcomes Measurement Division of the Department of Quantitative Health Sciences helped me throughout my doctoral research. Nina Deng,

EdD provided me with the bootstrapping software that I used in Chapter IV and gave me much psychometric advice, as did Matthias Rose, MD, PhD. Mikel Strom, MSLIS extracted the data that allowed me to submit my systematic review for publication. Dan Frendl (future MD/PhD) was my physician “on call” with whom I shared many conversations about PRO measurement. My longtime colleague Jakob Bjorner, MD, PhD kindly spent a semester teaching me about the confirmatory factor analysis and item response theory methods that I used in Chapter III. I also would like to thank my classmates in the CPHR program for their support, especially those who shared my first and second year classes: Carol Curtin, MSW (and soon-to-be PhD), Ally Hajduk, PhD, Lisa Lines, PhD, Alan Rathbun, PhD, and Sanouri Ursprung, PhD.

I owe much to the clinicians and staff at the Auerbach Breast Center and the Vernon Cancer Center at Newton-Wellesley Hospital for their care during the past three years. I especially would like to thank Thomas Grady, MD, surgeon, for reminding me that the difference between absolute risk and relative risk was a concept that applied to my life as well as to research.

Finally, I would like to thank my mother, sister, brother-in-law, and friends for their constant support throughout this process. Special thanks go to my nephews Greg, Eric and Chris; it was fun being a student with you and commiserating about the unfairness of having homework on the weekends. I hope that life provides you with many opportunities to follow your dreams too.

ABSTRACT

Background: Patient reports of pain and function are used to inform the need for and timing of total knee replacement (TKR) and evaluate TKR outcomes. This dissertation compared measurement properties of commonly-used patient surveys in TKR and explored ways to develop more efficient knee-specific function measures.

Methods: 1,179 FORCE-TJR patients (mean age=66.1, 61% female) completed questionnaires before and 6 months after TKR. Patient surveys included the knee-specific Knee injury and Osteoarthritis Outcome Score (KOOS) and Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) and generic SF-36 Health Survey. Tests of KOOS and WOMAC measurement properties included evaluations of scaling assumptions and reliability. Item response theory methods were used to calibrate 22 KOOS function items in one item bank; simulated computerized adaptive tests (CAT) then were used to evaluate shorter function scores customized for each patient. Validity and responsiveness of measures varying in attributes (knee-specific versus generic, longer versus shorter, CAT versus fixed-length) were compared.

Results: KOOS and WOMAC scales generally met tests of scaling assumptions, although many pain items were equally strong measures of pain and physical function. Internal consistency reliability of KOOS and WOMAC scales exceeded minimum levels of 0.70 recommended for group-level comparisons across sociodemographic and clinical subgroups. Function items could be calibrated in one item bank. CAT simulations indicated that reliable knee-specific function scores could be estimated for most patients with a 55-86% reduction in respondent burden, but one-third could not achieve a reliable (≥ 0.95) CAT score post-TKR because the item bank did not include enough items measuring high function levels. KOOS and WOMAC scales were valid and responsive. Short function scales and CATs were as valid and responsive as longer KOOS and WOMAC function scales. The KOOS Quality of Life (QOL) scale and SF-36 Physical Component Summary discriminated best among groups evaluating themselves as improved, same or worse at 6 months.

Conclusions: Results support use of the KOOS and WOMAC in TKR. Improved knee-specific function measures require new items that measure higher function levels. TKR outcomes should be evaluated with a knee-specific quality of life scale such as KOOS QOL, as well as knee-specific measures of pain and function and generic health measures.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ADL - Activities of Daily Living

CAT - Computerized Adaptive Test

CFA - Confirmatory Factor Analysis

CFI - Comparative Fit Index

DIF - Differential Item Functioning

EFA - Exploratory Factor Analysis

FORCE-TJR - Function and Outcomes Research for Comparative Effectiveness in Total Joint Replacement

GRM - Graded Response Model

IIF - Item Information Function

IRT - Item Response Theory

KOOS - Knee injury and Osteoarthritis Outcome Score

KOOS-PS - KOOS-Physical Function Short Form

OA - Osteoarthritis

PRO - Patient-Reported Outcomes

QOL - Quality of Life

PCS - SF-36 Physical Component Summary

RV - Relative Validity

RMSEA - Root Mean Square Error of Approximation

SET - Self-Evaluated Transition

SF-36 - SF-36 Health Survey

TJR - Total Joint Replacement

TKR - Total Knee Replacement

WOMAC - Western Ontario and McMaster Universities Osteoarthritis Index

PREFACE

Publications related to this study but not presented in detail in this thesis are:

> Gandek B. Measurement properties of the Western Ontario and McMaster Universities Osteoarthritis Index: A systematic review. Arthritis Care & Research 2014 Jul 21. [Epub ahead of print]

Chapter II of this dissertation is under preparation as:

> Gandek B, Ware JE. Tests of data quality, scaling assumptions and reliability of the KOOS and WOMAC in total knee replacement: Results from the FORCE-TJR cohort.

WOMAC® is a registered trademark of Nicholas Bellamy.
SF-36® and SF-12® are registered trademarks of the Medical Outcomes Trust.
EQ-5D™ is a trademark of the EuroQol Group.

CHAPTER I: INTRODUCTION

The burden of osteoarthritis (OA) is an important public health concern. The National Arthritis Data Workgroup estimated that in 2005 nearly 27 million US adults had clinically-defined osteoarthritis, a 28% increase in prevalence since 1995[1]. The prevalence of OA is expected to increase even further due to an aging population, with the CDC projecting a 40% increase in doctor-diagnosed cases of arthritis from 2005 to 2030[2]. Treatment of OA is focused on reducing joint pain and stiffness, preserving and improving joint mobility, limiting joint damage, reducing disability, and improving health-related quality of life[3]. Expert consensus recommendations for conservative treatment of OA include a combination of pharmacological and non-pharmacological therapies[3-5]. However, when medical management of knee and hip osteoarthritis is no longer successful, consideration of total joint replacement (TJR) is recommended[3, 6].

Total joint replacement is one of the most common and costly Medicare surgical procedures. TJR also is becoming more frequent among younger adults[7]; nearly 45% of TJR patients were younger than 65 in 2008[8] and more than 50% are projected to be below age 65 by 2016[9]. Across all age groups, rates of TJR surgery are expected to increase exponentially in the next few decades. Total knee replacement (TKR) is the most common type of TJR surgery; annual numbers of TKRs are projected to increase from 670,000 in 2012[10] to an estimated 3.5 million in 2030[11].

Because TKR is performed to restore function and to reduce pain and other symptoms, the Osteoarthritis Research Society International (OARSI), OMERACT, and other clinician groups recommend that patient-reported measures of pain, function, and global assessment, along with radiographic measures of joint damage, be used to inform the need for and timing of TKR[6] and to evaluate TKR outcomes[12-14]. While TKR reduces pain and improves function substantially on average[15], there are significant variations in patient-reported outcomes (PROs) that have been associated with both patient (e.g., age, gender, BMI, comorbid conditions)[15-17] and provider (e.g., implant device, hospital and surgeon volume, post-operative care)[18-20] characteristics. These variations underscore the importance of obtaining reliable and valid information from patients in evaluating TKR outcomes.

PRO Questionnaires Used in Knee OA and Total Knee Replacement

The most widely-used joint-specific PRO questionnaire used in knee OA and TKR research is the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC®)[21-23], which was published nearly 30 years ago and is recommended by OARSI and other clinical groups for OA studies[12]. It contains 24 items that measure the impact of a specific joint on stiffness, pain during activities, and difficulties with function in activities of daily living (ADL). Shorter 7 and 8-item versions of the 17-item WOMAC function scale have been proposed but are not widely used[24-26]. The 42-item Knee injury and Osteoarthritis Outcome Score (KOOS)[27] was developed as an extension of the WOMAC. The KOOS contains all 24 WOMAC items, additional pain and symptom items, and items about knee-related function in sport/recreation and knee-related quality of life. A shorter 7-item function scale (KOOS-PS) has been constructed from the KOOS function items[28]. In addition to joint-specific questionnaires, generic (general) health surveys, which are not specific to any disease, age, or treatment group, are also used in OA studies. The generic health questionnaire that is most widely-used in TKR research is the SF-36® Health Survey[29], which is scored as an eight-scale profile as well as physical and mental health summary measures[30, 31].

Collection of PRO Data in National TJR Registries

As policy makers and payers place more emphasis on understanding the patient's perspective of their health, interest in collecting PRO data from TJR patients in national registries has increased. The UK mandates that its Department of Health collect PRO data for all TJR patients before and after surgery, using the joint-specific Oxford Hip and Knee Scores[32, 33] and the generic EuroQol EQ-5D™[34]; the UK may link payments to PRO data in the future[18]. Many other registries also collect both a joint-specific and generic measure[35]; for example, the New Zealand joint registry administers both the Oxford Hip and Knee Scores and EQ-5D, while the Swedish registry collects data for the EQ-5D and a joint-specific visual analogue pain scale[36]. While the Oxford Hip and Knee Scores contain 12 items each and thus are attractive due to relatively low respondent burden, they do not provide separate pain and function scores and thus pain relief cannot be assessed independently of functional improvement[37]. Questionnaires that measure pain and function separately, such as the WOMAC, are more widely-used in knee replacement randomized clinical trials[38] and studies assessing pain after TKR[39].

In the US, a CMS Technical Expert Panel (TEP) has been established to make recommendations on using PROs to evaluate TJR outcomes in performance-based measurement systems; the TEP has recommended considering using KOOS and HOOS as joint-specific measures[40]. The US also has one of the largest registries collecting PRO data, the Function and Outcomes Research for Comparative Effectiveness in Total Joint Replacement (FORCE-TJR) cohort (Franklin, AHRQ P50 HS018910-04). Based at the University of Massachusetts Medical School, FORCE-TJR is enrolling more than 30,000 diverse TJR patients of more than 130 orthopedic surgeons throughout the United States[41, 42]. FORCE-TJR administers the KOOS and SF-36 to TKR patients prior to surgery, 6 months and 1 year after surgery, and annually thereafter. Because all WOMAC items (version LK3.0) are included in the KOOS, it is possible to score the WOMAC and the short WOMAC function scales using FORCE-TJR data, in addition to the short function scale (KOOS-PS) constructed from the KOOS. Thus, FORCE-TJR allows for analysis of the KOOS, KOOS-PS, WOMAC, short WOMAC function scales, and the SF-36. It therefore provides an ideal database for comparative research into the measurement properties of many of the most widely-used PRO instruments in TKR and for evaluating implications of the measurement properties for use of these instruments.

Current Research Topics in PRO Measurement in Total Knee Replacement

A number of topics related to PRO measurement in TKR warrant further investigation, particularly before a performance-based measurement system for TKR is established in the US. These topics are discussed below.

Evaluation of KOOS and WOMAC measurement properties among TKR patients in the US using classical psychometric methods

Joint-specific questionnaires such as the KOOS and WOMAC were developed using classical test theory, in which certain assumptions apply; for example, reliability is assumed to be the same at all levels of a scale, and the distance between each item response score generally is assumed to be equivalent. Measurement properties of the KOOS and WOMAC primarily have been evaluated using classical test methods; for example, the internal consistency reliability of KOOS and WOMAC scales generally has been assessed using Cronbach’s coefficient alpha. Widely-accepted guidelines for PRO measures recommend that an instrument’s measurement properties be re-examined whenever it is used in a new culture and patient population[43, 44]. The WOMAC was developed in Canada[23], and initial testing of the KOOS was primarily conducted in

Sweden along with small tests with ACL reconstruction patients in the US[45]. While the measurement properties of the KOOS have been evaluated among patients with knee OA in Europe[27, 46-50] and Asia[51, 52], they have not been examined among knee OA and TKR patients in the US. There also is limited research published on the reliability and validity of the WOMAC among knee OA patients in the US[53-56]. In addition, no published studies have evaluated KOOS or WOMAC measurement properties for different sociodemographic groups, which is particularly important in a country as diverse as the US. Particularly because interest is growing in the US in using the KOOS as a PRO-based performance measure for TKR[40], its measurement properties should be examined for patients who differ in characteristics such as age, gender, education and income.

In particular, assumptions underlying the construction and scoring of KOOS and WOMAC scales should be evaluated. KOOS and WOMAC scales are scored using Likert's classical method of summated ratings, in which scores for all items in a scale are simply added to calculate a scale score[57]. The simplicity of this method is based on a number of assumptions that should be tested prior to scoring a scale; such tests were used extensively in construction and evaluation of the SF-36 Health Survey, for example[58]. However, these tests have not been applied to the KOOS or WOMAC, with the exception of very small studies of the KOOS in Iran[59] and the WOMAC in Singapore[60]. Tests of scaling assumptions may also provide a new perspective on the high (>0.70) correlations observed in multiple studies[21] between the KOOS pain and function scales and between the WOMAC pain and function scales.

Development of shorter knee-specific function measures using modern psychometric methods

The KOOS and WOMAC both contain a 17-item function scale that measures difficulty in performing activities of daily living due to a joint problem. Research has found that the reliability of this scale is consistently higher than the minimum value recommended for group-level comparisons[21] and that there is redundancy in its item content[61, 62]. These findings indicate that this function scale could be shortened, thereby reducing respondent burden, and still meet reliability standards for clinical research.

In the past few decades, modern psychometric methods such as item response theory (IRT) have increasingly been used to evaluate and improve PRO measures[63, 64]. IRT models are used to create item banks, which consist of a set of items measuring the same construct and parameters that describe the items' measurement properties[65]. An item bank provides information about the relative difficulty of each item in the bank and how well the items as a whole cover the full range of measurement. IRT models do not assume that item responses are equidistant and thus yield information that improves the scoring of items on an interval scale. In addition, item banks allow improved fixed-length measures to be constructed, by selecting the best subset of items from the full item bank based on information about the items' measurement properties. Item banks also provide information needed to implement computerized adaptive tests (CATs). CATs administer only the most informative items to each individual respondent, thus leading to shorter, less redundant, and more precise measurement[66].

Despite their advantages, modern psychometric methods have not been applied extensively to the KOOS or WOMAC function items. Those analyses that have been conducted have used a specific type of model, the one parameter Rasch model, with inconsistent results[28, 61, 62, 67-69]. Two parameter IRT models are often seen as better for use with polytomous items that have ordered response categories (such as the KOOS and WOMAC items), because they allow the item discrimination parameter (the item slope) to vary across items, and thus the model is thought to discriminate better between patients[70, 71]. In contrast, the item slope parameter is held constant across all items in the Rasch model, thereby requiring that all items have the same discrimination. IRT models could be particularly helpful in the development of shorter and more parsimonious measures of joint-specific function, by providing a new perspective on which of the KOOS and WOMAC function items are most informative.

Comparison of the validity and responsiveness of PRO measures used in TKR

No studies have compared the validity and responsiveness of the KOOS, WOMAC, KOOS-PS and short WOMAC function scales at the same time. Conducting these comparative analyses can provide additional guidance as to the relative strengths and weaknesses of joint-specific measures used in TKR. In addition, KOOS and WOMAC validity analyses have primarily evaluated correlations among scales; for example, the validity of the KOOS has been demonstrated by showing that its scales have high (e.g., KOOS Function and SF-36 Physical Functioning) or low (e.g., KOOS Function and SF-36 Mental Health) correlations with other scales. While informative, this type of analysis provides limited information about whether a scale is a stronger or weaker measure of a given construct than another scale. To fully evaluate validity, alternative forms of scales need to be compared in relation to external criteria. Such comparisons are often done using tests of known groups validity, which compare the relative efficiency of scales in detecting differences between groups known to differ at a point in time, or in detecting change known to have occurred over time. Tests of known groups validity were used extensively in the development of the SF-36[72, 73], but to my knowledge, have not been applied to the KOOS or WOMAC.

Specific Aims

This dissertation will evaluate the measurement properties of the KOOS and WOMAC and the short function scales derived from these questionnaires, using data from total knee replacement (TKR) patients enrolled in FORCE-TJR. In addition, the KOOS function items will be evaluated to test whether they can be calibrated using item response theory, and properties of the resulting item bank and simulated computerized adaptive tests (CATs) will be evaluated. Psychometric analyses will be conducted using both classical and modern methods, which will provide complementary perspectives on joint-specific measurement in TKR. Specific aims of this dissertation research are to:

Aim 1 (Chapter II). Evaluate the measurement model and reliability of the KOOS and WOMAC among TKR patients using classical psychometric methods.

This paper will evaluate the KOOS in terms of its data quality, tests of scaling assumptions, internal consistency reliability, and floor and ceiling effects (percentage of patients with worst and best scores, respectively). Parallel analyses will be conducted for the WOMAC. Because widespread use of the KOOS and the WOMAC assumes that their psychometric properties apply across diverse populations, analyses also will be conducted for groups differing in sociodemographic characteristics such as age, gender and education and in clinical characteristics such as body mass index.

Aim 2 (Chapter III). Use modern psychometric methods to evaluate measurement properties of KOOS function items and calibrate the items on a common metric.

This paper will use item response theory methods to determine if KOOS function in ADL and function in Sport/Recreation items define a unidimensional construct and if these items can be calibrated in a single function item bank. In addition to providing information about the range of function that is measured by the KOOS and how best to measure the difficulty level of individual function items, the item bank will be used to conduct computerized adaptive test (CAT) simulations using data from FORCE-TJR patients prior to and following TKR. CAT simulations will provide information as to which items are the most informative in TKR and how function might be measured more precisely with fewer items.

Aim 3 (Chapter IV). Evaluate the validity and responsiveness of the KOOS, KOOS-PS, WOMAC, short WOMAC function scales, new IRT-based and CAT-based scores developed in Aim 2, and SF-36 among TKR patients.

This paper will use pre- and post-TKR data to evaluate the validity and responsiveness of the KOOS in comparison to the WOMAC, the short function scales derived from the KOOS and the WOMAC, and the SF-36. In addition, the validity and responsiveness of the new item bank that was developed using all 22 KOOS function items, along with computerized adaptive test (CAT) scores calculated from the item bank, will be evaluated. While the paper will follow approaches that have been used in previous KOOS analyses (e.g., examination of scale correlations), it also will evaluate the relative performance of all measures using tests of known groups validity. By conducting a variety of cross-sectional and longitudinal tests, for which there are strong hypotheses as to the results that would be expected for valid measures, this paper will advance understanding of the comparative performance of joint-specific measures.

A long-term goal of the FORCE-TJR registry is to develop more practical patient-reported measures for use in comparative effectiveness research and clinical practice. Results of this dissertation will advance understanding of the conceptualization and measurement of joint-specific outcomes in TKR and thus inform future development of a parsimonious and comprehensive measurement system.

CHAPTER II: TESTS OF DATA QUALITY, SCALING ASSUMPTIONS AND RELIABILITY OF THE KOOS AND WOMAC IN TOTAL KNEE REPLACEMENT

Abstract

Objective: To evaluate measurement properties of the Knee injury and Osteoarthritis Outcome Score (KOOS) and Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) in US total knee replacement (TKR) patients.

Methods: 1,179 FORCE-TJR patients (43% below age 65) completed surveys before and 6 months after TKR. KOOS and WOMAC scales were evaluated for data completeness, tests of assumptions underlying item and scale scoring, internal consistency reliability, and floor and ceiling effects. Analyses were replicated for 29 sociodemographic and clinical subgroups and paper versus electronic administration.

Results: Item-level missing data was generally low (0.1-3.0%). Scale scores could be calculated for 97-100% of patients. Most tests of scaling assumptions were satisfied at both time points, across all subgroups, and for both methods of administration. However, many KOOS Symptom items had high correlations with the KOOS Pain and Activities of Daily Living (ADL) scales and many KOOS and WOMAC Pain items had high correlations with the KOOS and WOMAC Function in ADL scales. Internal consistency reliability exceeded 0.70 for all scales for the total sample and all subgroups with one subgroup exception. Floor and ceiling effects (percent with worst and best scores) generally were low, but ceiling effects were higher for the WOMAC Stiffness and Pain scales than the KOOS Symptoms and Pain scales at 6 months.

Conclusions: Measurement properties of the KOOS and WOMAC support their use

among US TKR patients. However, interpretation of their Pain scales is confounded because many pain items were equally strong measures of pain and function.

Introduction

Patient reports of pain and function are key indicators in determining the need for total knee replacement (TKR) and assessing TKR outcomes[3]. The most frequently used joint-specific patient-reported measure of these domains is the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), which was developed to measure pain, stiffness and function in relation to an index joint[22]. Subsequently, other joint-specific measures have been developed that include items intended to be more relevant for younger and more active patients. Notable among these is the Knee injury and Osteoarthritis Outcome Score (KOOS), which includes all 24 WOMAC items and 18 additional items[45].

Initial development and testing of the KOOS was conducted among patients with knee injuries in Sweden[74] and the US[45]. Subsequent psychometric analyses in Europe[27, 46-50, 75-80], Asia[51, 52, 59, 81], Northern Africa[82] and the US[83] primarily have studied ACL or meniscus injury patients, although some studies have evaluated patients with knee osteoarthritis (OA)[27, 46-52]. Overall, these studies found that the KOOS met internal consistency and test-retest reliability standards for group-level comparisons and demonstrated satisfactory construct validity. However, to my knowledge, no psychometric analysis of the KOOS has been published for knee OA patients in the US. In addition, previous analyses have not examined measurement properties for subgroups of particular interest such as those below age 65, who represent an increasing proportion of TKR patients[9].

KOOS and WOMAC scales are scored using Likert's method of summated ratings, in which scores for items within a scale are simply summed (with or without imputation for missing data) to derive a scale score[57]. However, the simplicity of this method is based on a number of assumptions which should be tested prior to scoring a scale; such scaling tests were used in the development of the SF-36 Health Survey, for example[58]. These tests have not been applied to the KOOS or WOMAC with the exception of studies of the KOOS in Iran[59] and WOMAC in Singapore[60], both of which had much smaller samples (n<70) than recommended for tests of 24-item (WOMAC) or 42-item (KOOS) questionnaires. Tests of scaling assumptions also may provide a new perspective on the high (>0.70) correlations between the pain and function scales which have been observed in multiple studies[21].

This paper evaluated the KOOS in terms of data quality, item-level tests of scaling assumptions, internal consistency reliability, scale-level correlations, and floor and ceiling effects, using data from US knee osteoarthritis patients obtaining TKR. Parallel analyses were conducted for the WOMAC, since the complete WOMAC (LK 3.0) is included in the KOOS. Because widespread use of these measures assumes that their psychometric properties apply across diverse populations, analyses also were conducted for groups differing in sociodemographic and clinical characteristics.

Methods

Patients

Data were from the Function and Outcomes Research for Comparative Effectiveness in Total Joint Replacement (FORCE-TJR) registry, which is enrolling joint replacement patients across the US[41]. Those who could not provide consent due to cognitive impairment or who had emergency surgery were ineligible. All questionnaires were self-administered via a web-based or scannable paper-pencil survey at the surgeon’s office or at home prior to surgery, 6 months post-surgery and annually thereafter. This paper analyzed data from 1,179 randomly selected patients with knee osteoarthritis who elected unilateral TKR at FORCE-TJR high volume orthopedic centers between May 2011 and August 2012.

Questionnaire

The FORCE-TJR questionnaire included the KOOS, SF-36 Health Survey, and items about back pain severity and pain in the non-surgical knee and each hip joint. Additional questions asked about height and weight, chronic conditions (modified Charlson comorbidity index[84]), sociodemographics, and other patient characteristics.

The KOOS contains 42 knee-specific items measuring pain, other symptoms, function in activities of daily living (ADL), function in sport/recreation, and knee-related quality of life (QOL). Because all 24 WOMAC items (version LK3.0) are included in the 42-item KOOS, the WOMAC Pain (k=5), Stiffness (k=2) and Function (k=17) scales could be scored from the KOOS items that were administered in FORCE-TJR[22]. The KOOS Function in ADL and WOMAC Function scales are identical, while the KOOS Pain and Symptoms scales augment the WOMAC Pain and Stiffness scales with 4 and 5 additional items, respectively. (KOOS and WOMAC item content is abbreviated in Tables 2.2 and 2.3; see www.koos.nu for more information). In FORCE-TJR, KOOS items were asked in reference to the surgical knee. Most KOOS items used a recall period of the “last week”; the first KOOS pain and all KOOS QOL items had no recall period. All WOMAC items used a “last week” recall period. Following the developer’s scoring algorithms, KOOS scales were scored so 0 was the worst possible and 100 the best possible score[85]. To be comparable with the KOOS, WOMAC scales also were scored so 0 and 100 were the worst and best possible scores, respectively.

Analysis

The following measurement properties of the KOOS and WOMAC were evaluated using baseline (pre-TKR) data: (1) completeness of item and scale-level data; (2) tests of scaling assumptions; (3) internal consistency reliability; and (4) scale-level statistics including inter-scale correlations and floor and ceiling effects. Analyses were conducted using Stata 11.2.

Data completeness

The extent of missing data reflects patients' comprehension and acceptance of a survey and sets a limit on reliability and validity. For each item, the percent of patients with missing data was examined. In addition, the percent of patients who answered all items within each scale and the percent for whom a scale score could be calculated were computed. KOOS scale score calculations used standard KOOS scoring algorithms in which a scale score is computed if at least 50% of items within a scale are answered[85]. The WOMAC Function scale was calculated if at most three items were missing and the WOMAC Stiffness and Pain scales were calculated if at most one item was missing, following the developer's recommendations[86].

Tests of scaling assumptions

These tests evaluated whether it was appropriate to derive a scale score simply by adding item scores, as well as the empirical basis for including a specific item in a specific scale[87]. First, to be included in a scale, an item should be substantially linearly related to the total score computed from all other items in a scale (test of item internal consistency). Second, to avoid weighting of items, items in each scale should contain approximately the same proportion of information about the construct being measured (test of equality of item-scale correlations). Third, items within each scale should have roughly equivalent variances, to enable scoring without item standardization. In addition to these traditional Likert scaling criteria, correlations of an item with its hypothesized scale and all other scales were examined to determine the appropriateness of using the item to score one particular scale as opposed to another (test of item discriminant validity), following the logic of the multitrait-multimethod approach developed by Campbell and Fiske[88, 89]. Items that correlate substantially with two or more scales confound scales and complicate their interpretation.

Tests of scaling assumptions were evaluated using a multitrait/multi-item matrix that correlates each item with all hypothesized item groupings (i.e., scale scores). The Pearson product-moment correlation between an item and its hypothesized scale was estimated as if the item was not in the scale score (corrected for overlap), to avoid inflating the item-scale correlation coefficient[90]. Item internal consistency was evaluated by examining the correlation of each item with its hypothesized scale; a correlation of 0.40 or higher is considered substantial and generally accepted as satisfactory[91]. The equality of item-scale correlations was tested by examining correlations between all items in a scale and the hypothesized scale score. Item variances were compared within each scale to evaluate their equivalence[92]. Finally, item discriminant validity was supported to the extent that the correlation between an item and its hypothesized scale was significantly ($p<0.05$) higher than the correlations between that item and all other scales[91].

Reliability

Internal consistency reliability was estimated using Cronbach's coefficient alpha[93]. A minimum internal consistency reliability of 0.70 has been suggested for group-level analyses, and a minimum of 0.90 or 0.95 for individual-level data[94]. In addition, the average inter-item correlation (item homogeneity) was calculated. The reliability of a scale increases with the number of items and with higher item homogeneity[95]. Because the number of items in KOOS and WOMAC scales vary, evaluating reliability but not item homogeneity may be misleading.

Scale-level correlations

Internal consistency reliability can be thought of as a correlation between a scale and itself[87]. To the extent that a correlation between two scales is less than a scale's reliability, the scale has unique reliable variance[94]. To determine the extent to which each KOOS and WOMAC scale measured a distinct construct, correlations between scales were evaluated in relation to reliability coefficients.

Floor and ceiling effects

The proportion of patients scoring at the worst/lowest (floor) and best/highest (ceiling) levels also was examined. A high floor or ceiling effect may attenuate the correlation between scales. In addition, the ability of an instrument to detect change over time is constrained by the percent of respondents at the floor or ceiling.

Subgroup and 6-month post-TKR analyses

Psychometric properties were examined at baseline for subgroups defined by gender, age, education, income, body mass index, number of comorbid conditions (out of 14), back pain severity, and frequency of non-surgical knee and ipsilateral hip pain. Because electronic data capture is increasingly common, properties also were evaluated by method of data collection (paper, Internet). Internal consistency reliability was reported for all subgroups; other results were reported by subgroup only if they differed notably from overall results. In addition, because most patients benefit from TKR, 6-month post-TKR data was examined to confirm that psychometric properties did not change in a clinically improved patient population; 6-month results were reported only if they differed notably from pre-TKR data.

Additional analysis of Pain and Function discriminant validity

The KOOS and WOMAC Pain and Function scales contain items that ask about pain and difficulty while doing the same activity (e.g., pain walking, difficulty walking). This content overlap has been proposed as a reason for the high correlation between their Pain and Function in activities of daily living scales[96]. To further evaluate this hypothesis, the WOMAC Function items were scored as two subscales, one containing the eight items that have content overlap with the WOMAC Pain scale (Function-Similar) and the other containing the remaining nine items (Function-Dissimilar), following the logic proposed by Stratford[97]. A multitrait/multi-item correlation matrix of the WOMAC Pain items with the two WOMAC Function subscales was examined. It was hypothesized that the Pain items would have higher correlations with the Function-Similar subscale than the Function-Dissimilar subscale, due to item content overlap.

Results

The mean age of the sample was 66.1 (standard deviation 9.7); 57% were age 65 or older. Sixty-one percent were female, and 89% were white. The highest level of education was high school graduate for 24%, and 4% had not graduated from high school. Twelve percent reported an annual household income below $25,000.

Data Completeness

The amount of item-level missing data was low, generally ranging from 1-3% at baseline and 6-month follow-up (Table 2.1). However, 5-6% of patients did not answer the Sport/Recreation items about running and jumping at 6 months. All response choices were endorsed at both time points. The percent of patients who answered all items within a scale was greater than 90% for all scales and increased as the number of items in the scale decreased. Scale scores could be calculated for more than 99% of patients for all scales except the KOOS Sport/Recreation scale, which could be calculated for 97-98% of patients.

Using the developer's scoring algorithms, 1,158 patients (98.2%) had computable scale scores for all five KOOS scales pre-TKR; subsequent KOOS analyses were limited to these 1,158 patients. For the WOMAC, 1,169 patients (99.2%) had computable scale scores for all three scales pre-TKR; this group was evaluated in subsequent WOMAC analyses.

Tests of Scaling Assumptions

At baseline (pre-TKR), item-hypothesized scale correlations were 0.40 or greater for all KOOS items, meeting the test of item internal consistency (Table 2.2). However, item-hypothesized scale correlations for Symptom items S1 (knee swelling) and S2 (grinding, clicking) were close to 0.40. Within each KOOS scale, item-hypothesized scale correlations were approximately equivalent, with the exception of a lower item-hypothesized scale correlation for item QOL1 relative to other QOL items. Item standard deviations generally were similar within each KOOS scale except for item QOL1, which was highly skewed. In addition, standard deviations for the non-stiffness Symptom items were high (1.26-1.53).

All KOOS Sport/Recreation and QOL items had item-hypothesized scale correlations that were significantly ($p<0.05$) greater than all corresponding item-other scale correlations; that is, they demonstrated item discriminant validity. However, there were problems with item discriminant validity for the KOOS Pain, ADL and Symptom scales. Several KOOS Pain items (P4, P5, P6, P9) and one ADL item (A6) did not correlate significantly higher with their hypothesized scale than with all other scales. In addition, one Pain item (P6, pain on stairs) had a higher correlation with the ADL scale than with the Pain scale. Two KOOS Symptom items (S6, stiffness in morning; S7, stiffness later in day) had higher correlations with the KOOS Pain scale than with the Symptoms scale; several other KOOS Symptom items (S1, S2) also had high correlations with the Pain scale. Symptom item S6 also had a higher correlation with the KOOS ADL scale than with the Symptoms scale.

WOMAC items met all tests of scaling assumptions, except for the WOMAC Pain item P6 (pain on stairs) which had a significantly higher correlation with the Function scale than the Pain scale (Table 2.3). In addition, two Pain items (P5, P9) did not have significantly higher correlations with the Pain scale than with the Function scale, and three Function items (A4, A6, A12) did not have a significantly higher correlation with the Function scale than with the Pain scale, demonstrating a lack of item discriminant validity.

Results from tests of scaling assumptions were comparable across sociodemographic and clinical subgroups for both the KOOS and WOMAC, with one exception. Among those younger than age 55, item-hypothesized scale correlations for the KOOS Symptom items were notably lower (range of 0.25-0.50, with four correlations below 0.40); however, this group only had 139 patients. In addition, tests of scaling assumptions were comparable for pre-TKR and 6 month post-TKR data, also with one exception. At 6 months, KOOS Symptom items S2 (grinding, clicking) and S3 (knee catch or hang up) had low correlations with the Symptoms scale, with item-scale correlations of 0.29 and 0.32, respectively.

Additional analysis of Pain and Function discriminant validity

The multitrait/multi-item correlation matrix of the WOMAC Pain items with the Function-Similar and Function-Dissimilar subscales showed that while all item-subscale correlations were moderate (r=0.49-0.67), four of the five Pain items (P6-P9) had significantly higher correlations with the Function-Similar subscale than with the Function-Dissimilar subscale (Table 2.4). This suggests that the high Pain and Function inter-scale correlation can be explained in part by the overlap in activities across the two scales. In addition, the Pain and Function-Similar scale correlation (r=0.80) was higher than the Pain and Function-Dissimilar scale correlation (r=0.71). Internal consistency reliability of the Function-Similar and Function-Dissimilar subscales was 0.90 and 0.92.

Reliability

Internal consistency reliability of all KOOS and WOMAC scales exceeded 0.70 at baseline and was 0.95 for the KOOS Function in ADL and WOMAC Function scales (Table 2.5). Reliability statistics for each KOOS and WOMAC scale were similar across subgroups; however, reliability was below 0.70 for the KOOS Symptoms scale among those younger than age 55. Item homogeneity (average inter-item correlation) was lowest for the KOOS Symptoms scale (0.31) and highest for the KOOS Sport/Recreation (0.65) and WOMAC Stiffness (0.64) scales, with homogeneity for the remaining KOOS and WOMAC scales ranging from 0.46-0.54.

Scale-level correlations

KOOS scales were moderately to highly correlated, as were the WOMAC scales (Table 2.6). The correlation between the KOOS Pain and Function in ADL scales was r=0.78 and the internal consistency reliability of the KOOS Pain scale was α=0.88, indicating that most of the reliable variance in the Pain scale was shared with the ADL scale. Similarly, most of the reliable variance in the WOMAC Pain scale was shared with the WOMAC Function scale. The correlation between the KOOS Symptoms and Pain scales (r=0.67) also approached the reliability of the KOOS Symptoms scale (α=0.74).

Floor and ceiling effects

Prior to TKR, floor (percent with the worst/lowest possible score) and ceiling (percent with the best/highest possible score) effects were low (≤6%) to negligible for most scales, although there was a notable floor effect for the KOOS Sport/Recreation scale (Table 2.7). At 6 months, floor and ceiling effects also were low to negligible for many KOOS and WOMAC scales. However, ceiling effects approached 10% for the KOOS Function in ADL and WOMAC Function scales. In addition, the WOMAC Stiffness scale had a higher ceiling effect at 6 months post-TKR than the KOOS Symptoms scale, and the WOMAC Pain scale had a higher ceiling effect than the KOOS Pain scale. Among those scoring at the ceiling of the WOMAC Pain scale at 6 months, 13% reported some pain (monthly, weekly or daily) on KOOS Pain item P1 (knee pain frequency).

Discussion

This comprehensive evaluation of scaling assumptions underlying scoring of the KOOS and WOMAC scales yielded results supporting their use among US knee osteoarthritis patients obtaining TKR. Favorable results included low rates of missing data, satisfactory results from most tests of scaling assumptions, internal consistency reliability estimates that exceeded recommended standards for group comparisons, and floor and ceiling effects that generally were low and followed the pattern expected for measurements before and after surgery. It also is notable that results from these evaluations were consistent across groups differing in age and gender, socioeconomic status, and clinical status. However, tests of item discriminant validity for the Pain and Function in activities of daily living measures and their high inter-scale correlations call into question the empirical basis for the conceptual distinction between the pain and function measures. These and other issues are discussed below, along with recommendations for future research.

Notable among the issues that should be addressed is the lack of discriminant validity for some items in the KOOS Pain and Function in ADL scales and in the WOMAC Pain and Function scales. This overlap calls into question whether the pain items should be interpreted as measures of pain during physical function. Similar results demonstrating confounding of the pain and function measures have been seen with techniques as diverse as Rasch analysis[61], item-level exploratory factor analysis[96], and scale-level confirmatory factor analysis[98]. The current analysis adds to this literature through formal tests of item discriminant validity, which showed that many Pain items had high correlations with both the Function and the Pain scales, particularly the five Pain items included in both the WOMAC and KOOS.

