

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI[®]

UNIVERSITY OF ALBERTA

THE VALIDITY OF EUROQOL SCORES OF PATIENTS WITH HIP OR KNEE
REPLACEMENT

BY

Barbara L. Spady



A thesis submitted to the Faculty of Graduate Studies and Research in partial
fulfillment of the requirements for the degree of Doctor of Philosophy

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY

EDMONTON, ALBERTA

Spring, 2000



National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisitions et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-60028-9

Canada

University of Alberta
Library Release Form

Name of Author: Barbara Louise Spady

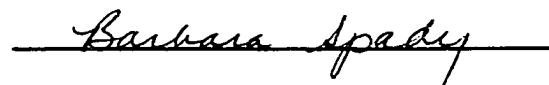
Title of Thesis: The Validity of EuroQol Scores of Patients with Hip or Knee
Replacement

Degree: Doctor of Philosophy

Year this Degree Granted: 2000

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly, or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as hereinbefore provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.




176 Quesnell Crescent
Edmonton, Alberta
T5R 5P3

Date: January 31/2000

University of Alberta

Faculty of Graduate Studies and Research

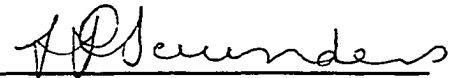
The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled **The Validity of EuroQol Scores of Patients with Hip or Knee Replacement** submitted by **Barbara Louise Spady** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy**.



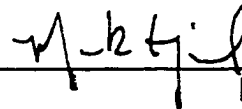
Dr. T. O. Maguire (Supervisor)



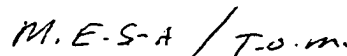
Dr. W. T. Rogers



Dr. L. D. Saunders



Dr. M. Gierl



Dr. M. E. Suarez-Almazor



Dr. S. A. Warren



Dr. A. I. Rothman

Date: January 27/2000

In memory of my Dad, Robert Conner

ABSTRACT

With an increasing demand on health care services and a limited supply of resources, the measurement of health outcomes in the evaluation of health services and technologies has become more common. Health related quality of life (HRQL) in combination with duration of life is used as a measure of benefit in cost-utility analysis. The EuroQol (EQ-5D), consisting of self-ratings on 5 dimensions of health, is used worldwide to measure HRQL in a wide variety of clinical populations. However, validity evidence supporting its use has been inconsistent. Consequently, the study purpose was to examine the validity of EQ-5D scores in 540 patients 26 to 89 years having hip or knee replacement surgery.

HRQL was measured pre- and 6 months post-surgery. The Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) and the Short-Form 36 (SF-36) were used as comparison instruments for assessing aspects of validity.

Four aspects of construct validity were examined: substantive, structural, external, and consequential. Although the development of the EQ-5D was not based directly on theory, the five EQ-5D dimensions reflect the broad areas measured by competing HRQL instruments: physical, social, and psychological functioning. Evidence supporting construct validity included dimensions relevant to the patient population, responsiveness, and convergent validity. Evidence adversely affecting construct validity included a lack of theoretical model, lack of clarity of items, items measuring more than one construct, the discrepancy

between societal and patient valued health states, the inconsistency of the scoring model with a multidimensional construct, and the potential social consequences of the effect of different weights on the calculation of quality-adjusted life-years (QALYs).

ACKNOWLEDGEMENTS

I thank each of my committee members for their contributions towards the completion of my dissertation. I especially thank my supervisor, Dr. Tom Maguire. His insight into the relevant measurement issues combined with a practical common-sense approach, and his broad knowledge and understanding of issues relevant to the health field were invaluable in the development and completion of the thesis. I am grateful to Dr. Todd Rogers, not only for his very thoughtful and thorough comments on various drafts of my thesis but also for his enthusiasm, support, and guidance throughout the program. I appreciate Dr. Maria Suarez-Almazor for our many constructive discussions and for involving me in other research projects. I thank Dr. Duncan Saunders and Dr. Mark Gierl for their support throughout my program. As well, I thank Dr. Ceinwen Cumming for the opportunity to explore many relevant aspects of quality of life through collaboration in other research projects.

I especially appreciate the contribution of my husband, Don. His valuable feedback, support, and encouragement were essential to the completion of this dissertation. And finally, thank you to my children, Christine, David, and Robert, and my mother, Muriel Conner, for their continuous support.

TABLE OF CONTENTS

	Page
Chapter I: Introduction	1
Overview	1
Purpose of the Study	4
Definition of Terms	5
Organization of the Thesis	8
 Chapter II: The Measurement of Health Related Quality of Life Using the EQ-5D	 9
Introduction to the Measurement of Health Related Quality of Life	9
Description of the EQ-5D	10
Section 1 Measurement of HRQL in the Target Population	11
Section 2 The Measurement of Preferences	13
Time-trade off method	15
Standard gamble method	16
The Conversion of EQ-5D Health States to EQ-5D Index Scores	17
The calculation of preference weights	17
The application of preference weights of an individual health state	19
The Calculation of QALYs	21
Use of the EQ-5D	21
 Chapter III: A Review of the Literature: the Reliability and Validity of EQ-5D Scores	 24
Reliability of the EQ-5D	24
Validity of the EQ-5D	25
Measurement Issues	28
Ceiling effect	28
EQ-5D scaling	29
Differences in societal and patient preferences	31
Contextual effects	32
Generalizability of EQ-5D valuations	33
Convergent and Discriminant Evidence	34
Known Group Differences	37
Responsiveness	39
Summary and Research Questions	41

Table of Contents (continued)

	Page
Chapter IV: Methods	44
Sample	44
Data Collection	44
Instruments	45
The Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC)	45
The SF-36	45
The EQ-5D	46
Matching WOMAC and SF-36 Dimensions with EQ-5D Dimensions	46
Data Analysis	50
Descriptive Statistics	50
Comparison of EQ-5D Index Scores Using Different Weights	51
Factor Analysis	52
Multitrait-multimethod	53
Responsiveness	54
Effect size	55
Repeated measures	55
The standard error of measurement	55
Quality-Adjusted Life Years (QALYs)	56
Chapter V: Results	58
Descriptive Statistics	58
Description of the Sample	58
Distribution of the EQ-5D Responses Pre- and Post-Surgery	62
Comparison of EQ-5D Distributions with SF-36 and WOMAC Subscales and Items	64
EQ-5D VAS and Index Score Distributions	72
Are EQ-5D Index Scores Reflective of Self-Reported HRQL of Respondents?	73
Relationship of Distribution to Scoring System	74
Comparison of Index Scores Using Different Weights	76
Factor Structure	82
Tests of Overall Goodness-of-Fit	82
Assessing Goodness-of-Fit of Individual Model Parameters	84
t-values	84
Standardized residuals	84
Modification indices	85
Confirmatory Factor Analysis using Essink-Bot et al.'s Factor Structure	85
Factor Analyses Using a Polychoric Correlation Matrix	86

Table of Contents (continued)

	Page
The Use of Factor Analysis in Determining the Factor Structure of the EQ-5D	88
Convergent and Discriminant Validity Evidence	89
Convergent Validity	90
Discriminant Validity	93
Responsiveness	95
Effect Size	97
Repeated Measures	97
Comparison of Changes in EQ-5D Dimensions with SF-36 and WOMAC Change Scores using the SEM	99
QALYs	102
Summary of Results	103
 Chapter VI: Construct Validity	 108
Substantive Component of Construct Validity	108
The EQ-5D Health State Descriptive System	109
Purpose and theoretical basis of the EQ-5D descriptive system	109
The nature and boundaries of the construct	111
Content representativeness and relevance	114
Preference Scores	120
Structural Component	127
External Component	131
Convergent and Discriminant Validity Evidence	132
Responsiveness	134
The Consequential Aspect of Validity	136
Messick's Conception of the Consequential Aspect of Validity	138
Values and Philosophical Theory Underlying the Measurement of the EQ-5D	140
The QALY Debate	142
How Can Validity Inquiry Assess the Potential Consequences of EQ-5D Use	145
Summary	151
 Chapter VII: Conclusions, Study Limitations, and Future Research	 152
Conclusions	152
Limitations of the Study	154
Future Research	155
 References	 157

LIST OF TABLES

Table	Title	Page
1.	Preference Weights for York Model	20
2.	Comparison of Dimensions of HRQL measures	27
3.	EQ-5D and WOMAC Items	47
4.	EQ-5D and SF-36 Physical Health Subscales	48
5.	EQ-5D and SF-36 Mental Health Subscales	49
6.	Coefficients for EQ-5D Index Scores	51
7.	Demographic Characteristics of Sample	59
8.	Descriptive Statistics Pre-surgery	60
9.	Reliability and Standard Errors of Measurement (SEM) for WOMAC and SF-36 Subscales	61
10.	Health States Used by 10 or More Respondents Pre- and Post-surgery	63
11.	Comparison of York TTO Weights and Pre- and Post-surgery Patient Regression Weights	77
12.	Pre-surgery Descriptive Statistics for Index Scores Using Different Weights	80
13.	Pearson Correlation Coefficients for Pre-surgery Index Scores Using Different Weights	81
14.	Fit Indices for Five Models Using Confirmatory Factor Analysis	83
15.	Polychoric Correlation Coefficients for EQ-5D Dimensions	87
16.	Factor Loadings Using EQ-5D Items and Principal Axis Factoring .	88
17.	Convergent and Discriminant Correlation Coefficients for Dimensions Measured by Three Measures Pre-surgery	91

List of Tables (continued)

Table	Title	Page
18.	Convergent and Discriminant Correlation Coefficients for Dimensions Measured by Three Measures Post-surgery	92
19.	Heterotrait-monomethod and Heterotrait-heteromethod Correlation Coefficients for Three Comparable Dimensions	95
20.	Comparison of Effect Size by Type of Surgery	96
21.	Repeated Measures Analysis of Covariance (Time by Joint) with Age as a Covariate	97
22.	Wilcoxon Signed Ranks Test for EQ-5D Dimensions	98
23.	Percent of Respondents Who Changed at Least One Level From Pre- to Post-surgery on each EQ-5D Dimension	99
24.	Comparison of % Improved on EQ-5D and SF-36 Compared with WOMAC for Comparable Constructs	101
25.	Descriptive Statistics for EQ-5D Index Scores and QALYs Using Different Weights	103

LIST OF FIGURES

Figure	Title	Page
1.	EQ-5D measurement and uses	11
2.	Distribution of responses on EQ-5D dimensions pre- and post-surgery	64
3.	Boxplots of WOMAC and SF-36 subscales for each level of EQ-5D item for comparable constructs	65
4.	Boxplots of SF-36 mental health and role emotional for each level of EQ-5D anxiety/depression	66
5.	Distribution of responses for WOMAC self-care items when EQ-5D = 1 (no problems)	67
6.	Distribution of WOMAC mobility items when EQ-5D mobility = 2 (some problems in walking about)	68
7.	Comparison of the distribution of responses for EQ-5D mobility and SF-36 'walking 1 block'	69
8.	The percentage of respondents reporting a problem on SF-36 and WOMAC activities for three levels of EQ-5D usual activities	71
9.	Distribution of EQ-5D index scores pre-and post-surgery	72
10.	SF-36 self-rated health for respondents with index scores equal to or less than 0	74
11.	Distributions of EQ-5D index scores pre-surgery using different Weights	79
12.	Matrix scatterplot of EQ-5D index scores using different weights. Patient weights are derived from pre-surgery data	82
13.	Scree plot based on EQ-5D items	87
14.	Joint by time interaction for EQ-5D index scores from pre- to post-surgery	98

CHAPTER I

Introduction

Overview

Health-related quality of life (HRQL) is a common and important component of health outcome measurement. Traditionally, evaluation of health interventions was determined by survival rates and life expectancy. However, these measures are now considered insufficient to capture the important health outcomes of interventions that are designed to improve health but that may also produce unintended side effects, thereby interfering with the quality of life post intervention. Further, with the increasing cost of health care and limited resources, health policy is often guided by economic analyses which compare the costs of health care interventions with its benefits. These 'cost-utility' analyses measure benefits in terms of the impact of the intervention on both length of life and quality of life. Cost-utility analysis results in a cost-utility ratio where the denominator reflects a change in health outcome and the numerator reflects the costs of that intervention (Gold et al., 1996).

Measures of HRQL used in cost-utility analysis produce a single index score, a number which summarizes the multiple dimensions which are used to measure HRQL. Rather than using a summative model, where each item is assumed to have equal weight when determining the single index score, the weighting system used is a preference weighting model in which the weights used are based on the relative desirability which people associate with various HRQL states. The summary score, obtained by aggregating across items now

weighted and representing each dimension of a health state, is then used as a measure of health outcome in cost-utility analysis.

The EuroQol (The EuroQol Group, 1990), the current version also known as the EQ-5D, is one such measure of HRQL that is being used in a wide variety of clinical populations. In response to the lack of a standardized HRQL instrument, the EuroQol was developed by the 'EuroQol Group', a multidisciplinary group of European researchers from England, Finland, the Netherlands, Norway, and Sweden. It was designed to be used alongside other health measures to enable the comparison of results in different groups, different countries, and different care settings (Nord, 1991a). Its purpose is twofold: 1. to describe HRQL, and 2. to provide a preference measure of HRQL (van Agt, Essink-Bot, Krabbe, & Bonsel, 1994).

The EQ-5D is a self-report measure used to describe the health status of populations or patient groups and to assess patient outcomes at a clinical level. Preferences are judgments about the desirability of various health outcomes. Respondents are asked to look at a subset of hypothetical health states and make judgements about the desirability of each state. Using scaling techniques, it is then possible to place the health states along a continuum. Modelling techniques are used to provide a set of weights which can be applied to health states, which have not necessarily been measured directly, to provide a preference measure for the particular health states of concern in the population of interest.

Despite the popularity of the EQ-5D, no single method for measuring

HRQL or preferences has been recognized universally as standard. Key measurement issues include which dimensions of HRQL should be measured, how preferences should be measured, and whose preferences should be used (Essink-Bot, Stouthard, & Bonsel, 1993; Selai & Rosser, 1995). For example, should preferences be derived from patients experiencing the health states, from health professionals, or from members of the general public (Gold et al., 1996)? Although it is generally accepted among health economists that preferences from the general population should be used rather than preferences from individual subgroups (Weinstein, Siegel, Gold, Kamlet, & Russell, 1996), it is also agreed that the community sample should be representative and informed and that the community have some understanding of the health states they are asked to value (Gold et al., 1996). Because there have been differences in societally derived preferences and those derived from patients experiencing the health states, and because there have been problems with obtaining representative samples for the measurement of preferences, it is important to understand the relationship between community and patient preferences.

Further, although the reliability of the EQ-5D HRQL scores and preferences is acceptable, evidence of validity is inconsistent and the usefulness of the EQ-5D is not well established. Several measurement issues affect the validity of EQ-5D HRQL and preference scores. These include: large ceiling effects; a lack of responsiveness (sensitivity to clinically relevant change in health status); failure of the preference scores to yield interval scaling; and the lack of generalizability of the preference scores due to contextual effects, low

response rates (Nord, 1991a), logical inconsistencies (Dolan & Kind, 1996), differences in population and patient derived preferences, and the difficulty in valuing the state of being dead, especially with ill populations (Selai & Rosser, 1995).

Although the EQ-5D has been less responsive than other HRQL measures (Hollingworth, Mackenzie, Todd, & Dixon, 1995; Jenkinson et al., 1997), few populations that experience large clinical changes in health status have been studied. The need for further research in the area of responsiveness of the EQ-5D is well documented (Brazier, Jones, & Kind, 1993; de Haan, Aaronson, Limburg, Hwer, & van Crevel, 1993; Essink-Bot, Krabbe, Bonsel, & Aaronson, 1997; Hurst et al., 1994; Hurst, Kind, Ruta, Hunter, & Stubbings, 1997; Kind, Dolan, Gudex, & Williams, 1998).

Purpose of the Study

The purposes of the EQ-5D are to assess individual quality of life and to assess preferences. The objective of the present study is to examine aspects of validity of the EQ-5D for these two purposes with patients who have experienced a large clinical change in health status. There are two parts to this examination: 1. an empirical investigation of the validity of EQ-5D scores in the context of total joint arthroplasty (TJA), a surgical procedure which involves both hip and/or knee replacement, and 2. a conceptual analysis of the EQ-5D as a measure of HRQL. TJA has been chosen because clinical changes are large (Laupacis et al., 1993), it is a common procedure, the waiting time for such a procedure is often used as an informal indicator of health care reform, and good data sets are available.

This research has indirect implications for health policy in providing evidence to assess the validity of EQ-5D scores for use as outcome measures in health care where the clinical changes brought about by a surgical intervention are large and clinically significant.

Definition of Terms

Before proceeding with a survey of the literature, it is first necessary to define the terms commonly found in the HRQL literature. These terms and their definitions are:

Cost-effectiveness analysis (CEA): a form of economic analysis which deals with the comparative analysis of alternative courses of action in terms of both their costs and consequences. Cost-effectiveness analysis results in a cost-effectiveness ratio. The numerator is a measure of cost; the denominator is a measure of health effect most relevant to the intervention under study, for example, number of life years saved.

Cost-utility analysis (CUA): a special form of cost-effectiveness analysis in which the measure of health effect is quality-adjusted life-years (QALYs) gained. Because CUA uses a common unit of measure, QALYs gained, it allows for comparisons across programs.

EQ-5D index score: a health status index in which the score is produced from the application of weights to a EQ-5D health state description. Weights are derived from the measurement of community preferences of hypothetical EQ-5D health states using either a visual analogue scale or a time trade off method.

Health-related Quality of Life (HRQL): used interchangeably with the term health status. Defined by Essink-Bot, a founding member of the EuroQol Group, as physical, psychological, and social functioning (Essink-Bot, van Royen, Krabbe, Bonsel, & Rutten, 1995).

Health state: the combination of one or more dimensions that describes the health-related quality of life of an individual. In the EQ-5D the dimensions include: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression.

Health status index: an aggregation of two or more dimensions of HRQL into a single number which summarizes the health of an individual or group of individuals.

Preference: judgments of the desirability of a particular health state. Preferences may be measured on an ordinal or cardinal (interval or ratio) scale and usually range from a value of 0 (death) to 1 (being healthy). There are two main types of preferences, valuations and utilities (definitions below).

Quality-Adjusted Life-Year (QALY): a single weighted measure of health outcome that combines morbidity and mortality. The QALY is calculated by multiplying the utility for a particular health state by the number of years a patient is in that health state and yields an equivalent number of years with full health. For example, three years in a health state valued at .67 is equivalent to two years of good health. QALYs are used in cost-utility analysis to relate the cost of an intervention to the number of QALYs gained by applying a treatment. "Cost per QALY gained" can be used to compare different treatments.

Standard Gamble (SG): a method of measuring preferences based directly on the axioms of von Neumann and Morgenstern utility theory in which judges are asked to compare life in a particular health state to a gamble with a probability p that perfect health is the outcome and $1-p$ that death is the outcome. The probability p is varied until the preference for the certain outcome is equal to the preference for the gamble; probability p is then a measure of preference (utility) for a particular health state.

Time Trade Off (TTO): a method of measuring health state preferences in which the respondent is asked to trade off life years in a state of less than perfect health (time t) for a shorter life span in a state of perfect health (time x). The ratio of the number of years in the perfect health state that is equivalent to the number of years in the less than perfect health state (x/t) is the measure of preference for that health state.

Utility: a type of preference measured on a cardinal scale such that the number represents the strength of a preference or desirability for a health state. Utilities are measured on a scale of 1 (being healthy) to 0 (death). The term utility is based on a normative model (von Neumann-Morgenstern utility theory) of how a rational individual ought to make decisions when faced with uncertain outcomes (von Neumann & Morgenstern, 1944).