The implications of these item-level discriminant validity tests also are apparent at a scale level. Correlations between the Pain and Function in activities of daily living scales approached the reliability of the Pain scale for both the KOOS and the WOMAC, suggesting that there is little unique reliable variance in the Pain scales. This might be expected due to the identical content describing physical activities across items in the Pain and Function scales, although the relatively high correlations between the WOMAC Pain items and Function-Dissimilar subscale indicates that the issue is broader than just the content of specific items. A pain scale that conceptualizes pain primarily in terms of its impact on physical activities will by definition be highly correlated with a physical function scale. A broader conceptualization of knee pain which encompasses its impact across a range of physical, emotional and role/social domains, as well as its severity, may result in greater discrimination between knee-specific measures of pain and physical function. Such an approach has been followed in development of the ICOAP, for example[99], and discussed in a recent review of pain measures used in TKR[39]. However, the nature of knee OA may be such that knee-specific measures of pain and function will be highly correlated, regardless of pain item content[100].

The KOOS Symptoms scale had relatively heterogeneous items with greater variability and lower average inter-item correlations, resulting in the lowest internal consistency reliability of the five KOOS scales, as has been seen in other studies[46, 52]. This pattern was even greater 6 months after TKR and among patients younger than age 55. Accordingly, many KOOS Symptom items did not demonstrate item discriminant validity. Although the WOMAC Stiffness scale performed better, it had a higher ceiling effect after TKR (16%) compared to 4% for the KOOS Symptoms scale. The pattern of results seen for the KOOS Symptoms scale is often observed for symptoms that largely vary independently. While the substantial correlation of the Symptoms scale with the Function scale supports the former scale’s validity, research should address whether this holds true for other data and other functional outcomes. Future analyses of the KOOS Symptoms scale may benefit from separate scoring and interpretation of its stiffness and non-stiffness components in addition to its overall scale score.

The KOOS Function in ADL and identical WOMAC Function scale consistently demonstrated sufficiently high reliability across subgroups (α=0.92-0.96) to support its use with individual patients. However, the homogeneity (average inter-item correlation) of the Function in ADL scale was comparable to that of other KOOS and WOMAC scales; thus, the comparatively higher reliability of the Function in ADL scale was in large part due to its length. For group-level comparisons, the practical implication of using a 17-item scale to measure function is that surveys are longer than needed for adequate measurement. Reducing the number of function items also could allow for inclusion of other items that may be important in evaluating TKR outcomes (e.g., psychological status) in studies without increasing overall respondent burden.

No differences were found in rates of missing data, tests of scaling assumptions, internal consistency reliability, or floor and ceiling effects for data collected over the Internet versus a traditional paper questionnaire (PQ). While some TKR patients may not be comfortable answering a survey electronically, PQ and Internet samples did not differ greatly in age (age ≥65 was 59% for PQ, 51% for Internet) or education (29% high school graduate or less for PQ, 26% for Internet). If data from TKR patients is to be routinely collected in a cost-effective manner in the future, electronic data capture will be required for at least part of the patient population. Results from this study, along with results from other KOOS[101] and WOMAC[102] studies and evaluation of PROMIS measures[103], are encouraging in terms of the quality of electronic data.

This study had a number of limitations. Data used in this analysis came from high volume orthopedic centers in the US only. While analyses should be replicated among TKR patients in other US settings, it is likely that results will be similar due to the diverse nature of FORCE-TJR patients. The types of analyses conducted in this paper also should be replicated in other countries, to increase understanding of the measurement properties of the KOOS internationally. In addition, the sample was 89% white and therefore analyses could not be replicated by race. While the racial distribution in the sample parallels current national TKR utilization, additional research is needed to evaluate the KOOS and WOMAC across racial and ethnic groups. Data for patients receiving bilateral knee replacement or other types of knee surgery were not examined, and results may not apply to knee OA patients with mild or moderate disease. Finally, other measurement properties of the KOOS and WOMAC, such as their validity and responsiveness, were not evaluated in this analysis; these are topics of subsequent chapters in this dissertation.

In summary, results of this analysis support use of the KOOS and WOMAC among knee osteoarthritis patients obtaining TKR in the US. However, interpretation of the KOOS Pain and Function in ADL scales, and the WOMAC Pain and Function scales, is confounded. Further research is needed to examine if these scales should be scored and interpreted as two separate measures or as a higher-order factor, using techniques such as confirmatory factor analysis. Item response theory also should be used to evaluate the relative difficulty of the KOOS pain and function items and the extent to which pain and function items that ask about the same activity may be redundant. Continued development and evaluation of pain measures that assess the impact of knee pain on domains in addition to physical function also is recommended.

Table 2.1: Percent missing item- and scale-level data and percent computable scales

	KOOS Scales					WOMAC Scales		
	Symptoms (k=7)	Pain (k=9)	ADL (k=17)	Sport (k=5)	QOL (k=4)	Stiffness (k=2)	Pain (k=5)	Function (k=17)
Pre-TKR (n=1,179)								
% of Missing Data per Item								
Minimum	0.7	0.6	0.5	0.9	0.5	0.7	0.6	0.5
Maximum	2.5	2.2	3.0	2.6	1.1	2.5	1.8	3.0
Median	2.1	0.8	0.8	1.7	0.7	1.6	0.8	0.8
% Scales w/Complete Data	92.0	93.6	90.8	95.6	97.8	97.3	96.7	90.8
% Computable Scales*	99.5	99.7	99.7	98.6	99.8	99.6	99.5	99.7
6-month Post-TKR (n=886)								
% of Missing Data per Item								
Minimum	0.1	0.4	0.3	1.6	0.4	0.1	0.4	0.3
Maximum	2.5	2.8	1.9	5.9	1.5	1.7	1.5	1.9
Median	1.5	0.9	0.7	2.8	1.1	0.9	0.9	0.7
% Scales w/Complete Data	93.2	91.9	91.3	91.8	96.7	98.3	96.7	91.3
% Computable Scales*	99.9	99.8	100.0	97.3	99.7	99.9	99.5	99.3

k=Number of items in scale.

* Percent for whom scale scores are computable using standard KOOS and WOMAC scoring algorithms.

Table 2.2: KOOS item means and standard deviations and correlations between items and scales, Pre-TKR (n=1,158)

				Item-Scale Correlation				
Item	**Abbreviated Item Content**	**Mean**	**SD**	**Symp-toms**	**Pain**	**ADL**	**Sport**	**QOL**
Symptoms								
S1	Swelling in knee[a]	2.33	1.29	0.40*	0.35[§]	0.27	0.21	0.30
S2	Grinding, clicking, noise[a]	2.59	1.26	0.41*	0.38[§]	0.31	0.26	0.35[§]
S3	Knee catch/hang up[a]	1.78	1.26	0.48*	0.40	0.32	0.23	0.30
S4	Straighten knee fully[b]	1.38	1.46	0.44*	0.37	0.26	0.19	0.17
S5	Bend knee fully[b]	1.78	1.53	0.47*	0.39	0.27	0.27	0.23
S6	Stiffness in morning	2.23	1.00	0.52*	0.59[†]	0.54[‡]	0.37	0.42
S7	Stiffness later in day	2.28	0.96	0.55*	0.58[‡]	0.54[§]	0.38	0.39
Pain								
P1	Frequency knee pain[c]	3.24	0.76	0.39	0.54*	0.44	0.31	0.42
P2	Pain twisting/pivoting	2.47	1.04	0.49	0.60*	0.53	0.47	0.46
P3	Pain straightening fully	1.78	1.11	0.61	0.69*	0.56	0.35	0.35
P4	Pain bending fully	2.15	1.15	0.61[§]	0.63*	0.52	0.40	0.39
P5	Pain walking on flat	1.88	0.95	0.40	0.64*	0.60[§]	0.33	0.45
P6	Pain up or down stairs	2.66	0.93	0.46	0.63*	0.64[‡]	0.48	0.51
P7	Pain at night in bed	1.69	1.06	0.43	0.62*	0.54	0.28	0.32
P8	Pain sitting or lying	1.51	0.96	0.48	0.69*	0.58	0.30	0.34
P9	Pain standing upright	1.95	0.96	0.45	0.67*	0.65[§]	0.38	0.48
Function in Activities of Daily Living								
A1	Descending stairs	2.42	0.97	0.41	0.55	0.63*	0.51	0.51
A2	Ascending stairs	2.38	0.98	0.39	0.57	0.66*	0.49	0.49
A3	Rising from sitting	2.17	0.96	0.46	0.58	0.72*	0.44	0.44
A4	Standing	1.85	0.95	0.41	0.63	0.70*	0.37	0.45
A5	Bending to floor	2.03	1.08	0.39	0.54	0.67*	0.41	0.41
A6	Walking on flat surface	1.72	0.92	0.40	0.63[§]	0.69*	0.37	0.46
A7	Getting in/out of car	1.98	0.90	0.39	0.61	0.76*	0.40	0.45
A8	Going shopping	2.14	0.97	0.39	0.61	0.72*	0.47	0.51
A9	Putting on socks/stockings	1.58	1.02	0.38	0.55	0.73*	0.36	0.38
A10	Rising from bed	1.76	0.98	0.47	0.60	0.76*	0.36	0.42
A11	Taking off socks/stockings	1.53	1.00	0.39	0.56	0.75*	0.36	0.38
A12	Lying in bed	1.61	1.00	0.44	0.62	0.69*	0.34	0.35
A13	Getting in/out of bath	1.74	1.11	0.38	0.55	0.72*	0.42	0.41
A14	Sitting	1.25	0.93	0.42	0.59	0.70*	0.30	0.35
A15	Getting on/off toilet	1.65	1.00	0.41	0.56	0.76*	0.39	0.41
A16	Heavy domestic duties	2.64	1.00	0.36	0.51	0.63*	0.55	0.46
A17	Light domestic duties	1.65	0.91	0.40	0.60	0.73*	0.43	0.51
Function in Sport and Recreation								
Sp1	Squatting	3.14	0.99	0.38	0.43	0.47	0.74*	0.44
Sp2	Running	3.47	0.84	0.32	0.36	0.39	0.79*	0.45
Sp3	Jumping	3.47	0.85	0.32	0.37	0.41	0.81*	0.44
Sp4	Twisting/pivoting	3.08	0.99	0.36	0.47	0.51	0.70*	0.46
Sp5	Kneeling	3.17	0.98	0.38	0.49	0.53	0.73*	0.47
Knee-specific Quality of Life								
QOL1	Awareness knee problem[d]	3.62	0.55	0.35	0.44	0.41	0.39	0.54*
QOL2	Modified life style[e]	2.77	1.00	0.34	0.42	0.45	0.44	0.67*
QOL3	Lack confidence in knee[e]	2.69	1.09	0.35	0.42	0.44	0.40	0.67*
QOL4	General difficulty knee	2.85	0.86	0.49	0.59	0.60	0.52	0.73*

* Item-hypothesized scale correlation corrected for overlap.

[†] Item-other scale correlation significantly ($p<0.05$) higher than item-hypothesized scale correlation.

[‡] Item-other scale correlation higher than item-hypothesized scale correlation.

[§] Item-other scale correlation not significantly ($p<0.05$) lower than item-hypothesized scale correlation.

Standard error=0.029.
Response options are 0=None, 1=Mild, 2=Moderate, 3=Severe, 4=Extreme except where noted:
[a] 0=Never, 1=Rarely, 2=Sometimes, 3=Often, 4=Always.
[b] 0=Always, 1=Often, 2=Sometimes, 3=Rarely, 4=Never.
[c] 0=Never, 1=Monthly, 2=Weekly, 3=Daily, 4=Always.
[d] 0=Never, 1=Monthly, 2=Weekly, 3=Daily, 4=Constantly.
[e] 0=Not at all, 1=Mildly, 2=Moderately, 3=Severely, 4=Extremely.

Table 2.3: WOMAC item means and standard deviations and correlations between items and scales, Pre-TKR (n=1,169)

				Item-Scale Correlation		
Item	**Abbreviated Item Content**	**Mean**	**SD**	**Stiffness**	**Pain**	**Function**
Stiffness						
S6	Stiffness in morning	2.22	1.00	0.64*	0.52	0.55
S7	Stiffness later in day	2.28	0.96	0.64*	0.53	0.54
Pain						
P5	Pain walking on flat surface	1.87	0.95	0.41	0.64*	0.60[§]
P6	Pain up or down stairs	2.65	0.93	0.46	0.57*	0.65[†]
P7	Pain at night in bed	1.68	1.06	0.45	0.62*	0.54
P8	Pain sitting or lying	1.50	0.96	0.48	0.70*	0.58
P9	Pain standing upright	1.95	0.96	0.45	0.67*	0.65[§]
Function						
A1	Descending stairs	2.41	0.97	0.43	0.54	0.63*
A2	Ascending stairs	2.37	0.98	0.41	0.58	0.66*
A3	Rising from sitting	2.17	0.96	0.54	0.57	0.72*
A4	Standing	1.85	0.95	0.44	0.66[§]	0.70*
A5	Bending to floor/pick up object	2.03	1.08	0.41	0.50	0.67*
A6	Walking on flat surface	1.72	0.92	0.44	0.68[§]	0.69*
A7	Getting in/out of car	1.98	0.90	0.46	0.59	0.76*
A8	Going shopping	2.14	0.97	0.42	0.62	0.71*
A9	Putting on socks/stockings	1.58	1.01	0.42	0.52	0.73*
A10	Rising from bed	1.76	0.98	0.58	0.58	0.76*
A11	Taking off socks/stockings	1.53	1.00	0.43	0.54	0.74*
A12	Lying in bed	1.61	1.00	0.47	0.64[§]	0.69*
A13	Getting in/out of bath	1.73	1.11	0.42	0.54	0.72*
A14	Sitting	1.25	0.93	0.45	0.60	0.70*
A15	Getting on/off toilet	1.64	1.00	0.46	0.56	0.76*
A16	Heavy domestic duties	2.63	1.00	0.40	0.48	0.63*
A17	Light domestic duties	1.65	0.91	0.45	0.61	0.73*

* Item-hypothesized scale correlation corrected for overlap.

[†] Item-other scale correlation significantly ($p<0.05$) higher than item-hypothesized scale correlation.

[§] Item-other scale correlation not significantly ($p<0.05$) lower than item-hypothesized scale correlation.

Standard error=0.029.

Response options are 0=None, 1=Mild, 2=Moderate, 3=Severe, 4=Extreme.

Table 2.4: Correlations of WOMAC Pain items with Function-Similar and Dissimilar subscales, Pre-TKR (n=1,169)

		WOMAC Scale		
Pain Item	**Abbreviated Item Content**	**Pain**	**Function Similar**[†]	**Function Dissimilar**[‡]
P5	Pain walking on flat surface	0.64*	0.61	0.56
P6	Pain up and down stairs	0.57*	0.67	0.58
P7	Pain at night in bed	0.62*	0.57	0.49
P8	Pain sitting or lying	0.70*	0.60	0.53
P9	Pain standing upright	0.67*	0.67	0.59

* Item-total correlation corrected for overlap. Standard error=0.029.

[†] Includes items A1, A2, A3, A4, A6, A10, A12, A14.

[‡] Includes items A5, A7, A8, A9, A11, A13, A15, A16, A17.

Table 2.5: Internal consistency reliability of KOOS and WOMAC scales, Pre-TKR

	KOOS						WOMAC			
	N	**Symp -toms**	**Pain**	**ADL**	**Sport**	**QOL**	**N**	**Stiff- ness**	**Pain**	**Func -tion**
Total Sample	1,158	0.74	0.88	0.95	0.90	0.81	1,169	0.78	0.84	0.95
*Homogeneity**	*1,158*	*0.31*	*0.46*	*0.53*	*0.65*	*0.54*	*1,169*	*0.64*	*0.51*	*0.53*
Gender										
Male	453	0.72	0.87	0.95	0.89	0.79	457	0.78	0.83	0.95
Female	705	0.74	0.89	0.95	0.91	0.82	712	0.78	0.84	0.95
Age										
<55	139	0.68	0.87	0.95	0.90	0.79	138	0.77	0.82	0.95
55-64	362	0.73	0.89	0.96	0.90	0.80	365	0.78	0.85	0.95
65-74	422	0.71	0.87	0.95	0.90	0.82	425	0.77	0.83	0.95
75+	234	0.74	0.89	0.94	0.89	0.78	240	0.79	0.82	0.94
Education[†]										
<= High School	325	0.74	0.89	0.96	0.93	0.82	327	0.82	0.84	0.96
Post-high school	337	0.77	0.89	0.95	0.91	0.80	342	0.79	0.84	0.95
College grad.	458	0.72	0.87	0.94	0.86	0.80	459	0.73	0.82	0.94
Income										
<$25,000	121	0.80	0.91	0.96	0.96	0.81	123	0.80	0.87	0.96
$25-45,000	216	0.73	0.89	0.96	0.89	0.83	217	0.82	0.85	0.96
>$45,000	655	0.73	0.87	0.94	0.88	0.82	659	0.76	0.82	0.94
BMI										
<25	152	0.76	0.89	0.94	0.90	0.77	153	0.80	0.82	0.94
25-29.9	356	0.74	0.88	0.95	0.89	0.82	362	0.78	0.82	0.95
30-34.9	332	0.73	0.87	0.95	0.89	0.80	332	0.76	0.84	0.95
≥35	301	0.74	0.89	0.95	0.91	0.81	306	0.79	0.85	0.95
Comorbid Conditions										
0	495	0.75	0.88	0.95	0.89	0.82	497	0.79	0.84	0.95
1	410	0.74	0.88	0.95	0.90	0.81	412	0.78	0.83	0.95
2+	253	0.74	0.89	0.95	0.91	0.79	260	0.77	0.83	0.95
Back Pain										
Never	560	0.73	0.88	0.94	0.89	0.80	567	0.79	0.83	0.94
Mild	266	0.73	0.86	0.94	0.88	0.80	269	0.71	0.80	0.94
Moderate	251	0.74	0.87	0.95	0.92	0.80	250	0.78	0.84	0.95
Severe	78	0.74	0.85	0.92	0.84	0.86	80	0.83	0.75	0.92
Non-Surgical Knee Pain										
None	311	0.75	0.89	0.95	0.91	0.79	315	0.83	0.85	0.95
Monthly/Weekly	312	0.74	0.86	0.94	0.87	0.84	314	0.74	0.81	0.94
Daily/Always	521	0.73	0.88	0.95	0.90	0.80	526	0.76	0.82	0.95
Ipsilateral Hip Pain										
None	693	0.73	0.88	0.95	0.90	0.81	701	0.79	0.83	0.95
Monthly/Weekly	258	0.73	0.88	0.95	0.89	0.79	261	0.77	0.84	0.95
Daily/Always	167	0.77	0.85	0.94	0.90	0.82	167	0.76	0.81	0.94
Survey Method										
Paper-pencil	889	0.74	0.88	0.95	0.90	0.79	899	0.77	0.83	0.95
Internet	269	0.75	0.89	0.96	0.90	0.86	270	0.82	0.86	0.95

* Average inter-item correlation.

[†] Education categories: <= High School=High school graduate or less; Post-high school=Some post-high school education but not 4-year college graduate; College graduate=4-year college graduate or higher.

Table 2.6: Correlations between scales and internal consistency reliability, Pre-TKR*

KOOS (n=1,158)						WOMAC (n=1,169)			
	Symptoms	**Pain**	**ADL**	**Sport**	**QOL**		**Stiffness**	**Pain**	**Function**
Symptoms	(0.74)					Stiffness	(0.78)		
Pain	0.67	(0.88)				Pain	0.58	(0.84)	
ADL	0.54	0.78	(0.95)			Function	0.60	0.77	(0.95)
Sport	0.42	0.51	0.55	(0.90)					
QOL	0.47	0.57	0.58	0.54	(0.81)				

* Cronbach's coefficient alpha on diagonal.

Table 2.7: KOOS and WOMAC scale-level descriptive statistics

	KOOS Scales						WOMAC Scales			
	N	**Symptoms**	**Pain**	**ADL**	**Sport**	**QOL**	**N**	**Stiffness**	**Pain**	**Function**
Pre-TKR										
Mean	1,158	48.6	46.3	52.8	18.3	25.4	1,169	43.6	51.8	52.9
SD		19.8	17.9	18.3	19.6	17.9		22.3	18.9	18.2
% Floor		0.7	1.3	0.7	28.9	14.4		6.1	1.3	0.7
% Ceiling		0.3	0.5	0.3	0.9	0.1		2.3	0.9	0.3
6-month Post-TKR										
Mean	859	73.9	80.0	81.6	48.0	63.3	878	71.1	83.9	81.7
SD		17.1	17.5	16.7	27.5	22.9		20.8	16.2	16.8
% Floor		0.0	0.0	0.0	4.8	0.8		1.0	0.1	0.0
% Ceiling		3.6	14.2	9.5	4.2	8.0		16.1	21.9	9.3

% Floor: percent with worst possible score, % Ceiling: percent with best possible score.
All scales are scored so 0=worst score, 100=best score.

CHAPTER III: DEVELOPMENT AND EVALUATION OF A KOOS FUNCTION ITEM BANK IN TOTAL KNEE REPLACEMENT PATIENTS

Abstract

Objective: To use item response theory (IRT) methods to calibrate the 22 function in Activities of Daily Living (ADL) and Sport/Recreation items from the Knee injury and Osteoarthritis Outcome Score (KOOS) questionnaire and conduct computerized adaptive testing (CAT) simulations with total knee replacement (TKR) patients.

Methods: 1,179 patients completed surveys before and 6 months after TKR. To represent different functional states, one survey per patient (pre- or post-TKR) was randomly selected for IRT analyses. IRT assumptions of unidimensionality, local independence, item monotonicity and differential item functioning were tested; items were calibrated using the graded response model. Real data CAT simulations were conducted on pre- and post-TKR data.

Results: IRT assumptions were supported. The item bank was most reliable -2.5 SD below to 1.7 SD above the combined pre/post-TKR sample mean. Sport items measured higher levels of function than ADL items but were less informative. In CAT simulations, a reliable score could be achieved for most patients in 3-8 items pre-TKR, but more items were needed post-TKR. One-third could not achieve a reliable (≥ 0.95) CAT score post-TKR because the item bank had few items at a high function level. Eight items accounted for most CAT administrations; most Sport items were rarely selected by CAT.

Conclusions: Knee-specific function items were unidimensional. CAT scores could be reliably estimated in less than 22 items, but additional items that measure higher levels

of function are needed for TKR. Use of IRT methods to test new items and evaluate new short function scales is recommended.

Introduction

Treatment of knee osteoarthritis (OA) is focused on reducing joint pain and stiffness, preserving and improving joint mobility, limiting joint damage, reducing disability, and improving health-related quality of life[3]. Therefore, patient reports of pain and function are central to the management of knee OA, including the decision to undergo total joint replacement when medical treatment is no longer successful[6] . The most frequently used joint-specific measure of these patient-reported outcomes (PRO) is the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), which contains 17 items measuring difficulty in function, along with 5 pain and 2 stiffness items[22, 23]. A body of research has found that the reliability of the WOMAC function scale is consistently higher than the minimum level required for group-level comparisons[21] and that there is redundancy in its item content[61, 62]. These findings indicate that the WOMAC function scale could be shortened, thereby reducing respondent burden and still meet group-level reliability standards needed for clinical research.

Several short function scales have been developed from the full-length WOMAC function scale. Methods used to select items for these 7 and 8-item scales have included patient ratings of item importance before and after total knee (TKR) and hip (THR) replacement[24]; patient and clinician ratings of item importance along with descriptive analyses of data from knee and hip OA patients[25]; and clinician recommendations and descriptive analyses of data from TKR and THR patients[26]. Only four items (ascending stairs, rising from sitting, walking on a flat surface, getting in/out of car) were selected for all three short function scales. None of these scales has been widely adopted.

While the WOMAC function scale is lengthy, at the same time it does not include items about activities that are more likely to be done by younger and more active OA patients. The 42-item Knee injury and Osteoarthritis Outcome Score (KOOS) questionnaire was developed in part to address this limitation of the WOMAC[45]. The KOOS includes the 17-item WOMAC Function in Activities of Daily Living (ADL) scale (version LK3.0), along with a 5-item joint-specific Function in Sport and Recreation scale that measures difficulty with more demanding physical activities[45, 104]. To reduce respondent burden, the KOOS-PS, a shorter form containing four ADL and three Sport items, was developed using Rasch analysis of KOOS data from 2,145 knee OA patients in five countries[28]. In a subsequent Rasch analysis, Franchignoni et al. corroborated that the KOOS ADL and Sport items could be combined into a single scale, but they were not able to replicate the KOOS-PS item selection process[67]. They concluded that additional work was needed to develop a robust KOOS short function scale.

The past two decades have seen increased use of Rasch and item response theory (IRT) models in PRO survey development. Rasch and IRT models can be used to calibrate item banks, which consist of a set of items measuring the same construct and parameters that describe the items' measurement properties[65]. Improved short forms can be constructed by selecting a subset of items from the bank based on the items' measurement properties. In addition to their use in developing short forms, item banks are also the foundation for computerized adaptive tests (CATs). Unlike fixed-length surveys such as the WOMAC and KOOS, CATs administer only the most informative items to each individual, resulting in more efficient and precise measurement[66].

Rasch models have been used in a number of studies to calibrate the WOMAC and KOOS function items[28, 61, 62, 67-69]. While these studies all have confirmed that the function items are unidimensional -- that is, that they measure a single underlying domain -- the studies differed in terms of the number of items that were excluded because they did not fit the Rasch model, as well as the values of the item calibrations. These inconsistencies in results have been attributed to the use of different Rasch software packages, differences in the samples examined, and differences in the criteria used to evaluate Rasch model fit[67]. Another possible explanation for these differences is that the Rasch model is not optimal for the WOMAC and KOOS function items. A separate class of IRT models are widely-used in the PRO field (such as in the NIH-sponsored PROMIS initiative[105]) but have not been used to analyze the WOMAC or KOOS function items, to the best of my knowledge. Unlike the Rasch model, in which item discrimination (or the model slope, α) is held constant across all items, these two parameter IRT models are often seen as better for use with polytomous items that have ordered response categories (such as the WOMAC and KOOS items), because they allow item discrimination to vary across items[70, 71]. Two parameter IRT models are particularly useful in CATs, which require a combination of items with steeper slopes (which are more discriminating over a narrower range) and items with less steep slopes (which are more useful in categorizing respondents over a wider range).

Because the KOOS function items have not been evaluated with IRT models, there is a gap in our understanding of their measurement properties. This paper will use IRT methods to determine if the KOOS function in ADL and Sport items define a unidimensional construct, and if these items can be calibrated in a function item bank. It then will use information from the item bank to conduct CAT simulations using data from knee OA patients prior to and after TKR, to better understand the performance of the KOOS function items in measuring outcomes for patients receiving surgical treatment. In addition, results of the CAT simulations will provide guidance as to which items might be most informative in a short fixed-length KOOS function scale.

Methods

Patients

Data for this analysis came from the Function and Outcomes Research for Comparative Effectiveness in Total Joint Replacement (FORCE-TJR) registry, which includes joint replacement patients from across the US[41]. Patients who could not provide consent due to cognitive impairment or who had emergency surgery were ineligible. All questionnaires were self-administered via a web-based or scannable paper-pencil survey at the surgeon’s office or at home prior to surgery, 6 months post-surgery and annually thereafter. This analysis included data from 1,179 TKR patients randomly selected from FORCE-TJR enrollees at high volume surgical centers between May 2011 and August 2012. Of these, data was available for 886 patients at the 6-month follow-up.

Questionnaire

TKR patients in FORCE-TJR completed the KOOS at baseline and 6 months. The KOOS contains 42 knee-specific items measuring pain, other symptoms, function in activities of daily living (ADL), function in sport/recreation (Sport), and knee-related quality of life (see Table 2.2 and www.koos.nu for more information). KOOS items were asked in reference to the surgical knee, and all ADL and Sport items had a recall period of the last week. ADL and Sport items used the same response scale (none--extreme). The KOOS was administered midway through the 140-item FORCE-TJR survey, after questions about health habits, surgical status, back pain and sociodemographics, the SF-36 Health Survey[31], and a modified Charlson comorbidity index[84]. The majority of respondents (76.9%) completed the paper-pencil version of the survey, with the remainder (23.1%) completing the survey via the Internet.

Analysis

IRT methods were used to calibrate the KOOS function in ADL and Sport items on a common metric, and the resulting item calibrations were applied to pre-TKR and post-TKR patient data to conduct real data CAT simulations. Prior to calibrating the items, assumptions of unidimensionality, local independence, and monotonicity of item response categories were tested, and differential item functioning (DIF) also was examined[71, 106, 107]. Unless otherwise indicated, analyses were performed in Stata 11.2.

Unidimensionality. IRT models assume that all items within an item bank are unidimensional; that is, that they measure a single underlying construct. To evaluate unidimensionality, both exploratory and confirmatory factor analyses of all 22 ADL and Sport items were conducted. Exploratory factor analysis (EFA), using a principal components model, was performed to evaluate the structure of the data without assuming any pre-specified assignment of items to factors. Because the KOOS items are ordinal, the factor analysis was based on the matrix of polychoric correlations among all 22 items[71]. The number of factors that were retained was determined by a number of criteria, including Kaiser's eigenvalue-greater-than-one rule and whether a minimum of 5% of the total variance was explained by the factor. If loadings on the first factor were substantial (≥0.70) and greater than loadings on the second factor, that supported calibrating all 22 KOOS items in a single item bank.

As a second test of unidimensionality, and following the approach often used in item bank development[71, 106], item-level categorical confirmatory factor analysis (CFA) was performed using Mplus software[108] and weighted least square methods for factor analysis of categorical data[109, 110]. Overall model fit was evaluated with the Comparative Fit Index (CFI) and the root mean square error of approximation (RMSEA) for categorical data, with commonly-used (CFI≥0.95, RMSEA≤.08) criteria used to judge model fit. Three CFA models were evaluated: (1) 1 factor (all 22 items); (2) 2 correlated factors (17 ADL items, 5 Sport items); and (3) a bifactor model (1 common factor of all 22 items, plus ADL and Sport group factors). If the 1-factor model fit sufficiently and all items had high (≥0.70) factor loadings, this supported evaluating the ADL and Sport items in one bank; it was anticipated that the 1-factor solution would fit the data, based on previous research[28, 67]. In the 2-factor solution, items were assigned to factors based on how they were presented to respondents, either as items measuring "physical function" (17 ADL items) or "physical function when being active on a higher level" (5 Sport items). A high correlation between the two factors supported analyzing all 22 items in a single bank. Finally, the bifactor model is another method used to determine if items are "sufficiently" unidimensional[71, 111]. In a bifactor model, each item loads on a common factor (all items) and on a group factor (a subset of items). If item loadings on the common factor were substantial and substantively larger than corresponding item loadings on the group factor, then the items were seen as sufficiently homogeneous.

If all 22 items were viewed as unidimensional, then any items with loadings <0.70 in the 1-factor CFA were eliminated from further analysis, as in previous studies[71, 107].

***Local independence*.** IRT models assume that there are not any significant associations among pairs of items once a respondent's overall level on a domain is controlled for; that is, that items are locally independent[71]. Local independence was evaluated in the final CFA model by examining residual correlations of all item pairs. A residual correlation of ≥0.20 was used to identify items with possible local dependence, as in previous analyses[71, 106]. IRT parameter estimates may be inflated if items that have local dependence are included in an item bank[71], and therefore a strategy to handle these items needed to be determined. One approach is to delete items showing local dependence from the bank. However, this approach was not followed because a primary purpose of this analysis was to better understand the measurement properties of all 22 ADL and Sport items. Therefore, another approach, which has been used in calibrating SF-36 physical functioning items which are locally dependent[107], was followed, in which locally dependent items were modelled separately during item calibration (see below).

***Item monotonicity*.** IRT models assume that, for each item, each response choice has a maximum probability of being selected over a unique interval of a scale[106]. The monotonicity of response choices was evaluated on an item-by-item basis, using nonparametric kernel-smoothing techniques and the TestGraf software[112, 113]. If a response option curve did not have a clear maximum, that response option was combined with an adjacent response option, for purposes of IRT modeling of the item.

***Differential Item Functioning (DIF)*.** IRT models assume that the probability of responding to an item in a certain way is the same for all people who are at an equivalent level of the underlying domain. If the likelihood of answering an item in a particular way is related to another factor (e.g., gender), DIF is present. DIF was examined for groups defined by age (<55, 55-64, 65-74, 75+), gender, and education (less than high school, high school graduate, some post-high school education, college graduate), using ordinal logistic regression models in which the item response was the dependent variable and the total sum score (of all items) and the group indicator were the independent variables[114, 115]. A significant effect of the group variable indicated uniform DIF, while a significant interaction effect (between the group variable and the sum score) indicated non-uniform DIF[116, 117]. The magnitude of DIF was evaluated with the coefficient of determination R^2 as developed by Nagelkerke[118] (ΔR^2); a $\Delta R^2 > 0.03$ (combined uniform and non-uniform DIF) indicated notable DIF as in previous analyses[107].

Item calibration. Items were calibrated with the graded response model (GRM)[119, 120], a unidimensional, polytomous IRT model, using the software program IRTPRO[121, 122]. The first GRM model excluded items that demonstrated local dependence. Item parameters were fixed for all items included in the first IRT model, and then a second GRM model was run to estimate item parameters for the remaining items. Item fit was assessed using S-X^2 fit statistics, which quantify the difference between observed and expected item response frequencies at various score levels[123, 124]. Items were calibrated to have a mean of 0 and standard deviation (σ) of 1 in the TJR population. To be consistent with KOOS scale scoring, in which a higher scale score indicates better function, items were recoded so a higher item score indicated better function.

After items were calibrated, slopes (measuring item discrimination) and threshold parameters (measuring item difficulty) were examined. The number of threshold parameters equals the number of response choices minus one. For GRM, a threshold parameter indicates the location along the latent trait continuum (or theta, θ) where the probability is 50% or more that a respondent selects a particular response choice or a higher response choice. Examination of threshold parameters indicates the extent to which item responses span the full measurement range and also indicates whether an item is harder or easier. Item slopes measure an item's discrimination, that is, how quickly the probability of endorsing a particular response choice increases at any given level of the latent trait[65]. Items with higher slopes provide more information about the likelihood of endorsing one response choice versus another along the latent continuum.

After item calibration, the item information function (IIF) was calculated and graphed for each item[125]. The IIF indicates the degree of precision that an item provides at each level of the latent trait; a higher level of information indicates lower measurement error[65]. The sum of the IIFs is the test information function [I(θ)], which measures the information provided by all items in a bank at each level of the latent trait. The test information function allows for estimation of the standard error (SE) and reliability of an item bank across all levels of theta, where $SE=1/\sqrt{I(\theta)}$ and $reliability=\sigma^2/(\sigma^2+SE^2)$. For example, if $\sigma=1$, a standard error of 0.32 is comparable to a reliability of 0.90; a standard error of 0.23 is comparable to a reliability of 0.95. Reliability was assessed across all levels of the latent trait, to determine the range of theta scores across which reliability was ≥0.95 (often recommended as a minimum level of reliability when measures are used with individual patients) and ≥0.90 (a less conservative reliability level that also has been used)[94].

Finally, for each FORCE-TJR patient, total item bank scores (i.e., theta scores) were estimated, based on the responses to all KOOS function items that the respondent answered and the item parameters (i.e., slope, difficulty) of those items. Theta scores were calculated for all patients prior to TKR and 6 months post-TKR

CAT Simulations**.** Real data CAT simulations were conducted using the pre-TKR and 6 month post-TKR data to approximate how well the item bank would perform in a CAT, using data from FORCE-TJR patients and CAT simulation software FIRESTAR[126, 127]. A simulated CAT selects the most informative items for each respondent, using actual data and pre-set stopping rules, to derive a CAT score for that respondent[66, 128]. Two CATs were run on the pre-TKR data, with different stopping rules: (1) CAT stopped when the standard error of the score (SE) was ≤0.32 (equal to a reliability of 0.90) or a maximum of 10 items were administered (to limit respondent burden); and (2) CAT stopped at SE≤0.23 (reliability=0.95) or a maximum of 10 items. Both CATs required that a minimum of 2 items be administered. The first item (start item) in the CAT was determined by FIRESTAR, based on the mean of the prior distribution[127]. Two similar CATs (SE≤0.32, SE≤0.23; maximum of 10 items) also were run on the post-TKR data. Results of the CAT simulations were examined to determine how often each item was selected by the CAT, as well as the number of items administered to each respondent and the percentage of CAT scores that met the criteria of SE≤0.32 and SE≤0.23.

Results

The mean age of the sample was 66.1 (SD=9.7); 57% were age 65 or older, while 12% were younger than age 55. Sixty-one percent were female. The majority (89.8%) were white, while 7.6% were black and 2.6% were another race. The highest level of education was high school graduate or less for 28%, while 39% were college graduates or had post-college education.

Descriptive statistics and creation of analytic dataset. Prior to TKR, most patients reported substantial impairment on the Sport items. All five Sport items were negatively skewed (range -1.00 to -2.16), and nearly 90% of patients reported “severe” or “extreme” difficulty running and jumping (Table 3.1). The high level of disability before TKR can be seen in the knee pain frequency item (P1), with more than 92% reporting knee pain “daily” or “always”. In contrast, most ADL items had mean values ranging from 2.5 to 3.5

(on a 1=never to 5=extreme scale) and little skewness. However, 6 months after TKR, many of the ADL items showed substantial positive skewness, with ≤5% of patients reporting “severe” or “extreme” difficulty in most ADL activities (Table 3.2). The Sport items had a more symmetrical distribution and little skewness 6 months after TKR.

While this favorable shift in the distribution of responses to the function items is to be expected after successful TKR, it has implications for calibrating item banks, in terms of having sufficient data for all items at all response levels. Many item by response choice cells had fewer than 25 patients prior to TKR, and many item by response choice cells had fewer than 10 patients in the 6 month post-TKR data. For that reason, an analytic dataset was created by randomly selecting either pre-TKR or 6-month post-TKR data for each patient; the resulting dataset contained pre-TKR items for 2/3 of the sample (n=786) and 6-month post-TKR items for the remaining 1/3 (n=393). The combined dataset had substantial numbers of patients in each item by response choice cell, and skewness was reduced in comparison to pre-TKR and post-TKR data (Table 3.3). This dataset was used to evaluate and calibrate the ADL and Sport items.

***Exploratory factor analysis*.** In the combined dataset, eigenvalues for the first two unrotated factors were 13.88 and 1.91, and the percentage of total variance explained was 63.1% and 8.7%, supporting extraction of two factors (Table 3.4). Loadings on the first unrotated factor ranged from 0.74 to 0.88 for the ADL items, and ADL item loadings were considerably higher on the first factor than on the second factor. For the Sport items, loadings on the first unrotated factor ranged from 0.69 to 0.78, and loadings on the second unrotated factor ranged from 0.40 to 0.62. Loadings were always higher on the first factor than on the second factor for all Sport items, but loadings were particularly high on the second factor for Sport items Sp2 (running) and Sp3 (jumping). Therefore, while the preponderance of evidence supported analyzing all 22 items in a single bank, particular attention was paid to Sport items Sp2 and Sp3 in subsequent analyses.

***Confirmatory factor analysis*.** Confirmatory factor analysis (CFA) of the ADL and Sport items demonstrated that all 22 items had high (0.77-0.94) loadings in the 1-factor model (Table 3.5), indicating that the 22 function items measured a unidimensional construct. These 22 items explained 72.6% of the variance in the data. The CFI for the 1-factor model approached 0.95, which indicated good model fit; the RMSEA (0.158) was high in relation to accepted standards, although RMSEA is often high in CFAs of PRO data[129]. There were 22 item pairs of residual correlations that were 0.20 or higher (see below, Local Independence). Fit of the 1-factor model was not notably improved by allowing ADL items with similar content (ascending stairs/descending stairs) to correlate. Model fit also was not notably improved in the 2-factor model (CFI=0.954, RMSEA=0.142); the ADL and Sport factors had a correlation of 0.764. Model fit was improved in the bifactor model (CFI=0.979, RMSEA=0.101). While the RMSEA of 0.10 was a bit higher than the criteria of 0.08, it was lower than in the 1- and 2-factor models and RMSEA is often slightly higher than conventional criteria in CFAs of ordinal PRO items[129]. In the bifactor model, factor loadings on the common factor were high (0.66-0.92) and 20 of the 22 items (all but A09 and A11) had a higher loading on the common factor than on a group (ADL or Sport) factor. Four items (A09, A11, Sp2, Sp3) had high loadings on a group factor. However, these four items also had very high (0.92-0.94) factor loadings in the 1-factor model, indicating that they also were strong measures of a unidimensional construct. Overall, the results indicated that the 22 ADL and Sport items were sufficiently unidimensional to be considered homogeneous for purposes of item bank calibration.

***Local independence*.** In the 1-factor model, there were 22 pairs of items (out of 231 total pairs) that had residual correlations of 0.20 or greater. All 22 item pairs included one or more of the items A09 (putting on socks), A11 (taking off socks), Sp2 (running), or Sp3 (jumping), which were the same items that had high loadings on the group factors in the bifactor CFA model (see above). The activities in items A09 and A11 are so comparable that most respondents may have provided similar answers to them. Similarly, while running (Sp2) and jumping (Sp3) are more distinct, it may be that both of these items were seen as high impact activities by TKR patients and thus responses to both items tended to be similar. When items A11 and Sp3 (i.e., the latter item in each item pair) were removed from the 1-factor CFA, only one residual correlation was ≥0.20 (A14 (sitting) with Sp2 (running) r=0.226; model CFI=0.958, RMSEA=0.130). Because there was only one high residual correlation out of 190 item pairs in the 20-item CFA (less than what might be expected by chance), and because there was no content reason for items about “sitting” and “running” to be locally dependent, items A14 and Sp2 were both included in the first IRT calibration model. However, to address issues of local dependence, items ADL11 and Sp3 were not in the first IRT model (see below).

Item monotonicity. Evaluation of trace curves for item response categories supported the monotonicity of most ADL and Sport items; each response choice had a maximum probability of being selected over a unique interval of its scale. An example is provided in Figure 3.1 for ADL04 (difficulty standing); there is a unique range across the ADL scale (on the x-axis, ranging from 17-84) for which each of the 5 response choices was most likely to be selected. However, monotonicity was not seen for some items. For ADL12 (lying in bed), there was not a unique range for which response choice 4 (“severe”) was most likely to be chosen, with the curve for response choice 4 totally subsumed under the curve for response choice 3 (“moderate”). Similarly, there was not a unique range for which response choice 2 (“mild”) was most likely to be selected for both Sport items 2 (running) and 3 (jumping) (Figure 3.1 continued, on the x-axis ranging from 5-25).

For purposes of IRT modeling, response choices were collapsed if a response curve did not have a clear maximum and a unique range for which it was most likely to be selected. Response choices collapsed were: “moderate” and “severe” for item ADL12 (lying in bed), and “mild” and “moderate” for Sport items 2 (running) and 3 (jumping).

Differential item functioning. Differential item functioning (DIF) was evaluated for groups differing in age, gender, and education. DIF was not found for any item in any test; the ΔR^2 (test of combined uniform and non-uniform DIF) was below 0.03 for all items in all tests (Table 3.6). Although no DIF was found using pre-specified and conventional criteria, women were significantly more likely than men to report difficulty kneeling at a given level of function (odds ratio OR=1.88 (95% CI 1.46, 2.43)) and were less likely than men to report difficulty twisting/pivoting on their surgical knee (OR=0.67 (0.52, 0.86)). College graduates also were more likely than other education groups to report difficulty descending stairs (OR=1.53 (1.15, 2.02)). From a content point of view, it is unclear why women would report less difficulty twisting/pivoting on their knee but more difficulty kneeling, or why higher education would be related to more difficulty climbing stairs. Furthermore, given the large number of DIF tests, these results are only slightly more than would be expected by chance.

Item bank calibration**.** The initial IRT model included 19 ADL and Sport items; it did not include items ADL11 (taking off socks), Sp3 (jumping), and ADL02 (ascending stairs). ADL11 and Sp3 were excluded due to issues of local dependence (as discussed above), while ADL02 was excluded due to content overlap with item ADL01 (descending stairs). All 19 items in the initial model showed satisfactory fit. Therefore, item parameters for these 19 items were fixed, and then item parameters for the remaining 3 items (ADL02, ADL11, Sp3) were estimated in a second IRT model that included all 22 items. Item fit (S-X^2) statistics are presented in Table 3.7; almost all items fit the model well.

Across all 22 ADL and Sport items, discrimination parameters (slopes) ranged from 1.72 to 3.66 (mean=2.70), with higher slopes for the ADL (slope=2.23-3.66) items than the Sport (slope=1.72-2.17) items (Table 3.7). Difficulty parameters for the ADL items ranged from -2.69 (threshold between “extreme” and “severe” difficulty sitting, item ADL14) to 1.55 (threshold between “none” and “mild” difficulty with heavy domestic duties, item ADL16); only 23 (34.3%) of the 67 ADL item thresholds were positive and only 5 of these thresholds were greater than 1.0. For the Sport items, difficulty parameters ranged from -0.68 (threshold between “extreme” and “severe” difficulty twisting/pivoting, item Sp4) to 2.64 (threshold between “none” and “mild” difficulty kneeling, item Sp5); 13 (72.2%) of the 18 Sport item thresholds were positive and 9 of these were greater than 1.0. Thus the ADL items primarily measured at a lower level of function and the Sport items primarily measured at a higher level.

The ADL items provided more information than the Sport items, with a mean maximum information value of 2.44 (range 1.42-3.66), compared to maximum information values ranging from 0.87 to 1.45 for the Sport items. (See column labeled I_{max} at Θ in Table 3.7; for example, the maximum information for item ADL02 was 2.34 at a theta (Θ) score of 0.4). However, the Sport items had their greatest information value at a higher range of the latent trait continuum than most ADL items. Sport items achieved their maximum item information at a theta score ranging from 0.6-0.9, and were the only items (aside from ADL16, heavy domestic duties) to provide information at a theta score higher than 2.0. In contrast, the maximum item information for most ADL items was provided at a negative theta score, and information for most ADL items dropped sharply above a theta score of 1.0. This pattern is illustrated for two items in Figure 3.2. Item ADL17 (light domestic duties) provided information starting at a theta score of about -2.7 but its information value dropped off sharply after a theta score of 1.0. In contrast, item Sp3 (jumping) did not provide information over most of the negative theta score range but did provide information throughout the positive score range. Item information curves for all 22 items are provided in Figure 3.5.

Evaluation of the test information function showed that the total item bank was most reliable (reliability≥ 0.95, or SE≤0.23) over the range of theta scores from -2.5 to 1.7 (Figure 3.3). Item bank reliability dropped sharply at a theta score greater than 1.7 and was below 0.90 at a theta score greater than 2.2. Thus, the total item bank was more reliable at lower levels of function. For this reason, the KOOS function items better matched the function levels of the sample prior to TKR, when 70.7% of the sample had a negative theta score (Figure 3.4). Six months after TKR, 85.3% of the sample had a positive theta score; 46.4% had a theta score ≥1.0 and 15.7% had a theta score ≥1.7.

The mean theta score (calculated from the 22-item bank) was -0.40 (SD=0.74) prior to TKR and 0.90 (SD=0.85) 6 months after TKR, where 0 is the combined (pre-TJR/post-TJR) sample mean and 1 is the standard deviation. Prior to TKR, theta score reliability was ≥0.90 for 99.2% of the patients and ≥0.95 for 98.6%. Six months after TKR, theta score reliability was ≥0.90 for 94.4% of the patients and ≥0.95 for 84.4%.

***CAT Simulations*.** When the CAT was instructed to stop once the SE for a score was ≤0.32 (equivalent to reliability≥0.90) or at a maximum of 10 items, patients answered an average of 3.4 items pre-TKR, with 97% achieving a SE ≤0.32 in 3 to 5 items (Table 3.8). More items were required post-TKR to achieve a SE ≤0.32, with a mean number of

5.0 items answered. However, 8.5% of patient scores did not reach a SE ≤0.32 post-TKR (reliability≥0.90), even after 10 items were administered.

When the higher reliability standard of achieving a SE≤0.23 (reliability≥0.95) or a maximum of 10 items was applied, patients answered a mean of 6.9 items pre-TKR, with most patients answering 6 to 8 items. However, the SE was higher than 0.23 for 4.2% of patients pre-TKR, even after they had answered 10 items, indicating that CAT scores for these patients did not achieve a reliability of 0.95. Post-TKR, patients answered a mean of 8.3 items, and 34% of patient scores did not have a SE≤0.23 (reliability of 0.95) even after answering 10 items. Patients whose scores had a reliability <0.95 post-TKR generally had relatively high function (median post-TKR CAT score=1.67, interquartile range=1.43, 2.03). In contrast, patients whose scores had a reliability of ≥0.95 post-TKR generally had lower function (median CAT score=0.63, interquartile range=0.15, 0.95).

In the pre-TKR data, three items (going shopping, light domestic duties, walking on flat surface) accounted for 74.7% of item administrations in the CAT using the SE≤0.32 stopping rule, with two other items (rising from bed, standing) accounting for another 17.7%. The same five items accounted for 69.7% of item administrations pre-TKR when the SE≤0.23 stopping rule was applied. There was more variation in item usage post-TKR, with five items (going shopping, getting in/out of a car, light domestic duties, rising from sitting, heavy domestic duties) accounting for 62.6% of item administrations using the SE≤0.32 stopping rule. Sport items accounted for less than 1% of administrations pre-TKR and 13% of the administrations post-TKR using either stopping rule. Five items (putting on socks, taking off socks, lying in bed, getting in/out of bath, sitting) each accounted for less than 1% of item administrations in all four CAT simulations.

In simulations where the CAT stopped once the SE was ≤0.32 (reliability≥0.90) or at a maximum of 10 items, the mean CAT score was -0.39 (SD=0.72) pre-TKR and 0.89 (SD=0.83) 6 months post-TKR. In CAT simulations where the CAT stopped once the SE was ≤0.23 (reliability≥0.95) or at a maximum of 10 items, the mean CAT score was -0.41 (SD=0.73) pre-TKR and 0.92 (SD=0.84) 6 months after TKR. Thus, on average, CAT scores were similar to theta scores for the full 22-item bank (mean theta score (SD) pre-TKR=-0.40 (SD=0.74); mean theta score (SD) 6 months after TKR=0.90 (SD=0.85)).

Discussion

This analysis demonstrated that the 22 KOOS function in ADL and Sport items defined a unidimensional construct and could be calibrated on a common metric using an IRT model. As would be expected from their content, the ADL items primarily measured at a lower level of function and the Sport items primarily measured at a higher level; the full item bank was more reliable at lower levels of function. The KOOS item bank could be used to successfully conduct CAT simulations, thereby achieving precise and much more efficient measurement of knee-specific function for most patients. Results of the simulations also indicated that a relatively few items accounted for the majority of CAT administrations. However, the simulations also showed the limitations of conducting a computerized adaptive test using the KOOS function items. These findings and related methodological issues are discussed below.

In the CAT simulations, a reliable function score could be achieved in 10 items or less for almost all patients prior to TKR and for a substantial proportion of patients 6 months after TKR. Prior to TKR, a reliable score (reliability ≥0.95 or ≥0.90) could be estimated for 96-99% of patients using CAT, with a 55-86% reduction in respondent burden compared to answering all 22 KOOS function items. However, the number of items needed to achieve a reliable score increased after TKR, and one third of patients could not achieve a function score with a reliability of 0.95 six months post-TKR, even after administration of 10 KOOS items. These patients tended to score at a higher level of function, where there were relatively few informative items to administer. As illustrated in Figure 3.4, 6 months after TKR nearly 25% of patients had a theta score of 1.5 or higher. Only six items -- ADL16 (heavy domestic duties) and all five Sport items -- had at least one item threshold above 1.5. Thus there was a ceiling effect post-TKR -- a high percentage of patients achieving the best possible score -- due to the mismatch between where a notable proportion of patients were located and where the KOOS items were located. This mismatch between people and items post-TKR would be even greater for the WOMAC, because the WOMAC does not include any Sport items.

While the KOOS Sport/Recreation scale was developed to add activities that were more difficult than the ADL items, the performance of the Sport items in the TKR data was mixed. The Sport items did measure a higher level of function than almost all of the ADL items. However, the Sport items had relatively low item information, did not discriminate well (i.e., had low slopes) and were not administered frequently in CAT simulations. Thus while the Sport items extended the measurement range, they did not fully meet the need for more difficult items to administer to TKR patients. Additional items that extend the measurement range to measure a higher level of function are needed in TKR. This can be done by developing new knee-specific items with more difficult item content[130] or by modifying the response choices of existing items to measure higher levels of function[131]; both of these approaches have been used to extend the measurement range of generic (general, not applying to any specific disease) items. These results may also have implications for studies of patients with milder knee OA who are not substantially impaired, as well as for studies of treatments that are anticipated to move knee OA patients into the positive end of the function spectrum.

Conclusions from this study about the Sport/Recreation items should not be generalized to patients with other knee disorders, however. KOOS Sport/Recreation items were selected to add specific activities that were affected by knee problems but not included in the WOMAC, as well as to extend the WOMAC's range of measurement. The KOOS initially was developed and tested on young and middle-aged patients with anterior cruciate ligament injuries, meniscus injuries and early knee osteoarthritis[45] and was only later applied to patients undergoing total knee replacement[27, 104]. Empirical evidence from this study showed that the Sport items raised the ceiling among TKR patients, but not as effectively as needed. However, these items may perform better in a younger, more active population of patients with early knee OA or other knee disorders. Evaluation of all 22 KOOS function items among other groups of knee patients, using IRT methods similar to those used in this study, is recommended.

Only two of the seven KOOS-PS items (rising from sitting, rising from bed) were selected frequently in the CAT simulations, raising the question as to whether the optimal set of items was selected for this KOOS short form. The process of selecting items for the KOOS-PS was in large part driven by Rasch DIF analyses, which resulted in nine of the 22 ADL and Sport/Recreation items being eliminated due to age and/or gender DIF and another six items being eliminated due to country DIF[28]. Underscoring the importance of replication in tests of DIF, none of the items that demonstrated age or gender DIF in the KOOS-PS development study[28] had notable age or gender DIF in this analysis. Franchignoni et al. also did not find age or gender DIF for 18 of the 22 KOOS function items, including all seven KOOS-PS items (DIF results for the other four function items were not reported)[67]. Researchers have cautioned against large-scale elimination of items from an item bank based on finding statistically significant DIF alone, without considering whether the DIF has substantive or clinical meaning[67, 70]. In addition, the variation in slopes between the ADL (2.23-3.66) and Sport (1.72-2.17) items suggests that the Rasch model (which sets all slopes to be equivalent) may not be appropriate for the KOOS function items. Therefore, IRT methods are recommended when constructing any new short function scales from the KOOS. CAT simulations also indicated that other items which are not in the KOOS-PS (e.g., ADL08 going shopping, ADL17 light domestic duties) may be good candidate items for a short form because they were selected frequently for the CAT.

The IRT analyses benefitted from use of a combined sample of pre-TKR and post-TKR data. As a result, the analyses included data from patients with widely varying function levels, ranging from patients with severe functional limitations (pre-TKR) to those with high functioning (post-TKR). If the analyses had been conducted with only pre-TKR data or only post-TKR data, the sample size would have been sparse in many item by response choice cells, which is not optimal for establishing stable IRT parameter estimates[120]. Item frequency distributions are often not described in Rasch and IRT analyses, and thus the potential impact of any lack of sample diversity is unaddressed. At a minimum, reporting of summary item descriptive statistics (e.g., skewness) is recommended.

This analysis has a number of limitations. First, patients were from high volume orthopedic centers in the US only. Similar analyses using IRT methods should be replicated in other patient groups, particularly with patients from other countries. Second, the calibrations presented in this article are centered (mean=0, SD=1) on a particular sample (2/3 pre-TKR, 1/3 post-TKR) and thus the absolute values of the threshold parameters reported here are not generalizable to other populations. Ultimately, establishment of a standard metric for calibrations of the KOOS function items would need to be based on a representative and well-defined patient population, as has been done in other diseases[106]. Third, data was collected by both paper-pencil and PC methods. Analyses were not conducted separately by method of administration because the PC sample (n=272) did not meet minimum sample size recommendations for IRT analyses[71]. However, a meta-analysis of PRO studies that compared paper-pencil and PC methods of data collection found that scale scores generally were equivalent across these methods of administration[132]. In addition, a recent study of PROMIS physical function items using IRT methods found no significant differences in item parameters (slopes and thresholds) between paper-pencil and PC data[133]. Thus it is unlikely that results of this analysis were notably impacted by the multiple methods used to collect data. Fourth, additional IRT modeling should be conducted before a knee-specific function CAT is used with TKR patients in real-time. In particular, other types of IRT models such as the generalized partial credit model (GPCM) might be applied to the data; although results for GRM and GPCM models are likely to be similar overall[65], they may vary for individual patients. Finally, the response scale for all items was keyed in the same direction (“None” to “Extreme” in the survey form), which makes it difficult to identify any possible straight-lining (respondents providing the same answer to all items). While it is common for function scales to present all response choices in the same order, the potential impact on item calibrations of even a small number of respondents who engage in straight-lining has not been studied, to the best of my knowledge. If new knee-specific function items are developed to address the lack of measurement at the higher end of the function spectrum, it is recommended that not all responses be keyed in the same direction.

In summary, this analysis has used IRT modeling, including evaluation of the underlying IRT assumptions of unidimensionality and local independence, to develop and evaluate a knee-specific function item bank based on KOOS function in ADL and Sport items. KOOS items could successfully be calibrated, and the resulting item bank was sufficiently reliable for individual administrations over a wide range of theta scores, although primarily at lower levels of function. In addition, CAT simulations suggested that while a function score can be estimated precisely and efficiently in many fewer than 22 items for those with severe OA, the item bank was less successful for patients at higher levels of knee-specific function. Furthermore, the CAT simulations raised questions as to whether the best items had been selected for the KOOS-PS. Additional research to extend the measurement range of the KOOS function items and to develop new KOOS short function scales is recommended.

Table 3.1: KOOS Pain and Function item descriptive statistics, pre-TKR (n=1,179)

					Percentage by Response Choice					
Item	**Abbreviated Content**	**Mean**	**SD**	**Skew-ness**	**1**	**2**	**3**	**4**	**5**	**Miss-ing**
Pain										
P1	Frequency knee pain*	4.24	0.75	-1.74	2.1	0.8	3.8	56.5	35.9	0.8
P2	Pain twisting/pivoting	3.48	1.03	-0.45	4.2	12.3	28.9	37.2	15.1	2.2
P3	Pain straighten fully	2.78	1.11	0.10	14.4	24.3	35.3	18.1	6.7	1.2
P4	Pain bending fully	3.15	1.15	-0.14	8.8	19.5	31.0	26.8	12.7	1.2
P5	Pain walking on flat	2.88	0.95	0.02	7.7	24.3	44.4	18.6	4.4	0.6
P6	Pain up or down stairs	3.66	0.93	-0.40	1.8	7.8	31.4	40.0	18.2	0.8
P7	Pain at night in bed	2.68	1.06	0.17	14.5	28.6	35.2	16.3	4.7	0.8
P8	Pain sitting or lying	2.50	0.96	0.28	15.1	34.7	37.1	9.8	2.7	0.7
P9	Pain standing upright	2.95	0.97	-0.01	6.7	22.7	42.4	21.1	5.3	1.8
Function in Activities of Daily Living										
A1	Descending stairs	3.41	0.97	-0.12	2.2	14.2	37.3	31.4	14.2	0.8
A2	Ascending stairs	3.37	0.98	-0.12	2.5	15.4	37.1	31.5	13.1	0.5
A3	Rising from sitting	3.17	0.96	-0.09	4.2	18.1	41.6	27.7	7.8	0.6
A4	Standing	2.85	0.95	-0.04	8.5	24.9	42.6	19.9	3.5	0.7
A5	Bending to floor	3.03	1.08	0.04	7.8	23.7	35.1	22.5	9.7	1.3
A6	Walking on flat surface	2.72	0.92	0.12	8.7	31.0	41.8	14.9	2.8	0.8
A7	Getting in/out of car	2.98	0.91	0.01	4.7	23.3	44.9	21.9	4.2	0.9
A8	Going shopping	3.14	0.97	-0.06	4.6	19.3	41.1	26.3	8.0	0.8
A9	Putting on socks	2.58	1.02	0.21	15.9	29.6	38.0	12.1	3.6	0.7
A10	Rising from bed	2.76	0.99	0.13	10.1	28.6	40.0	16.2	4.2	0.8
A11	Taking off socks	2.52	1.00	0.26	16.5	32.1	35.5	11.5	3.1	1.3
A12	Lying in bed	2.60	1.00	0.25	13.9	32.0	36.6	12.8	3.7	0.9
A13	Getting in/out of bath	2.73	1.12	0.26	14.2	27.4	33.8	14.0	7.6	3.1
A14	Sitting	2.24	0.93	0.45	22.5	39.4	29.4	6.4	1.5	0.8
A15	Getting on/off toilet	2.64	1.00	0.26	12.0	34.4	34.0	15.4	3.6	0.7
A16	Heavy domestic duties	3.64	1.00	-0.50	2.9	9.1	28.7	37.8	20.3	1.3
A17	Light domestic duties	2.65	0.91	0.20	9.6	33.1	42.2	11.8	2.8	0.6
Function in Sport and Recreation										
Sp1	Squatting	4.13	0.99	-1.11	2.1	4.7	15.7	30.9	45.0	1.6
Sp2	Running	4.48	0.83	-2.11	2.2	0.8	6.1	27.3	61.0	2.6
Sp3	Jumping	4.48	0.85	-2.16	2.4	1.2	5.1	27.8	61.0	2.5
Sp4	Twisting/pivoting	4.07	0.99	-1.00	2.4	4.2	18.1	33.2	40.5	1.7
Sp5	Kneeling	4.17	0.97	-1.11	1.9	4.4	15.7	30.5	46.6	0.9

Response options are 1=None, 2=Mild, 3=Moderate, 4=Severe, 5=Extreme except where noted:
* 1=Never, 2=Monthly, 3=Weekly, 4=Daily, 5=Always.

Table 3.2: KOOS Pain and Function item descriptive statistics, 6 months post-TKR (n=886)

					Percentage by Response Choice					
Item	**Abbreviated Content**	**Mean**	**SD**	**Skew-ness**	**1**	**2**	**3**	**4**	**5**	**Miss-ing**
Pain										
P1	Frequency knee pain*	2.52	1.36	0.14	35.7	12.1	16.5	28.0	4.5	3.2
P2	Pain twisting/pivoting	1.83	0.96	1.05	45.1	30.0	14.9	5.2	1.2	3.6
P3	Pain straighten fully	1.53	0.82	1.66	63.2	22.7	9.6	2.2	0.8	1.5
P4	Pain bending fully	2.05	1.08	0.87	38.2	31.2	18.3	7.8	3.0	1.5
P5	Pain walking on flat	1.36	0.66	1.90	71.4	18.6	6.6	1.0	0.1	2.2
P6	Pain up or down stairs	2.04	0.94	0.66	32.0	38.4	20.5	6.5	0.9	1.7
P7	Pain at night in bed	1.71	0.84	1.03	49.5	31.4	14.4	2.6	0.4	1.7
P8	Pain sitting or lying	1.55	0.74	1.28	57.3	30.2	9.4	1.3	0.2	1.5
P9	Pain standing upright	1.55	0.78	1.31	59.4	26.3	11.1	1.8	0.2	1.2
Function in Activities of Daily Living										
A1	Descending stairs	2.05	0.97	0.67	33.9	35.3	21.7	6.6	1.2	1.2
A2	Ascending stairs	1.92	0.91	0.76	37.7	36.3	18.6	4.8	0.7	1.9
A3	Rising from sitting	1.91	0.89	0.82	37.3	38.9	17.9	3.8	0.9	1.2
A4	Standing	1.49	0.77	1.54	64.6	22.6	9.5	1.9	0.2	1.1
A5	Bending to floor	1.76	0.94	1.22	49.2	30.9	11.8	5.3	1.2	1.7
A6	Walking on flat surface	1.35	0.65	1.92	73.1	17.8	6.8	0.8	0.1	1.3
A7	Getting in/out of car	1.87	0.83	0.72	36.5	41.2	17.5	2.7	0.4	1.7
A8	Going shopping	1.70	0.90	1.18	52.6	27.1	14.0	3.6	0.8	1.9
A9	Putting on socks	1.80	0.92	1.09	46.0	32.5	14.9	4.0	1.2	1.3
A10	Rising from bed	1.62	0.83	1.32	56.0	27.9	11.5	2.6	0.6	1.5
A11	Taking off socks	1.70	0.87	1.27	50.8	31.5	12.1	3.0	1.0	1.6
A12	Lying in bed	1.67	0.84	1.17	52.1	31.0	12.2	3.0	0.4	1.2
A13	Getting in/out of bath	1.67	0.95	1.46	56.6	24.9	10.1	4.9	1.5	2.1
A14	Sitting	1.42	0.69	1.76	67.3	22.7	7.3	0.8	0.3	1.6
A15	Getting on/off toilet	1.62	0.80	1.10	54.2	29.2	13.1	1.9	0.2	1.3
A16	Heavy domestic duties	2.36	1.15	0.56	26.1	31.2	23.5	11.3	5.2	2.7
A17	Light domestic duties	1.49	0.76	1.58	64.5	22.6	10.0	1.2	0.4	1.2
Function in Sport and Recreation										
Sp1	Squatting	2.93	1.26	0.16	13.4	26.3	25.5	17.7	14.4	2.6
Sp2	Running	3.38	1.31	-0.31	9.6	15.6	22.2	21.9	24.3	6.4
Sp3	Jumping	3.37	1.32	-0.34	11.1	13.0	23.6	21.5	24.2	6.6
Sp4	Twisting/pivoting	2.47	1.24	0.46	26.8	26.8	23.2	13.7	7.3	2.4
Sp5	Kneeling	3.25	1.24	-0.09	8.3	20.2	27.7	19.9	20.4	3.6

Response options are 1=None, 2=Mild, 3=Moderate, 4=Severe, 5=Extreme except where noted:
* 1=Never, 2=Monthly, 3=Weekly, 4=Daily, 5=Always.

Table 3.3: KOOS Pain and Function item descriptive statistics, combined pre- and post-TKR sample (n=1,179)

					Percentage by Response Choice					
Item	**Abbreviated Content**	**Mean**	**SD**	**Skew-ness**	**1**	**2**	**3**	**4**	**5**	**Miss-ing**
Pain										
P1	Frequency knee pain*	3.67	1.28	-1.06	13.4	4.4	8.1	47.1	25.0	2.0
P2	Pain twisting/pivoting	2.95	1.29	-0.11	18.2	17.1	23.9	26.3	11.4	3.1
P3	Pain straighten fully	2.38	1.17	0.42	29.5	24.1	27.8	12.5	4.7	1.4
P4	Pain bending fully	2.80	1.25	0.10	19.0	22.7	26.2	21.0	9.8	1.3
P5	Pain walking on flat	2.39	1.12	0.24	28.3	22.1	32.6	13.1	2.8	1.2
P6	Pain up or down stairs	3.16	1.20	-0.26	11.6	16.8	27.7	30.1	12.7	1.0
P7	Pain at night in bed	2.37	1.09	0.39	25.6	29.3	29.3	11.4	3.2	1.3
P8	Pain sitting or lying	2.19	0.99	0.51	28.4	33.5	28.2	6.7	2.0	1.3
P9	Pain standing upright	2.52	1.13	0.18	23.5	22.6	33.0	15.2	3.9	1.8
Function in Activities of Daily Living										
A1	Descending stairs	2.99	1.18	-0.06	12.6	20.1	31.9	23.6	10.6	1.2
A2	Ascending stairs	2.93	1.19	-0.04	14.4	20.1	31.5	22.9	9.8	1.4
A3	Rising from sitting	2.77	1.11	0.06	14.8	24.8	33.6	20.3	5.8	0.8
A4	Standing	2.43	1.10	0.21	25.5	24.3	32.7	13.7	2.7	1.0
A5	Bending to floor	2.64	1.21	0.25	20.5	26.5	26.0	17.8	7.4	1.8
A6	Walking on flat surface	2.29	1.06	0.34	29.2	26.2	31.3	10.2	2.0	1.1
A7	Getting in/out of car	2.66	1.01	0.14	13.6	28.9	37.5	15.2	3.6	1.3
A8	Going shopping	2.69	1.18	0.06	20.8	20.5	31.9	19.9	5.8	1.1
A9	Putting on socks	2.35	1.05	0.39	24.4	30.7	31.0	9.8	2.9	1.1
A10	Rising from bed	2.41	1.09	0.31	24.9	27.1	31.0	12.4	3.2	1.4
A11	Taking off socks	2.27	1.04	0.49	26.8	32.1	28.2	8.5	2.6	1.7
A12	Lying in bed	2.30	1.04	0.45	26.0	31.6	29.9	8.6	2.8	1.1
A13	Getting in/out of bath	2.40	1.19	0.46	28.1	25.4	26.0	12.0	5.7	2.8
A14	Sitting	1.97	0.94	0.74	36.6	35.1	21.3	4.6	1.2	1.2
A15	Getting on/off toilet	2.32	1.07	0.41	26.6	30.3	28.1	11.2	2.6	1.2
A16	Heavy domestic duties	3.26	1.20	-0.34	10.4	14.9	26.7	30.7	15.3	2.0
A17	Light domestic duties	2.29	1.04	0.36	27.5	28.9	31.6	9.1	2.1	0.8
Function in Sport and Recreation										
Sp1	Squatting	3.74	1.23	-0.69	6.1	11.9	18.2	27.0	34.9	1.9
Sp2	Running	4.12	1.14	-1.27	4.7	5.9	11.3	25.3	48.6	4.2
Sp3	Jumping	4.13	1.14	-1.33	5.1	4.7	11.6	25.3	49.1	4.2
Sp4	Twisting/pivoting	3.58	1.30	-0.60	9.8	12.0	17.9	28.2	29.9	2.1
Sp5	Kneeling	3.88	1.15	-0.76	3.6	10.3	19.0	27.0	38.3	1.9

Response options are 1=None, 2=Mild, 3=Moderate, 4=Severe, 5=Extreme except where noted:
* 1=Never, 2=Monthly, 3=Weekly, 4=Daily, 5=Always.

Table 3.4: Exploratory factor analysis - Unrotated factor loadings for KOOS ADL and Sport items, combined TKR sample (n=1,167)

				Factor Loading		
Item	**Abbreviated Content**	**Mean**	**SD**	**Factor 1**	**Factor 2**	**Uniqueness**
Activities of Daily Living						
A1	Descending stairs	2.99	1.18	0.809	0.092	0.337
A2	Ascending stairs	2.93	1.19	0.831	0.052	0.307
A3	Rising from sitting	2.77	1.11	0.864	-0.065	0.249
A4	Standing	2.43	1.10	0.819	-0.114	0.316
A5	Bending to floor	2.64	1.21	0.791	-0.131	0.357
A6	Walking on flat	2.29	1.06	0.817	-0.125	0.317
A7	Getting in/out of car	2.66	1.01	0.834	-0.177	0.273
A8	Going shopping	2.69	1.18	0.854	-0.044	0.269
A9	Putting on socks	2.35	1.05	0.758	-0.312	0.328
A10	Rising from bed	2.41	1.09	0.826	-0.265	0.248
A11	Taking off socks	2.27	1.04	0.777	-0.321	0.293
A12	Lying in bed	2.30	1.04	0.762	-0.245	0.359
A13	Getting in/out of bath	2.40	1.19	0.792	-0.140	0.354
A14	Sitting	1.97	0.94	0.740	-0.347	0.332
A15	Getting on/off toilet	2.32	1.07	0.818	-0.209	0.287
A16	Heavy domestic	3.26	1.20	0.806	0.159	0.325
A17	Light domestic duties	2.29	1.04	0.875	-0.074	0.230
Sport/Recreation						
Sp1	Squatting	3.74	1.23	0.767	0.426	0.230
Sp2	Running	4.12	1.14	0.693	0.621	0.135
Sp3	Jumping	4.13	1.14	0.691	0.622	0.137
Sp4	Twisting/pivoting	3.58	1.30	0.775	0.396	0.243
Sp5	Kneeling	3.88	1.15	0.747	0.397	0.284
Eigenvalue				13.88	1.91	
% total variance explained				63.1	8.7	

Table 3.5: Confirmatory factor analysis - Factor loadings for KOOS ADL and Sport items in three models, combined TKR sample (n=1,167)

	Abbreviated			1 factor	2 factor		Bifactor		
Item	Content	Mean	SD		ADL	Sport	Common	ADL	Sport
Activities of Daily Living									
A1	Descending stairs	2.99	1.18	0.863	0.875		0.905	-0.126	
A2	Ascending stairs	2.93	1.19	0.879	0.889		0.921	-0.105	
A3	Rising from sitting	2.77	1.11	0.853	0.863		0.852	0.180	
A4	Standing	2.43	1.10	0.853	0.861		0.864	0.112	
A5	Bending to floor	2.64	1.21	0.795	0.807		0.777	0.262	
A6	Walking on flat	2.29	1.06	0.863	0.870		0.872	0.121	
A7	Getting in/out of car	2.66	1.01	0.855	0.864		0.828	0.289	
A8	Going shopping	2.69	1.18	0.880	0.890		0.893	0.112	
A9	Putting on socks	2.35	1.05	0.920	0.924		0.658	0.698	
A10	Rising from bed	2.41	1.09	0.853	0.861		0.809	0.342	
A11	Taking off socks	2.27	1.04	0.937	0.940		0.679	0.708	
A12	Lying in bed	2.30	1.04	0.788	0.800		0.740	0.356	
A13	Getting in/out of bath	2.40	1.19	0.805	0.815		0.782	0.269	
A14	Sitting	1.97	0.94	0.798	0.808		0.737	0.390	
A15	Getting on/off toilet	2.32	1.07	0.849	0.857		0.809	0.324	
A16	Heavy domestic	3.26	1.20	0.818	0.835		0.856	0.018	
A17	Light domestic duties	2.29	1.04	0.866	0.879		0.872	0.163	
Sport/Recreation									
Sp1	Squatting	3.74	1.23	0.798		0.887	0.727		0.490
Sp2	Running	4.12	1.14	0.926		0.956	0.709		0.659
Sp3	Jumping	4.13	1.14	0.926		0.957	0.712		0.661
Sp4	Twisting/pivoting	3.58	1.30	0.815		0.927	0.781		0.414
Sp5	Kneeling	3.88	1.15	0.770		0.858	0.703		0.485
Model Fit Statistics									
CFI				0.943	0.954			0.979	
RMSEA				0.158	0.142			0.101	
# of item pairs w/residual correlations >0.20				22	7			0	
Correlation between factors				-	0.764			-	

Table 3.6: Tests of differential item functioning, combined TKR sample (n=1,179)

		Age		Gender		Education	
Item	**Abbreviated Content**	**ΔR²**	**p**	**ΔR²**	**p**	**ΔR²**	**p**
Activities of Daily Living							
A1	Descending stairs	0.003	0.7047	0.000	0.7607	0.027	0.0000
A2	Ascending stairs	0.009	0.1181	0.005	0.0444	0.006	0.1833
A3	Rising from sitting	0.007	0.2457	0.001	0.5968	0.004	0.3799
A4	Standing	0.005	0.5043	0.002	0.3016	0.004	0.3987
A5	Bending to floor	0.009	0.1029	0.004	0.0878	0.004	0.3137
A6	Walking on flat surface	0.008	0.1787	0.012	0.0010	0.003	0.4955
A7	Getting in/out of car	0.005	0.4379	0.001	0.4988	0.002	0.7919
A8	Going shopping	0.004	0.5191	0.003	0.1651	0.005	0.2360
A9	Putting on socks/stockings	0.012	0.0303	0.011	0.0015	0.006	0.1454
A10	Rising from bed	0.004	0.5265	0.002	0.3399	0.002	0.7517
A11	Taking off socks/stockings	0.006	0.2778	0.007	0.0145	0.002	0.6219
A12	Lying in bed	0.002	0.9109	0.010	0.0027	0.004	0.3937
A13	Getting in/out of bath	0.007	0.2471	0.012	0.0010	0.002	0.7866
A14	Sitting	0.012	0.0315	0.000	0.7943	0.005	0.2623
A15	Getting on/off toilet	0.010	0.0647	0.003	0.1974	0.002	0.7728
A16	Heavy domestic duties	0.008	0.1810	0.006	0.0249	0.001	0.8260
A17	Light domestic duties	0.012	0.0327	0.002	0.2804	0.001	0.9144
Sport/Recreation							
Sp1	Squatting	0.014	0.0160	0.006	0.0249	0.006	0.1667
Sp2	Running	0.004	0.5718	0.002	0.2775	0.006	0.1482
Sp3	Jumping	0.008	0.1676	0.004	0.0807	0.002	0.6818
Sp4	Twisting/pivoting	0.004	0.6166	0.015	0.0002	0.002	0.6448
Sp5	Kneeling	0.016	0.0057	0.020	0.0000	0.006	0.1382

Groups are: age (<55, 55-64, 65-74, 75+); gender (male, female); education (less than high school, high school graduate, some post-high school education, college graduate).