Valuation: a type of preference measured on an interval or ordinal scale and measured under certainty, where there is no risk taken into consideration when valuing a health state.

Visual Analogue Scale (VAS): a type of rating scale in which an individual indicates their subjective values for one or more stimuli (in HRQL measurement, the stimuli are health states). The EQ-5D VAS, also referred to as a thermometer, is a 20 cm scale anchored by 0 (worst imaginable health state) and 100 (best imaginable health state).

Organization of the Thesis

The thesis is organized as follows. The next chapter begins with an overview of HRQL measures and then describes EQ-5D. In Chapter III the literature is reviewed as it relates to the reliability and validity of EQ-5D scores. The chapter concludes with the research questions. Chapter IV deals with the methods used in the present study, including the sample and materials and data analysis. Chapter V presents the results of the study. In Chapter VI study findings are discussed and integrated with the relevant literature and a conceptual analysis to assess the construct validity of the inferences of EQ-5D scores. Chapter VII closes with conclusions from the study and suggestions for future research.

CHAPTER II

The Measurement of Health-Related Quality of Life Using the EQ-5D

Introduction to The Measurement of Health-Related Quality of Life

Although the concept health-related quality of life has no one standard definition, most researchers agree that HRQL is a multidimensional construct which assesses three broad areas of health: physical, psychological, and social (Patrick & Erickson, 1993; Schipper, Clinch, & Olweny, 1996; Osoba, 1991; Torrance, 1986; Ware, 1987; Wood-Dauphinee, 1992). The measurement of HRQL is used for many different purposes: evaluation of the treatment of individual patients, the monitoring of HRQL of population subgroups, the evaluation of new therapies, and the allocation of health care resources (Kind, Gudex, Dolan, & Williams, 1994). Which HRQL tool is used depends on its applicability to the purpose. Because there is no 'gold standard' in HRQL measures and no one HRQL measure that satisfies all criteria, researchers often use a combination of HRQL measures. There are three broad classes of HRQL measures: condition-specific, generic, and preference-based (Bennett & Torrance, 1996).

Condition-specific measures may relate to a disease, a population of patients such as the elderly, or to a certain function or problem such as mobility. Items are designed to be responsive to particular changes that are relevant to a particular patient population. For example, a tool used with cancer patients will likely measure aspects of pain and symptoms of cancer treatment, such as nausea, whereas, for patients with rheumatoid arthritis, items related to joint pain,

mobility, and self-care would be more relevant. The advantage of specific tools is their responsiveness to change; the disadvantage is that it is difficult to compare outcomes across different populations. As well, they may not capture a more global quality of life.

Generic measures or non-condition specific measures are designed to apply to a wide variety of populations and contain a broad spectrum of items which usually measure physical, social, and emotional well being. They are generally not as responsive as specific instruments but can be used to compare relative impacts of health programs across various populations.

Preference-based measures produce a single index score which can be used in economic evaluation. Preference-based measures require a weighting system to be applied to a health state. This process requires a number of steps to derive the single index score: the measurement of HRQL of the target population, the measurement of preferences, and the assignment of preferences to the health states of individuals (Patrick & Erickson, 1993).

Description of the EQ-5D

The EQ-5D is both a generic measure and a preference-based measure. It consists of three sections: the measurement of self-rated HRQL, the measurement of HRQL preferences, and demographic information (See Figure 1). Usually only the first section, i.e., the measurement of self-rated HRQL, is used in clinical populations to obtain a measure of health outcome. If a preference-based measure of HRQL is required, the self-rated HRQL can then be converted to a preference measure.

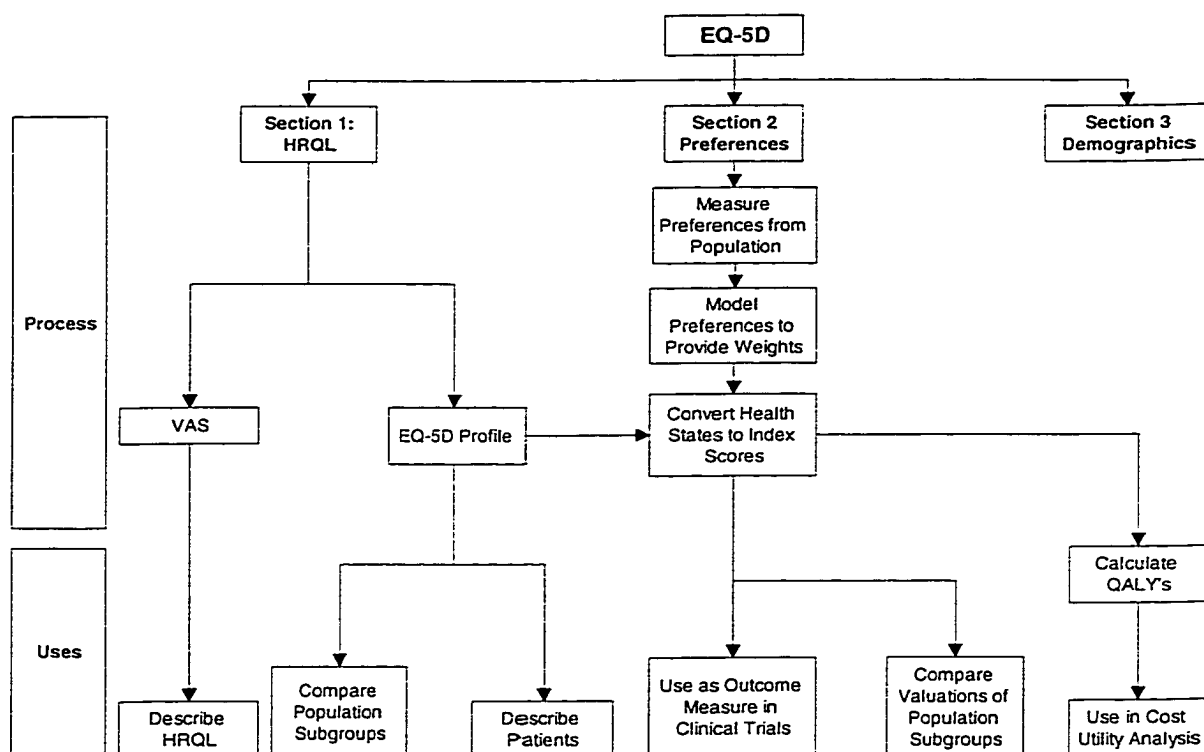


Figure 1. EQ-5D measurement and uses.

Section 1: Measurement of HRQL in the Target Population

In the first section of the EQ-5D two methods are used to measure HRQL: 3-point rating scales, and a 20 cm. 'thermometer' or visual analogue scale (VAS). First, respondents are presented with five items each measuring one dimension: mobility, self care, usual activities, pain/discomfort, and anxiety/depression. Descriptions of three levels of health are then presented for each dimension: level 1 (no problems), level 2 (some problems), and level 3 (inability or extreme problems). For example, the dimension 'usual activities' (work, study, housework, family or leisure activities) has the following three levels:

- 1 I have no problems with performing my usual activities;

- 2 I have some problems in performing my usual activities; and
- 3 I am unable to perform my usual activities.

Respondents are asked to place a check mark in the box next to the health level for each of five health state dimensions. The scores on each health dimension are then combined to produce the composite health state for that person. For example, a health state of 11221 would mean:

- 1 No problems in walking about
- 1 No problems with self-care
- 2 Some problems with performing usual activities
- 2 Moderate pain or discomfort
- 1 Not anxious or depressed.

There are 243 health state descriptions, but not all are plausible. For example, it would be highly unlikely to score a 3 on mobility (confined to bed) and a 1 on self-care (no problems in self-care). Profile scores have been used to describe patients (Wolfe & Hawley, 1997), compare population subgroups (Brazier et al., 1993), and to assess change in each dimension (Hurst et al., 1997). In addition to indicating their health level for each dimension, respondents are asked to rate their general health 'today' on a 3-point scale (better, much the same, or worse) compared to their general level of health over the past 12 months.

The second approach to measuring HRQL requires respondents to rate their current health status on a VAS with 0 (worst imaginable health state) and 100 (best imaginable health state) as endpoints. VAS scores have been used to

describe health status (Sculpher, Dwyer, Byford, & Stirrat, 1996) and as a self-valuation of health as a comparison measure with societally derived preferences (Hurst et al., 1994; Hurst et al., 1997).

Section 2: The Measurement of Preferences

The second section of the EQ-5D is designed to measure preferences; this process is referred to as 'valuation' or 'valuing of health states'. Preferences are numerical judgments of the desirability of a set of outcomes. More specifically, in HRQL measurement, preferences are a set of numbers assigned to health states. The numbers usually range from a value of 0 (being dead) to 1 (being healthy) and are interpreted as a measure of the strength of an individual's preference for a particular outcome (Torrance, 1986; Torrance & Feeny, 1989). There are two main types of preferences, valuations and utilities. Valuations are preferences measured under certainty, where there is no risk taken into consideration when valuing a health state. The term 'utility' arises out of utility theory, which is a prescriptive or normative theory of how people acting rationally ought to make decisions under uncertainty. Although, technically, the term utility is reserved for methods based on utility theory where uncertainty is incorporated, the terms utility, value, valuation, and preference are used interchangeably in the literature (Torrance, 1986).

EQ-5D preference scores are usually referred to as valuations and are usually measured in population surveys. Respondents selected from the general population are asked to make value judgements about sixteen hypothetical health states presented in two groups of eight on two consecutive pages. The

two sets of eight health states are presented in boxes on either side of a VAS, with the same endpoints as the one used previously. Thus the valuation in the case of the EQ-5D can range from 0 (worst imaginable health state) to 100 (best imaginable health state). The logically best health state (11111) and logically worst health state (33333) are repeated on both pages to provide a check on consistency and to serve as aids in maintaining a constant perceptual framework (The EuroQol Group, 1990; Kind et al., 1994). Thus, fourteen different health states are presented.

The criteria used in choosing these fourteen health states were that they are likely to occur in practice and that they represent a wide range of severity (The EuroQol Group, 1990). Respondents are asked to value these health states by drawing a line from the box containing the state to the point indicating on the VAS how good or bad these states would be for a person like them. They are asked to imagine that the health state will last for one year and that what happens after that is not known and should not be taken into account. The preference for each state is the value associated with its placement on the scale. Following this valuation process, respondents are asked to rate the state of 'dead' by drawing a line across the thermometer at the point where they would locate that state. Although there are fourteen core health states on the EQ-5D questionnaire, additional health states formed by other combinations of levels are possible. Researchers have measured up to 45 health states in any one study, using subsets of up to 16 health states for any one respondent to value (The EuroQol Group, 1990; Johnson, Coons, Ergo, & Szava-Kovats, 1998). The state

of unconscious has also been valued.

Time-trade off method. Although the EQ-5D was designed to measure preferences using the VAS, a second method, the time-trade off (TTO) method (Dolan, 1997), has frequently been used. Respondents first rate a set of 15 health states on a VAS with endpoints of 100 (best imaginable state) and 0 (worst imaginable state). However, in the TTO method, each state is to be regarded as lasting for 10 years without change, followed by death. The states are then valued by the TTO method, using a double sided board, one for states regarded by the respondent as better than being dead, the other for states regarded as worse than dead. The method differs depending on whether the states are valued as better than dead or worse than dead.

For states better than dead, respondents are asked to select a length of time (x) in the 11111 state that they regard as equivalent to 10 years in the target state. For example, if you were in the health state 11221 (no problems in mobility or self-care, some problems in usual activities and moderate pain/discomfort and no problems in depression/anxiety) for 10 years, how many years in the state of perfect health (11111) would be equivalent? The score for the target state would be $x/10$. For states worse than dead, the choice is between dying immediately and spending a length of time $(10-x)$ in the target state followed by x years in the 11111 state. The score is derived by the formula $-x/(10-x)$. Since states worse than dead are theoretically unbounded, Dolan transformed the valuations for states worse than dead using $(x/10) - 1$, where x represented the number of years spent in full health. Thus, the transformed

scores have a lower limit of -1 . It should be noted, though, that there are slight variations in the way TTO preferences have been obtained and consequently, there is no standard TTO method of deriving EQ-5D preferences (M. Buxton, personal communication, July 31, 1998).

Standard gamble method. Although the VAS and the TTO are the commonly used methods to measure preferences using the EQ-5D health state descriptions, the standard gamble (SG) has also been used to obtain preferences. The SG is the classical method of measuring utilities (Bennett & Torrance, 1996). The respondent is presented with two choices and asked to select the more preferred. One alternative offers the respondent an outcome with certainty (choice A), while the other alternative offers a gamble (choice B) with specified probabilities for the occurrence of two outcomes, the preferred state, p , and the least preferred state, $1-p$. For example, a person with severe cardiac disease and poor HRQL is given two choices: living in the present health state (Choice A) or taking a gamble on a treatment, such as bypass surgery (Choice B), which would lead to a better quality of life but also has a risk or probability $1-p$ of death. The two living health states would be specified as lasting the same length of time, usually an age-specific normal life expectancy for the individual and terminating in death. The probability p in choice B is varied systematically until the respondent is indifferent between choice A and B. The probability p at the indifference point is the utility of choice A for the specified duration. Because the SG mirrors decision making people have to make in real life, and also produces cardinal values, it is the preferred method for cost-utility

analysis.

Which method is best for measuring EQ-5D preferences is controversial. The choice of method depends on the purpose, feasibility, and the use to which the data will be put. The EuroQol Group used the VAS in designing the EuroQol because it was more feasible than the other methods for postal surveys. The TTO was initially developed as a proxy for the SG because of the difficulty using the standard gamble (SG) technique (D. Feeny, personal communication, March 17, 1999). Both methods require interviewing. The VAS differs from the TTO and SG in that no reference to decision making or uncertainty is made. Unlike the VAS, the TTO and the SG methods both require people to sacrifice one thing they value, life expectancy and certainty, respectively, for a gain in quality of life (Dolan, Gudex, Kind, & Williams, 1996a). In studies which have compared TTO and VAS valuations using EQ-5D health states, TTO valuations have been both higher and lower than VAS valuations (Brooks, 1996). In a random sample of the United Kingdom (U. K.) population, Dolan, Gudex, Kind, and Williams (1996b) compared variations of the TTO and SG and found TTO valuations higher than SG utilities.

The Conversion of EQ-5D Health States to EQ-5D Index Scores

The calculation of preference weights. Various modelling techniques have been developed and used to provide a set of weights for the estimation of valuations for those health states not directly measured (Brooks, 1996). That is, while up to 16 health states are valued by a respondent, there are additional health states that could have been valued (see p. 14). The model currently used

is a linear regression (Dolan, 1997). The regression model regresses the EQ-5D preference scores, using either the VAS or the TTO method, on a set of dummy variables (Dolan, 1997). As each EQ-5D dimension has three levels, two dummy variables are used for each dimension: M2 and M3 (mobility), SC2 and SC3 (self-care), UA2 and UA3 (usual activity), PD2 and PD3 (pain/discomfort) and AD2 and AD3 (anxiety/depression). Using 'mobility' as an example, if the score is 1, M2 and M3 would have a value of 0. For a score of 2, M2 = 1 and M3 = 0. For a score of 3 M2 = 0, and M3 = 1. An additional term in the model, N3, is a dichotomous variable indicating whether any of the dimensions are at level 3.

The regression model is:

$$Y = \beta_0 + \beta_1 M2 + \beta_2 M3 + \beta_3 SC2 + \beta_4 SC3 + \beta_5 UA2 + \beta_6 UA3 + \beta_7 PD2 + \beta_8 PD3 + \beta_9 AD2 + \beta_{10} AD3 + \beta_{11} N3 + \epsilon,$$

where Y is one minus the value given to a particular health state on a zero to one scale, β_0 is the constant term, β_1 to β_{10} are the coefficients or preference weights corresponding to the dummy variables described above, β_{11} is the coefficient corresponding to the N3 term, and ϵ is the residual. The resulting score can be thought of as a disutility composite with larger values indicating lower quality of life.

Preference weights have been obtained in various countries, including the Netherlands (Essink-Bot et al., 1993), Sweden (Brooks, Jendteg, Lindgren, Persson, & Bjork, 1991), the U. K. (Gudex, Dolan, Kind, & Williams, 1996; Dolan, 1997), Norway (Nord, 1991a), Spain (Badia, Fernandez, & Sequra, 1995), Germany (M. Buxton, personal communication, July 31, 1998), and the United

States (U. S.) (Johnson et al., 1998). As well, weights derived from an Alberta survey are currently being calculated (J. Johnson, personal communication, June, 1999). Although health states have been ranked using different preference weights, there is no international agreement as to which preference weights should be used as standard. Because the York surveys included a large ($n = 3395$) representative population sample, the York TTO valuations (Dolan, 1997) are the currently recommended and most commonly used weights in clinical research (M. Buxton, personal communication, July 31, 1998).

The application of preference weights to an individual health state. The preference weights can then be applied to each of the 243 possible health states. If the purpose is for a cost-utility analysis, health states would be measured for a target population before and after an intervention, for example, pre- and post-surgery. The health states for each individual would be converted to preference scores, referred to as EQ-5D index scores, and used in the calculation of quality-adjusted life-years (QALYs) for use in cost-utility analysis, discussed in the next section.

In the following example, TTO weights (Dolan, 1997) derived from a random sample of 3395 respondents from the U. K. are used (Table 1).

Table 1

Preference Weights for York Model

EQ-5D	Level 2	Level 3
Mobility	0.069	0.314
Self-Care	0.104	0.214
Usual Activity	0.036	0.094
Pain/Discomfort	0.123	0.386
Anxiety/Depression	0.071	0.236
Constant Term	0.081	
N3	0.269	

The estimated valuations are referred to as EQ-5D index scores and are produced from the subtraction of the disutility composite from 1 (representing perfect health). The algorithm for calculating the index score from the model is as follows:

$$1 - [\beta_0 + \beta_1M2 + \beta_2M3 + \beta_3SC2 + \beta_4SC3 + \beta_5UA2 + \beta_6UA3 + \beta_7PD2 + \beta_8PD3 + \beta_9AD2 + \beta_{10}AD3 + \beta_{11}N3].$$

For example, applying the York beta weights to the health state 11221, the estimated value for health state 11221 would be:

$$1 - [.081 + .069*0 + .314*0 + .104*0 + .214*0 + .036*1 + .094*0 + .123*1 + .386*0 + .071*0 + .236*0 + .269*0] = .760$$

Using this method EQ-5D index scores can be calculated for all 243 health states.

The Calculation of QALYs

Once EQ-5D index scores are obtained, they can be used as measures of health outcome or can be combined with life-years. Because of its utility and ease of implementation, the QALY is the method of choice (Gold et al., 1996). The QALY is a measure of health outcome that simultaneously takes into account HRQL and life years. The QALY is based on the idea that each individual in a lifetime moves along a path through different health states over time until death and that if an extra year of healthy life-expectancy is worth one, then an extra year of poorer quality life-expectancy is worth less than one (Williams, 1993). In the calculation of the QALY, index scores are multiplied by the number of life years remaining in that health state. An implication of using the QALY is that an extra year of healthy life-expectancy is regarded as equally valuable to everyone, regardless of illness or disability, in policy decisions. Cost-utility analysis requires that preferences be placed on a continuum between 1 (optimal health) and 0 (death) and that changes on the continuum be followed for the duration of survival. Because cost-utility analysis treats all gains as equal, a requirement of the preferences is that they have interval scale properties (Torrance, 1996).

Use of the EQ-5D

Since the EQ-5D was designed to enable comparisons across different countries, it was simultaneously released in the Dutch, English, Finnish, Norwegian, and Swedish languages. Subsequently, it has been translated into German, French, Spanish, Catalan, Italian, Danish, and Greek. In addition to

members of the EuroQol Group, users include researchers in France, Australia, the U. S., Japan, Hungary, Germany, Russia, Thailand, Italy, South Africa, and Canada (Brooks, 1996). As an indication of its acceptability in the field of health economics, the Panel on Cost-Effectiveness in Health and Medicine commissioned by the U. S. Department of Public Health recommended the use of QALYs to measure health effectiveness and listed the EQ-5D as one of a number of commonly used health-state classification systems (Weinstein et al., 1996). The EQ-5D has been used in cost-utility analyses in orthopedic patients (James, St Leger, & Rowsell, 1996), cancer patients (Norum, Angelsen, Wist, & Olsen, 1996; Uyl-de Groot, Hagenbeek, Verdonck, Lowenberg, & Rutten, 1995; Uyl-de Groot et al., 1997; Vellenga et al., 1996), and in patients with dystonia (Gudex, Hawthorne, Butler, & Duffey, 1997).

A search using Medline, Cinahl, Psychlit, Embase, and Healthstar from 1990 to the present produced 94 published articles with EuroQol or EQ-5D in the title or abstract. An additional 26 published articles were found by consulting reference lists of articles generated by the search and obtaining a reference list from the EuroQol Group. Over half of these articles, 65, were published over the last three years (1997-99).

Despite the growing popularity of the EQ-5D, issues about the construct validity of EuroQol scores remain. For example, as pointed out earlier, the use of different systems of weights used to calculate EQ-5D index scores which are subsequently used to calculate QALYs in cost-utility analysis raises questions about the validity of inferences drawn from the cost-utility analysis. In the

following chapter, the evidence will be reviewed as it relates to the reliability and validity of the EQ-5D.

CHAPTER III

A Review of the Literature: the Reliability and Validity of EQ-5D Scores

Reliability of the EQ-5D

Reliability is the degree of consistency of test scores and more specifically is the ratio of true score variance to observed score variance (Crocker & Algina, 1986). Five studies have examined test-retest reliability of EQ-5D scores (Brazier, Walters, Nicholl, & Kohler, 1996; Dolan et al., 1996a; Gudex et al., 1996; Hurst et al., 1997; van Agt et al., 1994). Of these studies, five examined reliability of EQ-5D index scores, one examined VAS scores, and one examined profile scores. Reliabilities were moderate to high and varied with sample size, type of population, time, and the statistical methods and types of scores used. Using the intraclass correlation (ICC), the reliability of the index scores varied from .73 to .85 in rheumatology patients after a 2 week interval (Hurst et al., 1997), to a mean ICC of .78 using VAS valuations (Gudex et al., 1996) and .73 using TTO index scores (Dolan et al., 1996a) in a population survey after a 10 week interval. The test-retest correlation of index scores measured 6 months apart in elderly women was .67 while the VAS score correlation was .53 (Brazier et al., 1996). Using generalizability theory, Van Agt et al. (1994) found that time accounted for .05% of the total variance in index scores in a sample of 208 respondents in a postal survey. The only study to examine the test-retest reliability of profile scores found scores did not change significantly in any domain over three months (Hurst et al., 1997). One challenge to measuring reliability in a dynamic construct is choosing the optimal time period between

measures.

Validity of the EQ-5D

Validity is an ongoing evaluative process of scientific inquiry, both quantitative and qualitative, by which we determine the degree of confidence we can place in the inferences we make based on a test score (Messick, 1989; Streiner & Norman, 1995). This judgment is based on empirical evidence plus theory and requires evaluation of the evidence and consequences of both test interpretation and test use (Messick, 1989). Traditionally, validity has been divided into three types: construct validity, criterion-related validity, and content validity. Over the years the concept of validity has evolved into a more unified view with construct validity as the foundation of validity inquiry. Most theorists accept that construct validity includes construct, content, and criterion-related evidence. However, the inclusion of an appraisal of the implications and social consequences of test interpretation and use as part of validity inquiry is controversial (Maguire, Hattie, & Haig, 1994; Shepard, 1997). Although the consequential aspect of validity is not referred to in the literature review, it is discussed as it relates to the application of EQ-5D scores in Chapter VI.

Because there is no gold standard of HRQL measurement, new instruments are often compared against older and more established HRQL instruments to assess aspects of validity. One of the most common generic HRQL profile measures is the Short-Form 36 (SF-36) (Ware & Sherbourne, 1992) or its shorter version, the SF-12 (Ware, Kosinski, & Keller, 1996). The SF-36 consists of 36 items, which are spread across eight subscales: physical

functioning, role limitations due to physical health problems, bodily pain, general health, vitality, social functioning, role limitations due to emotional problems, and mental health. The SF-12 consists of one or two items from each of the eight SF-36 dimensions for a total of 12 items. Although the SF-36 is widely used, it is not preference-based, and does not provide a health status index score; thus, it cannot be used to calculate QALYs. The EQ-5D has also been compared with other generic HRQL instruments such as the Rosser Index (Rosser & Kind, 1978), the Nottingham Health Profile (NHP) (Hunt, McEwen, & McKenna, 1985), and the Dartmouth Primary Care Cooperative Information Project (COOP) (Nelson, Wasson, Johnson, & Hays, 1996), and also various condition-specific instruments, depending on the specific patient population examined. For example, the Health Assessment Questionnaire (HAQ) (Ramey, Fries, & Singh, 1996), an assessment of functional status, has commonly been used in patients with rheumatic diseases, such as rheumatoid arthritis. Table 2 provides a comparison of the content of commonly used generic and condition specific instruments that have been used as comparison instruments with the EQ-5D. One of the complexities in examining the validity of the information yielded by the EQ-5D is that the EQ-5D measures two constructs, HRQL and preference. As well, the EQ-5D uses a combination of scoring methods with different uses being made of the scores. The profile score, the VAS, and the preference scores can all be used as a measure of health outcome. Both the societally derived preference and the self-reported VAS can be used as a measure of preference.

Table 2

Comparison of Dimensions of HRQL Measures

	HRQL Instruments								
	EQ-5D ^a	SIP ^b	NHP ^c	RI ^d	QWB ^e	COOP ^f	SF-36 ^g	HAQ ^h	HUI ⁱ
Self-care	X	X		X	X		X	X	X
Mobility	X	X	X	X	X	X	X	X	X
Pain	X		X	X	X	X	X	X	X
Social	X	X	X	X	X	X	X		
Role	X	X	X	X	X	X	X	X	X
Leisure	X	X	X	X	X				
Emotional	X				X	X	X		X
Energy		X	X		X		X		
Sensation									X
Cognition		X							X
Symptoms					X				
Communication		X			X		X		
General Health									
Perception	X					X	X		
Change in									
Health	X					X	X		

^{a,d,i} Preference-based

^bSIP: Sickness Impact Profile

^cNHP: Nottingham Health Profile

^dRI: Rosser Index

^eQWB: Quality of Well-Being Scale

^fCOOP: Dartmouth Primary Care Cooperative Information Project

^gSF-36: Short-Form 36

^hHAQ: Health Assessment Questionnaire

ⁱHUI: Health Utilities Index

Further, as well as the more commonly used York TTO weights (Hurst et al., 1997; Jenkinson et al., 1997; Jenkinson, Stradling, & Petersen, 1998; MacDonagh et al., 1997 ; Wolfe & Hawley, 1997), researchers have used various other weights to calculate EQ-5D index scores, such as U. K. FROME weights (Brazier et al., 1996; James et al., 1996), VAS weights (Hurst et al., 1994), and Spanish weights (Badia, Diaz-Prieto, Rue, & Patrick, 1996). Finally, while most recent studies use the current EQ-5D, earlier studies used a six-dimension version (Brazier et al., 1993; Brooks et al., 1991; de Haan et al., 1993; Essink-Bot, Bonsel, & van der Maas, 1990; The EuroQol Group, 1990; Nord, 1991a; Nord, Richardson, & Macarounas-Kirchmann, 1993; O'Hanlon, Fox-Rushby, & Buxton, 1994; Selai & Rosser, 1995).

The sources of evidence used to assess validity are organized under the following headings: measurement issues, convergent and discriminant evidence, known group differences, and responsiveness.

Measurement Issues

A number of findings related to measurement issues have challenged the validity of EQ-5D scores: large ceiling effects, failure of the preference scores to yield interval scaling, differences in population and patient derived preferences, contextual effects, and the lack of generalizability of the preference scores due to low response rates (Nord, 1991a) and logical inconsistencies (Dolan & Kind, 1996).

Ceiling effect. For use as an HRQL outcome measure it is essential that a tool have an adequate distribution of responses so as to be able to assess

relevant clinical differences in patients outcomes. Extreme skewness with a large ceiling effect is a characteristic of EuroQol data both in population surveys and in ill populations. In the context of the EQ-5D, ceiling effect refers to the fact that there is little variability possible at the high quality end of the continuum. In a population study of general practitioner (GP) patients Brazier et al. (1993) examined possible ceiling effects of the EuroQol (six-dimension version) by comparing the distributions of EuroQol scores with those of the SF-36 subscales. They found that only 10 of the 243 possible EuroQol health states accounted for 95% of the self-rated health states. EuroQol data were skewed with over 95% at the ceiling (scores of 1) for the functional dimensions of self-care, mobility, main activity, and family/leisure in contrast to 37 to 72% for the functional dimensions as measured by the SF-36. For the anxiety/depression dimension, 81% were at the ceiling while only 2% were at the ceiling for mental health on the SF-36. In a study of migraine patients, Essink-Bot et al. (1997) found that 70 to 80% of the patients scored at the ceiling using the EQ-5D.

EQ-5D scaling. According to the EuroQol developers, the capacity to yield a single index preference score for any health state is the most important property of the EuroQol (The EuroQol Group, 1990). While acknowledging the multidimensional nature of health status, the researchers worked from the philosophy that health status can be modelled on a unidimensional continuum (Kind et al., 1994). However, only one study has assessed the factor structure of the EQ-5D. In a sample of migraine sufferers and non-migraine sufferers, Essink-Bot et al. (1997) combined the EQ-5D items with subscale scores from

other HRQL instruments and then factor analyzed the combined set. Two analyses were completed, the first, using the EQ-5D, the NHP, and the SF-36, and the second, the EQ-5D, the SF-36, and The Dartmouth Primary Care Cooperative Information Project (COOP), a measure of functional status used for screening, assessing and monitoring patient function. Both analyses yielded two factors, physical health and mental health. The anxiety/depression EQ-5D item loaded on the mental health dimension and self care, mobility, usual activities, and pain/discomfort loaded on the physical health dimension. Results are difficult to interpret due to the combination of items and subscales from different measures and the lack of competing models and other studies with which to compare results.

A requirement of the EQ-5D scoring model is that index scores are measured on an interval scale. This demand arises from the use made of index scores in the estimation of QALYs. Although the VAS was designed to measure preferences on a interval scale, Brooks (1996), in a review of the literature on behalf of the EuroQol Group, concluded that evidence did not support the VAS method of valuation as one that provides interval properties. Nord (1991a) found that respondents perceived the numbers on the scale as percentages of fitness with the possible consequence of bad health states being placed closer together even if their degrees of badness actually differed very much. As a consequence, he concluded that intervals may have to be weighted more the closer they are to the bottom of the scale.

Differences in societal and patient preferences. The measurement of EQ-5D community preferences assumes that societal based preferences are reflective of the self-reported valuations of health states that are actually experienced by patients. It has been suggested that population derived weights should be adjusted based on weights obtained from populations that are experiencing adverse health states (Rosser & Kind, 1978). Little work has been done on the relationship between preference weights derived from the public and those derived from ill populations with less than optimal health states. In a pilot study where valuations were obtained from acutely ill inpatients, Selai and Rosser (1995) found that patients valued the health states as having a higher valuation than did the general population. As well, the mean state of 'being dead' was valued more highly than previous valuations obtained from population surveys. A limitation of the study was the small sample size of 23 respondents. In a study with intensive care patients, Badia et al. (1996) reported that the worst health states were rated higher and some patients rated the better states lower than preferences derived from a healthy population.

In three other studies, the relationship between the valuation for a patient's self-reported health status and the patient's self-reported VAS score was examined. Norum et al. (1996) reported a significant correlation between the EQ-5D VAS and index scores (TTO) in patients with Hodgkin's Disease; however, they did not report the value of the correlation coefficient. In two samples of rheumatoid arthritis outpatients, correlations between self-reported VAS scores and EQ-5D index scores (TTO) varied from .54 (n =55) (Hurst et al.,

1994) to .57 (n = 1372) (Wolfe & Hawley, 1997). However, Wolfe and Hawley found gaps in the distribution of scores, heteroskedasticity, discrepancy between the two measures, and a number of patients with EQ-5D health states valued worse than 0 (death).

Contextual effects. While developing the EQ-5D, researchers examined the effect of variations in measurement methods on the valuations. Consistency in the valuation scores was adversely affected by different methods of instruction, the length of the rating scale, and the number of states to be valued (Kind et al., 1994). There is some evidence of a contextual effect where valuations for health states may be dependent on the states with which they are compared. The inclusion of 'dead' is controversial and many respondents have not completed this part. However, since death is a health outcome, it is felt necessary to include this state (Kind et al., 1994). In a pilot study removal of the state "dead" from a set of eight health states produced an average reduction of 4.5 points in valuations for all remaining states (Nord, 1991a). Dolan (1996) varied the duration in hypothetical health states, using time periods of one month, one year, and 10 years and found that poor states of health became more intolerable the longer they lasted. In a survey study designed to compare valuations by Norwegian subjects to British, Dutch, and Swedish subjects and to address issues of validity and feasibility, the six dimension EuroQol was administered to eight random samples of people aged 18 and 84 years of age (Nord, 1991a). Forms were varied as to number of rating tasks, wording, and anonymity. Inconsistencies occurred much more often between states that were

on separate pages than between states on the same page.

Generalizability of EQ-5D valuations. To use valuations in cost-utility analysis requires that valuations be based on a representative sample of the population. However, survey response rates in various countries have varied from 26% (Johnson et al., 1998; Nord, 1991a) to 83% (Brazier et al., 1993). The higher response rate was achieved by using only the first section of the Euroqol, indicating some difficulty with the task of valuing health states. Generalizability of valuations has been questioned due to high levels of non-response and unsuccessful response (based on more than two missing valuations) in the Dutch postal surveys (Essink-Bot et al., 1993). The majority of respondents with successful responses tended to be relatively healthy, middle class, and younger (Essink-Bot et al.). In a Norwegian survey, factors which affected the response rate were age (subjects over 70 years had a 16% response rate), and number of health states to be valued (Nord, 1991). The ill, the aged, and the less educated have been underrepresented in most surveys. Older age groups have been associated with higher valuations and increased missing data (Brazier et al., 1993). Other factors significantly affecting valuations include social class, education, home ownership, and illness in family members (Gudex et al., 1996).

In the measurement of valuations, logical inconsistencies have also been problematic. Logical inconsistencies occur when an expected ordinal relationship between certain pairs of states does not occur. For example, state 13221 should be logically worse than state 12221. However, a health state of 12111 can not be assumed to be rated with a higher or lower valuation than the

health state 11211. Dolan and Kind (1996) hypothesized that logical inconsistencies could be due to factors such as the difficulty of the task of rating a complex multidimensional description on a unidimensional scale, framing effects, respondent confusion, and respondent misinterpretation. Researchers have often excluded respondents from the analysis if they were logically inconsistent (Dolan & Kind, 1996), thus little is known about the reasons for logical inconsistencies. As well, since logical inconsistency is higher in older and less educated populations (Dolan & Kind, 1996), a bias in health state valuations could occur.

Convergent and Discriminant Evidence

Convergent evidence is used to examine the relationship between the scores obtained from the measure of interest with other measures of the same construct whereas discriminant evidence examines relationships of scores obtained from similar but different constructs. Studies are not easily comparable as both the earlier six-dimension EuroQol and the current five-dimension version have been used. In a study comparing the six-dimension EuroQol with the SF-36, Brazier et al. (1993) found, as hypothesized, that patients responding with a health problem on the EuroQol had significantly lower mean SF-36 scores (lower HRQL) for all dimensions, with the most marked differences for dimensions tapping similar domains of health. Similar results were found in a study using the same method with the five-dimension EQ-5D and the SF-12 (Johnson & Coons, 1998). As hypothesized, relationships were stronger between the mental dimension of the SF-12 and the anxiety/depression dimension of the EQ-5D, and

the physical dimension of the SF-12 and the self-care, mobility, usual activities, and pain/discomfort dimensions of the EQ-5D. Relationships between less comparable dimensions were not as strong. In a Dutch study of migraine sufferers, Essink-Bot et al. (1997) reported that 'anxiety/depression' (EQ-5D) correlated best with 'feelings' (COOP), and 'usual activities' (EQ-5D) with 'daily activities' and 'social activities' (COOP). However, they failed to point out that 'self-care' (EQ-5D) correlated more highly with 'daily activities'(COOP) and with 'social activities' (COOP) and that other high correlations between different dimensions were observed. Using the Rosser Index, James et al. (1996) reported positive correlations between the EQ-5D 'mobility' and 'pain' scores and similar dimensions in the Rosser index, but statistics were not reported. In a study comparing the EQ-5D with the SF-36 and NHP, Chetter et al. (1997) found that convergent correlation coefficients were higher than those measuring different dimensions for pain, mobility, self-care and depression. However, EQ-5D usual activities correlated more highly with SF-36 pain than with comparable SF-36 subscales. Finally, in rheumatoid arthritis patients, Hurst et al. (1997) compared scores in three EQ-5D domains, self-care, pain/discomfort, and anxiety/depression, with three condition specific scales: the HAQ, the Pain-VA, a pain visual analogue scale, and the Hospital Anxiety and Depression Scale (Zigmond & Snaith, 1983), respectively. Using t-tests to compare mean scores between pairs of EQ-5D levels, there was a significant deterioration in scores as patients scores moved from level 1 to 3 in the EQ-5D. Patients were stratified into 5 groups according to function. Using non-parametric statistics, results

showed that with decreasing function, the proportion of patients reporting some or severe problems in each of the five EQ-5D dimensions increased progressively.

In a comparison of EuroQol index scores with the SF-36 mental and physical health summary scores, Brazier et al. (1993) found positive correlations ranging from .48 to .60. In a similar analysis, the EQ-5D VAS was positively correlated with both the physical ($r = .55$) and mental ($r = .41$) dimensions of the SF-12 (Johnson & Coons, 1998). In a cross sectional study with multiple sclerosis patients, Rothwell, McDowell, Wong, and Dorman (1997) reported that the EQ-5D VAS correlated significantly with four subscales of the SF-36 ($r = .42$ -.57) but did not correlate with two disability scales.

Two studies compared EQ-5D index or VAS scores with condition-specific measures. In a pilot study Busschbach, Horikx, van den Bosch, Brutel de la Riviere, and de Charro (1994) used the EQ-5D VAS as a measure of patient valuation and compared scores with the NHP and three single index measures before and after lung transplantation; results showed the EQ-5D gave lower values than the other measures for the better health states. However, a small sample size ($n = 6$) and the use of retrospective data decreases the validity of the findings. In their study of rheumatology patients in which the EQ-5D was compared with condition-specific measures, Wolfe and Hawley (1997) reported a correlation of .60 between a global severity VAS and the EQ-5D VAS. Significant correlations between the EQ-5D index score and the EQ-5D VAS with condition-specific measures of pain, depression, anxiety, and a disability index ranged from

.47 to .69, and were always stronger with the EQ-5D index score than with the EQ-5D VAS.