ΔR²= Nagelkerke’s coefficient of determination.

Table 3.7: Item statistics for the function (ADL and Sport) item bank, combined TKR sample

Item	Abbreviated Content	Skewness	CFA	Slope (SE)	Step1 (SE)	Step2 (SE)	Step3 (SE)	Step4 (SE)	S-X^2	I_{max} at Θ	% CAT Util.
A1	Descending stairs	-0.06	0.863	2.67	-1.43	-0.51	0.47	1.40	0.236	2.00	0.25
				0.10	0.05	0.03	0.03	0.05		0.5[‡]	5.26
A2	Ascending stairs	-0.04	0.879	2.89	-1.46	-0.56	0.39	1.26	0.679	2.34	0.61
				0.14	0.06	0.04	0.05	0.07		0.4	6.10
A3	Rising from sitting	0.06	0.853	3.12	-1.71	-0.75	0.24	1.23	0.724	2.63	1.38
				0.15	0.07	0.04	0.04	0.07		-0.8	10.91
A4	Standing	0.21	0.853	3.24	-2.09	-1.08	-0.08	0.72	0.474	2.89	7.77
				0.16	0.09	0.04	0.04	0.05		0.0	6.97
A5	Bending to floor	0.25	0.795	2.56	-1.67	-0.81	0.00	0.98	0.747	1.91	0.00
				0.12	0.07	0.04	0.04	0.07		-0.8	0.25
A6	Walking on flat surface	0.34	0.863	3.27	-2.22	-1.26	-0.24	0.57	0.205	2.92	17.66
				0.16	0.10	0.05	0.04	0.05		-0.2	3.87
A7	Getting in/out of car	0.14	0.855	3.29	-1.95	-0.99	0.14	1.27	0.083	2.88	2.53
				0.16	0.08	0.04	0.04	0.07		-1.0	12.51
A8	Going shopping	0.06	0.880	3.66	-1.64	-0.74	0.15	0.89	0.520	3.66	28.78
				0.18	0.06	0.04	0.04	0.06		0.2	19.70
A9	Putting on socks	0.39	0.920	2.23	-2.38	-1.40	-0.21	0.86	0.476	1.42	0.22
				0.11	0.11	0.06	0.04	0.06		-1.5	0.13
A10	Rising from bed	0.31	0.853	3.23	-2.01	-1.12	-0.14	0.74	0.975	2.84	9.96
				0.16	0.08	0.04	0.04	0.06		-0.1	1.15
A11	Taking off socks	0.49	0.937	2.54	-2.31	-1.42	-0.32	0.72	0.189	1.83	0.22
				0.10	0.09	0.05	0.03	0.04		-1.5	0.02
A12	Lying in bed*	0.45	0.788	2.54	-2.26	-0.28	0.76		0.988	1.74	0.22
				0.13	0.10	0.04	0.06			-0.1	0.07
A13	Getting in/out of bath	0.46	0.805	2.74	-1.80	-1.09	-0.22	0.64	0.282	2.19	0.00
				0.14	0.07	0.05	0.04	0.06		-1.2	0.00
A14	Sitting	0.74	0.798	2.62	-2.69	-1.77	-0.72	0.36	0.001	1.93	0.37
				0.13	0.14	0.07	0.04	0.05		-1.8	0.00
A15	Getting on/off toilet	0.41	0.849	3.17	-2.12	-1.19	-0.26	0.67	0.808	2.74	0.61
				0.16	0.09	0.05	0.04	0.05		-1.2	0.09
A16	Heavy domestic duties	-0.34	0.818	2.67	-1.20	-0.16	0.76	1.55	0.554	2.07	0.52
				0.13	0.05	0.04	0.06	0.08		0.9	7.38
A17	Light domestic duties	0.36	0.866	3.60	-2.16	-1.28	-0.25	0.63	0.193	3.44	28.28
				0.18	0.09	0.05	0.04	0.05		-1.3	12.15
Sp1	Squatting	-0.69	0.798	1.83	-0.55	0.40	1.24	2.20	0.311	1.03	0.15
				0.10	0.05	0.05	0.08	0.12		0.8	3.33
Sp2	Running[†]	-1.27	0.926	1.72	-0.04	1.01	2.43		0.053	0.87	0.05
				0.10	0.05	0.08	0.14			0.7	1.10
Sp3	Jumping[†]	-1.33	0.926	1.81	-0.02	1.02	2.34		0.003	0.97	0.10
				0.09	0.04	0.05	0.10			0.7	2.45
Sp4	Twisting/pivoting	-0.60	0.815	2.17	-0.68	0.24	0.97	1.70	0.541	1.45	0.25
				0.11	0.04	0.05	0.07	0.09		0.9	4.48
Sp5	Kneeling	-0.76	0.770	1.74	-0.44	0.54	1.52	2.64	0.036	0.92	0.07
				0.10	0.05	0.06	0.09	0.15		0.6	2.09

* Severe/moderate response choices collapsed. † Moderate/mild response choices collapsed.

‡ Maximum item information was at theta values of -0.5 as well as theta value of 0.5.

CFA: Factor loading in one-factor confirmatory factor analysis.

Step1-Step4: Item difficulty thresholds. Unless response choices were collapsed, step1/threshold 1 is extreme/severe; step 2/threshold 2 is severe/moderate; step 3/threshold 3 is moderate/mild; step 4/threshold 4 is mild/none.

S-X^2: p-value for S-X^2 fit statistic.

I_{max} at Θ: Maximum item information (upper number) at a particular value of theta Θ (lower number).

% CAT Util.: % of item administrations in pre-TKR (upper number) and post-TKR (lower number) CAT administrations using stopping rule of SE≤0.32 or maximum of 10 items.

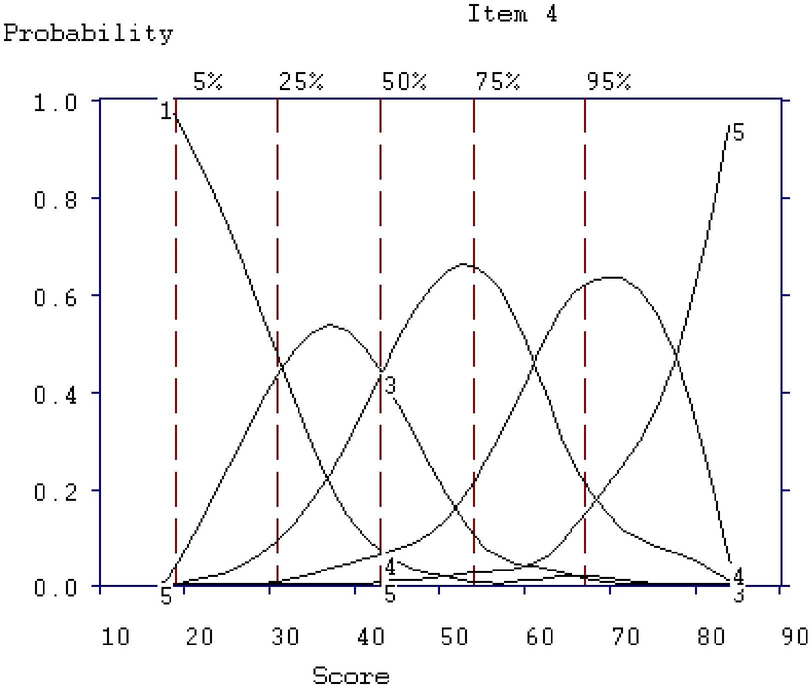
Table 3.8: Number of items administered and item characteristics in simulated CATs

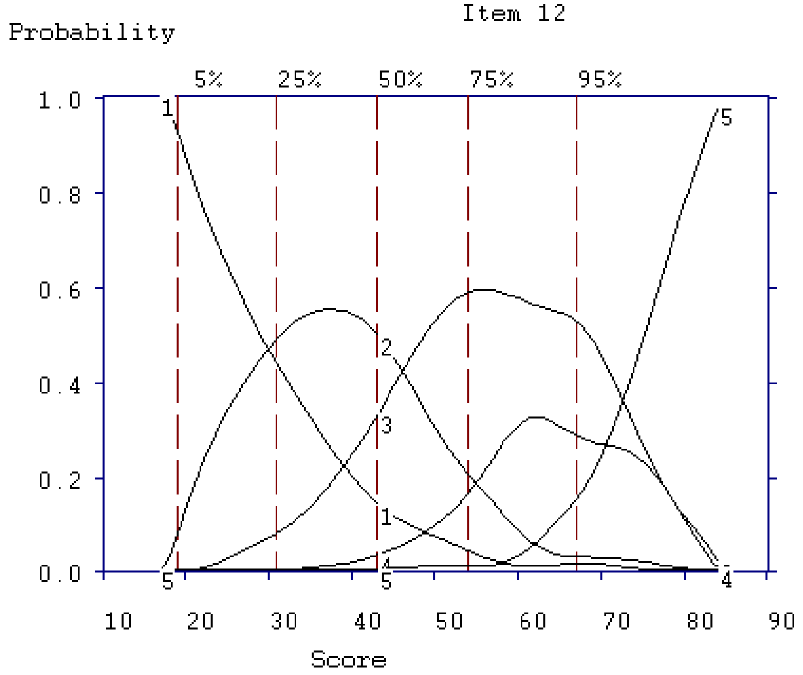
		SE ≤ 0.32*		SE ≤ 0.23†	
		Pre-TKR	**Post-TKR**	**Pre-TKR**	**Post-TKR**
Number of Items Administered					
1		0.0	0.0	0.0	0.0
2		0.0	0.0	0.0	0.0
3		71.3	28.6	0.0	0.0
4		20.2	32.7	0.0	0.0
5		5.4	10.9	4.3	0.0
6		1.3	5.1	41.9	16.8
7		0.6	5.8	30.2	18.5
8		0.3	3.6	12.3	16.5
9		0.1	2.4	4.8	8.7
10		0.9	10.9	6.5	39.5
Average number of items per patient		3.45	5.02	6.91	8.30
% patients with SE ≤ 0.32		99.1	91.5		
% patients with SE ≤ 0.23				95.8	65.7
% of Times Item Administered					
Activities of Daily Living					
A1	Descending stairs	0.25	5.26	0.34	5.67
A2	Ascending stairs	0.61	6.10	1.69	6.76
A3	Rising from sitting	1.38	10.91	4.53	9.00
A4	Standing	7.77	6.97	13.65	8.29
A5	Bending to floor	0.00	0.25	0.10	2.51
A6	Walking on flat surface	17.66	3.87	13.15	5.08
A7	Getting in/out of car	2.53	12.51	9.47	10.02
A8	Going shopping	28.78	19.70	14.39	11.91
A9	Putting on socks/stockings	0.22	0.13	0.11	0.14
A10	Rising from bed	9.96	1.15	14.28	9.14
A11	Taking off socks/stockings	0.22	0.02	0.21	0.04
A12	Lying in bed	0.22	0.07	0.18	0.05
A13	Getting in/out of bath	0.00	0.00	0.50	0.19
A14	Sitting	0.37	0.00	0.28	0.03
A15	Getting on/off toilet	0.61	0.09	11.82	4.01
A16	Heavy domestic duties	0.52	7.38	0.52	6.27
A17	Light domestic duties	28.28	12.15	14.25	7.80
Sport/Recreation					
Sp1	Squatting	0.15	3.33	0.14	3.26
Sp2	Running	0.05	1.10	0.02	0.68
Sp3	Jumping	0.10	2.45	0.10	2.58
Sp4	Twisting/pivoting	0.25	4.48	0.18	4.51
Sp5	Kneeling	0.07	2.09	0.07	2.05

Stopping rules: * SE≤0.32 (reliability≥0.90) or maximum number of items=10; † SE≤0.23 (reliability≥0.95) or maximum number of items=10. Minimum number of items=2 in all CATs.

Figure 3.1: Monotonicity of item response curves, combined sample (n=1,179)

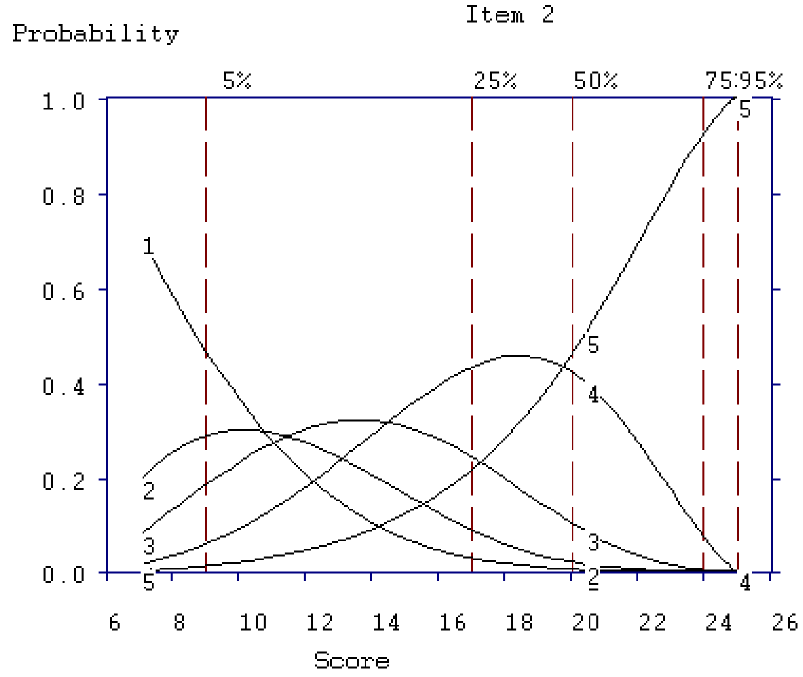
Probability of selecting each item response for Items AD04 (Standing) and ADL12 (Lying in bed), in relation to ADL scale score (on horizontal axis)

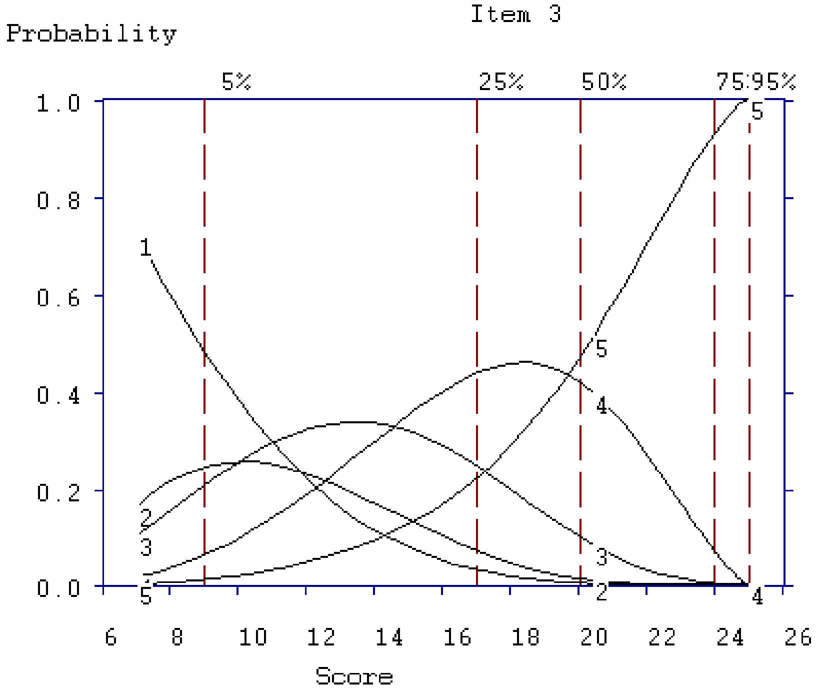




Response options are 1=None, 2=Mild, 3=Moderate, 4=Severe, 5=Extreme.

Figure 3.1: Monotonicity of item response curves (continued)
Probability of selecting each item response for Sport items Sp2 (Running) and Sp3 (Jumping), in relation to Sport/Recreation scale score (on horizontal axis)

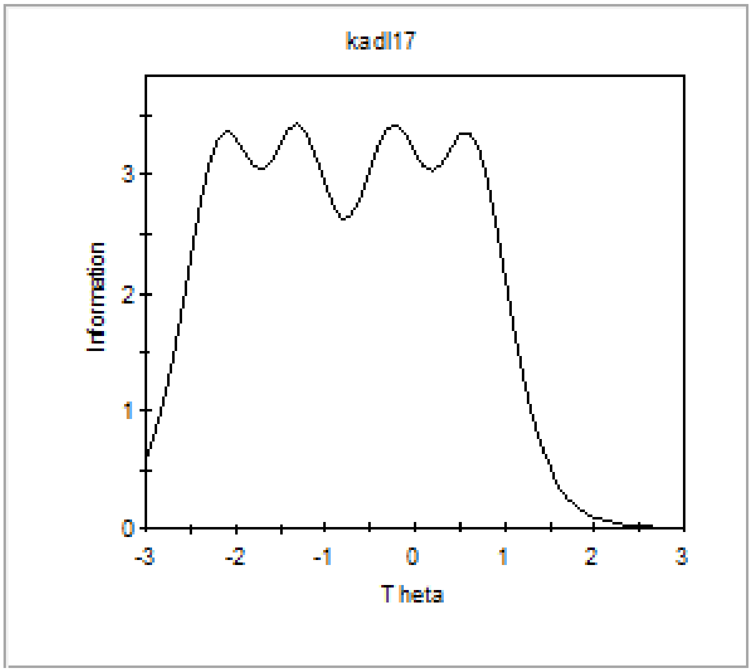




Response options are 1=None, 2=Mild, 3=Moderate, 4=Severe, 5=Extreme

Figure 3.2: Sample ADL and Sport item information functions

Item ADL17 - Light domestic duties



Item Sport 3 - Jumping

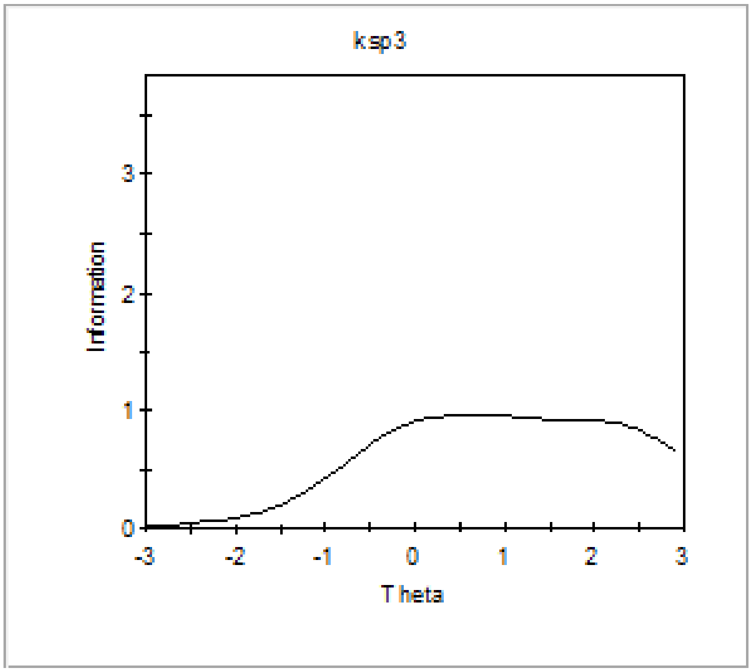


Figure 3.3: ADL and Sport 22-item bank - Total test information and standard error

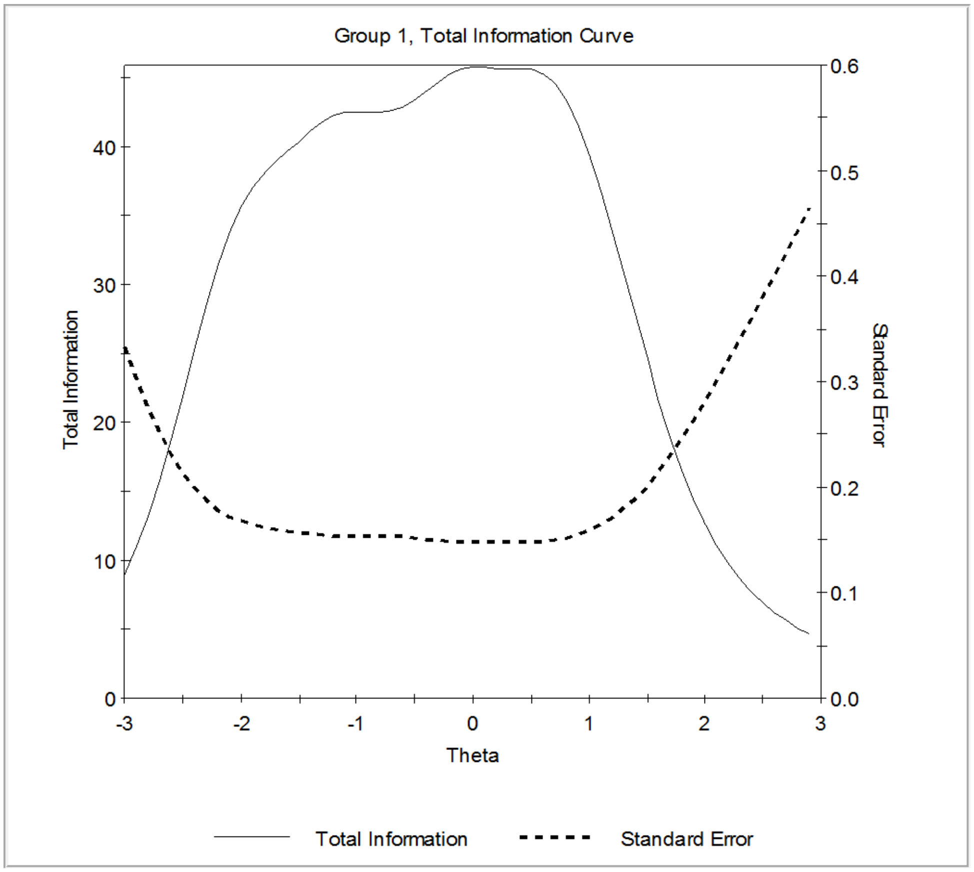
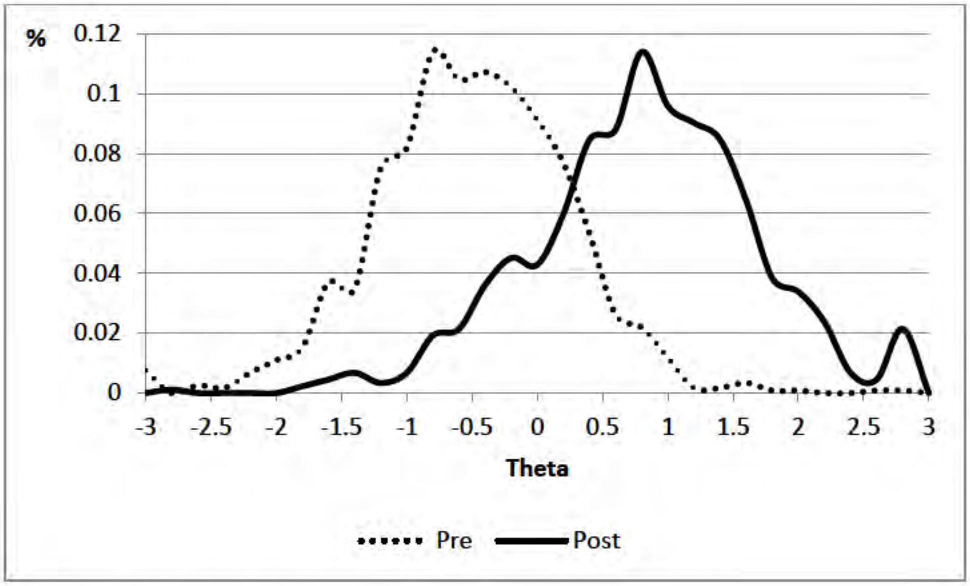
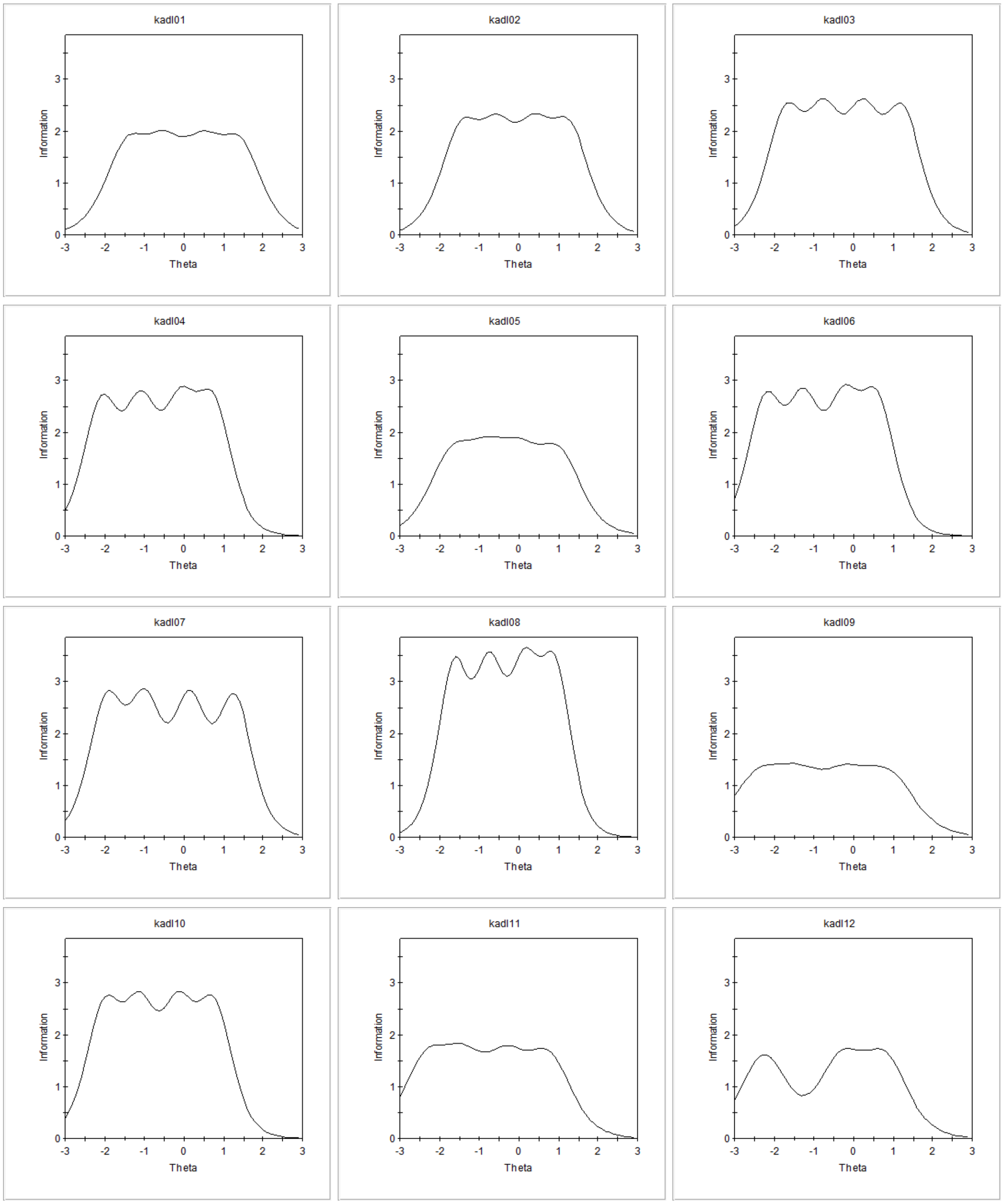


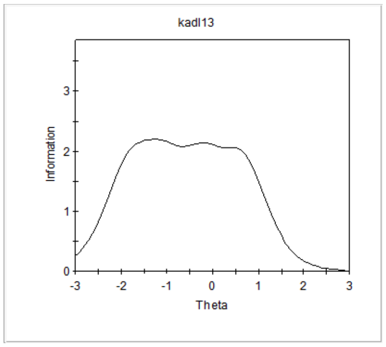
Figure 3.4: Distribution of patients by theta score, pre and post-TKR

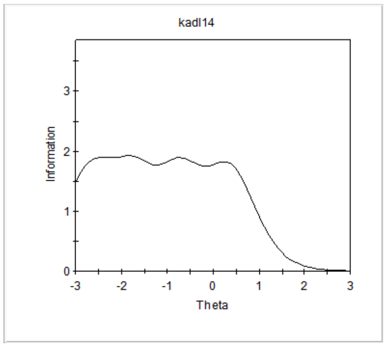


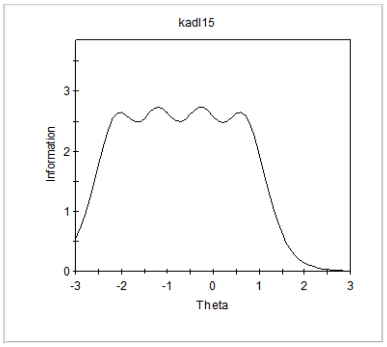
Sample size: pre-TKR n=1,179; 6 months post-TKR n=886

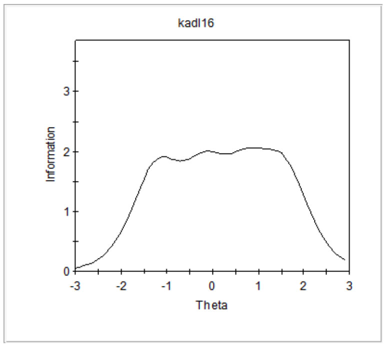
Figure 3.5: All ADL and Sport item information functions

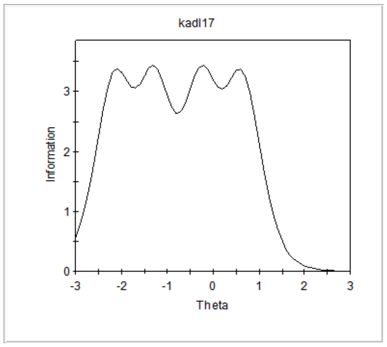


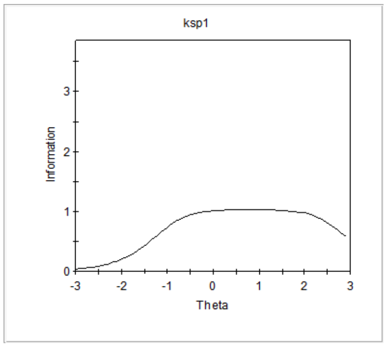

kadl13
Information
0
1
2
3
-3
-2
-1
0
1
2
3
Theta

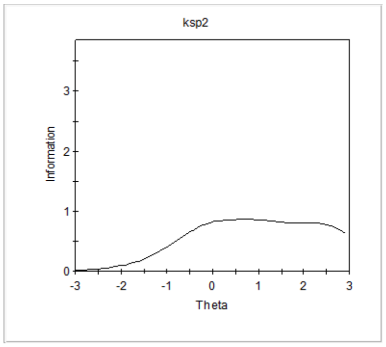

kadl14
Information
0
1
2
3
-3
-2
-1
0
1
2
3
Theta

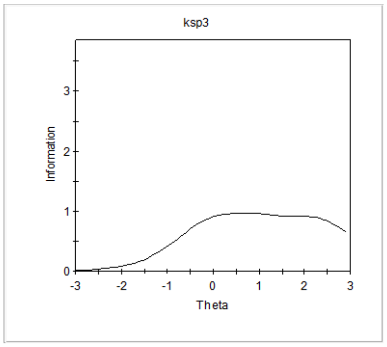

kadl15
Information
0
1
2
3
-3
-2
-1
0
1
2
3
Theta

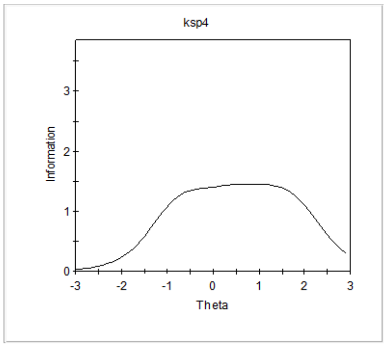

kadl16
Information
0
1
2
3
-3
-2
-1
0
1
2
3
Theta

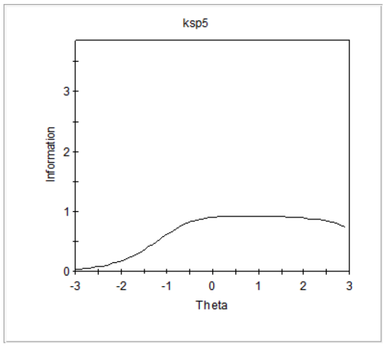

kadl17
Information
0
1
2
3
-3
-2
-1
0
1
2
3
Theta


ksp1
Information
0
1
2
3
-3
-2
-1
0
1
2
3
Theta


ksp2
Information
0
1
2
3
-3
-2
-1
0
1
2
3
Theta


ksp3
Information
0
1
2
3
-3
-2
-1
0
1
2
3
Theta


ksp4
Information
0
1
2
3
-3
-2
-1
0
1
2
3
Theta


ksp5
Information
0
1
2
3
-3
-2
-1
0
1
2
3
Theta

CHAPTER IV: VALIDITY AND RESPONSIVENESS OF THE KOOS IN COMPARISON WITH OTHER PATIENT-REPORTED OUTCOME MEASURES IN TOTAL KNEE REPLACEMENT

Abstract

Objective: To evaluate validity and responsiveness of the Knee injury and Osteoarthritis Outcome Score (KOOS) and other patient-reported outcome measures before and after total knee replacement (TKR).

Methods: Pre-TKR and 6-month post-TKR data from 1,143 patients was used to compare measures varying in attributes (knee-specific versus generic, longer versus shorter, computerized adaptive test (CAT) versus fixed-length) including KOOS, WOMAC, 7- and 8-item KOOS and WOMAC function scales, 3 to 10-item CAT function scores, and the SF-36. Validity was evaluated using ANOVA to compare pre-TKR scores for groups known to differ in knee pain, assistive device use and comorbid conditions, and to compare change scores (post- minus pre-TKR) for groups rating 6-month outcomes as better, same or worse. Responsiveness also was evaluated with effect sizes and standardized response means.

Results: Before TKR, KOOS scales discriminated between known groups as hypothesized. At 6 months post-TKR, the KOOS Quality of Life (QOL) scale discriminated significantly ($p<0.05$) better than all other knee-specific scales among better/same/worse outcome groups. All fixed-length function scales had similar validity in discriminating between post-TKR outcome groups. KOOS Pain and Symptom scales discriminated better than WOMAC Pain and Stiffness scales. The SF-36 Physical

Component Summary (PCS) discriminated as well as KOOS QOL among post-TKR outcome groups, although PCS had a smaller effect size.

Conclusions: KOOS was valid and responsive in TKR, and KOOS QOL was more responsive than other KOOS and WOMAC scales. Knee-specific short function scales and CATs were as valid and responsive as longer KOOS and WOMAC function scales.

Introduction

Important attributes of a patient-reported outcome (PRO) questionnaire include its conceptual and measurement model, reliability, validity and responsiveness[43]. Previous analyses in this dissertation have evaluated the conceptual and measurement models underlying the Knee injury and Osteoarthritis Outcome Score (KOOS) and its reliability. These analyses supported the developer's conceptual and measurement models and found that KOOS scales met scoring assumptions and were reliable across multiple sociodemographic and clinical groups (see Chapter II). This paper will evaluate the validity and responsiveness of the KOOS in relation to other PRO measures used in studies of knee osteoarthritis (OA) and total knee replacement (TKR).