Known Group Differences

The known group difference procedure is used to examine the ability of an instrument to discriminate between groups that are expected to differ on their health status. Using the 6-dimensional EuroQol in a postal survey of patients from General Practitioner lists, Brazier et al. (1993) examined hypotheses predicting different patterns of health between groups based on age, sex, socioeconomic status, use of health services, and diagnosis. Expected significant differences were found in most of the dimensions for age, social class, use of services (except for mobility), and physician visit. However, expected differences were not found on the basis of the diagnosis of a chronic health problem in the following dimensions: mobility, self-care, main activity, and family/leisure activity. As well, using the index scores, they found expected differences for age, gender, socioeconomic status, and chronic health problems. Johnson et al. (1998), in a U.S. population survey, found similar differences with the EQ-5D VAS. Using chi square analysis, they also found differences on all five EQ-5D dimensions between groups based on employment status, education, income, and medical problems. Groups based on age and marital status differed on all dimensions except anxiety/depression, and gender differences were found only on anxiety/depression with females significantly more likely to report some level of problem than males.

In a preliminary analysis using a representative sample of the Dutch population, Essink-Bot et al. (1995) compared the ability of four instruments, the EQ-5D, the SF-36, the COOP and the NHP, to discriminate between migraine sufferers and non-migraine sufferers. Using chi-square tests, they found significant differences between groups for three EQ-5D dimensions: usual activities, pain/discomfort and anxiety/depression. In a more extensive analysis of a larger sample, Essink-Bot et al. (1997) concluded that the SF-36 discriminated the best between migraine groups. They also reported significant differences between groups on all five EQ-5D dimensions. However, the sample size was large ($n = 1011$), leading to differences that, while statistically significant, were clinically questionable. For example, the mean 'self-care' scores, which were statistically significant, differed by only 0.02. Hollingworth et al. (1995) used the EQ-5D to differentiate knee patients from those in the general population. They found that percentages of knee patients reporting a problem in mobility, usual activities, and pain/discomfort were noticeably higher than in the general population; however, no statistical tests were reported to support this claim.

Brazier et al. (1996) compared the EQ-5D, the SF-36, and the Disability Survey in elderly women. Significant differences were detected by all three instruments based on groups who had recently visited their physician, had hospital inpatient stay, or had any long-standing illness. The SF-36 and the Disability Survey detected age differences where the EQ-5D did not. The reported VAS scores were reportedly higher than the EQ-5D index scores but

statistical analysis was not carried out. Finally, using a post test only design, Sculpher et al. (1996) compared the health status of two groups requiring surgery for menorrhagia and found no significant differences between groups using the EQ-5D VAS and the SF-36.

Responsiveness

An important aspect of validity in any instrument used for evaluative purposes is the instrument's responsiveness or sensitivity to clinically relevant change. There has been limited study of the ability of the EQ-5D to respond to clinically relevant change. Results are mixed and comparisons are complicated due to differences in the EQ-5D scores compared, time between measurements, and the context in which they were compared.

In a longitudinal study of men undergoing transurethral resection of the prostate, MacDonagh et al. (1997) found significant improvement over a 12 month period in EQ-5D VAS and index scores, and in the EQ-5D dimensions of usual activities, pain/discomfort, and anxiety/depression. Badia et al. (1996) had intensive care patients recovering in a step down unit retrospectively rate their health state prior to admission to intensive care and rate their current health state before leaving the step down unit. Using chi-square analysis to assess changes in the five EQ-5D categories, they found that the patients worsened significantly in VAS scores as well as mobility, self-care, and pain/discomfort. However, a limitation of this study was the retrospective measurement of prior health state and the lack of a stipulated time frame for the prior health state. As well, neither study hypothesized changes in specific dimensions. Finally, in a cost-utility

analysis of orthopedic patients measured pre-operatively and 6 months post discharge, the EQ-5D index score was statistically significantly different over time (James et al., 1996).

In the few studies that have compared responsiveness of the EQ-5D with other instruments, the EQ-5D has been less responsive than the SF-36 (Hollingworth et al., 1995; Jenkinson et al., 1997; Jenkinson et al., 1998), and condition specific measures, such as the American Urological Association symptom score (Jenkinson et al., 1997) and the Patient Generated Index (Jenkinson et al., 1998). In patients undergoing magnetic resonance imaging of the knee, Hollingworth et al. (1995) found that EQ-5D index scores measured at baseline and at 6 months were capable of detecting health improvements but the VAS was not sensitive to change. In a pilot study of construct validity of the EQ-5D scores with rheumatology outpatients over three months, Hurst et al. (1994) compared the correlations of EQ-5D index change scores and EQ-5D VAS change scores with condition-specific change scores (HAQ, and measures of pain, depression and anxiety). Correlations were relatively low, ranging from .10 to .52 for EQ-5D VAS scores and from .10 to .38 using EQ-5D index scores. However, the group as a whole was relatively stable with median EQ-5D change scores of zero. As well, the sample size was small ($n = 55$) with resulting large standard errors. In a subsequent study with a larger sample ($n = 233$), Hurst et al. (1997) found that change scores using EQ-5D index scores and the EQ-5D VAS were significantly correlated with self-reported change in rheumatoid arthritis and the condition-specific measures used in the 1994 study. Changes

over 3 months in profile scores in self-care, mobility, usual activities, and pain/depression were positively related to three categories of self-reported change (same, better, worse) in rheumatoid arthritis.

Summary and Research Questions

In summary, although the EQ-5D is increasingly being used as a preference-based measure of HRQL, there are a number of issues which affect the validity of inferences based on EQ-5D scores. Although test-retest reliability is acceptable, empirical evidence provides weak support for construct validity. An assessment of validity is complicated by the combination of scoring methods, the use of different weights, and the two versions of the EuroQol. As well, few studies have been designed explicitly to assess validity issues.

The validity of the EQ-5D as a preference-based HRQL measure is threatened by large ceiling effects in patient and population groups, a lack of research on the dimensionality of the EQ-5D, a lack of evidence supporting the interval scaling requirement, differences in societal and patient preferences, contextual effects on the measurement of valuations, and the lack of representative samples for the measurement of valuations due to poor response rates and logical inconsistencies. Because there is no gold standard in HRQL measure, the EQ-5D has been compared with older and commonly used generic instruments, such as the SF-36, and condition specific instruments. Validity evidence has included convergent and discriminant evidence, known group differences, and responsiveness.

Findings have moderately supported convergent and discriminant validity evidence for EQ-5D profile scores, although results are not consistent. In studies that have compared index scores and VAS scores with other measures of physical and mental health status, correlations have been moderate to high, ranging from .42 to .69. Although these findings generally support the convergent evidence of EQ-5D index and VAS scores, comparisons are difficult because of the use of different measures which are not necessarily measuring comparable constructs. Therefore, interpretability of these findings is questionable. In the studies that have examined known group differences, most of the evidence has supported the ability of the EQ-5D profile scores to differentiate between groups but with less ability than the SF-36.

The usefulness of the EQ-5D as a preference-based measure in evaluation studies is particularly dependent on its ability to be sensitive to relevant clinical changes. In the few studies that have examined responsiveness of EQ-5D dimensions, evidence has supported the ability of the EQ-5D to be sensitive to dimensions that are relevant to the patient's illness. In studies which have compared the responsiveness of the EQ-5D to other instruments, the EQ-5D has generally been less responsive. Because generic HRQL instruments do not include condition-specific items, they have less ability to detect clinically relevant change in health status. The lack of sensitivity to change in clinical populations may be due to the ceiling effects, broadness of levels, lack of relevance of the items to the population, and small effect sizes. Few studies

have examined the responsiveness in intervention studies where a significant clinical change would be likely to occur.

The present study addressed several aspects of construct validity. Empirical evidence and conceptual analysis were used to evaluate the construct validity of the interpretation of EQ-5D scores. One technique for assessing validity in HRQL instruments is to compare aspects of validity evidence of the EQ-5D with instruments with known psychometric properties. Two instruments were used for comparison: a disease specific instrument designed for patients with osteoarthritis, The Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) (Bellamy, Buchanan, Goldsmith, Campbell, & Stitt, 1988), and a widely used generic instrument, the SF-36. The research questions were:

1. What is the distribution of EQ-5D responses compared with the WOMAC and the SF-36?
2. What is the relationship between the EQ-5D self-reported VAS and the EQ-5D valuations using TTO and VAS derived valuations?
3. What is the factor structure of the EQ-5D in a sample of patients undergoing TJA?
4. What is the convergent and discriminant evidence of the EQ-5D?
5. Is the EQ-5D, measured by the EQ-5D VAS, the EQ-5D profile scores and the EQ-5D valuation scores, responsive to change from pre to post-surgery in patients undergoing TJA?
6. What are the potential consequences of applying different weights to the health states?

CHAPTER IV

Methods

Sample

The sample consisted of data from a prospective study of 540 patients undergoing either hip or knee replacement surgery at the Walter MacKenzie Health Sciences Center or Royal Alexandra Hospitals in 1997. The target population was all patients in the Edmonton area who were placed on a waiting list for elective primary total joint arthroplasty (TJA) by their respective orthopedic surgeons from January 1996 to January 1997. Eighty-five per cent of the patients contacted agreed to participate in the study. Access to the data was permitted by the Principal Investigators, Dr. Maria Suarez-Almazor and Dr. Donald Voaklander.

Data Collection

HRQL data were collected pre-surgery (86.7% within 31 days of surgery and 13.3% from 32 to 67 days pre-surgery) and 6 months post-surgery. A nurse and a physical therapist collected the pre-surgery data and the same physical therapist collected the post-surgery data. The data were collected during home interviews. The following three HRQL questionnaires were administered in the order listed: the WOMAC (a disease specific instrument), the SF-36 (a generic measure), and the EQ-5D (profile measure and the VAS). In addition, demographic information -age, gender, place of residence, and education- , and information about the use of an aid for walking was collected. The time required to administer the HRQL measures and collect the additional information was

approximately 20 minutes. The admission diagnosis and type of surgery were obtained from a chart review.

Instruments

The Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC)

The WOMAC is a disease specific health status questionnaire which consists of three subscales: pain, stiffness, and physical function. It was designed to measure function and symptoms in patients with osteoarthritis of the hip and/or knee. The instrument consists of 24 questions (5 pain, 2 stiffness, and 17 physical function) with degree of severity or difficulty measured on a 5 point Likert scale (0 = none, 1 = mild, 2 = moderate, 3 = severe, 4 = extreme). Subscale scores are derived from the summation of items for each dimension. Subscale scores were transformed to a 0 -100 scale with the higher score reflecting better function. The WOMAC items, developed from interviews with arthritis patients, were designed to capture quality of life components relevant to arthritis patients; therefore, it was expected to be the most responsive measure to surgical intervention (Bellamy et al., 1988).

The SF-36

The SF-36 is a widely used generic HRQL measure consisting of 36 Likert items. The number of scale points vary from 2 (e.g., yes or no, for the role physical subscale) to 6 (e.g., none to very severe for the bodily pain subscale). Eight subscales scores are derived from the summation of item scores and transformed to a 0 to 100 scale with higher numbers representing better health. The physical health subscales include physical functioning, role-physical, bodily

pain, and general health. The mental health subscales include vitality, social functioning, role-emotional, and mental health. The SF-36 also includes two aggregate scores, the physical component summary (PCS) and the mental component summary (MCS). The PCS and MCS scores were aggregated according to standard SF-36 scoring algorithms (Ware, Kosinski, & Keller, 1994) which produce standard scores with a mean of 50 and a standard deviation of 10; the norms are based on general population surveys in the U. S.

The EQ-5D

The EQ-5D was described in Chapter II. Consequently, other than to note that, like the SF-36, the EQ-5D is a generic measure of HRQL, the EQ-5D is not described here.

Matching WOMAC and SF-36 dimensions with EQ-5D dimensions

To determine comparable constructs, the content of the WOMAC and SF-36 items within each subscale was compared with the content of the five EQ-5D dimensions. Each WOMAC and SF-36 subscale was matched to the EQ-5D dimension which measured a conceptually similar construct. Tables 3-5 show the WOMAC and SF-36 items in columns under their respective subscales matched with the EQ-5D dimension with which they fit most closely. Table 3 includes items from the three WOMAC subscales. Table 4 includes items from the four physical health subscales and Table 5 includes items from the four mental health subscales. Although the EQ-5D, the WOMAC, and the SF-36 all measure aspects of HRQL, dimensions were not directly comparable.

Table 3

EQ-5D and WOMAC Items

EQ-5D	WOMAC (Function)
Mobility	What degree of difficulty do you have with - Descending Stairs Ascending Stairs Rising from sitting Standing Bending to floor Walking on flat Getting in/out of car ^b Going shopping ^b Putting on socks/stockings ^a Rising from bed ^a Taking off socks/stockings ^a Lying in bed Getting in/out of bath ^a Sitting Getting on/off toilet ^a Heavy domestic duties ^b Light domestic duties ^b
Self-Care	
Usual Activities	
Pain/discomfort	WOMAC (Pain) How much pain do you have-- Walking on a flat surface? Going up or down stairs? At night while in bed? Sitting or lying? Standing upright? WOMAC (Joint Stiffness) How severe is your stiffness after wakening in the morning? How severe is your stiffness after sitting, lying, or resting later in the day?
Depression/Anxiety	

^a Also fits under self-care

^b Also fits under usual activities

Table 4

EQ-5D and SF-36 Physical Health Subscales

EQ-5D	Physical Functioning	Role Physical	Bodily Pain	General Health
Mobility	<p><i>The following items are about activities you might do during a typical day. Does your health now limit you in these activities? If so, how much?</i></p> <p>Vigorous activities, e.g. running, lifting heavy objects, participating in strenuous sports.</p> <p>Moderate activities, e.g. moving a table, pushing a vacuum cleaner, bowling or playing golf.^a</p> <p>Climbing several flights of stairs.</p> <p>Climbing one flight of stairs.</p> <p>Bending, kneeling or stooping.</p> <p>Walking more than a mile.</p> <p>Walking several blocks.</p>			
Self-Care Usual Activities	Bathing or dressing yourself.	<p><i>During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities as a result of your physical health?</i></p> <p>Cut down the amount of time you spent on work or other activities.</p> <p>Accomplished less than you would like.</p> <p>Were limited in the kind of work or other activities.</p> <p>Had difficulty performing the work or other activities (e.g., it took extra effort).</p>		
Pain/Discomfort			<p>How much bodily pain did you have during the past 4 weeks?</p> <p>Extent pain interfered with normal work.^a</p>	
Other				<p>In general would you say your health is:</p> <p>How true or false is each of the following statements for you?</p> <p>I seem to get sick a little easier than other people.</p> <p>I am as healthy as anybody I know.</p>

^a Also fits under usual activities

Table 5

EQ-5D and SF-36 Mental Health Subscales

EQ-5D	SF Vitality	SF Social Functioning	SF Role Emotional	SF Mental Health
Mobility Self-Care Usual Activities		During the past 4 weeks, to what extent has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbors, or groups? During the past 4 weeks, how much of the time has your physical health or emotional problems interfered with your social activities (like visiting with friends, relatives, etc.)?		
Pain/Discomfort Anxiety/Depression	How much of the time during the past 4 weeks- did you feel full of pep? -Did you have a lot of energy? -Did you feel worn out? -Did you feel tired?		During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities as a result of any emotional problems (such as feeling depressed or anxious?) Cut down the amount of time you spent on work or other activities. ^a Accomplished less than you would like? ^a Didn't do work or other activities as carefully as usual. ^a	How much of the time during the past 4 weeks- have you been a very nervous person? Have you felt so down in the dumps that nothing could cheer you up? Have you felt calm and peaceful? Have you felt downhearted and blue? Have you been a happy person?

^aAlso fits under usual activities

For example, the physical function dimension of the WOMAC consists of 17 items measuring different aspects of function that are conceptualized as three separate dimensions in the EQ-5D: mobility, self-care, and usual activities. Items within subscales which also fit under a second or third EQ-5D dimension were noted. For example, in Table 4 under SF-36 physical functioning, the item measuring moderate activities also fits under EQ-5D usual activities.

Data Analysis

Descriptive Statistics

Descriptive statistics, including measures of central tendency and dispersion, were calculated to describe and assess the distributions of scores (EQ-5D VAS, EQ-5D index scores, the WOMAC, and the SF-36) in this sample. Responsiveness of evaluative instruments may be affected by ceiling effects, where respondents with the best score (EQ-5D level 1) may have some impairment. Consequently, it would not be possible to fully detect improvement by the HRQL measure. To assess the degree of ceiling effect of the EQ-5D, the distributions of EQ-5D profile scores pre-surgery were compared with those of the WOMAC and the SF-36 scores. For subscales measuring more than one construct, analysis of specific items was used to compare distributions of responses to comparable items. For example, the distribution of responses for EQ-5D self-care (washing or dressing) was compared with the distribution for self-care items within the WOMAC functioning subscale and the SF-36 physical functioning subscale.

Comparison of EQ-5D Index scores Using Different Weights

To assess the validity of the preference scoring model, the score distributions and correlations between the EQ-5D VAS and the EQ-5D index scores, using VAS 1 year weights, VAS 10 year weights and TTO YORK weights (with and without the N3 term) (Dolan, 1997) were compared using pre-surgery data. The weights are reported in Table 6.

Table 6

Coefficients for EQ-5D Index Scores

Dimension		TTO ^a	VAS 1 year ^b	VAS 10 year ^c
Mobility	Level 2	0.069	0.052	0.071
	Level 3	0.314	0.151	0.182
Self-Care	Level 2	0.104	0.073	0.093
	Level 3	0.214	0.138	0.145
Usual Activity	Level 2	0.036	0.045	0.031
	Level 3	0.094	0.095	0.081
Pain/discomfort	Level 2	0.123	0.096	0.084
	Level 3	0.386	0.187	0.171
Anxiety/Depression	Level 2	0.071	0.063	0.063
	Level 3	0.236	0.140	0.124
N3		0.269	0.183	0.215
Constant		0.081	0.113	0.155

^a Dolan, 1997; ^b Dolan, 1996; ^c Dolan, personal communication, 1999.

VAS 1 year and 10 year weights were both derived from members of the general public in the U. K. using the visual analogue scale method of valuing health states. VAS 1 year weights were derived from interviews with 234

subjects who were asked to value health states lasting a duration of 1 year and followed by a health state which was unknown and should not be taken into account (Dolan, 1996). VAS 10 year weights were derived from interviews with 3288 respondents where respondents were asked to value health states that would last for a duration of 10 years followed by death (Gudex et al, 1996). These weights were not published and were obtained from Dolan (personal communication, May 5, 1999). Because the issue of which weights to use has not been resolved, this analysis was exploratory. High positive correlations would support the validity of societally derived valuations as reflective of the HRQL of patients experiencing the health states.

Patient derived weights were calculated using both pre- and post-surgery data by regressing self-reported VAS scores on dummy variables in a similar manner in which regression weights were calculated for the EQ-5D regression model (see Chapter II). Two dummy variables were created for each EQ-5D dimension and an N3 term was used (value=1 if any of the dimensions were at level 3). For example, for anxiety/depression, AD2 and AD3 were the dummy variables. For a respondent with a score of 3 on anxiety/depression, AD2=0 and AD3=1. The N3 term would equal 1. Differences in patient derived weights and TTO weights were assessed with t-tests.

Factor Analysis

To assess the factor structure of the EQ-5D, a combination of SF-36 and EQ-5D items were used. Confirmatory factor analysis (CFA) using LISREL (Joreskog & Sorbom, 1996) was used to test two models based on the

theoretical conception of HRQL as either unidimensional (Model I, a general health factor) or two dimensional (Model II, a physical and mental health factor). Although the EQ-5D health state descriptive system is assumed to be multidimensional (five dimensions), the conversion of the health state to an index score is based on a unidimensional conception of health. In contrast, the SF-36 is conceptualized as consisting of two underlying health dimensions, a physical and mental health dimension (McHorney, Ware, & Raczek, 1993). The two factor model involved the assignment of EQ-5D self-care, mobility, usual activities, and pain/discomfort and the SF-36 physical health items to the physical health factor (see Table 4) and the EQ-5D depression/anxiety and the SF-36 mental health items to a mental health factor (see Table 5).

A number of fit indices are available to assess the fit between the hypothesized model and the sample data. Based on a review by Gierl and Rogers (1996), and a discussion of fit indices by Byrne (1989, pp. 54-57), three fit indexes were used to assess each model tested by CFA: the chi-square statistic, the root mean square residual (RMR), and the root mean square error of approximation (RMSEA).

Multitrait-Multimethod

Discriminant and convergent validity evidence was assessed by the multitrait-multimethod (MTMM) approach using both pre- and post-surgery data (Campbell & Fiske, 1959). Based on the fit of WOMAC and SF-36 subscales with the EQ-5D items (Tables 3-5), and the results of correlational analyses using the EQ-5D in previous studies (Chetter et al., 1997; Essink-Bot et al., 1995),

several hypotheses were made for the correlations between EQ-5D dimensions and WOMAC and SF-36 subscales. Correlations $> .50$ were expected between components measuring similar constructs: EQ-5D mobility with WOMAC functioning and SF-36 physical functioning; EQ-5D usual activities with WOMAC functioning and SF-36 social functioning, role physical, and physical functioning; EQ-5D pain/discomfort with WOMAC pain and SF-36 bodily pain; and EQ-5D anxiety/depression with SF-36 mental health and role emotional. Correlations between $.40$ and $.50$ were expected between similar constructs that fit less well, or had a small number of items within the subscale that corresponded to the EQ-5D dimension: EQ-5D self-care with WOMAC functioning and SF-36 physical functioning; EQ-5D usual activities with SF-36 bodily pain, EQ-5D pain/discomfort with WOMAC stiffness, and EQ-5D anxiety/depression with SF-36 vitality (as energy level and fatigue are symptoms of both depression and physical health problems). Finally, correlations between WOMAC and SF-36 items, while not part of the validity inquiry of the EQ-5D, are part of a multitrait-multimethod analysis and significant positive values were hypothesized as follows: WOMAC Function with SF-36 physical functioning, and SF-36 role physical; and WOMAC pain with SF-36 bodily pain.

Responsiveness

Relevant clinical changes were expected in each dimension of the EQ-5D; past research has shown that improvements in pain relief, mobility, social interactions, and psychological well being have been observed in patients undergoing TJA, with most of the improvement occurring by three months

(Laupacis et al., 1993). Therefore, it was expected that there would be a significant improvement in HRQL, measured by the EQ-5D VAS, the EQ-5D profile scores, and the EQ-5D index scores from pre- to post-surgery. There is evidence that age may be related to physical functioning and degree of improvement following TJA (Jacobsson, Rehnberg, & Djerf, 1991) and that hip replacement surgery patients would show greater improvement than knee surgery patients (Rissanen et al., 1997).

Effect size. Pre – post TJA surgery effect sizes were compared among the three instruments. The effect size was calculated by dividing the difference in pre-surgery and post-surgery mean scores by the standard deviation of the pre-surgery score. Effect sizes were interpreted as small (.2), medium (.5), and large (.8) using Cohen's (1988) conventions.

Repeated measures. A 2 X 2 (joint-by-time) ANCOVA with repeated measures on the second factor and age as the covariate was used to test the hypothesis of an improvement in EQ-5D index and VAS scores from pre- to post-surgery. The Wilcoxon signed rank test was used to test the responsiveness of EQ-5D dimensions. The level of significance used was .05.

The standard error of measurement. The standard error of measurement (SEM) is the standard deviation of errors of measurement associated with scores from a particular group of respondents. A 95% confidence band (2SEM) was used around individual scores to provide 'reasonable limits' for estimating the true score (Gulliksen, 1950). The SEM is a function of the both the reliability of the test score and the standard deviation of scores. Using pre-surgery data,

Cronbach's alpha was calculated as measure of reliability for the SF-36 and WOMAC subscales. The SEM was calculated by taking the square root of one minus the reliability and multiplying the result by the standard deviation of each subscale pre-surgery. Change scores were calculated for each individual by subtracting the pre- subscale score from the post- subscale score. Individual change scores for WOMAC and SF-36 subscales were categorized into three categories: within 2 SEM of 0 (no change in HRQL), >2SEM (better HRQL) and <2SEM (worse HRQL). Given that the EQ-5D items have only three broad levels, each level differing qualitatively, a change of one level was used to represent a substantive change in that item. Crosstabs were then used to compare the percentage of individuals who changed in the EQ-5D items and SF-36 subscales compared to those who changed on the WOMAC subscales. Because the WOMAC was designed specifically for osteoarthritis patients, it was used as the 'gold standard' by which to assess responsiveness. For mental health, the SF-36 was used as the 'gold standard'. The percent of individuals who improved on the EQ-5D items was also compared with the percent who improved in comparable WOMAC and SF-36 items.

Quality-Adjusted Life-Years (QALYs)

EQ-5D index scores were designed primarily for use in the calculation of QALYs in cost-utility analysis. Therefore, QALYs were calculated by multiplying life expectancy (Statistics Canada, 1990-1992), or the average number of years of life remaining for persons who have attained a given age, by the EQ-5D index score for each respondent. A mean change score, or QALY gained, was

calculated by subtracting the pre- from the post-surgery QALYs. This score was used to provide an estimate of the difference in change in HRQL with surgical versus non-surgical treatment over the remaining lifetime of the individual. To assess the potential consequences of using different weighting systems, QALYs gained were compared using different weights (TTO, VAS 10 year, VAS 1 year, sum, and patient derived) and TTO weights using a model with no N3 term.

CHAPTER V

Results

Descriptive Statistics

Description of the Sample

Of the 540 subjects scheduled for one joint replacement, 27 withdrew from the study, 38 had their surgery cancelled, and 3 patients died. Because larger sample sizes were used where possible to increase the stability of results, sample sizes varied depending on the analysis. Analyses assessing change were based on the subjects with HRQL data for all three measures both pre- and post- surgery (n = 436). Missing data for individual SF-36 and WOMAC items were imputed with mean subscale scores according to the scoring manuals for the SF-36 (Ware et al., 1994) and WOMAC (Bellamy, 1995).

Demographic data, admission diagnosis, and type of surgery are presented in Table 7. The sample of 540 respondents consisted of 58.9% females (n = 318) and 41.1% males (n = 222) ranging in age from 26 to 89 years of age (M = 67.8; SD = 10.8). On admission 85.4% of the respondents had a diagnosis of osteoarthritis and 48.7% of respondents used aids to walking: 38.8% used one cane or crutch; 4.2% used 2 canes or crutches and 5.7% used a walker.

Table 7
Demographic Characteristics of Sample

	Percent
Gender	
Female	58.9
Male	41.1
Age	
Over 75 years	25.7
50 – 75 years	67.4
Less than 50	6.9
Diagnosis	
Osteoarthritis	85.4
Rheumatoid Arthritis	5.2
Other/unknown	9.4
Joint Replacement	
Hip	45.6
Knee	54.4
Used Aids to Walk	48.7
One cane or crutch	38.8
Two canes or crutches	4.2
Walker	5.7
Residence	
House or Apartment	93.9
Seniors' Complex	5.7
Nursing Home	0.4
Education	
High School or Better	58.5

Pre-surgery descriptive statistics for the HRQL measures are presented in Table 8. The pre-surgery mean EQ-5D index score was 0.38 and mean EQ-5D VAS score was 54.24.

Table 8

Descriptive Statistics Pre-surgery

	N	Mean	SD	Min	Max
EQ-5D					
Index	536	0.38	0.31	-0.48	1.00
VAS	536	54.24	20.25	0.00	100.00
WOMAC					
Pain	539	43.35	17.21	0.00	100.00
Stiffness	539	39.73	21.04	0.00	100.00
Function	537	40.88	16.65	0.00	95.59
SF-36 ^a					
PF	540	20.39	17.64	0.00	95.00
RP	540	11.05	23.87	0.00	100.00
BP	540	29.05	17.38	0.00	100.00
GH	540	60.93	20.66	0.00	100.00
VT	540	41.01	20.72	0.00	90.00
SF	539	51.90	28.58	0.00	100.00
RE	533	53.35	44.31	0.00	100.00
MH	539	67.69	19.93	4.00	100.00
PCS	533	25.56	7.06	7.17	52.79
MCS	533	49.46	11.80	17.55	75.47

^a SF-36 Subscales and Summary Scores: PF physical functioning; RP role physical; BP bodily pain; GH general health; VT vitality; SF social functioning; RE role emotional; MH mental health; PCS physical component summary; MCS mental component summary.

Mean WOMAC and SF-36 subscales measuring pain and physical function and pain were low, particularly SF-36 role physical (RP; 11.05). In contrast, mean SF-36 mental health subscale scores (MH; 67.69) reflected a moderately high level of mental health. Mean SF-36 PCS scores were low (25.56), while mean SF-36 MCS scores (49.46) were close to mean U. S. population values (Ware et al., 1994).

Reliability coefficients and SEM's for the WOMAC and SF-36 subscales are shown in Table 9.

Table 9

Reliability and Standard Errors of Measurement (SEM) for WOMAC and SF-36 Subscales

		Reliability		SEM	
		α	SD	SEM	2SEM
WOMAC	Pain	0.81	17.49	7.59	15.19
	Stiffness	0.76	21.33	10.52	21.03
SF-36 ^a	Function	0.93	17.24	4.45	8.90
	PF	0.84	17.63	6.98	13.97
	RP	0.78	23.87	11.26	22.51
	BP	0.75	17.38	8.64	17.27
	GH	0.71	20.66	11.19	22.39
	VT	0.76	20.73	10.13	20.26
	SF	0.77	28.60	13.58	27.15
	RE	0.87	44.32	15.74	31.47
	MH	0.80	19.93	8.92	17.85

^a SF-36 Subscales: PF physical functioning; RP role physical; BP bodily pain; GH general health; VT vitality; SF social functioning; RE role emotional; MH mental health.

Reliability estimates for the WOMAC are similar to those reported in a validation study of the WOMAC (Bellamy, 1995). The estimates of SEM for the SF-36 are similar to those published for the U. S. population (Ware et al., 1994).

Distribution of EQ-5D Responses Pre- and Post-surgery

The number of unique EQ-5D health states used by respondents was 47 pre-surgery and 52 post-surgery. Table 10 shows the health states used by 10 or more respondents pre- and post-surgery. Pre-surgery, 13 of these health states described 83.3% of the subjects, while post-surgery, 11 described 78.3% of the subjects. A ceiling effect in health state descriptions was not evident pre-surgery. Only 0.2% of respondents scored a 11111 pre-surgery compared with 12.3% post-surgery.

Distributions of EQ-5D Profile Scores pre- and post-surgery are presented in Figure 2. Distributions generally showed an improvement in function from pre- to post-surgery. It was noted that two EQ-5D dimensions showed a substantial percent of respondents scoring a 1 (no problem) pre-surgery: 51.8% for self-care and 46.3% for anxiety/depression. A majority (96.3%) of respondents reported a 2 on EQ-5D mobility pre-surgery (some problems in walking about). Only 0.4% scored a 3 on self-care and 0.9% scored a 3 on mobility pre-surgery. Ninety-one percent of the respondents reported a problem on usual activities and 99.3% on pain/discomfort, reflecting a substantial level of problem pre-surgery.

Table 10.

Health States Used by 10 or More Respondents Pre- and Post-surgery

Pre-surgery Health state description	n	Percent	Post-surgery Health state description	n	Percent
Common to Pre- and Post Surgery					
Pre-Surgery			Post-Surgery		
21221	85	15.9	21221	64	14.1
21222	64	11.9	21222	57	12.5
22222	55	10.3	22222	38	8.4
22221	50	9.3	22221	18	4.0
21121	22	4.1	21121	33	7.3
21122	10	1.9	21122	11	2.4
Frequently Used Pre-Surgery			Frequently Used Post-Surgery		
22232	45	8.4	11111	56	12.3
21232	27	5.0	11121	48	10.5
22332	25	4.7	11221	11	2.4
21231	21	3.9	11122	10	2.2
22331	16	3.0	11211	10	2.2
22231	15	2.8			
22321	11	2.1			

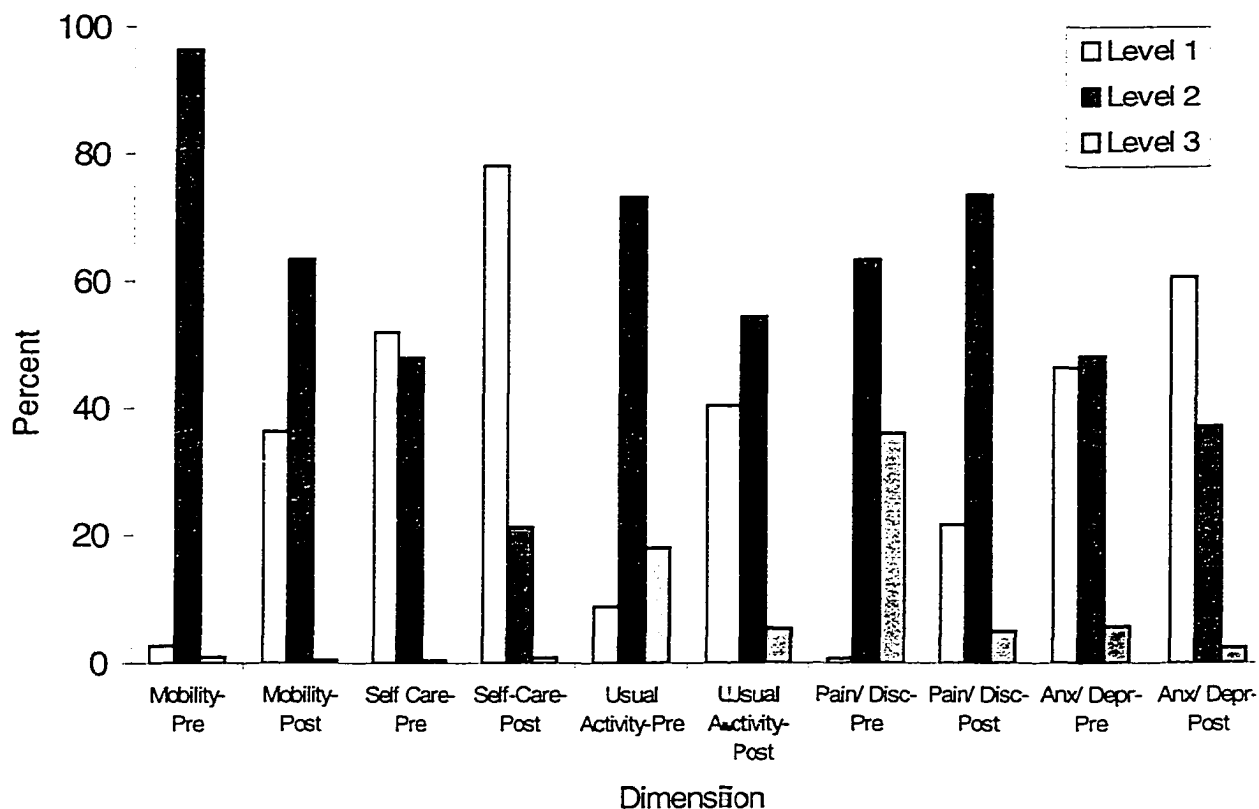


Figure 2. Distribution of responses on EQ-5D dimensions pre- and post-surgery.

Comparison of EQ-5D Distributions with SF-36 and WOMAC Subscales and Items

To assess possible ceiling effects and the adequacy of EQ-5D response levels, distribution of responses pre-surgery were compared between the EQ-5D dimensions and SF-36 and WOMAC subscales and items measuring similar constructs. Similar patterns of distributions in comparable items and subscales were expected between the three measures. As seen in Figures 3 and 4, although there was a wide range of SF-36 and WOMAC responses within each EQ-5D level, for most comparisons the median values for SF-36 and WOMAC

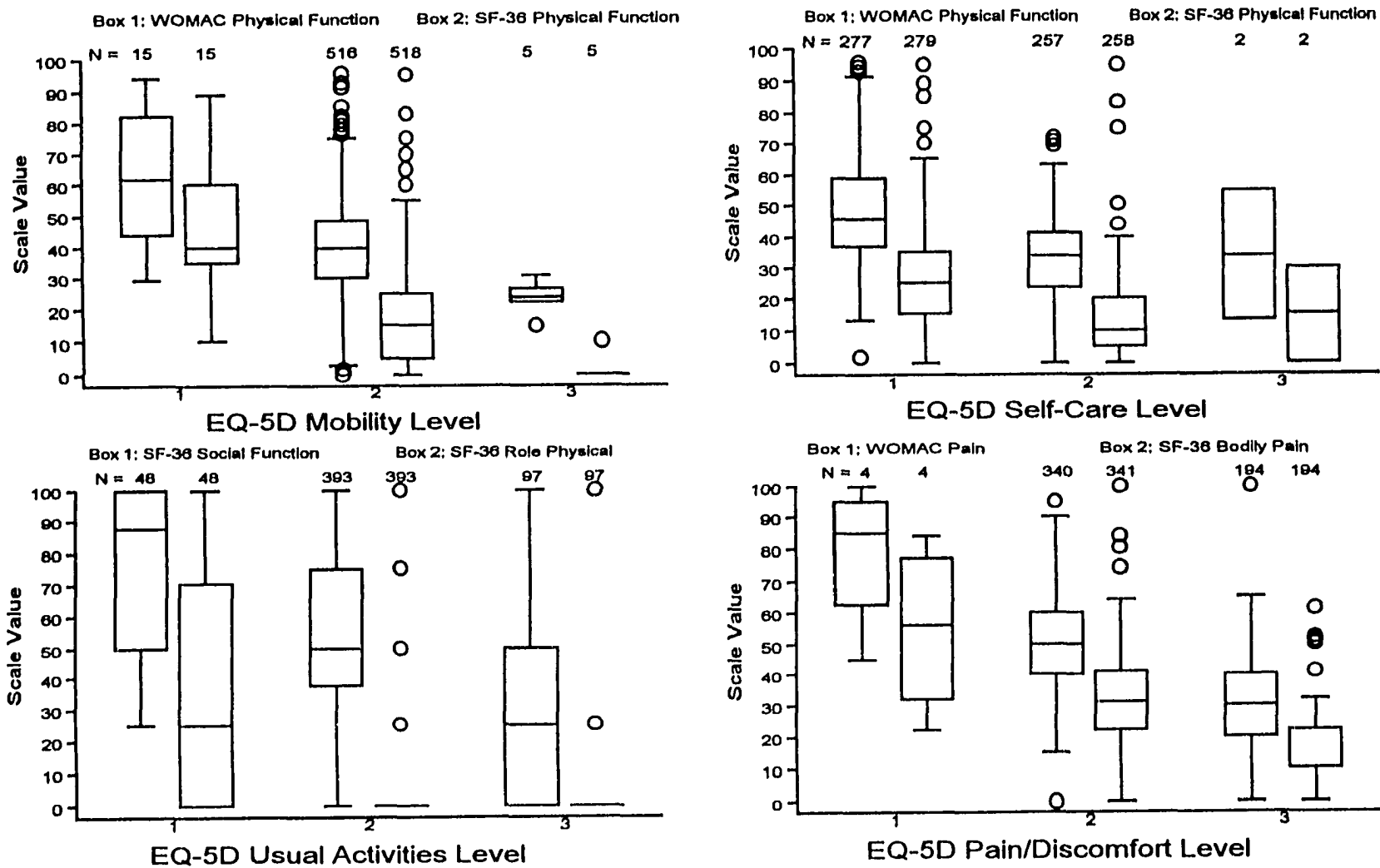


Figure 3. Boxplots of WOMAC and SF-36 subscales for each level of EQ-5D item for comparable constructs. Boxplots are read from left to right.