Tests of validity and responsiveness provide information that is useful in interpreting the meaning of the quantitative scores (i.e., scales) that are derived from a questionnaire. Validity indicates the extent to which a scale measures what it is intended to measure. Responsiveness indicates the ability of a scale to detect change over time. Some have argued that from a psychometric perspective, the responsiveness of a PRO scale is best evaluated in relation to another variable, such as the change in clinician ratings or the patient's own rating of change[134, 135]. This type of anchor-based method of validation allows the meaning of a scale to be interpreted in relation to other measures[136, 137]. In this sense, responsiveness can be seen as longitudinal validity.

Convergent and discriminant validity of the KOOS have been demonstrated through examination of correlations between KOOS and SF-36 Health Survey scales in studies of knee OA[27, 46, 47, 49, 51, 52] and of ACL and meniscus injuries[77-79, 81-83]. While showing that scales from different questionnaires (e.g., KOOS Function and SF-36 Physical Functioning) have a high correlation provides some information about their validity, to fully evaluate validity, alternative forms of scales need to be compared in relation to external criteria. This is often done using tests of known groups validity, which compare the statistical efficiency of scales in detecting differences between groups known to differ at a point in time, or in detecting differences in change known to have occurred over time[138]. Tests of known groups validity provide additional information that is useful in interpreting the meaning of an individual scale or comparing the relative validity of multiple scales. These tests were used extensively in development of the SF-36[72, 73], but to my knowledge, have not been used with the KOOS.

In addition to the KOOS, a number of other questionnaires have been used to evaluate patient-reported outcomes in knee OA and TKR. Foremost among the joint-specific questionnaires is the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), which is included in its entirety in the KOOS. Both the KOOS and the WOMAC include a 17-item scale that measures the degree to which function in activities of daily living (ADL) is difficult due to a specific joint. A number of studies have found that this scale contains redundant items and could be shortened[61, 62]. Several 7 or 8-item joint-specific function scales have been derived from the 17-item WOMAC Function scale[24-26], but the relative performance of these shorter scales has not been compared. Similarly, a 7-item KOOS function scale, which combines items from the KOOS Function in ADL and Function in Sport/Recreation scales, has been derived[28], but its performance in relation to other short function scales also has not been evaluated. Thus, while there is consensus that knee-specific function can be measured with fewer than 17 items, there is no consensus as to what the best method is for doing so.

In addition to joint-specific questionnaires, generic (general) health surveys that are not specific to any disease or treatment also are often used in knee OA and TKR studies, most notably the SF-36 Health Survey[29, 30]. Many TKR studies administer both a knee-specific questionnaire (e.g., WOMAC, KOOS) and a generic questionnaire (e.g., SF-36), to include scales that are more specific to the impact of knee problems along with scales that are more sensitive to the impact of comorbidity and allow outcomes to be compared across conditions. However, as a result up to 27 or 32 items may be used to measure function, if both the WOMAC or KOOS and the SF-36 are administered, which greatly increases respondent burden. While knee-specific measures such as KOOS and WOMAC have been shown to be more responsive to TKR than the generic SF-36[56, 139], other research has found that knee-specific and generic function measures were equally sensitive to the severity of knee problems[140]. The extent to which knee-specific measures of function provide different information than generic measures of function is an area of ongoing research.

While a multiplicity of knee-specific and generic PRO instruments are used in TKR, comprehensive information on their comparative validity and responsiveness is lacking. Therefore, this paper will evaluate the validity and responsiveness of the KOOS in comparison to the WOMAC, the short function scales derived from the KOOS and the WOMAC, and the SF-36, using pre-TKR and post-TKR data from a national joint replacement registry. In addition, the validity and responsiveness of a new knee-specific function item bank that was developed from all 22 KOOS function items and computerized adaptive test (CAT) scores calculated from the item bank (see Chapter III) will be evaluated. While the paper will follow approaches used in previous KOOS studies (e.g., examination of scale correlations and post-TKR effect sizes), it also will evaluate the relative performance of all measures using tests of known groups validity. To increase the generalizability of results, multiple tests of known groups validity will be conducted; multiple tests of known groups validity also were conducted when the validity of the physical and mental component scores from the SF-36 was initially evaluated, for similar reasons[73]. By conducting a variety of cross-sectional and longitudinal tests, for which there are strong hypotheses as to the results that would be expected for valid measures, this paper will advance understanding of the comparative performance of the KOOS and other PRO measures, thereby informing their use in future studies of knee OA and TKR.

Methods

Patients

Data came from the Function and Outcomes Research for Comparative Effectiveness in Total Joint Replacement (FORCE-TJR) registry, which includes joint replacement patients from across the US[41]. Patients who could not provide consent due to cognitive impairment or who had emergency surgery were ineligible. All questionnaires were self-administered via a web-based or scannable paper-pencil survey at the surgeon’s office or at home prior to surgery, 6 months post-surgery, and annually thereafter. This analysis included data from 1,179 TKR patients randomly selected from FORCE-TJR enrollees at high volume surgical centers between May 2011 and August 2012. Of these, data was available for 886 patients at the 6-month follow-up.

Questionnaire

The FORCE-TJR questionnaire included the KOOS and SF-36 Health Survey. Patients first answered a number of questions about their TKR surgery, assistive device use and sociodemographic characteristics, followed by the SF-36, a modified Charlson comorbidity index[84], a back pain severity item, the KOOS, and additional items on pain in joints other than the surgical knee.

KOOS and KOOS-PS. The KOOS contains 42 knee-specific items that measure pain (number of items k=9), other symptoms (k=7), function in activities of daily living (ADL) (k=17), function in sport and recreation (Sport) (k=5), and knee-related quality of life (QOL) (k=4)[45, 50]. (Abbreviated item content for all KOOS items is provided in Table 2.2). All KOOS items have 5-point response scales and were answered in reference to the surgical knee. Most KOOS items have a recall period of the last week; one Pain frequency and all QOL items do not have a specific recall period.

A score for each KOOS scale was calculated by summing its component items and then transforming the sum so that 0 was the worst possible and 100 was the best possible score, following the developer's scoring algorithms[85]. A scale score was calculated as long as at least 50% of the items in the scale were answered; the mean score of all non-missing items within the scale was imputed for any missing item-level data. Assumptions underlying KOOS scale scoring were confirmed in the FORCE-TJR data (see Chapter II). In addition, the KOOS-PS was constructed from four KOOS ADL and three KOOS Sport items[28]. Unlike the KOOS, the KOOS-PS is scored so 0 is the best possible and 100 is the worst possible score; the KOOS-PS is the only scale evaluated in this paper in which a higher score indicates poorer functioning.

WOMAC and short WOMAC Function scales. Because all 24 WOMAC items (version LK3.0) are included in the 42-item KOOS[50], the WOMAC Pain (k=5), Stiffness (k=2) and Function (k=17) scales could be scored from the KOOS items that were administered in FORCE-TJR[22]. To be consistent with KOOS scoring, all WOMAC scales were scored so 0 was the worst possible and 100 was the best possible score. Following the developer's scoring algorithms, the WOMAC Function scale was calculated if at most three items were missing, and the WOMAC Stiffness and Pain scales were calculated if at most one item was missing[86].

As noted previously, the content of the 17 items in the WOMAC Function (LK3.0) scale is identical to that of the 17 items in the KOOS Function in ADL scale; the scales differ only in the number of items that need to be answered to calculate a scale score (9 for KOOS, 14 for WOMAC). Because there was so little missing data in the FORCE-TJR dataset and because the validity and responsiveness analyses were limited to patients who had scores for all KOOS and WOMAC scales, the psychometric performance of these two scales was identical in this paper. Therefore, to simplify the narrative, these two scales will be described as the "KOOS/WOMAC Function in ADL" scale in some parts of the text, which indicates that the same results were found for the KOOS Function in ADL scale and the WOMAC Function scale.

In addition, short WOMAC function scales were constructed from subsets of the 17 WOMAC function items, including a 7-item scale developed by Liebs[24], an 8-item scale developed by Tubach[25] and a 7-item scale developed by Whitehouse[26]. For all of these short function scales, 0 was the worst possible and 100 was the best possible score.

IRT function bank and CAT scores. Previously in this dissertation (see Chapter III), a function item bank was developed that contained all 22 KOOS ADL and Sport items, following procedures used in previous IRT calibration studies[106] and more recently in PROMIS[71]. In summary, a graded response model was used to calibrate the items, after evaluation of underlying IRT assumptions of item bank unidimensionality, local independence, and item monotonicity. The resulting IRT calibration parameters (item slopes and item difficulty thresholds) then were used to calculate a total item bank score (i.e., theta score) at both the pre-TKR and 6-month post-TKR time points for all patients, based on their responses to all 22 ADL and Sport items.

The item bank also was used to conduct real-data simulations of computerized adaptive tests (CAT)[106, 141]. A simulated CAT selects the most informative items for each patient, using data that has already been collected from the patient and pre-set stopping rules, to derive a CAT score[66, 128]. CATs minimize respondent burden while optimizing test precision, resulting in more efficient measurement[66]. In previous analyses (see Chapter III), two CAT scores were calculated from patient responses to the ADL and Sport items at each time point (pre-TKR and 6 months post-TKR) using two stopping rules: (1) CAT stopped when the standard error (SE) of an individual patient score was ≤0.23 (equal to a reliability ≥0.95) or a maximum of 10 items were administered; or (2) CAT stopped when SE≤0.32 (reliability≥0.90) or at a maximum of 10 items. As shown in Chapter III, all simulated CAT scores were estimated with 3 to 10 items. The mean number of items used in CAT simulations was: (a) pre-TKR, reliability≥0.95=6.9 items; (b) post-TKR, reliability≥0.95=8.3 items; (c) pre-TKR, reliability≥0.90=3.4 items; and (d) post-TKR, reliability≥0.90=5.0 items. All theta and CAT measures were scored to have a mean of 0 and standard deviation of 1 in a combined pre-TKR and post-TKR sample.

SF-36 Health Survey. Unlike the KOOS and WOMAC which are joint-specific, the SF-36 Health Survey is a generic measure of health status, containing 36 items that are not specific to any diagnosis or treatment[30, 31]. Generic measures are expected to reflect both the impact of a primary condition (e.g., knee osteoarthritis in TKR) along with all other comorbid conditions that a patient may have. The SF-36 (Version 2, standard 4-week recall) was scored as eight scales, including scales primarily measuring physical health (Physical Functioning (PF), Role Limitations due to Physical Health (RP) and Bodily Pain (BP)), scales primarily measuring emotional health (Mental Health (MH) and Role Limitations due to Emotional Problems (RE)), and scales that have been shown to be strong to moderate measures of both physical and emotional health (General Health (GH), Vitality (VT) and Social Functioning (SF))[72]. In comparison with the KOOS and the WOMAC, the SF-36 Physical Functioning scale asks about limitations in physical activities due to health in general (versus the KOOS and WOMAC which ask about joint-specific difficulty in function) and the SF-36 Bodily Pain scale asks about pain throughout the body (versus joint-specific pain in the KOOS and WOMAC). All SF-36 scales were scored so 0 was the worst possible and 100 was the best possible score, following the developer's scoring algorithms[142, 143]. Summary Physical (PCS) and Mental (MCS) Component Scores, which are often used in TKR studies, were calculated from all eight scales using the developer's scoring algorithms[73]; PCS and MCS were scored so 50 was the mean and 10 was the standard deviation in the US general population[143]. Reliability and validity of the SF-36 in TKR patients has been established in numerous studies[29, 144].

Analysis

Construct validity, or the extent to which a scale is related to measures in a manner consistent with theory, was evaluated with correlations among measures prior to

TKR and with tests of known-groups validity using cross-sectional (pre-TKR) and longitudinal data. Responsiveness of the scales to TKR also was assessed with the effect size and standardized response mean.

Concurrent validity was evaluated by examining product-moment correlations among measures of the conceptually related (e.g., KOOS pain, SF-36 pain) and conceptually different (e.g., KOOS pain, SF-36 mental health) domains at baseline (pre-TKR). Two-tailed tests were used to determine significant ($p<0.05$) differences between correlations. A number of hypotheses about the magnitude of the correlations were established based on scale content and results from previous studies:

- The correlation between the KOOS Pain and Function in ADL scales was hypothesized to be high ($r>0.70$), as was the correlation between the WOMAC Pain and Function scales[145].

- The correlation between the KOOS Pain and SF-36 Bodily Pain scales was hypothesized to be moderate (0.40-0.70), as was the correlation between the KOOS/WOMAC Function in ADL and SF-36 Physical Functioning scales[27, 46].

- All KOOS and WOMAC scales were hypothesized to have low (<0.40) correlations with the SF-36 Mental Health (MH) scale because joint-specific impairment was hypothesized to be only weakly related to mental health[27].

- The short WOMAC function scales (Liebs, Tubach, Whitehouse) and 17-item WOMAC Function scale all were hypothesized to have similar correlations with other scales, because the short scales contain WOMAC Function items only.

- Because the KOOS-PS contains both ADL and Sport items, the correlation of the KOOS-PS with other scales were hypothesized to be between the correlation of the KOOS Function in ADL scale and the correlation of the KOOS Sport scale.

- IRT theta and CAT scores were hypothesized to have stronger correlations with the KOOS Function in ADL scale than with other knee-specific scales, because the IRT theta and CAT scores primarily contain ADL items. The IRT theta and CAT scores were hypothesized to have moderate correlations with the KOOS Sport scale.

- The remaining knee-specific scales (KOOS Symptoms, KOOS QOL, WOMAC Stiffness) were hypothesized to have higher correlations with other knee-specific scales than with generic SF-36 scales because the latter are not joint-specific.

Cross-sectional known groups validity. All measures were compared in terms of how well they discriminated according to conceptually-related external variables at a point in time, using baseline (pre-TKR) data. Known groups were defined that were expected to show differing patterns of relationships with knee-specific and generic measures:

- *Surgical knee pain*. The KOOS includes one item about the frequency of pain in the surgical knee. It was hypothesized that patients reporting more frequent knee pain would have worse scores on all knee-specific scales, particularly knee-specific pain scales. (To avoid criterion-measure confounding, this item was not included in calculating the KOOS pain scale for this known group comparison only). It also was hypothesized that knee pain would be weakly related to generic SF-36 scales, except for the SF-36 Bodily Pain scale for which a strong relationship was hypothesized.

- *Assistive walking device*. Patients who used a walking device (cane, crutch or wheelchair) prior to TKR were hypothesized to have poorer scores on all knee-specific scales and on SF-36 scales measuring physical health but not mental health[47, 146].

- *Comorbidity*. Comorbid condition groups (0, 1, 2+ conditions) were defined using a 14-item self-reported condition checklist based on the Charlson index[147]; conditions were: COPD, connective tissue disease, diabetes, cancer, liver disease, peripheral vascular disease, kidney disease, ulcer disease, AIDS, paralysis, heart attack/congestive heart failure/CABG, procedure to unblock neck blood vessels, stroke, and other (non-osteoarthritis). It was hypothesized that there would be a weaker relationship between the number of comorbid conditions and all knee-specific scales, in comparison with generic SF-36 scales.

In all known group validity tests, the total item bank (theta) score calculated from all 22 function items was hypothesized to perform better than the KOOS and WOMAC function scales and the CAT scores, because theta is scored using IRT parameters for all 22 function items and thus has greater precision[128]. CAT scores (calculated with 3 to 10 items) were hypothesized to perform better than the 7 and 8 item fixed-length function scales (short WOMAC function scales, KOOS-PS), because CATs use the most informative items for each patient to derive a more precise score using an IRT metric.

The relative performance of the scales in discriminating among groups was compared using one-way analysis of variance (ANOVA) and relative validity (RV) statistics[72, 138]. For each known group validity test, an ANOVA was conducted for each scale, with levels of the known group as the dependent variable and the scale as the independent variable, holding sample size constant across ANOVAs. The ANOVA F-statistic indicates how strongly a scale discriminates between groups, thus providing information about the scale's validity. To facilitate comparisons across scales, RV statistics were calculated. RV statistics express in proportional terms the empirical validity of a scale relative to the most valid scale. Within each known group test, the scale with the highest F-statistic has an RV of 1.0 and all other scales have an RV less than 1.0, based on the ratio of the F-statistic for that scale to the F-statistic for the best performing scale. 95% confidence intervals for RV statistics were estimated using empirical bootstrap[148, 149]; this allowed statistically significant differences between RV statistics to be identified within each known group validity test.

***Longitudinal known groups validity*.** Validity also was assessed longitudinally, using hypotheses about the performance of change scores (6 month post-TKR minus pre-TKR scores) in relation to external markers of change. For this purpose, overall changes (better/same/worse) were rated by patients six months after TKR. To distinguish different domains of outcomes, patients were asked to compare their status at six months post-TKR to their status before surgery in four areas: capability to do *everyday physical activities* (e.g., walking, climbing stairs, sports), ability to accomplish their *daily work role* (including work at home and in the workplace), feeling bothered by *emotional problems*, and *general health*; the verbatim content of these self-evaluated transition (SET) items is provided in Figure 4.1. For each SET item, patients were classified into four levels (much better, somewhat better, same, worse), as in previous analyses[73]. The four SET items then were used in four separate known groups validity analyses, in which the known groups were the patient's rating of their change in physical activities, daily work role, emotional problems, or general health at six months.

As in the cross-sectional known groups tests of validity (see above), ANOVA models were used to test longitudinal known groups validity, with the 4-level SET item as the independent variable and the change score for each scale (6 months minus pre-TKR) as the dependent variable. Relative validity (RV) statistics were used to compare the responsiveness of all scales in distinguishing among *much better/better/same/worse* groups for each SET item. It was hypothesized that:

- Scales that measured similar constructs as a SET item (e.g., KOOS Function in ADL scale in test of the physical activities SET; KOOS QOL scale in test of the daily work role SET) would have higher RVs than other scales in that SET test.
- Knee-specific measures would be more responsive than generic SF-36 measures, because the SET items asked patients to compare their status at 6 months to their status "before your joint surgery", plus two SET items (physical activities, daily work role) included specific attribution to change "because of your joint surgery".
- SF-36 measures previously shown to be most valid for physical outcomes (PF, RP and BP scales, PCS) would be more responsive than other SF-36 measures.

Effect sizes and standardized response means. The degree of patient-reported change in pain and function after surgery is important in evaluating the success of TKR[21]. In the TKR literature, responsiveness of scales to TKR is often assessed by examining the effect size (ES; observed change score divided by the standard deviation of the baseline score)[150] and the standardized response mean (SRM; observed change score divided by the standard deviation of the change score)[151]. Pre-TKR and 6-month post-TKR data were used to calculate these statistics for all measures.

All analyses were performed using Stata Version 11.2 (StataCorp, Irving, TX).

Results

Data was available for n=1,179 patients at baseline (pre-TKR). The mean age of the sample was 66.1 (SD=9.7); 57% were age 65 or older and 12% were younger than age 55. Sixty-one percent were female. The majority (89.8%) were white, while 7.6% were black and 2.6% reported another race. The highest level of education was high school graduate or less for 28%, while 39% were college graduates or had post-graduate education. Of the n=1,179 patients, 6 month post-TKR data was available for n=886. Six-month data was not available for the remaining 293 patients at the time the dissertation data was made available from the FORCE-TJR database, primarily because the patients were not yet scheduled to complete their 6-month post-TKR survey. In the n=886 sample, 59% were age 65 or older, 10% were younger than age 55, and 62% were female. Mean pre-TKR KOOS Pain scores for the n=1,179 and n=886 samples were 46.4 and 47.4 (p=0.20 for difference between samples), and the mean KOOS ADL scores for the n=1,179 and n=886 samples were 52.9 and 54.0 (p=0.18), respectively.

Missing data and scale reliability. The amount of missing item- and scale-level data at baseline (pre-TKR) was low. Before TKR, the mean percentage of missing data per KOOS item was 1.27% (range 0.51-3.05% missing per item), while the mean percent of missing data per SF-36 item was 0.88% (range 0.25-2.12% missing per item). Scale scores could be calculated for more than 99% of patients for all measures except the SF-36 PCS and MCS, the KOOS Sport/Recreation scale, and KOOS-PS, which could be scored for 98.9%, 98.6% and 94.8% of patients, respectively. Of the 1,179 patients at baseline, 26 were missing one or more of the KOOS or WOMAC scales and another 10 were missing one or more of the SF-36 measures. To maintain a constant sample for all baseline analyses, the sample was restricted to the n=1,143 patients who had scores for all KOOS, WOMAC, short WOMAC function, IRT theta, CAT, and SF-36 measures. An additional 41 patients were missing the KOOS-PS at baseline, so all cross-sectional analyses of the KOOS-PS were based on data from n=1,102 patients. Mean KOOS Function in ADL and Sport scores did not differ significantly for patients with and without KOOS-PS at baseline (mean KOOS ADL score=52.9 versus 50.7 (p=0.46), mean KOOS Sport score=18.3 versus 20.6 (p=0.45), for groups with and without KOOS-PS scores).

Six month post-TKR data was available for 886 of the 1,179 patients. Change scores could be calculated for more than 98% of these patients for all measures except the SF-36 PCS/MCS, KOOS Sport/Recreation scale and KOOS-PS, for whom change scores could be calculated for 97.5%, 96.2% and 94.8% of patients, respectively. Of the 886 patients for whom data was available at 6 months, 49 were missing change scores for one or more of the KOOS or WOMAC scales and another 17 were missing change scores for one or more of the SF-36 measures. To maintain a constant sample for all responsiveness analyses, the sample was restricted to the n=820 patients who had change scores for all KOOS, WOMAC, short WOMAC function, IRT theta, CAT, and SF-36 measures. An additional 31 patients were missing a change score for the KOOS-PS, so all longitudinal analyses of the KOOS-PS were based on data from n=789 patients.

Reliability statistics are provided for baseline (pre-TKR) data in Table 4.1, along with descriptive statistics. Reliability was calculated using Cronbach's coefficient alpha for all scales except the IRT and CAT scores. For IRT and CAT scores, the reliability of each individual patient score was computed based on the standard error of the patient score; the mean reliability across all patients is reported in Table 4.1. Reliability of all scales exceeded the minimum level of 0.70 recommended for group-level analyses[94]; reliability statistics were similar at baseline and 6 months (data not reported).

Concurrent validity. Correlations at baseline (pre-TKR) of all knee-specific scales with each other and with the SF-36 measures are in Table 4.2. As hypothesized, correlations of the KOOS Function in ADL and KOOS Pain scales (r=0.78) and the WOMAC Function and WOMAC Pain scales (r=0.77) were high. The KOOS Pain and WOMAC Pain scales had moderate correlations with the SF-36 Bodily Pain (BP) scale (r=0.66 and 0.63, respectively); these correlations were significantly (p<0.05) higher than correlations of the KOOS Pain and WOMAC Pain scales with the SF-36 Physical Functioning (PF) scale (r=0.49 and 0.50). However, the KOOS Function in ADL and WOMAC Function scales had a significantly (p<0.05) lower correlation with the SF-36 PF scale (r=0.57) than with the SF-36 BP scale (r=0.64). In addition, correlations of the KOOS Sport scale were nearly identical with the SF-36 PF (r=0.44) and BP (r=0.42) scales. Correlations of all KOOS and WOMAC scales were much lower with the SF-36 Mental Health scale (r=0.18-0.34) as hypothesized, demonstrating discriminant validity.

As would be expected, the 7- and 8-item short WOMAC Function scales (Liebs, Tubach, Whitehouse) and 17-item WOMAC Function scale all had similar correlations with other scales. All three short WOMAC Function scales also had significantly (p<0.05) lower correlations with the SF-36 PF scale (r=0.53-0.56) then with the SF-36 BP scale (r=0.62-0.64). The KOOS-PS had high correlations with the KOOS ADL (r=-0.89) and KOOS Sport (r=-0.71) scales, and the correlation of the KOOS-PS with another scale was consistently between the correlations of the KOOS Function in ADL scale and KOOS Sport scale with that scale. The IRT theta function score and CAT function scores had high correlations with the KOOS ADL scale (r=0.88-0.98) but only moderate correlations with the KOOS Sport scale (r=0.48-0.63), reflecting that the IRT item bank and thus the CAT scores mostly contained ADL items.

Among the remaining knee-specific scales, the WOMAC Stiffness scale had a higher correlation with the other WOMAC scales (Pain, Function) than with the generic SF-36 scales. The KOOS Symptoms scale generally had a higher correlation with the other KOOS scales than with the SF-36 scales, although it had a moderate correlation with the SF-36 BP scale (r=0.46). The KOOS QOL scale had moderate (r=0.47-0.59) correlations with the other KOOS scales but also with the SF-36 PF (r=0.50), BP (r=0.53) and Role Physical (RP, r=0.48) scales and the SF-36 PCS (r=0.50).

Cross-sectional known groups validity. Means and standard deviations for all scales along with F-statistics and relative validity (RV) statistics are presented for known groups defined by knee pain frequency, use of an assistive device, and number of comorbid conditions in Tables 4.3-4.5. ANOVA F-statistics and RV statistics for all known groups are summarized in Table 4.6. As expected, scores on all measures became monotonically worse as known group status declined (e.g., the mean KOOS Symptoms score was 66.9 for the group with knee pain “less than daily” compared to 39.6 for the group with knee pain “always”). Results for each known group comparison are summarized below.

As hypothesized, the knee-specific pain scales were most valid at discriminating between groups defined by the frequency of *surgical knee pain*. While the KOOS Pain scale (RV=1.00) had a higher relative validity than the WOMAC Pain scale (RV=0.91), RVs for these two scales were not significantly different (see overlapping confidence intervals (CIs) for their RVs in Table 4.3). Most of the knee-specific function scales had moderate RVs (RV=0.52-0.61); however, the RV for the KOOS Sport scale was lower (RV=0.27) than the RV for all other function measures. RVs for other KOOS and WOMAC scales (KOOS Symptoms, KOOS QOL, WOMAC Stiffness) were moderate

(RV=0.43-0.52). Generic SF-36 scales and summary measures (excluding Bodily Pain, RV=0.76) only weakly (RV=0.05-0.30) discriminated between knee pain groups, as hypothesized.

The generic SF-36 Physical Functioning (PF) scale was best at discriminating between groups who did and did not use an *assistive device* (cane, crutch or wheelchair) (RV=1.00), followed by the SF-36 PCS (RV=0.97) and Role Physical (RV=0.80) scale; the upper bound of the 95% CIs for the latter measures included the value of 1.00 indicating that it cannot be concluded that they discriminated less well than PF. All knee-specific function measures discriminated moderately between assistive device groups (RV=0.53-0.67), with the exception of the KOOS Sport scale (RV=0.28); however, the upper bound of the 95% CI for these function measures did not include 1.00, indicating that none of the knee-specific scales were as valid as the SF-36 PF scale in discriminating between assistive device groups. RVs for other KOOS and WOMAC scales (Symptoms, Stiffness, Pain, QOL) were low to moderate (RV=0.11-0.41).

The SF-36 General Health (GH) scale (RV=1.00) was the most valid measure in discriminating between groups with different numbers of *comorbid conditions* (0, 1, 2+), as might be expected since it is an overall evaluation of health. RVs of all other generic and all knee-specific measures were low in relation to the GH scale. Of note, all knee-specific and generic function scales had similar (RV=0.16-0.24) RVs. In addition, the KOOS Symptoms and QOL scales did not discriminate significantly (p>0.05) between comorbid condition groups, while the KOOS Pain, KOOS Sport, WOMAC Stiffness and WOMAC Pain scales discriminated significantly (p<0.05) but not strongly so. In contrast, all SF-36 measures demonstrated stronger discrimination across groups differing in numbers of comorbid conditions (p<0.01).

***Longitudinal known groups validity*.** Mean change scores and standard deviations for all scales along with F-statistics and RV statistics are presented in Tables 4.7-4.10 for groups defined by self-evaluated transition (SET) items about changes in capability to do *everyday physical activities*, to accomplish *daily work roles* (at home or at work), in *emotional problems*, and in *general health*. ANOVA F statistics and RV statistics for all SET item groups are summarized in Table 4.11.

Mean change scores for all measures generally were monotonically less favorable as self-evaluated transitions went from better to worse. However, the magnitude of the mean change scores differed for knee-specific and generic measures, particularly for the worst group, in all SET tests. Among those who rated themselves as worse 6 months after surgery, mean change scores on all knee-specific measures showed improvement. In contrast, mean change scores for the SF-36 generally remained stable or declined among those who rated themselves as worse 6 months after surgery. For example, the mean change score on the KOOS Function in ADL scale was 12.4 (0.7 SD unit improvement) for the group who rated themselves as “less” capable in doing everyday physical activities. In contrast, the mean change score on the SF-36 Physical Functioning scale for patients who rated their capability to do everyday physical activities as “less” was 3.4 points or 0.15 SD, far below the minimum value that is viewed as an important change[152-154]. In addition, only the generic SF-36 scales showed average declines for the “worse” groups. Mean change scores in the “worse” groups were negative for many SF-36 scales; most notably, those who rated their overall health as “worse” 6 months after surgery had a decline of -12.8 points (more than -0.9 SD) on the SF-36 General Health (GH) scale. The one exception to this pattern was for the SF-36 Bodily Pain (BP) scale, where patients in the “worse” groups had a notable improvement of 0.3-0.6 SD units across the four SET items.

It was hypothesized that change scores for scales that were most closely conceptually related to a SET item would be most responsive in that SET test. All KOOS and WOMAC scales were responsive in all known groups tests of SET items. However, the KOOS QOL scale was the most responsive (RV=1.00) of all measures (both knee-specific and generic) for three of the four SET items (*physical activities*, *daily work role*, *emotional problems*) and also had the highest RV statistic of all knee-specific measures for the fourth SET item (*general health*). For the *physical activities* SET item, the upper bound of the knee-specific IRT theta and CAT function scores approached 1.00, indicating that these measures were nearly as strong as the KOOS QOL scale in responding to the overall change in ability to do physical activities. However, RVs for the fixed-length knee-specific function scales were only moderate (RV=0.43-0.57), and these scales were significantly worse than the KOOS QOL scale in responding to the SET item about overall change in physical capabilities. Similarly, for the SET item about changes in *daily work role*, the upper bound of the 95% CI for all other KOOS scales and all WOMAC scales was below 1.00, indicating that the KOOS QOL scale was significantly more responsive than all other knee-specific scales in relation to the patient’s rating of change in their ability to do everyday work.

The performance of the SF-36 Physical Component Summary (PCS) was not significantly different from that of the KOOS QOL scale for all four SET items; the upper bound of the 95% CI for PCS included 1.00 in three SET comparisons and PCS was the most valid measure in the fourth (*general health*). While the KOOS QOL scale was the most valid measure for the SET item about *emotional problems*, many SF-36 measures (Physical Functioning, Role Physical, Bodily Pain, Vitality, PCS) had RVs that were not significantly different from 1.00. However, RVs for the SF-36 Mental Health scale and the overall Mental Component Summary (MCS) were low (RV=0.29-0.35), indicating that patient answers to the SET item about changes in emotional problems since joint surgery were more related to changes in their physical health than their general mental health. The SF-36 PCS had the highest RV for the SET *general health* item (RV=1.00), but RVs for the SF-36 Bodily Pain scale, KOOS QOL scale and IRT theta score also did not differ significantly from 1.00.

Relative validity statistics did not differ notably for any of the fixed-length knee-specific function measures (KOOS/WOMAC Function in ADL, KOOS Sport, KOOS-PS, short WOMAC scales) in almost all of the longitudinal validity tests, indicating that no knee-specific function measure was better than the others. However, IRT theta (RV=0.67-0.74) and CAT (RV=0.53-0.77) scores consistently had the highest RVs of all function measures across all four SET tests. Furthermore, the KOOS-PS consistently had the lowest RVs (RV=0.33-0.39). In addition, RVs for the KOOS Pain scale were consistently higher than RVs for the WOMAC Pain scale, and RVs for the KOOS Symptoms scale were consistently higher than RVs for WOMAC Stiffness scale.

Effect sizes and standardized response means. Six months after TKR, the KOOS QOL scale had the highest effect size (ES=1.99), followed by the KOOS Pain scale (1.80) in tests of responsiveness (Table 4.12). The ES for the KOOS Pain scale was slightly higher than that for the WOMAC Pain scale (1.63). Effect sizes for IRT theta and CAT function scores (1.72-1.79) were somewhat higher than the ES for all KOOS and WOMAC function scales including the KOOS-PS and short WOMAC function scales (1.36-1.69). Effect sizes were lower for the SF-36, with the highest ES for the Bodily Pain (1.31) and Physical Functioning (1.04) scales and the PCS (1.08).

Standardized response means (SRM) were similar for the KOOS/WOMAC Function in ADL, short WOMAC function, IRT theta and CAT function scales (SRM=1.41-1.59), but lower for the KOOS-PS (1.25) and Sport (1.07) scales. SRM for the KOOS (1.51) and WOMAC (1.47) Pain and KOOS QOL scales (1.46) were similar to those for the function scales. Standardized response means were lower for the SF-36, with the highest SRM for the Bodily Pain (1.06) and Physical Functioning (0.96) scales and the PCS (1.00).

Discussion

This study demonstrated the validity and responsiveness of the KOOS and other knee-specific measures among total knee replacement patients using various methods and criteria. Most hypotheses in tests of concurrent validity and cross-sectional and longitudinal known groups validity were supported, with a few exceptions discussed below. However, knee-specific KOOS and WOMAC measures (with the exception of the KOOS QOL scale) were not more valid than generic SF-36 measures of pain, physical functioning and overall physical health (Physical Component Summary, PCS) in discriminating between groups who rated their outcomes as better, the same or worse at 6 months. The implications of these findings and other results from this study for the measurement of patient-reported outcomes in TKR are discussed below.

Measurement of knee-specific function. A major issue addressed in this paper concerned the relative validity of the different knee-specific function measures that have been proposed to measure function in ADL. How well can knee-specific function be estimated with fewer than the 17 ADL items in the KOOS/WOMAC Function in ADL scale? Almost all of the fixed-length function measures (excluding the KOOS

Sport/Recreation scale) had high correlations with each other, had similar performance in cross-sectional and longitudinal tests of relative validity (RV), and had similar responsiveness statistics. Thus, it could not be concluded that any of the fixed-length short function scales was less valid than the 17-item function scale used in the KOOS and WOMAC.

CAT scores performed significantly better than the short WOMAC function scales (Liebs, Tubach, Whitehouse) and the KOOS-PS in most longitudinal tests of known groups validity. CAT scores consistently had higher RVs than the short function scales (Liebs, Tubach, Whitehouse, KOOS-PS) in all longitudinal validity tests, even though all of these measures were of similar length; CAT scores were estimated in a mean of 3 to 8 items and maximum of 10 items (see Chapter III) and the KOOS-PS and short WOMAC function scales each had 7 or 8 items. These trends suggest that a short knee-specific function measure that targets items to each patient's function level, as a CAT does, may be more useful than a similar length scale that administers the same items to all patients.

In addition, the KOOS-PS was notably weaker than the other short function measures in tests of longitudinal known groups validity and responsiveness. These findings, in addition to the relatively high rate of missing KOOS-PS data because its scoring algorithm requires that all 7 KOOS-PS items be answered, suggest that the KOOS-PS is not a preferred measure for use in TKR.

***Performance of KOOS Function in Sport/Recreation scale in TKR*.** The KOOS Sport/Recreation scale did not discriminate well among known groups in cross-sectional tests of validity, which may have been due to the low functional level of patients prior to TKR and the resulting lack of variation in Sport scores. Six months after TKR, the Sport scale performed as well as the other function measures in longitudinal known groups validity analyses, and the Sport scale had an equivalent effect size as the KOOS Function in ADL scale. However, the Sport scale had a much higher standard deviation after TKR than before TKR, and therefore its standardized response mean was low in comparison with its effect size. The higher post-TKR variation in Sport scores may reflect differences in trajectories of functional recovery. However, some variation also may reflect patient preferences for engaging in Sport activities. In an early KOOS study, Roos noted that the activities included in the Sport scale (squatting, running, jumping, twisting, kneeling) were very important to only about 50% of TKR patients[27]. Thus, as opposed to ADL activities such as walking that all patients would find relevant, some patients may not attempt many Sport activities after TKR, and their overall functional improvement may not be reflected in the Sport items. Additional items that ask about more difficult activities but which are more applicable to the broader TKR population may need to be developed and tested, to measure higher levels of function after TKR.

***Discriminant validity of KOOS Pain and Function in ADL scales*.** Previous studies have found that the WOMAC Pain and Function scales have high correlations and thus have questioned the extent to which these scales measure distinct constructs[61, 96, 98]. The high correlations have been attributed in part to content overlap, since items about the same activities (e.g., pain walking, difficulty walking) are included in both scales [97]. Because the KOOS Function in ADL scale is the same as the WOMAC Function scale, and because the 9-item KOOS Pain scale includes all 5 WOMAC Pain items, it is not surprising that issues that were raised previously for the WOMAC surfaced again in this study for the KOOS. The KOOS Pain and KOOS Function in ADL scales had a high correlation, and the KOOS Function in ADL scale had a significantly higher correlation with the SF-36 Bodily Pain scale than with the SF-36 Physical Functioning scale, indicating a lack of discriminant validity of the ADL scale. However, in tests of relative validity, the KOOS Pain scale had significantly higher relative validity than the ADL scale in the test involving the frequency of surgical knee pain, and the ADL scale had higher relative validity than the Pain scale in the test involving use of an assistive device. Thus, while the KOOS Pain and Function in ADL scales are highly related, the scales also performed differently in relation to other criteria, lending some support to their distinctiveness. These results also underscore the importance of looking at validity in relation to external criteria, not just in terms of scale correlations.

***Comparison of KOOS and WOMAC Pain scales*.** The 9-item KOOS Pain and 5-item WOMAC Pain scales had a high correlation (r=0.94) and their relative validity was not significantly different across cross-sectional tests. However, RV statistics were higher for the KOOS Pain scale (RV=0.51-0.62) than the WOMAC Pain scale (RV=0.36-0.49) in longitudinal validity tests, and the KOOS Pain scale had a higher effect size and standardized response mean than the WOMAC scale. Thus, it may be worthwhile to administer the KOOS Pain scale instead of the WOMAC Pain scale, in spite of the slightly higher respondent burden of the KOOS measure.

***Comparison of KOOS Symptoms and WOMAC Stiffness scales*.** The KOOS Symptoms scale adds 5 items about joint symptoms to the WOMAC Stiffness scale. The KOOS scale contains relatively heterogeneous item content, and the KOOS and WOMAC scales only had a moderately high correlation (r=0.72). In longitudinal validity comparisons of better, same and worse outcome groups, the KOOS scale had higher RVs (RV=0.38-0.46) than the WOMAC scale (RV=0.22-0.29), although many of the RVs were low. This finding suggests that it may be worthwhile to administer the KOOS

Symptoms scale instead of the WOMAC Stiffness scale, but this depends on the extent to which joint symptoms are of interest in a particular study.

***KOOS Quality of Life scale*.** The KOOS QOL scale was not hypothesized to perform better than the other KOOS scales in most tests of validity and responsiveness, with the exception of the longitudinal known groups validity test involving the daily work role SET item. However, KOOS QOL was the knee-specific scale with the highest relative validity in all four longitudinal validity tests of SET items, and was the scale with the highest RV of all knee-specific and generic measures in three of the four SET tests. This scale also had the highest effect size 6 months after TKR. Unlike other KOOS and WOMAC scales which focus on the physical impact of a knee problem, the 4-item KOOS QOL scale asks about its cognitive (*awareness of knee problem*), emotional (*troubled by knee problem*), functional (*modification of life style due to knee problem*) and overall (*general difficulty with knee*) impact. Healing from TKR takes place in the mind as well as the knee, as patients adapt both physically and psychologically to a new way of life post-surgery. The KOOS QOL scale does not provide clinicians with a direct measure of whether TKR surgery reduced knee pain and improved function; rather, it provides a more holistic evaluation of the broader life impact of knee problems.

***Comparison of knee-specific and SF-36 measures*.** As in previous studies[27, 155], this study found that knee-specific measures of function and pain had higher effect sizes and standardized response means than generic measures 6 months after TKR; in particular, the KOOS QOL scale had an effect size that was nearly twice that of the most responsive SF-36 measure. However, in tests of longitudinal validity with groups who rated themselves as better, the same or worse 6 months after TKR, generic (SF-36) and knee-specific (KOOS, WOMAC) function and pain scales were equally valid in discriminating between groups. Furthermore, the best SF-36 measure (PCS) was as valid as the most valid knee-specific scale (KOOS QOL) in these longitudinal validity tests. While there may be conceptual or other reasons to include a knee-specific function scale in studies of knee OA and TKR, these results suggest that for measuring function at least, the SF-36 may be sufficient, or that a much shorter knee-specific function measure could be administered along with the SF-36.

It is notable that even patients who rated their status as “worse” 6 months after TKR improved on average on all knee-specific scales. In contrast, mean scores for the “worse” groups generally declined or remained stable on the generic SF-36 measures. This difference between generic and knee-specific results for the “worse” group may reflect the impact of comorbid (non-knee) conditions on health (leading to lower generic change scores) even if there was an actual knee-specific improvement, or it may reflect other factors. From a measurement perspective, however, these results suggest that knee-specific measures are not sufficient by themselves to fully understand patient outcomes after TKR and should be supplemented with generic measures such as the SF-36. Administering fewer knee-specific function items appears to be a more effective way to reduce respondent burden than excluding generic measures.

Study limitations. This study had a number of limitations. First, TKR patients were from high volume orthopedic centers in the US only. Analyses should be replicated with patients from other countries and from lower volume US orthopedic centers. Second, the criteria used to establish the known groups were based on patient self-report; additional analyses using clinician reports to define knee OA severity groups or to rate patient change after TKR should be conducted. Third, only 6-month post-TKR data was available, and results should be re-examined using data collected 1 or more years after surgery to see if longitudinal validity and responsiveness findings are consistent. Fourth, some significant differences in relative validity statistics may not have been detected. The power to detect significant differences in relative validity between two measures is related to a number of factors, including sample size, the correlation between the measures, and the magnitude of the F-statistic for the more valid measure[148]. While the sample for this study was relatively large, the power for some validity tests was below 0.80 and some significant differences between measures may not have been identified. Failure to do so, however, would not change the major conclusions of this paper. Finally, this study was limited to TKR patients. Tests of known groups validity that compare the KOOS and other questionnaires should be conducted for patients with milder knee OA and other knee disorders. Results of this study may not apply to these other patient populations.

Conclusion. In summary, this study demonstrated that the KOOS is valid and responsive in TKR patients in the US. Due to its high responsiveness and ability to discriminate among groups differing in self-evaluated ratings of change post-TKR, use of the KOOS QOL scale in all studies of TKR is encouraged, even if the entire KOOS is not used in the study. Furthermore, the KOOS Symptoms and Pain scales appear to have advantages over their WOMAC counterparts in terms of their validity. Knee-specific function in ADL can be measured with fewer than 17 items, either using computerized adaptive tests or fixed-length short forms, without a reduction in validity or responsiveness. KOOS Sport scores had wide variation after TKR, and other methods for measuring higher levels of function that are more applicable to the entire TKR patient population should be pursued. Finally, to comprehensively understand the health of patients after TKR, generic measures such as the SF-36 are needed in addition to knee-specific measures.

Table 4.1: Summary of measures compared and score interpretation

Questionnaire/Scale	k	Mean*	SD*	Reliab.[†]	Lowest Score	Highest Score
KOOS						
Symptoms	7	48.6	19.8	0.74	Extreme stiffness and other knee symptoms	No stiffness or other knee symptoms
Pain	9	46.4	18.0	0.88	Constant knee pain, extreme pain in activities	No knee pain, no knee pain doing activities
Function in Activities of Daily Living (ADL)	17	52.8	18.3	0.95	Extreme difficulty doing ADL due to knee	No difficulty doing ADL due to knee
Function in Sport/ Recreation	5	18.4	19.6	0.89	Extreme difficulty in sport activities due to knee	No difficulty in sport activities due to knee
Knee-Specific Quality of Life (QOL)	4	25.4	18.0	0.81	Extreme problems with quality of life due to knee	No problems or changes in lifestyle due to knee
KOOS-PS (scored negatively)	7	49.0	14.6	0.86	No difficulty with sport or ADL activities due to knee	Extreme difficulty with sport/ADL due to knee
WOMAC						
Stiffness	2	43.5	22.3	0.78	Extreme knee stiffness on wakening and during day	No knee stiffness on wakening or during day
Pain	5	51.7	18.9	0.84	Extreme knee pain while doing activities	No knee pain while doing activities
Function (same items as KOOS ADL scale)	17	52.8	18.3	0.95	Extreme difficulty doing ADL due to knee	No difficulty doing ADL due to knee
Short WOMAC Function						
Liebs	7	48.0	18.7	0.89	See WOMAC Function	See WOMAC Function
Tubach	8	49.9	18.6	0.90	See WOMAC Function	See WOMAC Function
Whitehouse	7	54.2	18.5	0.89	See WOMAC Function	See WOMAC Function
IRT-based KOOS ADL/Sport						
IRT Theta score	22	-0.41	0.74	0.98	Extreme difficulty with ADL/sport due to knee	No difficulty with ADL or sport due to knee
CAT $R_{TT}\geq0.95$[‡]	3-10	-0.41	0.72	0.95	See IRT Theta score	See IRT Theta score
CAT $R_{TT}\geq0.90$[‡]	3-10	-0.40	0.71	0.92	See IRT Theta score	See IRT Theta score
SF-36 Health Survey						
Physical Functioning (PF)	10	38.6	22.1	0.87	Limited in all activities including bathing/dressing	Not limited in vigorous activities due to health
Role Physical (RP)	4	43.5	27.3	0.93	Extreme problems with role due to physical health	No problems with role due to physical health
Bodily Pain (BP)	2	36.0	18.4	0.77	Very severe and extremely limiting pain	No pain or limitations due to pain
General Health (GH)	5	70.0	18.8	0.77	Evaluates health as poor	Rates health as excellent
Vitality (VT)	4	51.8	20.8	0.82	Tired and worn out all of the time	Full of energy all of the time
Social Functioning (SF)	2	67.0	27.6	0.83	Health interferes with social activities extremely	No interference of health with social activities
Role Emotional (RE)	3	73.8	28.6	0.92	Extreme problems with role due to emotions	No problems with role due to emotional problems
Mental Health (MH)	5	73.4	18.9	0.85	Feels nervous and depressed all of the time	Feels calm, peaceful and happy all of the time
Physical Component Summary (PCS)	35	33.2	8.4	0.92	Limited in self-care, work, daily activities; severe pain/fatigue; health poor	No limits in physical, work, daily activities; no pain; high energy; health excellent
Mental Component Summary (MCS)	35	51.7	11.9	0.92	Frequent mental distress, limited in activities due to emotional problems	Frequent positive affect, not limited in activities due to emotional problems

k=number of items.
Mean/SD/reliability data are for pre-TKR patients with scores for all scales (n=1,143), except for KOOS-PS (n=1,102).
* All measures scored so 0=worst possible/100=best possible score, except KOOS-PS (0=best/100=worst poss ble), IRT and CAT ADL/Sport (FORCE-TJR pre/post mean=0, SD=1), SF-36 PCS/MCS (US general population mean=50, SD=10).
† Cronbach's alpha for all scales, except IRT theta/CAT scores where reliability=mean reliability across all patient scores.
‡ CAT $R_{TT}\geq0.95$=CAT stopped when SE≤0.23 or at 10 items; CAT $R_{TT}\geq0.90$=CAT stopped at SE≤0.32 or at 10 items.

Table 4.2: Correlations of knee-specific and SF-36 measures, pre-TKR (n=1,143)

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
KOOS															
1) Symptoms	(.74)														
2) Pain	0.67	(.88)													
3) ADL	0.55	0.78	(95)												
4) Sport	0.43	0.51	0.55	(.89)											
5) QOL	0.47	0.57	0.59	0.54	(.81)										
6) **KOOS-PS**	-0 54	-0.71	-0 89	-0.71	-0.57	(.86)									
WOMAC															
7) Stiffness	0.72	0.65	0.61	0.42	0.45	-0.58	(.78)								
8) Pain	0.57	0.94	0.77	0.45	0 54	-0.67	0 58	(.84)							
9) ADL	0.55	0.78	1.00	0.55	0 59	-0.89	0 61	0.77	(.95)						
Short WOMAC Function															
10) Liebs	0.53	0.76	0.94	0.55	0 60	-0.85	0 58	0.76	0 94	(.89)					
11) Tubach	0.53	0.76	0.97	0.56	0 60	-0.86	0 58	0.76	0 97	0.96	(.90)				
12) Whitehouse	0.54	0.76	0.97	0.51	0 56	-0.87	0 61	0.76	0 97	0.93	0 97	(.89)			
IRT-based KOOS ADL/Sport															
13) IRT Theta	0.56	0.78	0.98	0.63	0 61	-0.91	0 61	0.77	0 98	0.94	0 96	0.95	(.98)		
14) CAT≥0.95*	0.52	0.76	0.95	0.51	0 58	-0.83	0 59	0.76	0 95	0.91	0 93	0.93	0 96	(.95)	
15) CAT≥0.90*	0.48	0.71	0.88	0.48	0 57	-0.76	0 52	0.72	0 88	0.83	0 86	0.83	0 90	0.95	(.92)
SF-36															
PF	0.38	0.49	0.57	0.44	0 50	-0.51	0 37	0.50	0 57	0.55	0 56	0.53	0 58	0.56	0.56
RP	0.36	0.49	0.53	0.42	0.48	-0.48	0 36	0.47	0 53	0.51	0 53	0.50	0 54	0.51	0.50
BP	0.46	0.66	0.64	0.42	0 53	-0.57	0 50	0.63	0 64	0.62	0 64	0.62	0 64	0.63	0.61
GH	0.19	0.29	0.31	0.18	0 26	-0.27	0.17	0.31	0 31	0.27	0 29	0.30	0 31	0.30	0.30
VT	0.30	0.41	0.46	0.31	0.40	-0.40	0 30	0.40	0.46	0.40	0.43	0.43	0.46	0.45	0.45
SF	0.37	0.49	0.53	0.32	0.45	-0.47	0 34	0.48	0 53	0.48	0 51	0.50	0 53	0.52	0.53
RE	0.28	0.38	0.43	0.20	0 31	-0.36	0 24	0.38	0.43	0.38	0.40	0.41	0.42	0.42	0.40
MH	0.24	0.33	0.34	0.18	0 30	-0.29	0 21	0.34	0 34	0.29	0 31	0.32	0 33	0.32	0.32
PCS	0.37	0.52	0.56	0.46	0 50	-0.51	0.40	0.51	0 56	0.55	0 56	0.52	0 58	0.55	0.54
MCS	0.26	0.36	0.38	0.18	0 31	-0.33	0 22	0.36	0 38	0.33	0 35	0.37	0 37	0.38	0.37

All measures scored so lower score=poorer health, except for KOOS-PS where lower score=better health.
KOOS-PS N=1,102. SE for all correlations=0.029 except for KOOS-PS correlations where SE=0.030.
Internal consistency reliability for knee-specific measures on diagonal.
* CAT≥0.95=CAT stopped when SE≤0.23 or at 10 items; CAT≥0.90=CAT stopped at SE≤0.32 or at 10 items.

Table 4.3: Relative validity tests for surgical knee pain groups, pre-TKR (n=1,137)

	Mean (Standard Deviation)			F statistic	RV (95% CI)
	Pain Less Than Daily (n=76)	Pain Daily (n=649)	Pain Always (n=412)		
KOOS					
Symptoms	66.9	52.1	39.6	100.62	0.43
	(20.0)	(17.9)	(18.4)		(0.34-0.56)
Pain*	75.6	54.0	38.2	232.33	**1.00**
	(17.6)	(15.6)	(15.9)		-
ADL	71.0	56.8	43.1	137.98	0.59
	(18.2)	(15.9)	(16.8)		(0.48-0.70)
Sport	32.3	21.3	11.0	61.72	0.27
	(26.7)	(19.6)	(14.9)		(0.18-0.37)
QOL	42.2	28.9	16.6	115.34	0.50
	(22.9)	(16.9)	(13.9)		(0.38-0.65)
KOOS-PS (-)	37.3	45.6	56.5	114.78	0.49
	(12.0)	(12.1)	(15.1)		(0.40-0.64)
WOMAC					
Stiffness	66.1	47.7	32.6	121.47	0.52
	(21.5)	(20.1)	(20.3)		(0.40-0.67)
Pain	76.3	55.9	40.3	211.57	**0.91**
	(18.4)	(15.7)	(16.5)		(0.84-1.00)
ADL	71.0	56.8	43.1	137.98	0.59
	(18.2)	(15.9)	(16.8)		(0.48-0.70)
Short WOMAC Function					
Liebs	67.9	51.7	38.5	134.01	0.58
	(19.7)	(16.4)	(16.8)		(0.46-0.68)
Tubach	69.2	53.9	40.0	142.01	0.61
	(19.0)	(16.1)	(17.0)		(0.49-0.73)
Whitehouse	72.6	58.2	44.4	137.02	0.59
	(17.9)	(16.1)	(17.1)		(0.48-0.70)
IRT-based KOOS ADL/Sport					
Theta score	0.34	-0.24	-0.80	142.12	0.61
	(0.76)	(0.62)	(0.70)		(0.50-0.72)
CAT $R_{TT}\geq0.95$	0.33	-0.26	-0.80	139.11	0.60
	(0.75)	(0.61)	(0.68)		(0.48-0.71)
CAT $R_{TT}\geq0.90$	0.29	-0.26	-0.75	120.47	0.52
	(0.75)	(0.61)	(0.68)		(0.41-0.64)
SF-36					
PF	52.4	40.7	32.4	36.13	0.16
	(21.4)	(21.6)	(21.1)		(0.09-0.23)
RP	62.4	46.9	34.3	50.71	0.22
	(25.4)	(26.1)	(26.4)		(0.14-0.30)
BP	58.8	39.7	25.8	176.81	0.76
	(20.1)	(16.6)	(14.3)		(0.59-0.93)
GH	77.0	71.1	67.0	11.76	0.05
	(17.6)	(18.1)	(19.6)		(0.02-0.09)
VT	67.5	53.8	45.5	46.63	0.20
	(18.9)	(19.3)	(21.3)		(0.13-0.29)
SF	80.1	70.5	58.7	34.49	0.15
	(23.4)	(26.0)	(28.7)		(0.09-0.22)
RE	81.2	77.1	67.4	17.81	0.08
	(23.9)	(26.5)	(31.3)		(0.04-0.13)
MH	81.1	75.0	69.4	18.05	0.08
	(16.0)	(17.9)	(20.1)		(0.03-0.13)
PCS	40.7	34.2	30.1	70.12	0.30
	(7.2)	(8.0)	(8.0)		(0.21-0.40)
MCS	55.4	52.9	49.1	17.44	0.08
	(9.8)	(11.3)	(12.8)		(0.04-0.13)

All measures scored so a lower score=poorer health, except for KOOS-PS where lower score=better health. KOOS-PS N=1,097. **Bold**=most valid scale or scale RV not significantly different from 1.00. All F-statistics p<0.0001.
* Frequency of knee pain item not included in KOOS Pain scale.

Table 4.4: Relative validity tests by assistive walking device use, pre-TKR (n=1,142)

	Mean (Standard Deviation)		F statistic	RV (95% CI)
	No Device (n=790)	Use Device (n=352)		
KOOS				
Symptoms	50.4 (18.8)	44.7 (21.3)	20.59	0.11 (0.04-0.22)
Pain	49.3 (17.1)	39.8 (18.1)	73.04	0.38 (0.24-0.59)
ADL	56.5 (16.9)	44.4 (18.5)	118.17	0.62 (0.44-0.90)
Sport	21.1 (19.6)	12.2 (18.1)	52.93	0.28 (0.15-0.46)
QOL	27.6 (17.8)	20.5 (17.5)	39.79	0.21 (0.11-0.34)
KOOS-PS (-)	46.2 (12.8)	55.3 (16.2)	100.14	0.53 (0.35-0.80)
WOMAC				
Stiffness	45.6 (21.8)	38.7 (22.8)	23.70	0.12 (0.05-0.24)
Pain	54.9 (17.9)	44.5 (19.3)	77.68	0.41 (0.25-0.61)
ADL	56.5 (16.9)	44.4 (18.5)	118.17	0.62 (0.44-0.90)
Short WOMAC Function				
Liebs	51.6 (17.5)	40.1 (18.8)	100.89	0.53 (0.35-0.76)
Tubach	53.7 (17.3)	41.4 (18.7)	117.05	0.62 (0.43-0.88)
Whitehouse	57.8 (17.1)	46.0 (18.9)	109.39	0.57 (0.40-0.86)
IRT-based KOOS ADL/Sport				
Theta score	-0.25 (0.67)	-0.76 (0.76)	127.45	0.67 (0.49-0.97)
CAT R_{TT}≥0.95	-0.27 (0.66)	-0.75 (0.74)	120.46	0.63 (0.45-0.90)
CAT R_{TT}≥0.90	-0.25 (0.65)	-0.72 (0.73)	116.35	0.61 (0.42-0.86)
SF-36				
PF	44.1 (20.9)	26.0 (19.3)	190.27	**1.00** -
RP	49.7 (26.0)	29.4 (24.8)	153.08	**0.80** (0.58-1.09)
BP	39.3 (17.7)	28.4 (17.4)	92.05	0.48 (0.32-0.72)
GH	73.8 (16.7)	61.4 (20.5)	116.63	0.61 (0.41-0.94)
VT	54.9 (19.8)	44.7 (21.2)	62.00	0.33 (0.18-0.52)
SF	72.8 (25.3)	53.7 (28.0)	130.74	**0.69** (0.46-1.02)
RE	78.4 (26.1)	63.6 (31.1)	69.57	0.37 (0.21-0.59)
MH	75.7 (18.2)	68.4 (19.5)	36.94	0.19 (0.09-0.37)
PCS	35.3 (7.8)	28.5 (7.9)	184.63	**0.97** (0.78-1.20)
MCS	53.3 (11.4)	48.1 (12.4)	47.38	0.25 (0.12-0.44)

All measures scored so a lower score=poorer health, except for KOOS-PS where lower score=better health. KOOS-PS N=1,101. **Bold**=most valid scale or scale RV not significantly different from 1.00. All F-statistics p<0.0001.

Table 4.5: Relative validity tests by number of chronic conditions, pre-TKR (n=1,143)

	Number of Comorbid Conditions			F statistic	RV (95% CI)
	0 (n=488)	1 (n=405)	2+ (n=250)		
KOOS					
Symptoms	48.9	48.8	47.8	0.29§	0.01
	(20.1)	(19.7)	(19.4)		(0.00-0.02)
Pain	47.8	46.6	43.4	5.04†	0.11
	(17.7)	(17.7)	(18.5)		(0.02-0.26)
ADL	54.7	53.3	48.3	10.57	0.22
	(17.9)	(18.5)	(17.9)		(0.08-0.45)
Sport	19.5	18.8	15.5	3.64‡	0.08
	(19.3)	(19.6)	(20.0)		(0.00-0.22)
QOL	26.1	25.9	23.5	1.90§	0.04
	(18.3)	(18.1)	(17.0)		(0.00-0.12)
KOOS-PS (-)	47.5	48.6	52.6	10.22	0.21
	(13.5)	(14.7)	(15.8)		(0.08-0.48)
WOMAC					
Stiffness	44.8	44.2	39.9	4.33‡	0.09
	(22.9)	(22.0)	(21.2)		(0.01-0.24)
Pain	53.2	51.9	48.5	5.16†	0.11
	(18.7)	(18.8)	(19.2)		(0.02-0.25)
ADL	54.7	53.3	48.3	10.57	0.22
	(17.9)	(18.5)	(17.9)		(0.08-0.45)
Short WOMAC Function					
Liebs	49.7	48.3	44.2	7.46*	0.16
	(18.6)	(18.7)	(18.4)		(0.04-0.35)
Tubach	51.7	50.3	45.6	9.28*	0.19
	(18.4)	(18.8)	(17.9)		(0.07-0.41)
Whitehouse	56.1	54.6	49.7	10.29	0.21
	(18.3)	(18.6)	(18.0)		(0.07-0.42)
IRT-based KOOS ADL/Sport					
Theta score	-0.32	-0.39	-0.60	11.54	0.24
	(0.72)	(0.75)	(0.72)		(0.10-0.48)
CAT R_{TT}≥0.95	-0.34	-0.41	-0.57	8.99*	0.19
	(0.71)	(0.74)	(0.71)		(0.06-0.38)
CAT R_{TT}≥0.90	-0.32	-0.40	-0.54	7.89*	0.16
	(0.69)	(0.74)	(0.70)		(0.05-0.36)
SF-36					
PF	40.4	39.9	32.8	11.16	0.23
	(21.2)	(23.2)	(21.1)		(0.08-0.42)
RP	46.5	43.6	37.4	9.38*	0.20
	(26.7)	(27.9)	(26.6)		(0.06-0.37)
BP	37.9	36.5	31.2	11.56	0.24
	(18.5)	(18.7)	(16.9)		(0.09-0.45)
GH	75.5	68.4	62.1	47.91	**1.00**
	(16.4)	(18.9)	(19.8)		-
VT	54.3	52.2	46.2	12.98	0.27
	(20.1)	(20.9)	(20.9)		(0.12-0.48)
SF	69.6	68.2	60.0	10.83	0.23
	(26.7)	(27.4)	(28.7)		(0.07-0.43)
RE	75.8	75.1	68.0	6.91†	0.14
	(27.9)	(27.8)	(30.3)		(0.03-0.33)
MH	74.8	74.0	70.0	5.72†	0.12
	(18.0)	(18.8)	(20.4)		(0.02-0.27)
PCS	34.5	33.1	30.6	18.11	0.38
	(8.1)	(8.6)	(8.2)		(0.19-0.62)
MCS	52.6	52.2	49.2	7.38*	0.15
	(11.4)	(12.0)	(12.6)		(0.03-0.32)

All measures scored so a lower score=poorer health, except for KOOS-PS where lower score=better health.
KOOS-PS N=1,102. **Bold**=most valid scale or scale RV not significantly different from 1.00.
All F-statistics p<0.0001 except for *p<0.001, † p<0.01, ‡ p<0.05, § p>0.05.

Table 4.6: Summary of cross-sectional known-groups validity tests, pre-TKR

	F-statistic			Relative Validity (95% CI)		
	Knee Pain	Device	Comorbid	Knee Pain	Device	Comorbid
KOOS						
Symptoms	100.62	20.59	0.29§	0.43 (0.34-0.56)	0.11 (0.04-0.22)	0.01 (0.00-0.02)
Pain	232.33	73.04	5.04†	**1.00** -	0.38 (0.24-0.59)	0.11 (0.02-0.26)
ADL	137.98	118.17	10.57	0.59 (0.48-0.70)	0.62 (0.44-0.90)	0.22 (0.08-0.45)
Sport	61.72	52.93	3.64‡	0.27 (0.18-0.37)	0.28 (0.15-0.46)	0.08 (0.00-0.22)
QOL	115.34	39.79	1.90§	0.50 (0.38-0.65)	0.21 (0.11-0.34)	0.04 (0.00-0.12)
KOOS-PS (-)	114.78	100.14	10.22	0.49 (0.40-0.64)	0.53 (0.35-0.80)	0.21 (0.08-0.48)
WOMAC						
Stiffness	121.47	23.70	4.33‡	0.52 (0.40-0.67)	0.12 (0.05-0.24)	0.09 (0.01-0.24)
Pain	211.57	77.68	5.16†	**0.91** (0.84-1.00)	0.41 (0.25-0.61)	0.11 (0.02-0.25)
ADL	137.98	118.17	10.57	0.59 (0.48-0.70)	0.62 (0.44-0.90)	0.22 (0.08-0.45)
Short WOMAC Function						
Liebs	134.01	100.89	7.46*	0.58 (0.46-0.68)	0.53 (0.35-0.76)	0.16 (0.04-0.35)
Tubach	142.01	117.05	9.28*	0.61 (0.49-0.73)	0.62 (0.43-0.88)	0.19 (0.07-0.41)
Whitehouse	137.02	109.39	10.29	0.59 (0.48-0.70)	0.57 (0.40-0.86)	0.21 (0.07-0.42)
IRT-based KOOS ADL/Sport						
Theta score	142.12	127.45	11.54	0.61 (0.50-0.72)	0.67 (0.49-0.97)	0.24 (0.10-0.48)
CAT R_{TT}≥0.95	139.11	120.46	8.99*	0.60 (0.48-0.71)	0.63 (0.45-0.90)	0.19 (0.06-0.38)
CAT R_{TT}≥0.90	120.47	116.35	7.89*	0.52 (0.41-0.64)	0.61 (0.42-0.86)	0.16 (0.05-0.36)
SF-36						
PF	36.13	190.27	11.16	0.16 (0.09-0.23)	**1.00** -	0.23 (0.08-0.42)
RP	50.71	153.08	9.38*	0.22 (0.14-0.30)	**0.80** (0.58-1.09)	0.20 (0.06-0.37)
BP	176.81	92.05	11.56	0.76 (0.59-0.93)	0.48 (0.32-0.72)	0.24 (0.09-0.45)
GH	11.76	116.63	47.91	0.05 (0.02-0.09)	0.61 (0.41-0.94)	**1.00** -
VT	46.63	62.00	12.98	0.20 (0.13-0.29)	0.33 (0.18-0.52)	0.27 (0.12-0.48)
SF	34.49	130.74	10.83	0.15 (0.09-0.22)	**0.69** (0.46-1.02)	0.23 (0.07-0.43)
RE	17.81	69.57	6.91†	0.08 (0.04-0.13)	0.37 (0.21-0.59)	0.14 (0.03-0.33)
MH	18.05	36.94	5.72†	0.08 (0.03-0.13)	0.19 (0.09-0.37)	0.12 (0.02-0.27)
PCS	70.12	184.63	18.11	0.30 (0.21-0.40)	**0.97** (0.78-1.20)	0.38 (0.19-0.62)
MCS	17.44	47.38	7.38*	0.08 (0.04-0.13)	0.25 (0.12-0.44)	0.15 (0.03-0.32)

All measures scored so a lower score=poorer health, except for KOOS-PS where lower score=better health.
Bold=most valid scale or scale RV not significantly different from 1.00.
All F-statistics p<0.0001 except for *p<0.001, † p<0.01, ‡ p<0.05, § p>0.05.

Table 4.7: **Relative validity tests for change in physical activities SET item (n=818)**

	Mean Change Score (Standard Deviation)				F statistic	RV (95% CI)
	Lot more capable (n=451)	More capable (n=207)	Same (n=77)	Less capable (n=83)		
KOOS						
Symptoms	30.6	22.4	14.5	8.9	33.28	0.45
	(21.9)	(19.4)	(22.3)	(22.7)		(0.28-.63)
Pain	38.4	30.0	23.6	14.0	42.97	0.58
	(19.9)	(19.0)	(23.5)	(19.9)		(0.40-.79)
ADL	32.6	26.3	20.5	12.4	37.40	0.51
	(17.0)	(17.8)	(18.3)	(19.3)		(0.35-.68)
Sport	37.4	22.8	15.5	11.1	41.96	0.57
	(25.4)	(24.5)	(26.3)	(24.6)		(0.38-.78)
QOL	46.1	31.1	23.3	12.0	73.90	**1.00**
	(22.6)	(21.6)	(22.9)	(22.2)		-
KOOS-PS (-)	-22.9	-16.2	-14.2	-9.2	27.51	0.37
	(14.2)	(14.1)	(15.0)	(15.6)		(0.24-.57)
WOMAC						
Stiffness	32.5	24.3	18.7	9.9	21.42	0.29
	(26.0)	(25.5)	(27.4)	(29.3)		(0.16-.44)
Pain	35.8	29.1	23.7	14.5	32.01	0.43
	(19.7)	(18.8)	(23.2)	(19.4)		(0.28-.61)
ADL	32.6	26.3	20.5	12.4	37.40	0.51
	(17.0)	(17.8)	(18.3)	(19.3)		(0.35-.68)
Short WOMAC Function						
Liebs	37.3	29.6	22.4	15.8	41.06	0.56
	(18.3)	(18.5)	(19.8)	(19.7)		(0.39-.76)
Tubach	34.8	27.3	20.9	13.4	39.42	0.53
	(18.1)	(17.9)	(20.1)	(21.1)		(0.37-.74)
Whitehouse	31.8	25.4	19.6	12.8	31.55	0.43
	(17.3)	(18.2)	(20.2)	(21.2)		(0.29-.62)
IRT-based KOOS ADL/Sport						
Theta score	1.53	1.11	0.84	0.52	54.27	0.73
	(0.78)	(0.77)	(0.71)	(0.78)		(0.57-.98)
CAT R_{TT}≥0.95	1.55	1.17	0.86	0.53	51.24	0.69
	(0.79)	(0.78)	(0.75)	(0.77)		(0.51-.92)
CAT R_{TT}≥0.90	1.51	1.10	0.88	0.47	51.71	0.70
	(0.80)	(0.76)	(0.72)	(0.79)		(0.50-.93)
SF-36						
PF	30.7	18.6	10.2	3.4	52.64	0.71
	(22.2)	(19.5)	(25.1)	(22.1)		(0.50-.99)
RP	32.7	18.7	9.5	1.9	44.96	0.61
	(28.1)	(24.6)	(28.4)	(23.2)		(0.42-.88)
BP	30.5	18.9	13.4	7.9	42.18	0.57
	(22.2)	(19.2)	(16.7)	(19.7)		(0.40-.85)
GH	3.6	-0.5	-0.3	-4.3	10.17	0.14
	(12.7)	(15.3)	(12.8)	(16.4)		(0.05-.25)
VT	12.6	6.3	5.1	-3.2	22.56	0.31
	(17.7)	(15.2)	(17.8)	(20.9)		(0.17-.46)
SF	17.6	9.3	5.8	0.8	15.72	0.21
	(24.7)	(23.7)	(22.1)	(28.3)		(0.11-.35)
RE	11.1	9.8	3.8	0.2	5.22*	0.07
	(26.5)	(26.2)	(24.7)	(25.5)		(0.02-.15)
MH	8.2	5.8	4.3	-0.3	7.70	0.10
	(16.0)	(14.0)	(13.7)	(17.3)		(0.03-.22)
PCS	12.6	6.8	4.1	1.4	65.79	**0.89**
	(8.7)	(7.4)	(8.2)	(7.9)		(.66-1.22)
MCS	2.2	1.9	1.2	-1.0	2.61†	0.04
	(10.2)	(9.5)	(7.5)	(10.7)		(0.00-.09)

All measures scored so lower score=poorer health except KOOS-PS where lower score=better health. KOOS-PS N=787. **Bold**=most valid scale or scale RV not significantly different from 1.00. All F-statistics p<0.0001 except *p<0.01, †p>0.05.

Table 4.8: Relative validity tests for change in daily work role SET item (n=815)

	Mean Change Score (Standard Deviation)				F statistic	RV (95% CI)
	Lot more able (n=415)	More able (n=216)	Same (n=108)	Less able (n=76)		
KOOS						
Symptoms	31.3	23.4	15.1	7.3	38.21	0.43
	(22.3)	(18.0)	(22.0)	(22.6)		(0.28-.62)
Pain	39.5	30.7	21.4	14.9	49.83	0.56
	(20.3)	(18.2)	(20.0)	(20.4)		(0.39-.75)
ADL	33.4	26.8	19.3	12.5	42.43	0.47
	(17.6)	(16.7)	(16.8)	(19.4)		(0.33-.64)
Sport	38.0	24.7	16.0	10.3	43.69	0.49
	(25.9)	(22.5)	(27.4)	(23.8)		(0.33-.68)
QOL	47.6	32.7	22.2	10.5	89.39	**1.00**
	(22.4)	(21.3)	(20.6)	(21.5)		-
KOOS-PS (-)	-23.4	-17.0	-13.0	-9.2	30.89	0.35
	(14.7)	(13.1)	(14.0)	(16.1)		(0.24-.53)
WOMAC						
Stiffness	33.9	23.7	18.4	9.5	26.09	0.29
	(26.3)	(25.1)	(26.1)	(28.7)		(0.18-.44)
Pain	36.6	29.9	21.3	16.5	34.28	0.38
	(20.2)	(18.0)	(20.4)	(20.2)		(0.25-.53)
ADL	33.4	26.8	19.3	12.5	42.43	0.47
	(17.6)	(16.7)	(16.8)	(19.4)		(0.33-.64)
Short WOMAC Function						
Liebs	38.0	30.9	21.2	15.9	46.25	0.52
	(18.9)	(17.0)	(18.7)	(19.6)		(0.35-.70)
Tubach	35.6	28.1	19.9	13.0	45.76	0.51
	(18.7)	(16.6)	(19.1)	(20.5)		(0.35-.69)
Whitehouse	32.6	26.1	18.3	12.4	38.30	0.43
	(17.9)	(17.0)	(18.4)	(21.2)		(0.28-.59)
IRT-based KOOS ADL/Sport						
Theta score	1.57	1.13	0.81	0.53	60.75	0.68
	(0.80)	(0.71)	(0.71)	(0.77)		(0.50-.87)
CAT R_{TT}≥0.95	1.58	1.19	0.85	0.53	56.51	0.63
	(0.82)	(0.73)	(0.72)	(0.77)		(0.46-.82)
CAT R_{TT}≥0.90	1.54	1.14	0.81	0.48	57.69	0.65
	(0.82)	(0.71)	(0.71)	(0.76)		(0.47-.84)
SF-36						
PF	31.2	19.5	13.8	0.3	55.27	0.62
	(21.8)	(20.5)	(23.8)	(22.4)		(0.44-.89)
RP	33.2	20.3	12.5	-0.9	46.35	0.52
	(28.9)	(25.1)	(24.1)	(22.8)		(0.35-.74)
BP	31.8	18.1	15.0	6.8	51.31	0.57
	(22.1)	(19.4)	(17.6)	(17.5)		(0.39-.87)
GH	3.8	-0.1	0.6	-6.2	12.65	0.14
	(12.6)	(15.2)	(13.7)	(15.8)		(0.06-.26)
VT	13.0	6.5	5.8	-5.4	27.72	0.31
	(17.7)	(16.4)	(16.8)	(18.0)		(0.20-.45)
SF	18.5	8.9	6.3	-0.7	19.67	0.22
	(25.5)	(23.3)	(21.0)	(25.9)		(0.12-.36)
RE	11.7	9.0	4.0	0.3	5.66*	0.06
	(26.8)	(24.8)	(25.9)	(27.2)		(0.02-.14)
MH	8.5	5.4	5.1	-1.8	10.40	0.12
	(16.1)	(15.1)	(12.5)	(15.3)		(0.05-.20)
PCS	12.9	7.3	5.4	0.2	69.59	**0.78**
	(8.7)	(7.8)	(7.5)	(7.4)		(.56-1.10)
MCS	2.4	1.5	1.0	-1.4	3.45†	0.04
	(10.4)	(9.3)	(8.3)	(9.8)		(0.01-.08)

All measures scored so lower score=poorer health except KOOS-PS where lower score=better health. KOOS-PS N=785. **Bold**=most valid scale or scale RV not significantly different from 1.00. All F-statistics p<0.0001 except *p<0.001, †p<0.05.