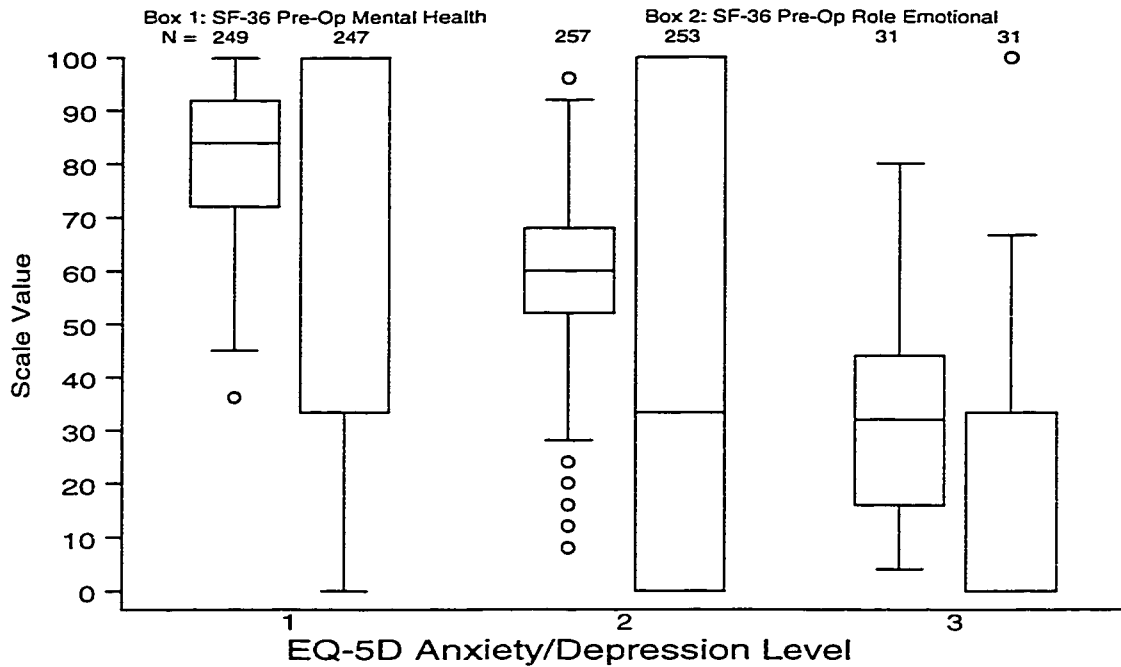


Figure 4. Boxplots of SF-36 mental health and role emotional for each level of EQ-5D anxiety/depression. Boxplots are read from left to right.

subscales decreased as the EQ-5D dimension scores for comparable dimensions increased (lower HRQL), as expected.

EQ-5D self-care and anxiety/depression were examined for the degree of ceiling effect. The distribution of responses for EQ-5D self-care was compared with the one 3-level SF-36 item measuring limitations in bathing or dressing and four 5-level WOMAC items measuring specific self-care activities. Of the 279 respondents scoring a 1 on EQ-5D self-care, 94.6% were 'not limited at all' or 'limited a little' on the SF-36 item. However, the responses of these 279 respondents on the WOMAC self-care items were much more varied: 75.8% had moderate to extreme difficulty in getting in and out of the bath, 65.8% had moderate to extreme difficulty in getting on or off the toilet, 67.7% had moderate

to extreme difficulty in taking off socks or stockings, and 70.9% had moderate to extreme difficulty in putting on socks or stocking (Figure 5).

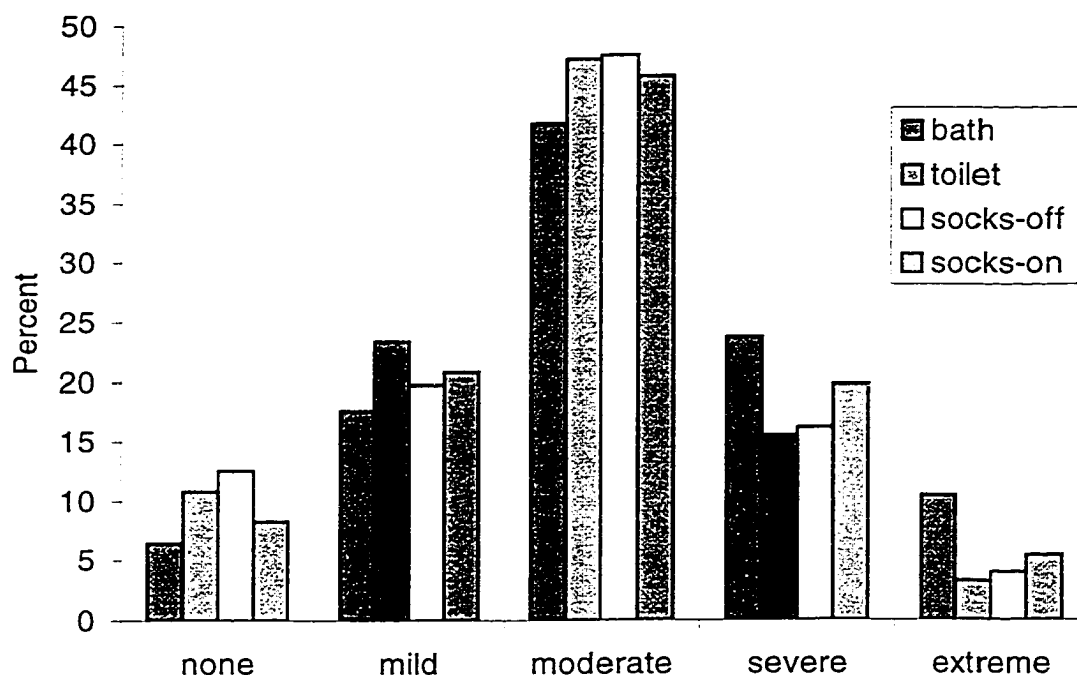


Figure 5. Distribution of responses for WOMAC self-care items when EQ-5D = 1 (no problems).

Of those respondents reporting severe to extreme difficulty with all of the above WOMAC items ($n = 76$), 22.4% reported a 1 (no problem) on EQ-5D self-care, 77.6% reported a 2, and 0% reported a 3 (unable). Comparing the scores of those same 76 respondents with the SF-36 item (washing or dressing), 7.9% were not limited at all, 46.1% limited a little, and 46.1% were limited a lot. Differences appear not only to be related to the number of levels, but also to the wording of levels. The wording of EQ-5D level 3 (unable to wash or dress

myself) may have prevented respondents from using that level, even with severe to extreme difficulty with aspects of self-care (WOMAC).

In contrast to EQ-5D self-care, a ceiling effect was not evident for anxiety/depression. The percentage of respondents (46.3) who scored a 1 on EQ-5D anxiety/depression was consistent with the high mean SF-36 mental health scores (Table 8). The distribution of responses for the EQ-5D anxiety/depression item followed a similar pattern to that of the SF-36 mental health and role emotional subscales (Figure 4), and appeared to reflect the degree of mental wellbeing in this population.

To assess whether the high percentage of responses on level 2 of EQ-5D mobility was congruent with the responses on the other HRQL measures, distributions of responses were compared with WOMAC items measuring walking, and ascending and descending stairs (Figure 6).

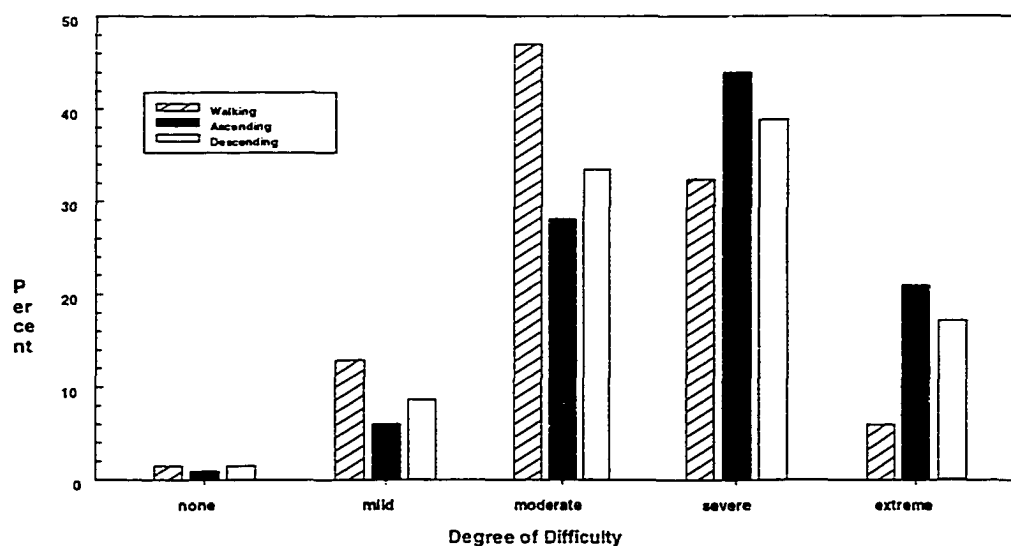


Figure 6. Distribution of WOMAC mobility items when EQ-5D mobility = 2 (some problems in walking about).

Of the respondents scoring a 2 on mobility, 38.5% had severe to extreme difficulty walking on a flat surface, 56.2% had severe to extreme difficulty descending stairs and 64.9% had severe to extreme difficulty ascending stairs (WOMAC). To further assess the ability of the EQ-5D 3-point mobility scale to adequately represent the underlying construct, the distributions of EQ-5D mobility were compared with the one SF-36 item (walking one block) most closely measuring the same construct as EQ-5D mobility (Figure 7).

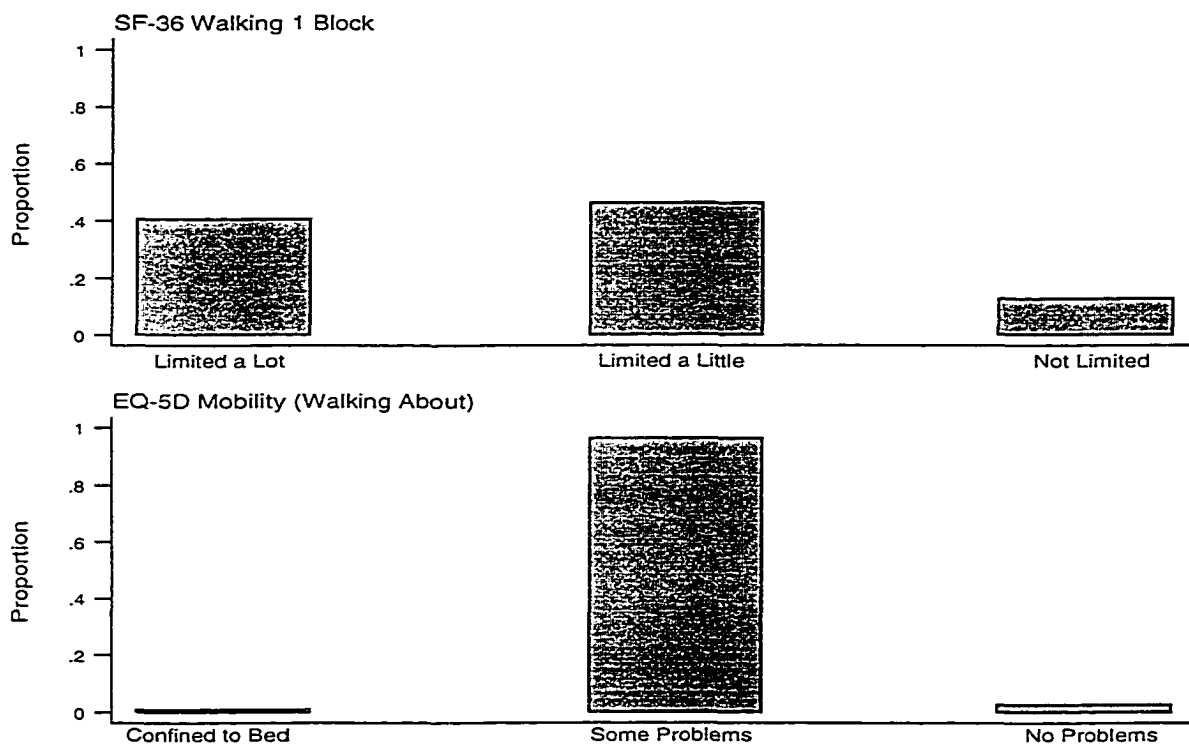


Figure 7. Comparison of the distribution of responses for EQ-5D mobility and SF-36 'walking 1 block'.

Respondents were more evenly distributed across all levels in the SF-36 item than with the EQ-5D mobility item. Although there was a wide range of dysfunction in mobility in these respondents, as in self-care, the extreme wording of level 3 (confined to bed) may have prevented respondents from using this level.

Ninety-one per cent of respondents reported a problem with usual activities (73.0% scored 2 and 18.0% scored a 3). Two SF-36 subscales directly measured either daily activities (role physical) or social activities (social functioning). As well, SF-36 physical functioning included one item measuring moderate activities, while WOMAC function included two comparable items (Tables 3-5). The high proportion of respondents reporting a problem with EQ-5D usual activities was consistent with the low mean (11.05) SF-36 role physical score seen in Table 8. This subscale included four items each coded 1 (yes) and 0 (no) for a problem on work or activities (Table 4). Of the 490 respondents reporting a problem (2 or 3) on the EQ-5D usual activities, 96.1% reported a problem on at least 2 out of 4 of these items. For the SF-36 social functioning subscale items, 68.6% of these respondents reported the extent to which problems interfered with social activities as moderate to extreme and 68.2% reported that problems interfered with social activities some of the time to all of the time. As well, 96.9% had moderate to extreme difficulty in shopping and 92.4% had moderate to extreme difficulty in getting in and out of a car (WOMAC). Because EQ-5D usual activities measured a variety of activities, interpretation of comparisons was difficult. The percentage of respondents reporting a problem

with various activities was consistent with the EQ-5D level for usual activities, reflecting the high level of dysfunction as measured by the SF-36 and the WOMAC (Figure 8). However, Figure 8 also shows that a substantial proportion of respondents who scored a 1 on usual activities reported problems with a variety of usual activities.

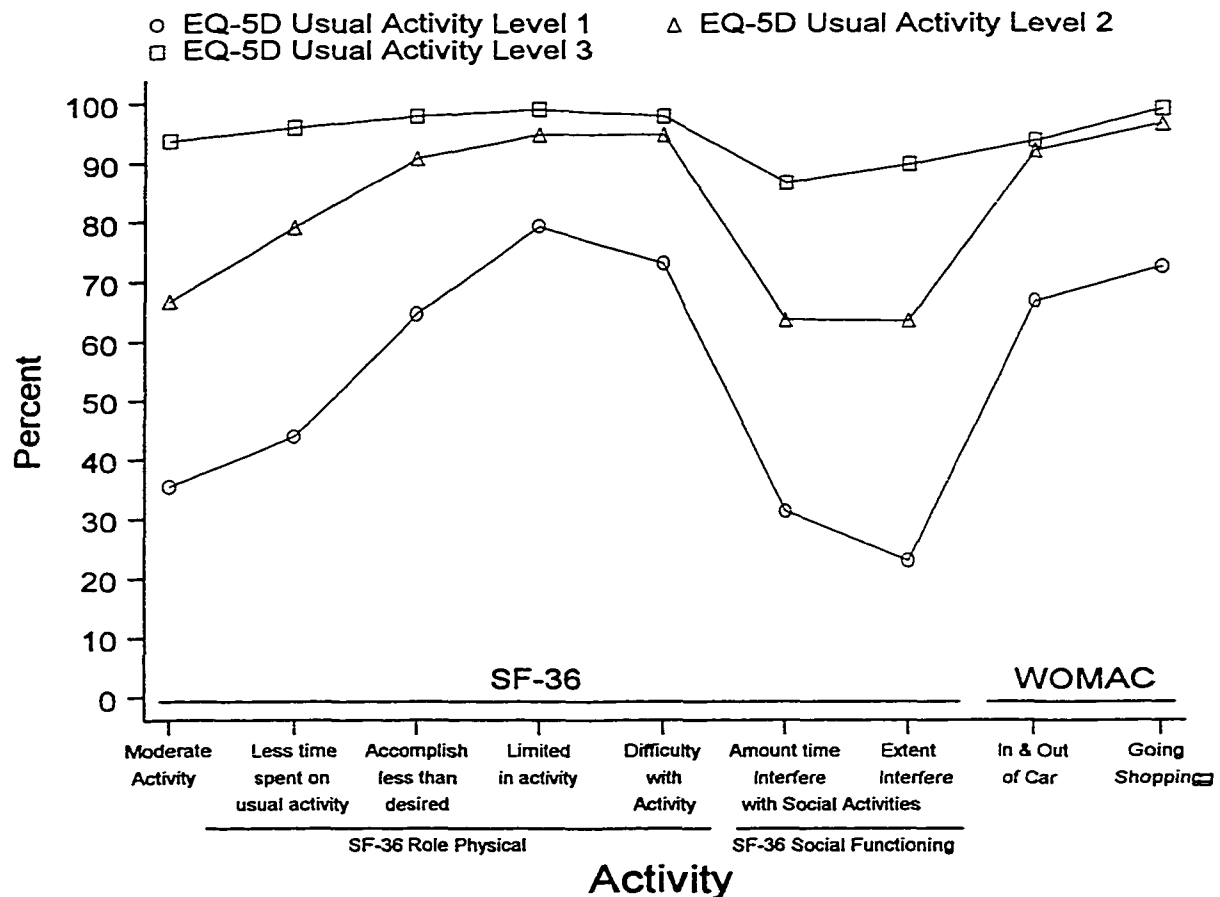


Figure 8. The percentage of respondents reporting a problem on SF-36 and WOMAC activities for three levels of EQ-5D usual activities.

The distribution of responses on EQ-5D pain, compared with those of the WOMAC pain subscale and SF-36 bodily pain subscale, followed a congruent

pattern with decreasing median pain subscale scores as EQ-5D levels moved from a 1 to a 3 (Figure 3).

EQ-5D VAS and Index Score Distributions

The distribution of the EQ-5D VAS scores was fairly symmetric both pre- and post-surgery with a slight negative skewness post-surgery. EQ-5D index scores ranged from -0.484 to 1.00 pre and post operatively (Figure 9).