Table 4.9: Relative validity tests for change in emotional problems SET item (n=792)

	Mean Change Score (Standard Deviation)				**F statistic**	**RV (95% CI)**
	Much less bothered (n=305)	Less bothered (n=103)	Same (n=327)	More bothered (n=57)		
KOOS						
Symptoms	30.2	28.6	20.8	15.3	14.03	0.38
	(22.0)	(23.9)	(20.7)	(26.1)		(0.20-.65)
Pain	39.2	34.5	27.9	20.8	23.11	0.62
	(20.1)	(21.4)	(19.9)	(23.1)		(0.38-.94)
ADL	33.3	28.3	24.6	16.8	19.96	0.54
	(16.8)	(19.9)	(17.9)	(21.0)		(0.31-.80)
Sport	35.4	31.0	24.4	14.7	15.33	0.41
	(26.2)	(26.8)	(26.0)	(26.1)		(0.23-.68)
QOL	46.2	35.3	32.6	14.3	37.28	**1.00**
	(23.9)	(23.5)	(22.7)	(25.8)		-
KOOS-PS (-)	-22.6	-20.3	-16.4	-12.8	12.13	0.33
	(14.2)	(15.6)	(14.6)	(17.0)		(0.18-.54)
WOMAC						
Stiffness	32.5	28.5	23.3	19.1	8.17	0.22
	(26.4)	(28.4)	(26.2)	(28.5)		(0.09-.40)
Pain	36.6	33.6	26.8	20.7	18.41	0.49
	(19.7)	(20.9)	(19.4)	(23.0)		(0.28-.78)
ADL	33.3	28.3	24.6	16.8	19.96	0.54
	(16.8)	(19.9)	(17.9)	(21.0)		(0.31-.80)
Short WOMAC Function						
Liebs	37.9	32.0	28.2	18.7	23.00	0.62
	(18.3)	(20.5)	(19.0)	(22.5)		(0.36-.91)
Tubach	35.4	29.8	25.8	17.8	20.93	0.56
	(18.1)	(20.3)	(19.3)	(21.6)		(0.33-.85)
Whitehouse	32.6	27.3	23.8	17.2	17.51	0.47
	(17.3)	(20.1)	(18.8)	(22.0)		(0.25-.71)
IRT-based KOOS ADL/Sport						
Theta score	1.54	1.26	1.08	0.70	27.65	**0.74**
	(0.76)	(0.87)	(0.80)	(0.87)		(.48-1.09)
CAT R_{TT}≥0.95	1.58	1.28	1.10	0.71	28.70	0.77
	(0.77)	(0.88)	(0.80)	(0.91)		(.49-1.12)
CAT R_{TT}≥0.90	1.51	1.24	1.07	0.69	24.49	0.66
	(0.79)	(0.90)	(0.80)	(0.87)		(0.40-.95)
SF-36						
PF	34.2	24.4	17.7	5.4	24.19	**0.65**
	(28.7)	(29.7)	(26.5)	(22.5)		(.38-1.01)
RP	31.2	23.8	19.3	8.3	28.12	**0.75**
	(23.2)	(17.8)	(20.7)	(17.4)		(.40-1.17)
BP	31.2	24.5	19.5	12.8	27.73	**0.74**
	(23.3)	(18.9)	(20.8)	(17.4)		(.45-1.18)
GH	4.2	2.3	0.1	-6.9	12.27	0.33
	(12.8)	(14.8)	(13.1)	(19.6)		(0.11-.61)
VT	14.5	10.5	5.3	-6.8	31.41	**0.84**
	(17.5)	(17.6)	(16.7)	(17.5)		(.52-1.34)
SF	19.2	14.4	8.9	-3.9	18.87	0.51
	(26.0)	(24.5)	(22.8)	(23.9)		(0.27-.86)
RE	13.8	10.4	7.1	-4.2	8.94	0.24
	(27.4)	(27.9)	(24.4)	(24.3)		(0.09-.45)
MH	8.6	10.6	4.9	-3.4	13.18	0.35
	(16.1)	(17.2)	(13.5)	(19.0)		(0.16-.72)
PCS	12.6	8.7	7.4	2.5	31.36	**0.84**
	(9.2)	(8.0)	(8.5)	(8.3)		(.51-1.28)
MCS	3.2	3.6	0.9	-4.0	10.89	0.29
	(10.3)	(10.7)	(8.8)	(10.7)		(0.12-.57)

All measures scored so lower score=poorer health except KOOS-PS where lower score=better health. KOOS-PS N=763. **Bold**=most valid scale or scale RV not significantly different from 1.00. All F-statistics are p<0.0001.

Table 4.10: Relative validity tests for change in general health SET item (n=814)

	Mean Change Score (Standard Deviation)				F statistic	RV (95% CI)
	Much better (n=290)	Better (n=223)	Same (n=252)	Worse (n=49)		
KOOS						
Symptoms	32.5 (20.8)	25.8 (22.1)	17.4 (21.1)	13.4 (25.7)	26.93	0.46 (0.28-.73)
Pain	40.8 (20.1)	31.1 (19.0)	25.7 (20.7)	21.6 (26.0)	30.05	0.51 (0.32-.79)
ADL	34.4 (17.1)	26.8 (15.7)	22.7 (19.3)	17.3 (22.8)	26.13	0.45 (0.25-.70)
Sport	39.3 (26.1)	27.7 (23.5)	20.1 (26.1)	17.6 (30.1)	28.94	0.50 (0.29-.79)
QOL	48.6 (23.7)	34.8 (19.6)	28.6 (23.6)	16.9 (30.7)	48.47	**0.83** (.53-1.26)
KOOS-PS (-)	-24.4 (14.1)	-17.4 (12.6)	-14.5 (15.8)	-16.3 (17.2)	22.56	0.39 (0.25-.73)
WOMAC						
Stiffness	35.0 (25.0)	25.6 (26.8)	20.1 (27.3)	17.3 (31.1)	16.58	0.28 (0.14-.49)
Pain	37.5 (19.8)	29.6 (18.8)	25.8 (20.5)	20.2 (25.8)	20.86	0.36 (0.19-.59)
ADL	34.4 (17.1)	26.8 (15.7)	22.7 (19.3)	17.3 (22.8)	26.13	0.45 (0.25-.70)
Short WOMAC Function						
Liebs	39.0 (18.6)	30.9 (16.2)	25.7 (20.5)	22.0 (24.3)	27.20	0.47 (0.28-.74)
Tubach	36.9 (18.2)	28.2 (16.0)	23.5 (20.9)	19.0 (23.4)	28.66	0.49 (0.29-.77)
Whitehouse	33.6 (17.4)	25.9 (16.3)	22.0 (20.4)	18.0 (22.8)	22.68	0.39 (0.21-.62)
IRT-based KOOS ADL/Sport						
Theta score	1.63 (0.79)	1.17 (0.66)	0.99 (0.84)	0.75 (0.98)	38.88	**0.67** (.41-1.02)
CAT $R_{TT}\geq0.95$	1.62 (0.81)	1.21 (0.70)	1.05 (0.85)	0.72 (0.99)	32.49	0.56 (0.33-.87)
CAT $R_{TT}\geq0.90$	1.57 (0.82)	1.18 (0.71)	1.01 (0.84)	0.67 (0.95)	31.02	0.53 (0.30-.82)
SF-36						
PF	33.0 (23.6)	21.7 (19.7)	16.1 (22.7)	4.5 (25.1)	38.62	0.66 (0.51-.86)
RP	35.5 (29.8)	23.8 (25.6)	14.6 (25.6)	2.0 (25.8)	37.95	0.65 (0.51-.85)
BP	34.9 (22.7)	21.4 (18.2)	16.0 (20.1)	7.2 (16.4)	52.68	**0.90** (.68-1.16)
GH	4.7 (13.1)	2.1 (13.4)	-0.4 (13.2)	-12.8 (16.7)	25.79	0.44 (0.24-.76)
VT	15.6 (19.1)	7.7 (14.0)	4.2 (16.5)	-4.8 (21.6)	31.65	0.54 (0.33-.84)
SF	20.4 (25.5)	12.6 (23.5)	5.9 (23.5)	1.0 (26.4)	19.93	0.34 (0.18-.56)
RE	12.9 (26.1)	9.5 (25.3)	6.1 (23.9)	-3.3 (36.4)	6.93*	0.12 (0.03-.29)
MH	9.3 (17.5)	7.0 (13.6)	3.5 (13.1)	-1.0 (19.7)	10.25	0.18 (0.06-.37)
PCS	13.9 (9.2)	8.6 (7.5)	6.1 (8.1)	0.9 (8.2)	58.42	**1.00** -
MCS	3.0 (10.5)	2.1 (9.1)	0.5 (8.6)	-2.1 (12.8)	5.71*	0.10 (0.02-.23)

All measures scored so lower score=poorer health except KOOS-PS where lower score=better health. N for KOOS-PS=784. **Bold**=most valid scale or scale RV not significantly different from 1.00. All F-statistics p<0.0001 except * p<0.001.

Table 4.11: Summary of longitudinal self-evaluated transition (SET) tests

	F-statistic				Relative Validity (95% CI)			
	Phys. Act.	Role	Emotional	Health	Phys. Act.	Role	Emotional	Health
KOOS								
Symptoms	33.28	38.21	14.03	26.93	0.45 (0.28-.63)	0.43 (0.28-.62)	0.38 (0.20-.65)	0.46 (0.28-.73)
Pain	42.97	49.83	23.11	30.05	0.58 (0.40-.79)	0.56 (0.39-.75)	0.62 (0.38-.94)	0.51 (0.32-.79)
ADL	37.40	42.43	19.96	26.13	0.51 (0.35-.68)	0.47 (0.33-.64)	0.54 (0.31-.80)	0.45 (0.25-.70)
Sport	41.96	43.69	15.33	28.94	0.57 (0.38-.78)	0.49 (0.33-.68)	0.41 (0.23-.68)	0.50 (0.29-.79)
QOL	73.90	89.39	37.28	48.47	**1.00** -	**1.00** -	**1.00** -	**0.83** (.53-1.26)
KOOS-PS (-)	27.51	30.89	12.13	22.56	0.37 (0.24-.57)	0.35 (0.24-.53)	0.33 (0.18-.54)	0.39 (0.25-.73)
WOMAC								
Stiffness	21.42	26.09	8.17	16.58	0.29 (0.16-.44)	0.29 (0.18-.44)	0.22 (0.09-.40)	0.28 (0.14-.49)
Pain	32.01	34.28	18.41	20.86	0.43 (0.28-.61)	0.38 (0.25-.53)	0.49 (0.28-.78)	0.36 (0.19-.59)
ADL	37.40	42.43	19.96	26.13	0.51 (0.35-.68)	0.47 (0.33-.64)	0.54 (0.31-.80)	0.45 (0.25-.70)
Short WOMAC Function								
Liebs	41.06	46.25	23.00	27.20	0.56 (0.39-.76)	0.52 (0.35-.70)	0.62 (0.36-.91)	0.47 (0.28-.74)
Tubach	39.42	45.76	20.93	28.66	0.53 (0.37-.74)	0.51 (0.35-.69)	0.56 (0.33-.85)	0.49 (0.29-.77)
Whitehouse	31.55	38.30	17.51	22.68	0.43 (0.29-.62)	0.43 (0.28-.59)	0.47 (0.25-.71)	0.39 (0.21-.62)
IRT-based KOOS ADL/Sport								
Theta score	54.27	60.75	27.65	38.88	0.73 (0.57-.98)	0.68 (0.50-.87)	**0.74** (.48-1.09)	**0.67** (.41-1.02)
CAT R_{TT}≥0.95	51.24	56.51	28.70	32.49	0.69 (0.51-.92)	0.63 (0.46-.82)	**0.77** (.49-1.12)	0.56 (0.33-.87)
CAT R_{TT}≥0.90	51.71	57.69	24.49	31.02	0.70 (0.50-.93)	0.65 (0.47-.84)	0.66 (0.40-.95)	0.53 (0.30-.82)
SF-36								
PF	52.64	55.27	24.19	38.62	0.71 (0.50-.99)	0.62 (0.44-.89)	**0.65** (.38-1.01)	0.66 (0.51-.86)
RP	44.96	46.35	28.12	37.95	0.61 (0.42-.88)	0.52 (0.35-.74)	**0.75** (.40-1.17)	0.65 (0.51-.85)
BP	42.18	51.31	27.73	52.68	0.57 (0.40-.85)	0.57 (0.39-.87)	**0.74** (.45-1.18)	**0.90** (.68-1.16)
GH	10.17	12.65	12.27	25.79	0.14 (0.05-.25)	0.14 (0.06-.26)	0.33 (0.11-.61)	0.44 (0.24-.76)
VT	22.56	27.72	31.41	31.65	0.31 (0.17-.46)	0.31 (0.20-.45)	**0.84** (.52-1.34)	0.54 (0.33-.84)
SF	15.72	19.67	18.87	19.93	0.21 (0.11-.35)	0.22 (0.12-.36)	0.51 (0.27-.86)	0.34 (0.18-.56)
RE	5.22†	5.66*	8.94	6.93*	0.07 (0.02-.15)	0.06 (0.02-.14)	0.24 (0.09-.45)	0.12 (0.03-.29)
MH	7.70	10.40	13.18	10.25	0.10 (0.03-.22)	0.12 (0.05-.20)	0.35 (0.16-.72)	0.18 (0.06-.37)
PCS	65.79	69.59	31.36	58.42	**0.89** (.66-1.22)	**0.78** (.56-1.10)	**0.84** (.51-1.28)	**1.00** -
MCS	2.61§	3.45‡	10.89	5.71*	0.04 (0.00-.09)	0.04 (0.01-.08)	0.29 (0.12-.57)	0.10 (0.02-.23)

All measures scored so a lower score=poorer health, except for KOOS-PS where lower score=better health.
KOOS-PS N=789. **Bold**=most valid scale or scale RV not significantly different from 1.00.
All F-statistics p<0.0001 except for * p<0.001, † p<0.01, ‡ p<0.05, § p>0.05.

Table 4.12: Responsiveness of knee-specific and SF-36 measures (n=820)

	Pre-TKR		6 month Post-TKR		Change Score			
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**ES**	**SRM**
KOOS								
Symptoms	49.2	19.8	74.1	16.9	24.9	22.7	1.25	1.10
Pain	47.6	18.1	80.1	17.2	32.5	21.5	1.80	1.51
ADL	54.0	18.2	81.8	16.5	27.9	18.8	1.53	1.49
Sport	19.0	19.4	48.1	27.2	29.0	27.1	1.49	1.07
QOL	26.7	18.5	63.4	22.6	36.8	25.2	1.99	1.46
KOOS-PS (-)	48.2	14.1	29.1	13.7	-19.1	15.2	1.36	1.25
WOMAC								
Stiffness	44.0	22.5	71.0	20.4	26.9	27.4	1.20	0.98
Pain	53.1	18.9	84.0	15.9	30.9	20.9	1.63	1.47
ADL	54.0	18.2	81.8	16.5	27.9	18.8	1.53	1.49
Short WOMAC Function								
Liebs	49.0	18.8	80.8	16.9	31.8	20.0	1.69	1.59
Tubach	51.0	18.6	80.5	17.1	29.5	19.9	1.58	1.48
Whitehouse	55.4	18.4	82.5	16.2	27.1	19.3	1.47	1.41
IRT-based KOOS ADL/Sport								
Theta score	-0.36	0.73	0.90	0.83	1.26	0.85	1.72	1.49
CAT $R_{TT}\geq0.95$	-0.37	0.72	0.92	0.82	1.29	0.86	1.79	1.50
CAT $R_{TT}\geq0.90$	-0.35	0.71	0.89	0.81	1.24	0.86	1.76	1.45
SF-36								
PF	40.1	22.1	63.1	24.3	23.0	23.9	1.04	0.96
RP	44.4	27.4	68.2	27.2	23.8	28.9	0.87	0.82
BP	37.0	18.2	60.8	22.9	23.7	22.3	1.31	1.06
GH	71.8	18.1	73.2	19.5	1.4	14.1	0.08	0.10
VT	53.4	20.7	62.1	20.0	8.7	18.1	0.42	0.48
SF	69.7	27.0	82.4	23.1	12.7	25.2	0.47	0.50
RE	76.1	27.7	85.1	22.3	9.0	26.4	0.32	0.34
MH	74.7	18.6	81.1	16.2	6.4	15.7	0.34	0.41
PCS	33.6	8.5	42.8	9.8	9.2	9.2	1.08	1.00
MCS	52.8	11.6	54.5	9.7	1.7	9.9	0.15	0.17

ES=Effect size; SRM=Standardized response mean.
All measures scored so a lower score=poorer health, except for KOOS-PS where lower score=better health.
KOOS-PS N=789.

Figure 4.1: Content of self-evaluated transition items in 6-month follow-up survey

Thinking about your everyday physical activities today (such as walking, climbing stairs, carrying groceries, or participating in sports); Compared to before your joint surgery, are you more or less capable now in your everyday physical activities because of your joint surgery?

- a lot more capable now
- somewhat more capable now
- about the same
- somewhat less capable now
- a lot less capable now

Thinking about your daily work at home or in the workplace; Compared to before your joint surgery are you more or less able to accomplish your work now because of your joint surgery?

- a lot more able to accomplish now
- somewhat more able to accomplish now
- about the same
- somewhat less able to accomplish now
- a lot less able to accomplish now

Compared to the time before your joint surgery, how often do you feel bothered by emotional problems, such as feeling anxious, depressed, or irritable now?

- feel this way a lot more often now
- feel this way somewhat more often now
- feel about the same
- feel this way somewhat less often now
- feel this way much less often now

Compared to the time before your joint surgery, how would you rate your health in general now?

- much better now than before surgery
- somewhat better now than before surgery
- about the same
- somewhat worse now than before surgery
- much worse now than before surgery

Source: FORCE-TJR Six-Month Follow-up Knee Survey. Survey Version Date January 10, 2012

CHAPTER V: DISCUSSION AND CONCLUSIONS

Results from this research support use of the KOOS and WOMAC among total knee replacement patients in the US. (Summary information for all knee-specific measures and selected SF-36 measures is presented in Table 5.1, including data on reliability, tests of scaling assumptions, floor and ceiling effects, validity, and responsiveness). Rates of missing KOOS and WOMAC data were low. Most tests of psychometric properties including scaling assumptions were met, and internal consistency reliability estimates exceeded recommendations for group-level comparisons. These results were consistent across groups differing in age and gender, socioeconomic status and clinical status, with few exceptions. Floor (percent with the worst measured score) and ceiling (percent with the best measured score) effects generally were low and followed expected patterns before and after TKR. KOOS and WOMAC scales demonstrated construct validity; hypothesized patterns of correlations among scales were observed, and knee-specific scales discriminated well between groups differing in knee pain frequency and did not discriminate as well as generic measures between groups differing in the number of comorbid conditions, as hypothesized. KOOS and WOMAC scales were responsive to TKR surgery in terms of statistics such as the effect size and in relation to patient self-evaluated ratings of change after TKR.

By examining numerous knee-specific and generic measures at the same time, this study also provided information that is useful in considering how to measure patient-reported outcomes in TKR more efficiently. Conclusions from these comparative analyses and implications for future measurement development are addressed below.

How well can knee-specific function be estimated with fewer than 17 items?

This study compared the 17-item Function in ADL scale used in both the KOOS and WOMAC with six different shorter function measures: 7 and 8-item short WOMAC function scales developed by Liebs, Tubach and Whitehouse; the 7-item KOOS-PS; a CAT score with a reliability≥0.95 or a maximum of 10 items; and a CAT score with a reliability≥0.90 or a maximum of 10 items. These comparisons indicated the following:

1. Fixed-length short function measures generally had similar validity as the longer Function in ADL scale.

The three short WOMAC function scales had extremely high (r=0.94-0.97) correlations with the KOOS/WOMAC Function in ADL scale. Correlations with the KOOS/WOMAC ADL scale were lower for the KOOS-PS (r=-0.89), but were still high. Thus, it was not surprising that relative validity statistics for all these measures generally were not notably different, within each cross-sectional and longitudinal known groups validity test. In addition, in support of their construct validity, the shorter and full-length function measures had similar correlations with scales measuring other constructs.

2. CAT scores performed better than fixed-length short function scales in tests of known groups validity.

Both CAT scores consistently had higher relative validity than fixed-length short function scales in all longitudinal validity tests. CAT scores with a reliability≥0.95 (a level recommended instead of a reliability≥0.90 when using a CAT with individual patients) were obtained in a mean of 7 to 8 items; similarly, the KOOS-PS and short WOMAC function scales each had 7 or 8 items. However, reliability of the KOOS-PS and short WOMAC function scales ranged from 0.86-0.90, well below the level of 0.95 recommended for use in individual patient monitoring and decision-making. These findings suggest that a short knee-specific function measure that targets items to each individual patient, as a CAT does, can be more useful than a similar length measure that administers the same items to all patients. Further research would need to be conducted with additional patient samples, and the item bank would need to be expanded (see point #5 below), prior to adoption of a KOOS-based function CAT. In addition, the issue of the equivalence and direct comparability of CAT scores and scores estimated from the same items administered to all respondents would need to be addressed. However, this is a promising area for future research.

3. Existing fixed-length short function scales may not contain optimal item content.

IRT models provided a test of how informative each KOOS Function in ADL and Sport item was in estimating function, and simulated CATs picked the most informative items for each patient. Items in the short WOMAC function scales (Liebs, Tubach, Whitehouse) only accounted for 40-56% of all CAT item administrations even for the CAT requiring the most items (CAT stopped when reliability was ≥0.95 or at 10 items), and the KOOS-PS items only accounted for 19-31% of CAT item administrations (Table 5.2). Two highly informative items were never (A17, light domestic duties) or were rarely (A8, going shopping) included in the existing fixed-length short function scales. A fixed-length short function scale that was created based on CAT item usage would have notably different item content than any existing fixed-length short function scale.

4. KOOS-PS had notably weaker performance than other short function scales.

The KOOS-PS has the lowest effect size of all function measures six months after TKR, along with the lowest relative validity in comparisons of groups rating their health as better, same or worse post-TKR. KOOS-PS items generally were not frequently selected in CAT simulations. In addition, there was a higher rate of missing data for the KOOS-PS compared to all other function scales. KOOS-PS scores were missing for 5.2% of patients before and 6 months after TKR, because the KOOS-PS scoring algorithm requires that all seven KOOS-PS items be answered in order for a score to be calculated. Altogether, these results indicate that the KOOS-PS is not a preferred function measure for use in TKR.

5. KOOS and WOMAC do not contain enough relevant items at high function levels.

CAT simulations showed that while a function score could be estimated precisely and efficiently with many fewer than the 22 function items included in the KOOS (or the 17 function items in the WOMAC) for those with severe OA, the item bank was less successful for patients with better levels of knee-specific function. One third of patients could not achieve a function score with a reliability of 0.95 six months post-TKR, even after administration of 10 items; these patients tended to score at higher levels. While the Sport/Recreation items were added to the KOOS in part to measure higher function levels, these items had lower item information functions and thus were not frequently selected in CAT simulations.

Additional knee-specific items need to be written to measure higher levels of function. One way to do this is to write items that ask about activities that are more difficult than those in the Function in ADL scale and are more applicable to an older and sicker patient population than the Sport items. For example, this approach was used to develop the PROMIS Physical Function item bank; the PROMIS bank includes items about more difficult domestic and recreational activities than those in widely-used physical functioning scales contained in measures such as the SF-36 and WOMAC[130]. A second approach is to ask about similar activities as in the Function in ADL scale but use a different set of responses that extend the range of measurement, such as an "easy-difficult" response continuum in which the best level of function is only achieved by patients who report ease in doing an activity, not just that the activity is not difficult. This approach is being evaluated by the developer of the SF-36 as a way to construct improved physical functioning measures[131] and has also been evaluated by others[156].

How useful is the KOOS Sport/Recreation scale among TKR patients?

While the KOOS Function in Sport/Recreation scale had good reliability and met all tests of scaling assumptions, more than one in four patients had the worst possible Sport score prior to TKR, and the mean pre-TKR Sport score was very low. As a result, the Sport scale only had moderate correlations with the SF-36 Physical Functioning scale prior to TKR and did not discriminate between known groups at baseline as well as other function measures. Six months after TKR, though, the Sport scale was as responsive as all other function scales.

However, results also suggest that the Sport scale may have limited applicability to the entire TKR population. There was a notable increase from pre-TKR to post-TKR (from 1.4% to 2.7%) in the percentage of patients for whom a Sport scale could not be calculated because fewer than half of the Sport items were answered. Most Sport items were not as informative as ADL items and were infrequently chosen in CAT simulations, even after TKR. In addition, there was considerable variation in Sport scores post-TKR, with the scale standard deviation increasing from a value of 19.4 pre-TKR to a value of 27.2 six months after TKR in the responsiveness analysis. While this may reflect differences in the trajectories of patient recovery, some of this variation may also reflect differences in patient preferences for sport activities. Some sport activities may not be meaningful to many TKR patients, who may not attempt them post-TKR. While CAT simulations show that more difficult items than those in the Function in ADL scale are needed to monitor TKR patients after surgery, the Sport items do not completely fill that gap in the TKR patient population.

Are the KOOS Pain and KOOS Function in ADL scales distinct?

Previous studies have questioned the extent to which the WOMAC Pain and Function scales are conceptually distinct, because they are highly correlated[61, 96, 98]. These high correlations were attributed at least in part to content overlap, since items about the same activities (e.g., pain walking, difficulty walking) are included in both scales[97]. High correlations also were seen in this study between the KOOS Pain and KOOS Function in ADL scales and between the WOMAC Pain and WOMAC Function scales. However, the Pain and Function scales performed somewhat differently in tests of known groups validity. Thus, this study lends some support to the distinctiveness of the Pain and Function measures. However, this research also suggests that the KOOS and WOMAC Pain scales should primarily be interpreted as measures of pain while doing physical activities.

Is the KOOS Pain scale preferable to the WOMAC Pain scale?

In some respects, the KOOS and WOMAC Pain scales are similar. As would be expected for two scales with five items in common, the scales were highly correlated (r=0.94). Both scales had good internal consistency reliability, and scale scores could be calculated for almost all patients at both time points. However, the four additional pain items in the KOOS Pain scale do not have content overlap with the Function in ADL items. As a result, the KOOS Pain scale had better item discriminant validity than the WOMAC Pain scale in multitrait scaling tests. The KOOS Pain scale also had a lower proportion of patients with the best possible score 6 months post-TKR (14% for KOOS versus 22% for WOMAC). The KOOS Pain scale also had higher relative validity than the WOMAC Pain scale in longitudinal known groups validity tests, and the KOOS Pain scale also had a larger effect size and standardized response mean. Although the KOOS Pain scale is longer than the WOMAC scale, the additional items notably enhanced scale performance. Thus, the KOOS Pain scale is recommended over the WOMAC Pain scale, in spite of the greater respondent burden of the KOOS Pain scale.

Is the KOOS Symptoms scale preferable to the WOMAC Stiffness scale?

It is less clear whether to recommend the KOOS Symptoms scale over the WOMAC Stiffness scale. The KOOS Symptoms scale is relatively heterogeneous and contains items about a variety of knee symptoms, while the WOMAC Stiffness scale only includes two items about stiffness. Both scales had similar internal consistency reliability, but for different reasons. Items in the KOOS Symptoms scale were not highly correlated with each other, but this scale has seven items which increased its internal consistency reliability since reliability is based on the average inter-item correlation and number of items in a scale. Items in the WOMAC Stiffness scale were highly correlated but the scale only contains two items. Six-months after TKR, the WOMAC Stiffness scale had a higher percentage of patients with the best possible score (16%) than the KOOS Symptoms scale (4%), indicating that a notable percentage of TKR patients experience symptoms post-TKR that are not captured by the WOMAC scale. In addition, the KOOS Symptoms scale better discriminated between groups with different self-rated longitudinal outcomes than the WOMAC Stiffness scale. These findings suggest that there may be advantages to using the KOOS Symptoms scale in studies of TKR. However, this scale’s relatively low item homogeneity, which is often seen in scales of symptoms that largely vary independently, indicate that it may benefit from separate scoring and interpretation of its stiffness and non-stiffness components in addition to its overall score. In addition, the WOMAC Stiffness scale may be preferable in some studies where only a brief measure of the key OA symptom of stiffness is needed.

Should the KOOS Quality of Life (QOL) scale be used in all TKR studies?

While the KOOS QOL scale was not particularly strong in discriminating between groups in cross-sectional tests, it stood out in the longitudinal validity tests as the knee-specific measure with the highest relative validity in discriminating among groups that self-evaluated their functioning or well-being as better, same or worse over time. It also had the largest effect size 6 months after TKR. The KOOS QOL scale also had good internal consistency reliability, little missing data, and met all tests of scaling assumptions. This scale has a broader conceptualization of the impact of knee problems than other KOOS and WOMAC measures, containing items about the cognitive, emotional, functional, and overall general QOL impact of a knee problem. Even if the entire KOOS is not used in favor of a shorter questionnaire such as the WOMAC, it is highly recommended that the 4-item KOOS QOL scale be included in all TKR studies.

Should both joint-specific and generic health measures be used in TKR?

Many TKR studies administer both a knee-specific questionnaire (e.g., WOMAC, KOOS) and a generic questionnaire (e.g., SF-36), to include measures that are more specific to the impact of knee problems as well as measures that allow for outcomes to be compared across different conditions[140]. However, this increases respondent burden and raises the issue as to whether both types of questionnaires are needed in studies of knee OA and TKR. As in previous studies[27, 155], this study found that knee-specific measures of function and pain had larger effect sizes and standardized response means than generic measures six months after TKR. However, KOOS, WOMAC and SF-36 function and pain scales had similar validity in discriminating among groups who rated their health as better, same or worse six months after TKR, with the best SF-36 measure (PCS) and best knee-specific scale (KOOS QOL) having equivalent relative validity.

It is notable that patients who rated their physical and role functioning outcomes as “worse” 6 months after TKR improved on average (by 0.4-0.8 SD) on all KOOS scales. In contrast, mean scores for the “worse” group generally declined or remained stable on the generic SF-36 measures, with the exception of the SF-36 Bodily Pain scale. The difference between generic and knee-specific results for the “worse” group may reflect the impact of comorbid (non-knee) conditions on health (leading to lower generic change scores) even if there was an actual knee-specific improvement, or it may reflect other factors. This discrepancy is a subject for future research. From a measurement perspective, however, these results suggest that knee-specific measures are not sufficient by themselves to fully understand patient outcomes after TKR and should be supplemented with generic measures such as the SF-36 or shorter SF-12® Health Survey[157]. PROMIS provides another set of generic measures that might be considered for use in TKR, although experience with PROMIS in osteoarthritis patients is limited[158]. In any event, administering fewer knee-specific function items appears to be a more effective way to lower respondent burden than eliminating generic measures from TKR studies altogether. In addition, these results suggest that better measurement of comorbidities and their impact is needed in evaluating TKR outcomes.

Study Limitations

This study had a number of limitations. First, TKR patients were from high volume orthopedic centers in the US only; the types of analyses reported in this study should be replicated for TKR patients from other countries, particularly countries in which English is not the primary language. Second, the IRT analyses in Chapter III used the graded response model, which is one type of IRT model used for classical categorical rating items such as those in the KOOS and WOMAC. Analyses might be replicated using other IRT models such as the generalized partial credit model, although it is unlikely that conclusions from the IRT modeling and CAT simulations would change based on use of a different IRT model[65]. Third, criteria used to establish the known groups in Chapter IV were based on patient self-report; additional analyses using clinician reports to define OA severity groups should be conducted. Fourth, only 6-month post-TKR data was available, and responsiveness of all knee-specific and generic measures should be re-examined one year or more after surgery. Finally, the patient population was limited to those with severe OA who were eligible for TKR. Replication of tests of scaling assumptions, reliability, known groups validity and responsiveness in patients with milder knee OA and other knee disorders is encouraged. Results of this study may not apply to these other patient populations.

Conclusion

In summary, this study found that the KOOS and WOMAC scales are reliable, valid and responsive among TKR patients in the US. Reliable knee-specific function scales that are shorter than the existing KOOS and WOMAC function scale can be developed, either as a CAT or a fixed-length short form, but new items that measure higher levels of function are required. TKR outcomes should routinely be evaluated with a knee-specific quality of life scale such as the KOOS QOL, as well as knee-specific measures of pain and function. A generic health measure such as the SF-36 also is needed to fully understand patient outcomes after TKR and to compare these outcomes with outcomes in other therapeutic areas.

Table 5.1: Summary of psychometric tests across measures

Scale	k	α*	% Disc. Validity†	% Floor‡		% Ceiling‡		SF-36 Correlation§	
				Pre-TKR	Post-TKR	Pre-TKR	Post-TKR	PF	BP
KOOS									
Symptoms	7	0.74	75	0.7	0.0	0.3	3.6	0.38	0.46
Pain	9	0.88	89	1.3	0.0	0.5	13.7	0.49	0.66
Function in ADL	17	0.95	99	0.7	0.0	0.3	9.4	0.57	0.64
Function in Sport	5	0.89	100	28.8	4.1	0.9	3.9	0.44	0.42
Quality of Life	4	0.81	100	14.5	0.6	0.1	8.2	0.50	0.53
KOOS-PS (-)	7	0.86	-	1.3	0.0	0.2	4.1	-0.51	-0.57
WOMAC									
Stiffness	2	0.78	100	6.2	0.7	2.2	15.6	0.37	0.50
Pain	5	0.84	70	1.3	0.0	0.9	21.7	0.50	0.63
Function	17	0.95	91	0.7	0.0	0.3	9.4	0.57	0.64
Short WOMAC Function									
Liebs	7	0.89	86	1.0	0.0	0.5	15.6	0.55	0.62
Tubach	8	0.90	94	0.8	0.0	0.6	14.1	0.56	0.64
Whitehouse	7	0.89	93	0.7	0.0	0.5	15.6	0.53	0.62
IRT Function									
IRT Theta	22	0.98	-	0.6	0.0	0.1	2.1	0.58	0.64
CAT R_{TT}≥0.95	3-10	0.95	-	0.7	0.0	0.1	2.2	0.56	0.63
CAT R_{TT}≥0.90	3-10	0.92	-	0.7	0.0	0.1	2.2	0.56	0.61
SF-36									
Physical Functioning (PF)	10	0.87	-	1.7	0.7	0.4	2.1	-	0.53
Bodily Pain (BP)	2	0.77	-	4.7	0.9	0.5	8.0	0.53	-
PCS	35	0.92	-	0.1	0.0	0.0	0.1	0.82	0.69
MCS	35	0.92	-	0.0	0.1	0.1	0.0	0.25	0.38

All measures scored so a lower score=poorer health, except for KOOS-PS where lower score=better health.

Abbreviations: k=number of items; CAT RTT≥0.95=CAT stopped at SE≤0.23 or maximum of 10 items; CAT RTT≥0.90=CAT stopped at SE≤0.32 or maximum of 10 items; PCS=Physical Component Summary; MCS=Mental Component Summary.

* Reliability is measured with Cronbach's coefficient alpha for all scales, except for IRT theta and CAT scores where reliability is the mean reliability across all patient scores.

† Percent of multitrait scaling tests in which item-hypothesized scale correlation is significantly ($p<0.05$) higher than item-other scale correlation.

‡ % Floor=% with worst possible score; % Ceiling=% with best possible score. Sample for all scales is n=1,143 pre-TKR and n=820 6 months post-TKR, except for n=1,102 pre-TKR and n=789 6 months post-TKR for KOOS-PS.

§ Correlation of scale with SF-36 PF and BP scales, pre-TKR. Sample is n=1,143 except for n=1,102 for KOOS-PS.