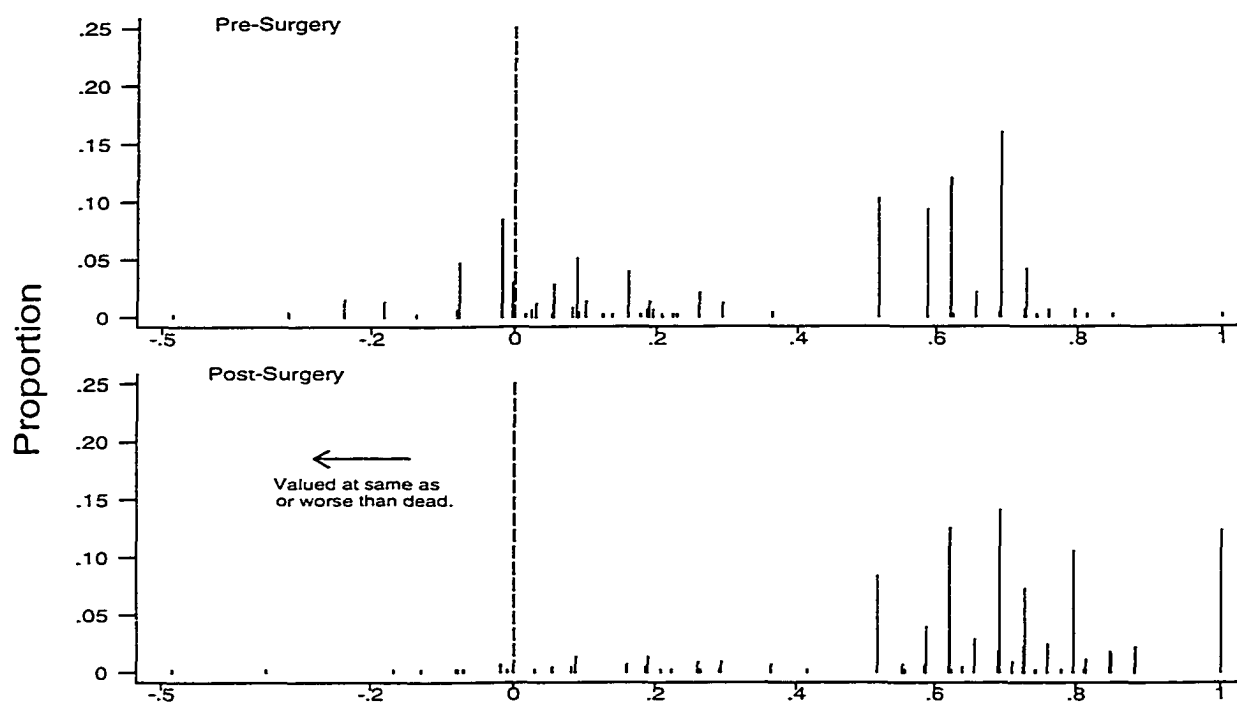


Figure 9. Distribution of EQ-5D index scores pre-and post-surgery.

The distribution pre-surgery was bimodal, similar to that found by Wolfe and Hawley (1997) in patients with rheumatic disorders. The distribution peaked at approximate values of 0 and 0.7 with no respondents scoring between 0.364 and 0.516. Post-surgery the distribution was negatively skewed and the gap remained with no scores between 0.416 and 0.516. Pre-surgery, 20.1% (n =

108) of respondents had index scores of 0 or < 0 , valued as 'the same as dead' or 'worse than dead', respectively, compared with 2.4% post-surgery.

Two questions arose from these findings. Do respondents with health states valued by society as 'dead' or 'worse than dead' rate their own health similarly? Does the bimodal distribution reflect the distribution of the underlying construct or is it an artifact of the scoring system?

Are EQ-5D Index Scores Reflective of Self-Reported HRQL of Respondents?

EQ-5D index scores were compared with two measures of self-rated health, the EQ-5D VAS and the SF-36 item measuring perception of health measured on a 5-point Likert scale from poor to excellent. Correlations of EQ-5D index scores with EQ-5D VAS scores were 0.45 (pre-surgery) and 0.60 (post-surgery) and with the SF-36 item, 0.32 and 0.48, respectively.

To determine whether community derived valuations were reflective of self-reported health, self-rated health from individuals with pre-surgery index scores 'valued' dead (0) or 'worse than dead' (< 0) was examined. All of the 108 respondents with a index score of 0 or < 0 had a level 3 on pain and a score of 2 or 3 on at least three other EQ-5D dimensions. Of the 108 respondents, 74.1% ($n = 80$) rated their health on the SF-36 question 1 as fair to excellent (Figure 10). EQ-5D VAS scores for these respondents ranged from 0 to 90 ($M = 40.67$, $SD = 19.85$).

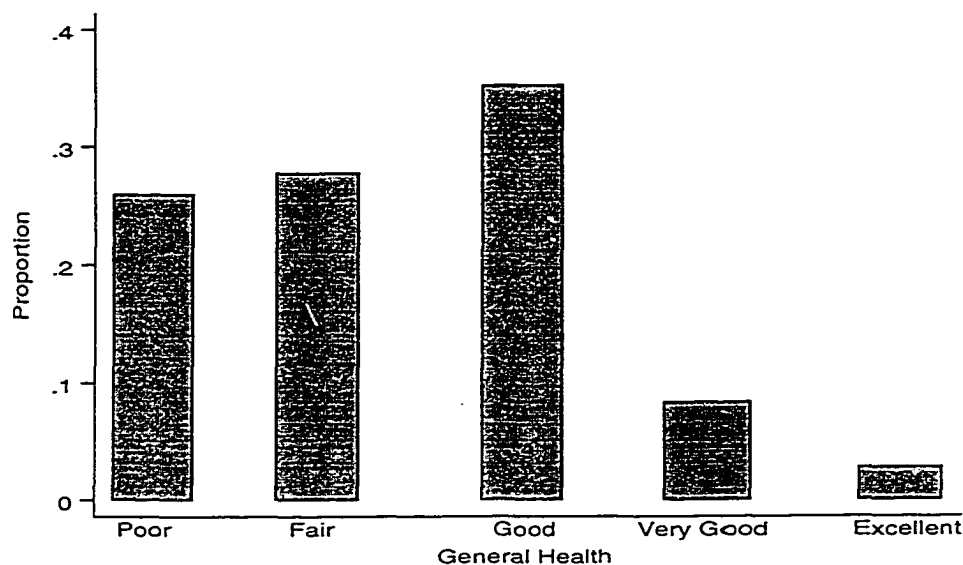


Figure 10. SF-36 self-rated health for respondents with index scores equal to or less than 0.

Relationship of Distribution to Scoring System

To understand the reasons for the bimodal distribution, two questions were explored: 1. Which health states have index scores between 0.416 and 0.516?; and 2. Are these health states clinically likely to occur? There are 5 theoretical health states with index values between .416 and .516: 13111, 11321, 12311, 11312, and 21311. None of these occurred either pre- or post-surgery. The health state 13111 would be implausible in most populations (unable to wash or dress oneself with no problems on any other dimension). The health state 11321 (no problems in mobility, self-care or anxiety/depression, unable to perform usual activities, and moderate pain/discomfort), although plausible, would be less likely to occur in patients with osteoarthritis because of the association of joint pain with movement. For example, of those patients with

a problem (2 or 3) on pain pre-surgery, 96.4% reported a problem on mobility. The remaining three health states (no pain/discomfort but unable to perform usual activities) would also be unlikely. All respondents in this study with a 3 (unable) on usual activities reported a problem on both pain and mobility.

As explained earlier, index scores are produced by the subtraction of disutility weights from 1 (representing perfect health). The N3 term comes into play when a respondent scores a 3 on at least one dimension. A level 3 on one or more dimensions represents an additional disutility of 0.269. The lowest possible index score without a 3 (i.e., 22222) is 0.516. The highest possible index score for a health state with a 3 on any one dimension is 0.556 for health state 11311 (unable to carry out usual activities with no problems on other dimensions). However, a health state with one three and 4 one's is unlikely. Because of the conceptual relationship between physical dimensions (mobility, self-care, usual activities and pain/discomfort), respondents are highly unlikely to be at a 3 on any dimension, without an accompanying problem on at least one other dimension. In this study, all respondents with a 3 on one dimension had a problem (2 or 3) on at least one other dimension, while 98.4% reported a problem on at least 2 other dimensions.

The presence of severe pain resulted in a particularly large number of index scores in the lower ranges. Pain/discomfort has the highest disutility weight and pain is often the cause of other functional impairments. This is particularly relevant in osteoarthritis where joint pain, exacerbated by movement, produces decreased functioning. Thirty-six per cent of respondents rated

themselves a 3 on pain/discomfort (extreme). Of these, 99.0% reported a problem with mobility and 97.9% experienced a problem (2 or 3) on at least 2 dimensions. Of the 99.3% of patients who reported a problem (2 or 3) with pain, 53.9% reported a problem with anxiety/depression. The maximum index score that a respondent with a level 3 on pain and no problems on any other dimension (11131) could have is .264 ($1 - 0.386 - 0.081 - 0.269$), but as noted, this is unlikely to occur. The highest index for those respondents who reported extreme pain/discomfort was 0.228 (11231). The lowest was -0.484 (32333). To further explore whether the bimodal distribution was an artifact of the scoring system, the distribution of index scores was examined using different weights and using a model without the N3 term.

Comparison of Index Scores Using Different Weights

Although the York TTO weights are the most frequently used to score the EQ-5D, there is no universally accepted set of weights. To examine the effect of different weights on the distribution of index scores, two additional sets of societal derived weights were used to calculate index scores (see Table 6): VAS weights using a 10 year duration (Dolan, personal communication, May 5, 1999), and VAS weights using a one year duration (Dolan, 1996).

As well, index scores were calculated using TTO weights with a model without an N3 term and also by using a simple summation of EQ-5D scores with a linear transformation to a 0 to 1 scale ($\text{Sum} - 5/10$). For example, for a health state of 22321, a sum of dimension scores would be 10. The index score (x)

would be $(10 - 5)/10 = .5$. For interpretability, scores were then reversed $(1 - x)$ so that 1 represented the best HRQL.

Finally, patient derived weights were calculated using both pre-surgery and post-surgery data (Table 11). Compared with the TTO weights, the patient derived weights tended to be lower except for usual activities (levels 2 and 3) and anxiety/depression (level 2). Using t-tests for independent samples to assess whether the patient-derived weights were significantly different from TTO weights, self-care, pain/discomfort, and anxiety/depression had significantly lower weights from TTO weights. As well, the N3 term was significantly lower for the patient derived weights.

Table 11

Comparison of York TTO Weights and Pre- and Post-surgery Patient Regression

Weights

EQ-5D	York TTO		Pre-surgery ^a		Post-surgery ^a	
	Level2	Level3	Level2	Level3	Level2	Level3
Mobility	0.069	0.314	0.049	0.252	0.045	0.106
Self-Care	0.104	0.214	0.035*	0.083	0.035*	0.075
Usual Activity	0.036	0.094	0.046	0.127	0.048	0.187
Pain/Discomfort	0.123	0.386	0.012	0.051*	0.081*	0.189*
Anxiety/Depression	0.071	0.236	0.072	0.129*	0.096	0.165
Constant	0.081		0.272*		0.139*	
N3	0.269		0.026*		0.015*	

^a Pre-surgery and post-surgery weights are patient-derived from pre- and post-surgery data.

*Significantly different ($p < .05$) from York TTO weights, using a t-test.

Figure 11 is a spike graph which shows the distribution of pre-surgery index scores using the different societal derived weights and the pre-surgery derived patient weights. As noted earlier, the distribution of index scores using the TTO weights was bimodal with a gap between the two distributions. The distributions using the 10 year and 1 year VAS weights were similar but compressed. The effect of removing the N3 term was to close the gap, but the bimodal distribution remained. Finally, the distributions using the sum and the pre-surgery patient derived weights were unimodal and more symmetric with less of a range.

Table 12 compares the measures of central tendency and variability of pre-surgery index scores using the different weights and a model without the N3 term. Lower and higher modal values are included for those distributions which were bimodal. Although means are not an appropriate measure of central tendency for bimodal distributions, they are commonly reported in the literature and are included for comparison purposes. All weights provided a maximum index score of 1, due to the presence of one case with a health state of 11111. Therefore, differences in range were due to the differences in minimum values produced by the various weights. The TTO weights (with and without the N3 term) resulted in the largest range and the lowest minimum values. The 10 year and 1 year VAS weights resulted in the next largest spread, with minimum values ranging from -.02 (10 year VAS weights) to .06 (1 year VAS weights). A simple transformed sum of patient scores and the patient derived weights resulted in minimum values ranging from .10 to .16.

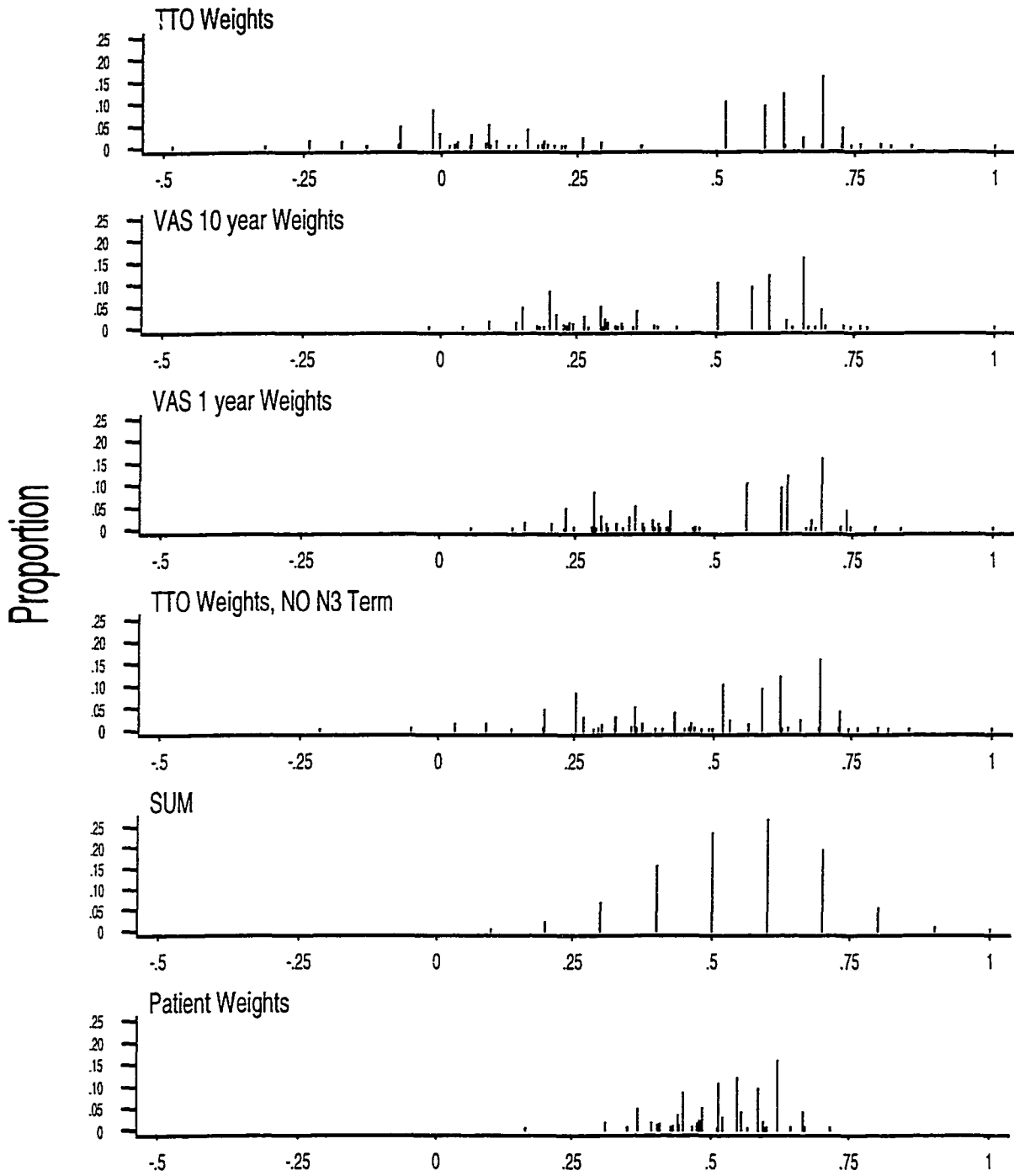


Figure 11. Distributions of EQ-5D index scores pre-surgery using different weights

Table 12

Pre-surgery Descriptive Statistics for Index Scores Using Different Weights

Weight	Lower Mode	Higher Mode	Single Mode	Mean	SD	Min	Max
TTO	-0.02	0.69	-	0.38	0.31	-0.48	1.00
VAS (10 year)	0.20	0.66	-	0.45	0.20	-0.02	1.00
VAS (1 year)	0.28	0.69	-	0.50	0.18	0.06	1.00
TTO (No N3)	0.25	0.69	-	0.49	0.19	-0.21	1.00
Sum	-	-	0.60	0.55	0.14	0.10	1.00
Pre Surgery ^a	-	-	0.62	0.53	0.09	0.11	1.00
Post Surgery ^a	-	-	0.69	0.55	0.13	0.16	1.00

Note. $n = 536$. Dashes indicate the mode was not calculated.

^a Pre- and post-surgery weights are patient-derived weights using pre- and post-surgery data.

Index scores using the TTO weights had the lowest mean (0.38) and highest standard deviation (.31). Means and standard deviations for the index scores using VAS weights and TTO weights with no N3 term were similar. Scores derived from the sum and patient derived weights had the highest mean values and lowest standard deviations.

Correlations between all index scores using the different weights were high (Table 13). Figure 12 shows a matrix scatterplot of the index scores derived from the following weights: TTO, VAS 10 year, VAS 1 year, TTO with no N3 term, sum, and the pre-surgery derived patients weights. The scatterplots are generally linear in nature. The scatterplots between the models with and without the N3 term shows the separation of points due to the N3 term.

Table 13

Pearson Correlation Coefficients for Pre-surgery Index Scores Using Different Weights

	TTO	VAS 10 year	VAS 1 year	No N3	Sum	Pre
TTO						
VAS 10 year	.99					
VAS 1 year	.99	.99				
TTO (No N3)	.97	.96	.97			
Sum	.87	.91	.92	.93		
Patient Pre ^a	.87	.91	.91	.90	.97	
Patient Post ^a	.92	.94	.95	.93	.96	.97

Note. $n = 536$.