Table 5.1: Summary of psychometric tests across measures (continued)

		RV Cross-Sectional#		RV Longitudinal#			
Scale	**k**	**Knee Pain**	**Device**	**Physical**	**Role**	**ES**	**SRM**
KOOS							
Symptoms	7	0.43 (0.34-0.56)	0.11 (0.04-0.22)	0.45 (0.28-.63)	0.43 (0.28-.62)	1.25	1.10
Pain	9	**1.00** -	0.38 (0.24-0.59)	0.58 (0.40-.79)	0.56 (0.39-.75)	1.80	1.51
Function in ADL	17	0.59 (0.48-0.70)	0.62 (0.44-0.90)	0.51 (0.35-.68)	0.47 (0.33-.64)	1.53	1.49
Function in Sport	5	0.27 (0.18-0.37)	0.28 (0.15-0.46)	0.57 (0.38-.78)	0.49 (0.33-.68)	1.49	1.07
Quality of Life	4	0.50 (0.38-0.65)	0.21 (0.11-0.34)	**1.00** -	**1.00** -	1.99	1.46
KOOS-PS (-)	7	0.49 (0.40-0.64)	0.53 (0.35-0.80)	0.37 (0.24-.57)	0.35 (0.24-.53)	1.36	1.25
WOMAC							
Stiffness	2	0.52 (0.40-0.67)	0.12 (0.05-0.24)	0.29 (0.16-.44)	0.29 (0.18-.44)	1.20	0.98
Pain	5	**0.91** (0.84-1.00)	0.41 (0.25-0.61)	0.43 (0.28-.61)	0.38 (0.25-.53)	1.63	1.47
Function	17	0.59 (0.48-0.70)	0.62 (0.44-0.90)	0.51 (0.35-.68)	0.47 (0.33-.64)	1.53	1.49
Short WOMAC Function							
Liebs	7	0.58 (0.46-0.68)	0.53 (0.35-0.76)	0.56 (0.39-.76)	0.52 (0.35-.70)	1.69	1.59
Tubach	8	0.61 (0.49-0.73)	0.62 (0.43-0.88)	0.53 (0.37-.74)	0.51 (0.35-.69)	1.58	1.48
Whitehouse	7	0.59 (0.48-0.70)	0.57 (0.40-0.86)	0.43 (0.29-.62)	0.43 (0.28-.59)	1.47	1.41
IRT Function							
IRT Theta	22	0.61 (0.50-0.72)	0.67 (0.49-0.97)	0.73 (0.57-.98)	0.68 (0.50-.87)	1.72	1.49
CAT $R_{TT}\geq0.95$	3-10	0.60 (0.48-0.71)	0.63 (0.45-0.90)	0.69 (0.51-.92)	0.63 (0.46-.82)	1.79	1.50
CAT $R_{TT}\geq0.90$	3-10	0.52 (0.41-0.64)	0.61 (0.42-0.86)	0.70 (0.50-.93)	0.65 (0.47-.84)	1.76	1.45
SF-36							
Physical Functioning (PF)	10	0.16 (0.09-0.23)	**1.00** -	0.71 (0.50-.99)	0.62 (0.44-.89)	1.04	0.96
Bodily Pain (BP)	2	0.76 (0.59-0.93)	0.48 (0.32-0.72)	0.57 (0.40-.85)	0.57 (0.39-.87)	1.31	1.06
PCS	35	0.30 (0.21-0.40)	**0.97** (0.78-1.20)	**0.89** (.66-1.22)	**0.78** (.56-1.10)	1.08	1.00
MCS	35	0.08 (0.04-0.13)	0.25 (0.12-0.44)	0.04 (0.00-.09)	0.04 (0.01-.08)	0.15	0.17

All measures scored so a lower score=poorer health, except for KOOS-PS where lower score=better health.

Abbreviations: k=number of items; RV=relative validity; ES=effect size; SRM=standardized response mean; CAT RTT≥0.95=CAT stopped at SE≤0.23 or maximum of 10 items; CAT RTT≥0.90=CAT stopped at SE≤0.32 or maximum of 10 items; PCS=Physical Component Summary; MCS=Mental Component Summary.

# Relative validity in relation to scale with highest F-statistic in known groups validity test (frequency of knee pain, use of assistive walking device, patient better/same/worse rating of 6 month change in physical activities and daily work role). **Bold**=most valid scale or scale RV not significantly different from 1.00. See Chapter IV for sample sizes.

Table 5.2: Percentages of item administrations (most to least) in simulated function CATs relative to content of short function scales

		CAT Utilization (%)*		Items in Short Function Scales†			
Label	**Scale/Item Content**	**Pre-TKR**	**Post-TKR**	**Liebs**	**Tubach**	**White-house**	**KOOS-PS**
Activities of Daily Living							
A8	Going shopping	14.39	11.91		●		
A10	Rising from bed	14.28	9.14			●	●
A17	Light domestic duties	14.25	7.80				
A4	Standing	13.65	8.29	●			
A6	Walking on flat surface	13.15	5.08	●	●	●	
A15	Getting on/off toilet	11.82	4.01		●		
A7	Getting in/out of car	9.47	10.02	●	●	●	
A3	Rising from sitting	4.53	9.00	●	●	●	●
A2	Ascending stairs	1.69	6.76	●	●	●	
A16	Heavy domestic duties	0.52	6.27				
A13	Getting in/out of bath	0.50	0.19				
A1	Descending stairs	0.34	5.67	●	●		
A14	Sitting	0.28	0.03			●	
A11	Taking off socks/stockings	0.21	0.04				
A12	Lying in bed	0.18	0.05				
A9	Putting on socks/stockings	0.11	0.14		●	●	●
A5	Bending to floor	0.10	2.51	●			●
Sport/Recreation							
Sp4	Twisting/pivoting	0.18	4.51				●
Sp1	Squatting	0.14	3.26				●
Sp3	Jumping	0.10	2.58				
Sp5	Kneeling	0.07	2.05				●
Sp2	Running	0.02	0.68				
Percent of total CAT item administrations‡							
Pre-TKR				42.9	55.5	43.5	19.4
Post-TKR				47.3	52.6	40.2	30.6

* Percent of time an item was selected in a simulated CAT with a stopping rule of SE≤0.23 (reliability≥0.95) or maximum number of 10 items. Each column sums to a total of 100% of person-item administrations.

† The ● indicates that the item is included in the fixed-length short function scale.

‡ Percent of all person-item administrations accounted for by items included in the short function scale.

.

BIBLIOGRAPHY

1. Lawrence RC, Felson DT, Helmick CG, Arnold LM, Choi H, Deyo RA*, et al*. Estimates of the prevalence of arthritis and other rheumatic conditions in the United States. Part II. Arthritis Rheum. 2008;58:26-35.

2. Hootman JM and Helmick CG. Projections of US prevalence of arthritis and associated activity limitations. Arthritis Rheum. 2006;54:226-9.

3. Zhang W, Moskowitz RW, Nuki G, Abramson S, Altman RD, Arden N*, et al*. OARSI recommendations for the management of hip and knee osteoarthritis, Part II: OARSI evidence-based, expert consensus guidelines. Osteoarthritis Cartilage 2008;16:137-62.

4. Zhang W, Moskowitz RW, Nuki G, Abramson S, Altman RD, Arden N*, et al*. OARSI recommendations for the management of hip and knee osteoarthritis, Part I: Critical appraisal of existing treatment guidelines and systematic review of current research evidence. Osteoarthritis Cartilage 2007;15:981-1000.

5. Zhang W, Nuki G, Moskowitz RW, Abramson S, Altman RD, Arden NK*, et al*. OARSI recommendations for the management of hip and knee osteoarthritis: Part III: Changes in evidence following systematic cumulative update of research published through January 2009. Osteoarthritis Cartilage 2010;18:476-99.

6. Gossec L, Paternotte S, Bingham CO,3rd, Clegg DO, Coste P, Conaghan PG*, et al*. OARSI/OMERACT initiative to define states of severity and indication for joint replacement in hip and knee osteoarthritis. J.Rheumatol. 2011;38:1765-9.

7. Losina E, Thornhill TS, Rome BN, Wright J, Katz JN. The dramatic increase in total knee replacement utilization rates in the United States cannot be fully explained by growth in population size and the obesity epidemic. J.Bone Joint Surg.Am. 2012;94:201-7.

8. Kurtz SM, Ong KL, Lau E, Widmer M, Maravic M, Gomez-Barrena E*, et al*. International survey of primary and revision total knee replacement. Int.Orthop. 2011;35:1783-9.

9. Kurtz SM, Lau E, Ong K, Zhao K, Kelly M, Bozic KJ. Future young patient demand for primary and revision joint replacement: national projections from 2010 to 2030. Clin.Orthop.Relat.Res. 2009;467:2606-12.

10. Agency for Healthcare Research and Quality. Healthcare Cost and Utilization Project. http://hcupnet.ahrq.gov/HCUPnet.jsp. Downloaded June 12, 2014.

11. Kurtz S, Ong K, Lau E, Mowat F, Halpern M. Projections of primary and revision hip and knee arthroplasty in the United States from 2005 to 2030. J.Bone Joint Surg.Am. 2007;89:780-5.

12. Bellamy N, Kirwan J, Boers M, Brooks P, Strand V, Tugwell P*, et al*. Recommendations for a core set of outcome measures for future phase III clinical trials in knee, hip, and hand osteoarthritis. Consensus development at OMERACT III. J.Rheumatol. 1997;24:799-802.

13. Dougados M, Leclaire P, van der Heijde D, Bloch DA, Bellamy N, Altman RD. Response criteria for clinical trials on osteoarthritis of the knee and hip: a report of the Osteoarthritis Research Society International Standing Committee for Clinical Trials response criteria initiative. Osteoarthritis Cartilage 2000;8:395-403.

14. Goldberg VM, Buckwalter J, Halpin M, Jiranek W, Mihalko W, Pinzur M*, et al*. Recommendations of the OARSI FDA Osteoarthritis Devices Working Group. Osteoarthritis Cartilage 2011;19:509-14.

15. Ethgen O, Bruyere O, Richy F, Dardennes C, Reginster JY. Health-related quality of life in total hip and total knee arthroplasty. A qualitative and systematic review of the literature. J.Bone Joint Surg.Am. 2004;86-A:963-74.

16. Santaguida PL, Hawker GA, Hudak PL, Glazier R, Mahomed NN, Kreder HJ*, et al*. Patient characteristics affecting the prognosis of total hip and knee joint arthroplasty: a systematic review. Can.J.Surg. 2008;51:428-36.

17. Vissers MM, Bussmann JB, Verhaar JAN, Busschbach JJV, Bierma-Zeinstra S, Reijman M. Psychological factors affecting the outcome of total hip and knee arthroplasty: A systematic review. Semin.Arthritis Rheum. 2012;41:576-88.

18. Baker PN, Deehan DJ, Lees D, Jameson S, Avery PJ, Gregg PJ*, et al*. The effect of surgical factors on early patient-reported outcome measures (PROMS) following total knee replacement. J.Bone Joint Surg.Br. 2012;94:1058-66.

19. Katz JN, Mahomed NN, Baron JA, Barrett JA, Fossel AH, Creel AH*, et al*. Association of hospital and surgeon procedure volume with patient-centered outcomes of total knee replacement in a population-based cohort of patients age 65 years and older. Arthritis Rheum. 2007;56:568-74.

20. Minns Lowe CJ, Barker KL, Dewey M, Sackley CM. Effectiveness of physiotherapy exercise after knee arthroplasty for osteoarthritis: systematic review and meta-analysis of randomised controlled trials. BMJ 2007;335:812-5.

21. Collins NJ, Misra D, Felson DT, Crossley KM, Roos EM. Measures of knee function. Arthritis Care Res. 2011;63 Suppl 11:S208-28.

22. Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt LW. 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. J.Rheumatol. 1988;15:1833-40.

23. Bellamy N. WOMAC: A 20-year experiential review of a patient-centered self-reported health status questionnaire. J Rheumatol 2002;29:2473-6.

24. Liebs TR, Herzberg W, Gluth J, Rüther W, Haasters J, Russlies M*, et al*. Using the patient's perspective to develop function short forms specific to total hip and knee replacement based on WOMAC function items. Bone Joint J. 2013;95-B:239-43.

25. Tubach F, Baron G, Falissard B, Logeart I, Dougados M, Bellamy N*, et al*. Using patients' and rheumatologists' opinions to specify a short form of the WOMAC function subscale. Ann.Rheum.Dis. 2005;64:75-9.

26. Whitehouse SL, Lingard EA, Katz JN, Learmonth ID. Development and testing of a reduced WOMAC function scale. J.Bone Joint Surg.Br. 2003;85:706-11.

27. Roos EM and Toksvig-Larsen S. Knee injury and Osteoarthritis Outcome Score (KOOS) - validation and comparison to the WOMAC in total knee replacement. Health.Qual.Life.Outcomes 2003;1:17.

28. Perruccio AV, Stefan Lohmander L, Canizares M, Tennant A, Hawker GA, Conaghan PG*, et al*. The development of a short measure of physical function for knee OA KOOS-Physical Function Shortform (KOOS-PS) - an OARSI/OMERACT initiative. Osteoarthritis Cartilage 2008;16:542-50.

29. Alviar MJ, Olver J, Brand C, Tropea J, Hale T, Pirpiris M*, et al*. Do patient-reported outcome measures in hip and knee arthroplasty rehabilitation have robust measurement attributes? A systematic review. J Rehabil. Med. 2011;43:572-83.

30. Ware JE,Jr and Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med.Care 1992;30:473-83.

31. Ware JE,Jr. SF-36 Health Survey update. Spine (Phila Pa.1976) 2000;25:3130-9.

32. Dawson J, Fitzpatrick R, Carr A, Murray D. Questionnaire on the perceptions of patients about total hip replacement. J.Bone Joint Surg.Br. 1996;78:185-90.

33. Dawson J, Fitzpatrick R, Murray D, Carr A. Questionnaire on the perceptions of patients about total knee replacement. J.Bone Joint Surg.Br. 1998;80:63-9.

34. EuroQol Group. EuroQol--a new facility for the measurement of health-related quality of life. Health Policy 1990;16:199-208.

35. Sedrakyan A, Paxton EW, Phillips C, Namba R, Funahashi T, Barber T*, et al*. The International Consortium of Orthopaedic Registries: overview and summary. J.Bone Joint Surg.Am. 2011;93 Suppl 3:1-12.

36. Rolfson O, Rothwell A, Sedrakyan A, Chenok KE, Bohm E, Bozic KJ*, et al*. Use of patient-reported outcomes in the context of different levels of data. J.Bone Joint Surg.Am. 2011;93 Suppl 3:66-71.

37. Franklin PD, Harrold L, Ayers DC. Incorporating patient-reported outcomes in total joint arthroplasty registries: challenges and opportunities. Clin.Orthop.Relat.Res. 2013;471:3482-8.

38. Riddle DL, Stratford PW, Bowman DH. Findings of extensive variation in the types of outcome measures used in hip and knee replacement clinical trials: a systematic review. Arthritis Rheum. 2008;59:876-83.

39. Wylde V, Bruce J, Beswick A, Elvers K, Gooberman-Hill R. The assessment of chronic post-surgical pain after knee replacement: A systematic review. Arthritis Care.Res.(Hoboken) 2013;65:1795-803.

40. Yale New Haven Health Services Corporation/Center for Outcomes Research & Evaluation. Summary of Technical Expert Panel (TEP) evaluation of measures: Total hip and/or total knee arthroplasty (THA/TKA) patient-reported outcome hospital and eligible-professional level performance measures. March 7, 2014.

41. Franklin PD, Allison JJ, Ayers DC. Beyond joint implant registries: a patient-centered research consortium for comparative effectiveness in total joint replacement. JAMA 2012;308:1217-8.

42. Ayers DC, Zheng H, Franklin PD. Integrating patient-reported outcomes into orthopaedic clinical practice: proof of concept from FORCE-TJR. Clin.Orthop.Relat.Res. 2013;471:3419-25.

43. Scientific Advisory Committee of the Medical Outcomes Trust. Assessing health status and quality-of-life instruments: attributes and review criteria. Qual.Life Res. 2002;11:193-205.

44. U.S. Department of Health and Human Services, Food and Drug Administration. Guidance for industry - Patient-reported outcome measures: Use in medical product development to support labeling claims. Rockville, MD: Food and Drug Administration 2009.

45. Roos EM, Roos HP, Lohmander LS, Ekdahl C, Beynnon BD. Knee Injury and Osteoarthritis Outcome Score (KOOS)--development of a self-administered outcome measure. J.Orthop.Sports Phys.Ther. 1998;28:88-96.

46. de Groot IB, Favejee MM, Reijman M, Verhaar JA, Terwee CB. The Dutch version of the Knee Injury and Osteoarthritis Outcome Score: a validation study. Health.Qual.Life.Outcomes 2008;6:16.

47. Goncalves RS, Cabri J, Pinheiro JP, Ferreira PL. Cross-cultural adaptation and validation of the Portuguese version of the Knee injury and Osteoarthritis Outcome Score (KOOS). Osteoarthritis Cartilage 2009;17:1156-62.

48. Monticone M, Ferrante S, Salvaderi S, Motta L, Cerri C. Responsiveness and minimal important changes for the Knee Injury and Osteoarthritis Outcome Score in subjects undergoing rehabilitation after total knee arthroplasty. Am.J.Phys.Med.Rehabil. 2013;92:864-70.

49. Ornetti P, Parratte S, Gossec L, Tavernier C, Argenson JN, Roos EM*, et al*. Cross-cultural adaptation and validation of the French version of the Knee injury and Osteoarthritis Outcome Score (KOOS) in knee osteoarthritis patients. Osteoarthritis Cartilage 2008;16:423-8.

50. Roos EM, Roos HP, Lohmander LS. WOMAC Osteoarthritis Index--additional dimensions for use in subjects with post-traumatic osteoarthritis of the knee. Osteoarthritis Cartilage 1999;7:216-21.

51. Nakamura N, Takeuchi R, Sawaguchi T, Ishikawa H, Saito T, Goldhahn S. Cross-cultural adaptation and validation of the Japanese Knee Injury and Osteoarthritis Outcome Score (KOOS). J.Orthop.Sci. 2011;16:516-23.

52. Xie F, Li SC, Roos EM, Fong KY, Lo NN, Yeo SJ*, et al*. Cross-cultural adaptation and validation of Singapore English and Chinese versions of the Knee injury and Osteoarthritis Outcome Score (KOOS) in Asians with knee osteoarthritis in Singapore. Osteoarthritis Cartilage 2006;14:1098-103.

53. Bombardier C, Melfi CA, Paul J, Green R, Hawker G, Wright J*, et al*. Comparison of a generic and a disease-specific measure of pain and physical function after knee replacement surgery. Med.Care 1995;33:AS131-44.

54. Ghanem E, Pawasarat I, Lindsay A, May L, Azzam K, Joshi A*, et al*. Limitations of the Knee Society Score in evaluating outcomes following revision total knee arthroplasty. J.Bone Joint Surg.Am. 2010;92:2445-51.

55. Johanson NA, Liang MH, Daltroy L, Rudicel S, Richmond J. American Academy of Orthopaedic Surgeons lower limb outcomes assessment instruments. Reliability, validity, and sensitivity to change. J.Bone Joint Surg.Am. 2004;86-A:902-9.

56. Lingard EA, Katz JN, Wright RJ, Wright EA, Sledge CB, Kinemax Outcomes Group. Validity and responsiveness of the Knee Society Clinical Rating System in comparison with the SF-36 and WOMAC. J.Bone Joint Surg.Am. 2001;83-A:1856-64.

57. Likert R. A technique for the measurement of attitudes. Archives of Psychology 1932;140:5-55.

58. McHorney CA, Ware JE,Jr, Lu JF, Sherbourne CD. The MOS 36-item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. Med.Care 1994;32:40-66.

59. Salavati M, Akhbari B, Mohammadi F, Mazaheri M, Khorrami M. Knee injury and Osteoarthritis Outcome Score (KOOS); reliability and validity in competitive athletes after anterior cruciate ligament reconstruction. Osteoarthritis Cartilage 2011;19:406-10.

60. Thumboo J, Chew LH, Soh CH. Validation of the Western Ontario and Mcmaster University osteoarthritis index in Asians with osteoarthritis in Singapore. Osteoarthritis Cartilage 2001;9:440-6.

61. Ryser L, Wright BD, Aeschlimann A, Mariacher-Gehler S, Stucki G. A new look at the Western Ontario and McMaster Universities Osteoarthritis Index using Rasch analysis. Arthritis Care Res. 1999;12:331-5.

62. Wolfe F and Kong SX. Rasch analysis of the Western Ontario MacMaster questionnaire (WOMAC) in 2205 patients with osteoarthritis, rheumatoid arthritis, and fibromyalgia. Ann.Rheum.Dis. 1999;58:563-8.

63. Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B*, et al*. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. Med.Care 2007;45:S3-S11.

64. Petersen MA, Groenvold M, Aaronson NK, Chie WC, Conroy T, Costantini A*, et al*. Development of computerized adaptive testing (CAT) for the EORTC QLQ-C30 physical functioning dimension. Qual.Life Res. 2011;20:479-90.

65. Embretson SE and Reise SP. Item Response Theory for Psychologists. Mahwah,NJ: Lawrence Erlbaum Associates 2000.

66. Wainer H, Dorans NJ, Eigor D, Flaugher R, Green BF, Mislevy RJ*, et al*. Computerized Adaptive Testing: A primer. Hillsdale, NJ: Lawrence Erlbaum Associates 2000.

67. Franchignoni F, Salaffi F, Giordano A, Carotti M, Ciapetti A, Ottonello M. Rasch analysis of the 22 Knee Injury and Osteoarthritis Outcome Score-Physical Function items in Italian patients with knee osteoarthritis. Arch.Phys.Med.Rehabil. 2012;.

68. Davis AM, Badley EM, Beaton DE, Kopec J, Wright JG, Young NL*, et al*. Rasch analysis of the Western Ontario McMaster (WOMAC) Osteoarthritis Index: results from community and arthroplasty samples. J.Clin.Epidemiol. 2003;56:1076-83.

69. Roorda LD, Jones CA, Waltz M, Lankhorst GJ, Bouter LM, van der Eijken JW*, et al*. Satisfactory cross cultural equivalence of the Dutch WOMAC in patients with hip osteoarthritis waiting for arthroplasty. Ann.Rheum.Dis. 2004;63:36-42.

70. McHorney CA and Monahan PO. Postscript: Applications of Rasch analysis in health care. Med.Care 2004;42:I73-8.

71. Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA*, et al*. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Med.Care 2007;45:S22-31.

72. McHorney CA, Ware JE,Jr, Raczek AE. The MOS 36-item Short-form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. Med.Care 1993;31:247-63.

73. Ware JE,Jr, Kosinski M, Bayliss MS, McHorney CA, Rogers WH, Raczek A. Comparison of methods for the scoring and statistical analysis of SF-36 health profile and summary measures: Summary of results from the Medical Outcomes Study. Med.Care 1995;33:AS264-79.

74. Roos EM, Roos HP, Ekdahl C, Lohmander LS. Knee injury and Osteoarthritis Outcome Score (KOOS)--validation of a Swedish version. Scand.J.Med.Sci.Sports 1998;8:439-48.

75. Bekkers JE, de Windt TS, Raijmakers NJ, Dhert WJ, Saris DB. Validation of the Knee Injury and Osteoarthritis Outcome Score (KOOS) for the treatment of focal cartilage lesions. Osteoarthritis Cartilage 2009;17:1434-9.

76. Comins J, Brodersen J, Krogsgaard M, Beyer N. Rasch analysis of the Knee injury and Osteoarthritis Outcome Score (KOOS): a statistical re-evaluation. Scand.J.Med.Sci.Sports 2008;18:336-45.

77. Monticone M, Ferrante S, Salvaderi S, Rocca B, Totti V, Foti C*, et al*. Development of the Italian version of the knee injury and osteoarthritis outcome score for patients with knee injuries: cross-cultural adaptation, dimensionality, reliability, and validity. Osteoarthritis Cartilage 2012;20:330-5.

78. Paradowski PA, Wito Ski D, K Ska R, Roos EM. Cross-cultural translation and measurement properties of the Polish version of the Knee injury and Osteoarthritis Outcome Score (KOOS) following anterior cruciate ligament reconstruction. Health.Qual.Life.Outcomes 2013;11:107.

79. van Meer BL, Meuffels DE, Vissers MM, Bierma-Zeinstra SM, Verhaar JA, Terwee CB*, et al*. Knee Injury and Osteoarthritis Outcome Score or International Knee Documentation Committee Subjective Knee Form: which questionnaire is most useful to monitor patients with an anterior cruciate ligament rupture in the short term? Arthroscopy 2013;29:701-15.

80. Vaquero J, Longo UG, Forriol F, Martinelli N, Vethencourt R, Denaro V. Reliability, validity and responsiveness of the Spanish version of the Knee Injury and Osteoarthritis Outcome Score (KOOS) in patients with chondral lesion of the knee. Knee Surg.Sports Traumatol.Arthrosc. 2014;22:104-8.

81. Salavati M, Mazaheri M, Negahban H, Sohani SM, Ebrahimian MR, Ebrahimi I*, et al*. Validation of a Persian-version of Knee injury and Osteoarthritis Outcome Score (KOOS) in Iranians with knee injuries. Osteoarthritis Cartilage 2008;16:1178-82.

82. Almangoush A, Herrington L, Attia I, Jones R, Aldawoudy A, Abdul Aziz A*, et al*. Cross-cultural adaptation, reliability, internal consistency and validation of the Arabic version of the Knee injury and Osteoarthritis Outcome Score (KOOS) for Egyptian people with knee injuries. Osteoarthritis Cartilage 2013;21:1855-64.

83. Engelhart L, Nelson L, Lewis S, Mordin M, Demuro-Mercon C, Uddin S*, et al*. Validation of the Knee Injury and Osteoarthritis Outcome Score subscales for patients with articular cartilage lesions of the knee. Am.J.Sports Med. 2012;40:2264-72.

84. Katz JN, Chang LC, Sangha O, Fossel AH, Bates DW. Can comorbidity be measured by questionnaire rather than medical record review? Med.Care 1996;34:73-84.

85. KOOS Scoring 2012. www.koos.nu. Downloaded March 24, 2013.

86. Bellamy N. WOMAC Osteoarthritis Index User Guide VIII. Queensland, Australia: University of Queensland 2007.

87. Ware JE,Jr and Gandek B. Methods for testing data quality, scaling assumptions, and reliability: the IQOLA Project approach. J.Clin.Epidemiol. 1998;51:945-52.

88. Campbell DT and Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychol.Bull. 1959;56:81-105.

89. Ware JE,Jr. Scales for measuring general health perceptions. Health Serv.Res. 1976;11:396-415.

90. Howard KI and Forehand GG. A method for correcting item-total correlations for the effect of relevant item inclusion. Educ Psychol Meas 1962;22:731-5.

91. Ware JE, Brook RH, Davies-Avery A, Williams K, Stewart AL, Rogers WH*, et al*. Model of Health and Methodology. Santa Monica, CA: RAND Corporation 1980.

92. Levy KJ. Some multiple range tests for variances. Educ Psychol Meas 1975;35:599-604.

93. Cronbach LJ. Coefficient alpha and the internal structure of tests. 1951;16:297-334.

94. Nunnally JC and Bernstein IH. Psychometric Theory. New York: Mc-Graw Hill 1994.

95. Tyler TA and Fiske DW. Homogeneity indices and text length. Educ Psychol Meas 1968;28:767-77.

96. Kennedy D, Stratford PW, Pagura SMC, Wessel J, Gollish JD, Woodhouse LJ. Exploring the factorial validity and clinical interpretability of the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC). Physiother.Can. 2003;55:160-8.

97. Stratford PW and Kennedy DM. Does parallel item content on WOMAC's pain and function subscales limit its ability to detect change in functional status? BMC Musculoskelet.Disord. 2004;5:17.

98. Pua YH, Cowan SM, Wrigley TV, Bennell KL. Discriminant validity of the Western Ontario and McMaster Universities Osteoarthritis index physical functioning subscale in community samples with hip osteoarthritis. Arch.Phys.Med.Rehabil. 2009;90:1772-7.

99. Hawker GA, Davis AM, French MR, Cibere J, Jordan JM, March L*, et al*. Development and preliminary psychometric testing of a new OA pain measure - an OARSI/OMERACT initiative. Osteoarthr.Cartil. 2008;16:409-14.

100. Gooberman-Hill R, Woolhead G, Mackichan F, Ayis S, Williams S, Dieppe P. Assessing chronic joint pain: lessons from a focus group study. Arthritis Rheum. 2007;57:666-71.

101. Gudbergsen H, Bartels EM, Krusager P, Waehrens EE, Christensen R, Danneskiold-Samsoe B*, et al*. Test-retest of computerized health status questionnaires frequently used in the monitoring of knee osteoarthritis: a randomized crossover trial. BMC Musculoskelet.Disord. 2011;12:190.

102. Bellamy N, Wilson C, Hendrikz J, Whitehouse SL, Patel B, Dennison S*, et al*. Osteoarthritis Index delivered by mobile phone (m-WOMAC) is valid, reliable, and responsive. J.Clin.Epidemiol. 2011;64:182-90.

103. Bjorner JB, Rose M, Gandek B, Stone AA, Junghaenel DU, Ware JE,Jr. Differences in method of administration of patient reported outcomes measures did not significantly impact score level, reliability or validity. J.Clin.Epidemiol. 2014;67:108-13.

104. Roos EM and Lohmander LS. The Knee injury and Osteoarthritis Outcome Score (KOOS): from joint injury to osteoarthritis. Health.Qual.Life.Outcomes 2003;1:64.

105. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S*, et al*. The Patient-reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. J.Clin.Epidemiol. 2010;63:1179-94.

106. Bjorner JB, Kosinski M, Ware JE,Jr. Calibration of an item pool for assessing the burden of headaches: An application of item response theory to the Headache Impact Test (HIT). Qual.Life Res. 2003;12:913-33.

107. Rose M, Bjorner JB, Becker J, Fries JF, Ware JE. Evaluation of a preliminary physical function item bank supported the expected advantages of the Patient-Reported Outcomes Measurement Information System (PROMIS). J.Clin.Epidemiol. 2008;61:17-33.

108. Muthén LK and Muthén BO. Mplus User’s Guide. Los Angeles, CA: Muthén & Muthén 1998-2010.

109. Bollen KA and Barb KH. Pearson's R and coarsely categorized measures. Am.Sociol.Rev. 1981;46:232-9.

110. Brown TA. Confirmatory Factor Analysis for Applied Research. New York: Guilford Press 2006.

111. MacDonald RP. Test Theory: A Unified Treatment. Mahwah, NJ: Lawrence Erlbaum 1999.

112. Ramsay J. Kernal smoothing approaches to nonparametric item characteristic curve estimation. Psychometrika 1991;56:611-30.

113. Ramsay J. TestGraf - A Program for the Graphical Analysis of Multiple Choice Test and Questionnaire Data. Montreal, Canada: McGill University 1995.

114. Swaminathan H and Rogers JH. Detecting differential item functioning using logistic regression procedures. 1990;27:361-70.

115. Zumbo BD. A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores. Ottawa, Ontario: Directorate of Human Resources Research and Evaluation, Department of National Defense 1999.

116. Camilli G and Shepard LA. Methods for Identifying Biased Test Items. London, U.K.: Sage Publications, Inc. 1994.

117. Holland PW and Wainer H. Differential Item Functioning. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc. 1993.

118. Nagelkerke NJ. A note on a general definition of the coefficient of determination. Biometrika 1991;78:691-2.

119. Samejima F. Graded response model. In: Handbook of Modern Item Response Theory. W. van der Linden and R. K. Hambleton, Eds. New York, NY: Springer-Verlag, 1997, pp: 85-100.

120. Bjorner JB, Chang CH, Thissen D, Reeve BB. Developing tailored instruments: item banking and computerized adaptive assessment. Qual.Life Res. 2007;16 Suppl 1:95-108.

121. Thissen D. The MEDPRO Project: An SBIR Project for a Comprehensive IRT and CAT Software System-IRT Software. 2010. http://www.psych.umn.edu/psylabs/catcentral/pdf%20files/cat09thissen.pdf

122. Cai L, Thissen D, du Toit S. IRTPRO 2.1 for Windows. Lincolnwood, IL: Scientific Software International 2011.

123. Orlando M and Thissen D. Likelihood-based item-fit indices for dichotomous item response theory models. Appl Psychol Meas 2000;24:50-64.

124. Orlando M and Thissen D. Further examination of the performance of S-X2,an item fit index for dichotomous item response theory models. Appl Psychol Meas 2003;27:289-98.

125. Muraki E. Information functions of the generalized partial credit model. Appl Psychol Meas 1993;17:351-63.

126. Choi SW. Firestar: Computerized Adaptive Testing (CAT) Simulation Program for Polytomous IRT Models. Appl.Psychol.Meas. 2009;33:644-5.

127. Choi S. FIRESTAR: Computerized Adaptive Testing (CAT) Simulation Program for Polytomous IRT Models Version 1.2.2. Chicago, IL: Northwestern University Feinberg School of Medicine 2009.

128. Ware JE,Jr, Kosinski M, Bjorner JB, Bayliss MS, Batenhorst A, Dahlöf CG*, et al*. Applications of computerized adaptive testing (CAT) to the assessment of headache impact. Qual.Life.Res 2003;12:935-52.

129. Cook KF, Kallen MA, Amtmann D. Having a fit: impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. Qual.Life Res. 2009;18:447-60.

130. Rose M, Bjorner JB, Gandek B, Bruce B, Fries JF, Ware JE,Jr. The PROMIS physical function item bank was calibrated to a standardized metric and shown to improve measurement efficiency. J Clin Epidemiol 2014;67:516-26.

131. Ware JE,Jr, Guyer R, Harrington M, Boulanger R. Standardizing the metric and increasing the efficiency of physical functioning outcomes measurement. Value Health. 2012;15:A476.

132. Gwaltney CJ, Shields AL, Shiffman S. Equivalence of electronic and paper-and-pencil administration of patient-reported outcome measures: a meta-analytic review. Value Health. 2008;11:322-33.

133. Bjorner JB, Rose M, Gandek B, Stone AA, Junghaenel DU, Ware JE,Jr. Difference in method of administration did not significantly impact item response: an IRT-based analysis from the Patient-Reported Outcomes Measurement Information System (PROMIS) initiative. Qual.Life Res. 2014;23:217-27.

134. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL*, et al*. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. BMC Med.Res.Methodol. 2010;10:22.

135. Hays RD and Hadorn D. Responsiveness to change: an aspect of validity, not a separate dimension. Qual.Life Res. 1992;1:73-5.

136. Ware JE,Jr and Keller SD. Interpreting general health measures. In: Quality of life and Pharmacoeconomics in clinical trials, 2nd Ed. B. Spilker, Ed. Philadelphia, PA: Lippincott-Raven Publishers, 1996, pp: 445-60.

137. Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR, Clinical Significance Consensus Meeting Group. Methods to explain the clinical significance of health status measures. Mayo Clin.Proc. 2002;77:371-83.

138. Kerlinger FN. Foundations of behavioral research. New York: Holt, Rinehart, and Winston 1964.

139. Escobar A, Quintana JM, Bilbao A, Azkárate J, Güenaga JI. Validation of the Spanish version of the WOMAC questionnaire for patients with hip or knee osteoarthritis. Clin.Rheumatol. 2002;21:466-71.

140. Kantz ME, Harris WJ, Levitsky K, Ware JE,Jr, Davies AR. Methods for assessing condition-specific and generic functional status outcomes after total knee replacement. Med.Care 1992;30:MS240-52.

141. Ware JE,Jr, Bjorner JB, Kosinski M. Practical implications of item response theory and computerized adaptive testing: a brief summary of ongoing studies of widely used headache impact scales. Med.Care 2000;38:II73-82.

142. Ware JE,Jr, Snow KK, Kosinski M, Gandek B. SF-36 Health Survey manual and interpretation guide. Boston, MA: The Health Institute 1993.

143. Ware JE,Jr, Kosinski M, Dewey JE. How to score Version 2 of the SF-36 Health Survey. Lincoln, RI: QualityMetric Incorporated 2000.

144. Veenhof C, Bijlsma JWJ, van den Ende CHM, Van Dijk GM, Pisters MF, Dekker J. Psychometric evaluation of osteoarthritis questionnaires: A systematic review of the literature. Arthritis Care Res. 2006;55:480-92.

145. Davis AM, Perruccio AV, Canizares M, Hawker GA, Roos EM, Maillefert JF*, et al*. Comparative, validity and responsiveness of the HOOS-PS and KOOS-PS to the WOMAC physical function subscale in total joint replacement for osteoarthritis. Osteoarthritis Cartilage 2009;17:843-7.

146. Hawker GA, Melfi CA, Paul JE, Green R, Bombardier C. Comparison of a generic (SF-36) and a disease specific (WOMAC) instrument in the measurement of outcomes after knee replacement surgery. J.Rheumatol. 1995;22:1193-6.

147. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J.Chronic Dis. 1987;40:373-83.

148. Deng N, Allison JJ, Fang HJ, Ash AS, Ware JE,Jr. Using the bootstrap to establish statistical significance for relative validity comparisons among patient-reported outcome measures. Health.Qual.Life. Outcomes 2013;11:89.

149. Henderson AR. The bootstrap: A technique for data-driven statistics using computer-intensive analyses to explore experimental data. 2005;359:1-26.

150. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Med.Care 1989;27:S178-89.

151. Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. Med.Care 1990;28:632-42.

152. Farivar SS, Liu H, Hays RD. Half standard deviation estimate of the minimally important difference in HRQOL scores? Expert Rev.Pharmacoecon Outcomes Res. 2004;4:515-23.

153. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Med.Care 2003;41:582-92.

154. Ware JE,Jr, Kosinski M, Bjorner JB, Turner-Bowker DM, Gandek B, Maruish ME. User's Manual for the SF-36v2® Health Survey. Lincoln, RI: QualityMetric Incorporated 2007.

155. Escobar A, Quintana JM, Bilbao A, Aróstegui I, Lafuente I, Vidaurreta I. Responsiveness and clinically important differences for the WOMAC and SF-36 after total knee replacement. Osteoarthritis Cartilage 2007;15:273-80.

156. Fisher WP,Jr, Eubanks RL, Marier RL. Equating the MOS SF36 and the LSU HSI Physical Functioning Scales. J.Outcome Meas. 1997;1:329-62.

157. Ware JE,Jr, Kosinski M, Keller SD. A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. Med.Care 1996;34:220-33.

158. Broderick JE, Schneider S, Junghaenel DU, Schwartz JE, Stone AA. Validity and reliability of patient-reported outcomes measurement information system instruments in osteoarthritis. Arthritis Care.Res.(Hoboken) 2013;65:1625-33.