^apatient-derived weights using pre- and post-surgery data

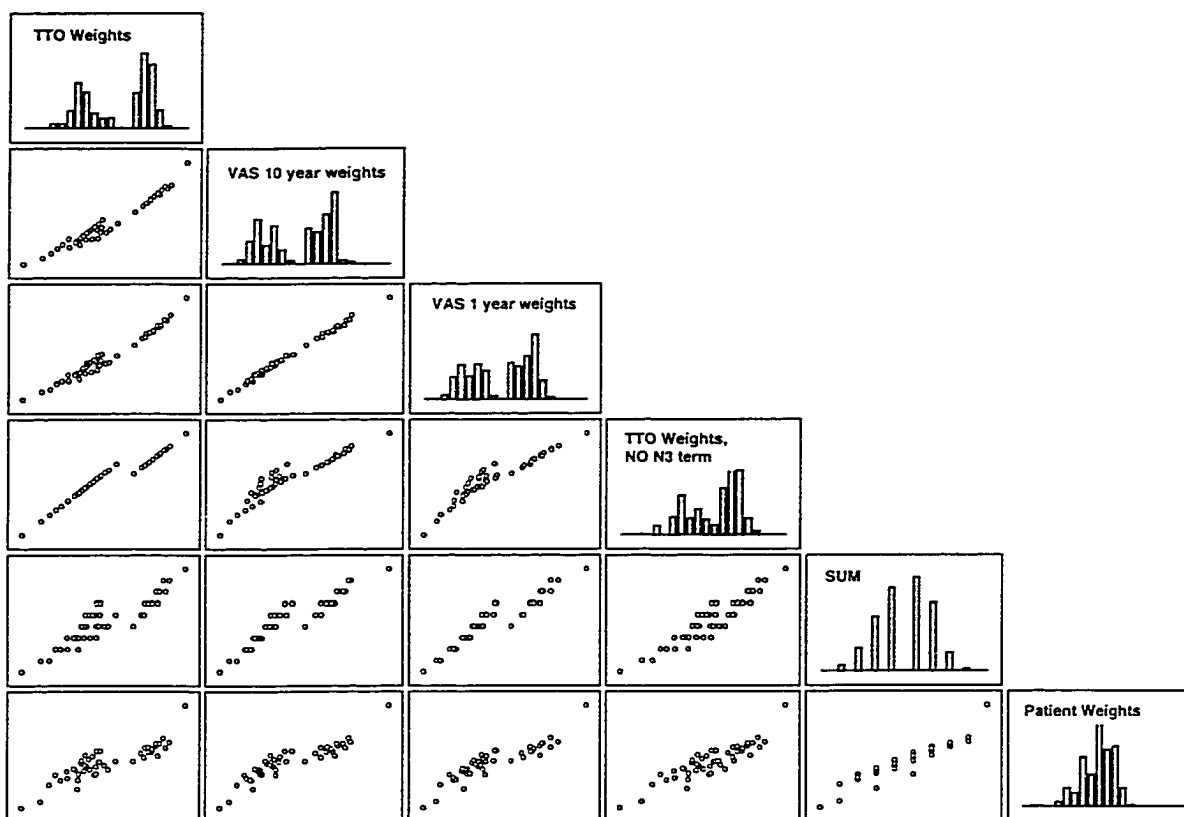


Figure 12. Matrix scatterplot of EQ-5D index scores using different weights.

Patient weights are derived from pre-surgery data. The scales for the horizontal and vertical axes differ.

Factor Structure

Tests of Overall Goodness of Fit

Two models were tested, a one factor model (Model I) and a two factor model (Model II) with correlated factors. Both models assumed uncorrelated error terms. Three fit indexes were used to assess each model, the chi-square statistic, the root mean square residual (RMR), and the root mean square error of approximation (RMSEA). The chi-square statistic was used as a test of the overall model fit, distributed with degrees of freedom equaling the difference

between the number of entries in the covariance matrix and the total number of coefficients estimated in the model. A non-significant chi-square ($>.05$) is desirable, with higher levels of significance indicating a better confirmation of model fit. Results (Table 14) indicated that the overall fit of both model I (χ^2 (740) = 5282.35) and model II (χ^2 (739) = 4359.46) was poor.

Table 14

Fit Indices for Five Models Using Confirmatory Factor Analysis

Model	χ^2	df	χ^2 ratio	RMR ^a	RMSEA ^b
Model I ^c	5282.35	740	7.14	0.09	0.10
Model II ^d	4359.46	739	5.90	0.10	0.09
Model III ^e	3907.63	559	6.99	0.12	0.10
Model IV ^f	226.36	61	3.71	0.06	0.07
Model V ^g	28.78	5	5.76	0.04	0.09

Note. All χ^2 values are statistically significant at the .05 level.

^aRMR = root mean square residual.

^bRMSEA = root mean square error of approximation.

^cModel I One factor using SF-36 and EQ-5D items.

^dModel II Two factor using SF-36 and EQ-5D items.

^eModel III Two factor using SF-36 items.

^fModel IV Two factor model using SF-36 subscales and EQ-5D items (Essink-Bot et al., 1997).

^g Model V One factor using EQ-5D items.

The Root Mean Square Residual (RMR) was 0.09 for Model I and 0.10 for Model II. Using .05 as a cutoff point for a good fit (Byrne, 1989), the results

indicate a poor fit. The RMSEA was 0.10 for Model I and 0.09 for Model II, indicated poor fitting models (using < 0.05 as a criteria for a good fitting model).

Assessing Goodness-of-Fit of Individual Model Parameters

t-values. The statistical significance of each parameter was examined by examining the 't-values', the coefficient estimates divided by their standard errors (Byrne, 1989). Non-significant parameters would be considered unimportant to the model. t-values for the parameters for Model I were all statistically significant ($t = > 1.96$) ranging from 6.12 to 17.94. t-values for parameters for Model II ranged from 1.67 to 19.72 and were all statistically significant ($t = > 1.96$) except the factor loading for SF-36 11c (I expect my health to get worse).

Standardized residuals. Ideally, standardized residuals should be within 2 standard deviations of zero for a good model fit. A positive residual indicates that the model underpredicts the covariance between two variables and a negative residual that it overpredicts the covariance. Standardized residuals for both models were all large ranging from absolute values of 2.59 to 18.17 for Model I and of 2.58 to 17.6 for Model II, indicating possible model misspecification (Bollen, 1989; Byrne, 1989).

The squared multiple correlation (R^2) for each observed (X) variable is an indication of the reliability of each variable in relation to its underlying factor structure (Byrne, 1989). R^2 's for Model I were low ranging from .02 to .47, an indication of poor model fit. R^2 's for Model II were low ranging from .07 to .55, an indication of poor model fit. R^2 's for the EQ-5D items ranged from 0.10 (mobility) to 0.49 (anxiety/depression).

Modification indices. Modification indices represent the expected drop in χ^2 if a particular parameter is set free and the model is reestimated. For Model I ten modification indices for theta-delta were > 100 , usually involving items measuring the same subscale. For Model II the largest two modification indices for the factor loadings were for SF-36 Q6 (101.58) and SF-36 Q11a (64.65). Question Q6 (During the past 4 weeks, to what extent has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbours, or groups) and Q 11a (I seem to get sick a little easier than other people) are both questions in which underlying mental and physical health factors are theoretically plausible. Modification indices for the EQ-5D items ranged from 0.02 to 19.29, with expected changes for the factor loadings ranging from 0.00 to 0.11, giving some support for the hypothesized factor structure for these items. To assess whether the poor fit was due to the addition of the EQ-5D items, a CFA was also done on only the SF-36 items (Model III). As shown in Table 14, the results ($\chi^2(559) = 3907.63$) showed an equally poor fit.

Confirmatory Factor Analysis using Essink-Bot et al.'s Factor Structure

Earlier it was noted that Essink-Bot et al. (1997) carried out two exploratory factor analyses using SF-36 subscale scores, EQ-5D items and one other HRQL measure. In each analyses, they found that three SF-36 subscales loaded on both the mental and physical factors: role physical, social functioning, and general health. This was in contrast to McHorney, Ware, and Raczek's findings (1993) on data from a U. S. population survey, who reported that vitality, social functioning, and general health all loaded on both factors. Vitality, social

functioning, and general health all include items which could theoretically be indicators of both physical and mental health (see Tables 4 and 5). However, the items in role physical all relate to problems with activities as a result of physical health. Although a theoretical basis for Essink-Bot et al.'s results, especially with regard to role physical, is not clear, it was decided to compare the structure that they found, using a sample of 496 migraine sufferers and a matched control group, with that in the present sample of patients with hip or knee replacement. Using their structure as the hypothesis, a confirmatory factor analysis was carried out (Model IV). The overall fit of the model ($\chi^2(61) = 226.36$) was poor (Table 14). Values for the RMSEA (.07) and RMR (.06), although lower than those in the previous models, also were indicative of a poor fit.

Factor Analyses using a Polychoric Correlation Matrix

An exploratory and confirmatory factor analysis (Model V) was done to assess the factor structure of the EQ-5D using only EQ-5D items. Because of the ordinal scaling and the skewed distribution of the EQ-5D items, a polychoric correlation matrix was used (Table 15). Polychoric correlations are an extension of tetrachoric correlations and provide an estimate of what the correlations between two ordered variables would be if the underlying bivariate distribution were normal (Bollen, 1989).

An exploratory factor analysis of the 5 EQ-5D items yielded 1 factor, based on the scree plot (Figure 13) and number of eigenvalues > 1.0 , explaining 50% of the common variance. Factor loadings ranged from .33 (anxiety/depression) to .85 (usual activities) (Table 16). Confirmatory factor

analysis resulted in an overall poor fit of the model ($\chi^2 (5) = 28.78$) (Table 14). While the RMR was within the cutoff value of .05, overall, the combination of results were not supportive of a good fit. A two factor model was not testable due to only one mental health item.

Table 15

Polychoric Correlation Coefficients for EQ-5D Dimensions

EQ-5D Dimensions	MO	SC	UA	PD
Mobility [MO]				
Self-Care [SC]	.3876			
Usual Activities [UA]	.5895	.5034		
Pain/Discomfort [PD]	.3249	.3544	.4786	
Anxiety/Depression [AD]	.2083	.2606	.2345	.3073

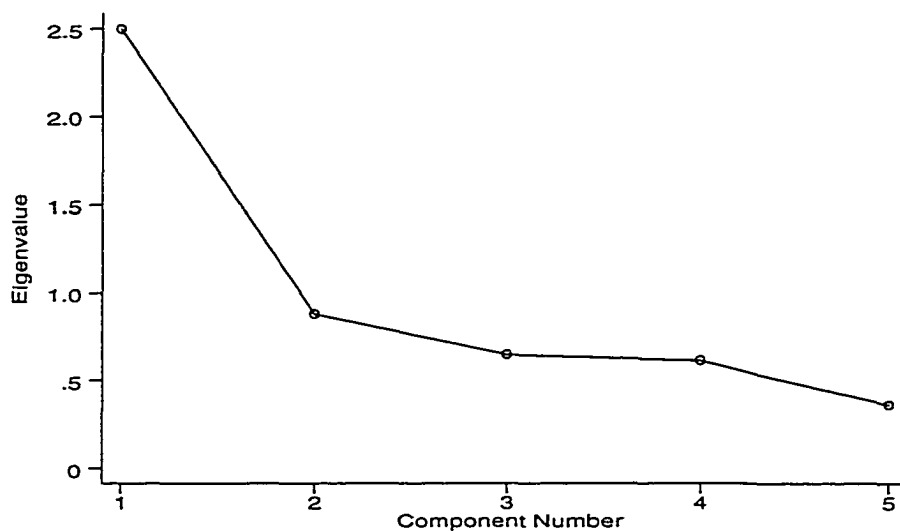


Figure 13. Scree plot based on EQ-5D items.

Table 16

Factor Loadings Using EQ-5D Items and Principal Axis Factoring

EQ-5D	Factor Loadings
Mobility	.67
Self-Care	.60
Usual Activities	.85
Pain/Discomfort	.56
Anxiety/Depression	.33

The Use of Factor Analysis in Determining the Factor Structure of the EQ-5D

Confirmatory factor analysis on the combined EQ-5D and SF-36 items for a one-factor and two-factor model resulted in a poor fit for both models. Items with large modification indices in the two factor model included items from the social functioning and general health subscales. These large modification indices were consistent with the findings of Essink-Bot et al. (1997) who found that the social functioning and general health subscales loaded on both factors. As well, McHorney, Ware, and Raczek (1993), using a principal components analysis with SF-36 subscales, found that the social functioning, general health and vitality subscales loaded on both the mental and physical health factors. Problems with interpreting factor analyses are increased by the multidimensional nature of some items such as the SF-36 item Q6, described above, which addresses both physical and emotional problems.

The use of the combined item factor analysis was based on the assumption that the factor structure of the SF-36 was a stable two factor structure. Because the factor analysis using only the SF-36 items was also a poor fit in the two dimensional model, it was concluded that the CFA using combined items was an inappropriate test of the factor structure of the EQ-5D. Finally, the model reported by Essink-Bot et al. (1997) was tested using CFA. Based on the purpose of the thesis, to assess the validity of interpretation of EQ-5D scores, no further models were tested with combined EQ-5D and SF-36 items.

The exploratory factor analysis using the five EQ-5D items yielded one factor, while the confirmatory factor analysis showed a poor model fit. However, interpretation is difficult, given the small number of items. It is questionable whether factor analysis is an appropriate test for the EQ-5D, given the small number of items, with each supposedly measuring one distinguishable dimension.

Convergent and Discriminant Validity Evidence

Based on the approach described by Campbell and Fiske (1959), the requirement for convergent validity evidence is that validity coefficients, or monotrait-heteromethod values (correlations between measures of the same construct using different measurement methods) should be significantly different from zero and sufficiently large to warrant consideration. A minimal value of $>.3$ was used as a cut-off in this study based on the commonly used value of $.3$ as a minimum factor loading (Gorsuch, 1983). Campbell and Fiske describe three

criteria for discriminant validity evidence: first, convergent validity coefficients should be higher than correlations between variables having neither trait nor method in common (i. e., a validity coefficient should be higher than the values lying in its column and row in the heterotrait-heteromethod triangle); second, convergent validity coefficients should be higher than correlations between measures of different constructs using the same method of measurement (heterotrait-monomethod values); and third, the same pattern of correlations between traits should be shown in both the heterotrait-monomethod and the heterotrait-heteromethod matrices.

Convergent Validity

All hypothesized correlations, as outlined in Chapter IV and highlighted in Tables 17 and 18, were in the expected direction. Unlike the SF-36 and WOMAC subscale scores, higher ED-5D scores represent poorer HRQL. Therefore, correlations will be discussed in absolute terms. While the correlations tended to be lower pre-surgery than post-surgery, the patterns were similar. Therefore, the results for pre-surgery will be discussed and differences between pre- and post-surgery discussed where relevant.

Of the 15 convergent validity coefficients between the EQ-5D items and similar constructs, 12 were $>.3$ and significantly correlated pre-surgery compared with all post-surgery. The three validity coefficients that were $.30$ or lower were between EQ-5D mobility and SF-36 physical functioning ($.23$), EQ-5D mobility and WOMAC functioning ($.20$), and EQ-5D usual activities and SF-36 role physical ($.30$). Corresponding correlations post-surgery were $.58$, $.59$, and $.64$,

Table 17

Convergent and Discriminant Correlation Coefficients^a for Dimensions Measured by Three Measures Pre-surgery

Measure	Dimension	EQ-5D					SF-36								WOMAC	
		MO	SC	UA	PD	AD	PF	RP	BP	GH	VT	SF	RE	MH	PA	ST
EQ-5D	Mobility [MO]															
	Self-care [SC]	.13														
	Usual Activities [UA]	.23	.34													
	Pain/Discomfort [PD]	.10	.24	.34												
	Anxiety/Depression [AD]	.06	.18	.15	.21											
SF-36	Physical Functioning [PF]	-.23	-.49	-.47	-.40	-.15										
	Role Physical [RP]	-.19	-.24	-.30	-.27	-.20	.43									
	Bodily Pain [BP]	-.15	-.33	-.46	-.56	-.24	.53	.38								
	General Health [GH]	-.10	-.24	-.17	-.21	-.34	.22	.18	.24							
	Vitality [VT]	-.16	-.33	-.36	-.37	-.40	.40	.31	.40	.43						
	Social Functioning [SF]	-.20	-.37	-.44	-.38	-.35	.52	.44	.54	.33	.49					
	Role Emotional [RE]	-.03	-.17	-.15	-.22	-.42	.22	.27	.24	.20	.27	.32				
	Mental Health [MH]	-.09	-.23	-.18	-.26	-.67	.24	.25	.30	.39	.47	.42	.50			
WOMAC	Pain [PA]	-.18	-.21	-.31	-.59	-.19	.40	.31	.59	.28	.32	.40	.18	.23		
	Stiffness [ST]	-.10	-.18	-.23	-.43	-.20	.30	.25	.46	.31	.29	.33	.22	.28	.58	
	Function [FU]	-.20	-.44	-.42	-.55	-.22	.61	.38	.64	.28	.38	.51	.28	.31	.75	.64

Note. n = 528; All correlations > .15 significant p < .01; Convergent Validity Coefficients are bold.

^a Spearman Correlation Coefficients

Table 18

Convergent and Discriminant Correlation Coefficients^a for Dimensions Measured by Three Measures Post-surgery

Measure	Dimension	EQ-5D					SF-36						WOMAC			
		MO	SC	UA	PD	AD	PF	RP	BP	GH	VT	SF	RE	MH	PA	ST
EQ-5D	Mobility [MO]															
	Self-care [SC]	.31														
	Usual Activities [UA]	.55	.37													
	Pain/Discomfort [PD]	.45	.19	.39												
	Anxiety/Depression [AD]	.29	.27	.34	.30											
SF-36	Physical Functioning [PF]	-.58	-.44	-.62	-.46	-.34										
	Role Physical [RP]	-.49	-.32	-.64	-.40	-.37	.64									
	Bodily Pain [BP]	-.49	-.32	-.54	-.62	-.39	.60	.58								
	General Health [GH]	-.31	-.30	-.39	-.32	-.46	.48	.51	.40							
	Vitality [VT]	-.45	-.32	-.53	-.44	-.52	.59	.61	.58	.58						
	Social Functioning [SF]	-.47	-.40	-.56	-.43	-.47	.64	.60	.60	.49	.60					
	Role Emotional [RE]	-.34	-.23	-.44	-.33	-.53	.43	.51	.42	.38	.46	.57				
Mental Health [MH]	-.28	-.25	-.37	-.32	-.72	.35	.45	.43	.54	.61	.57	.56				
WOMAC	Pain [PA]	-.44	-.23	-.42	-.50	-.35	.49	.43	.61	.36	.46	.42	.38	.38		
	Stiffness [ST]	-.42	-.15	-.33	-.42	-.26	.43	.40	.49	.28	.39	.36	.31	.31	.60	
	Function [FU]	-.59	-.39	-.58	-.51	-.33	.74	.60	.65	.43	.57	.56	.42	.41	.71	.65

Note. n = 449; All correlations > .15 significant p < .01; Convergent Validity Coefficients are bold.

^a Spearman Correlation Coefficients

respectively. The low correlations associated with mobility pre-surgery were most likely due to little variation in mobility (96.3% of respondents scored a 2 on EQ-5D mobility pre-surgery). Ten convergent coefficients were hypothesized to be greater than .50. Of these, three met the criteria pre-surgery, and nine post-surgery. Five coefficients were hypothesized to be between .40 and .50. All met the criteria pre-surgery and four were greater than .40 post-surgery.

To examine specific hypothesized correlations, correlations between EQ-5D anxiety/depression and the SF-36 mental health, role emotional, and vitality subscales ranged from 0.40 (vitality) to 0.67 (mental health) and were higher post-operatively (.52 to .72). Correlations between EQ-5D usual activities and three related SF-36 subscales and the WOMAC functioning subscale were not as high pre-surgery as hypothesized, ranging from .30 to .47, but were higher post-surgery, ranging from .56 to .64. EQ-5D pain/discomfort was correlated moderately with SF-36 bodily pain (0.56), WOMAC pain (0.59), and WOMAC stiffness (0.43). Moderate correlations were as hypothesized between EQ-5D self-care and SF-36 physical functioning (0.49) and WOMAC functioning (0.44). Correlations between related SF-36 and WOMAC subscales were moderate ranging from .38 to .74, using both pre- and post-surgery data.

Discriminant Validity

The first criterion for discriminant validity, that convergent validity coefficients be higher than correlations between different traits using different methods, was generally met. The few moderate to high correlations between variables having neither trait nor method in common were between constructs

that were related conceptually, for example, EQ-5D pain/discomfort with SF-36 physical functioning (.40) and WOMAC function (.55). As well, many correlations between related variables measuring different traits were moderate to high post-surgery, for example between EQ-5D usual activities and SF-36 vitality (.53).

The second criterion for discriminant validity was not met. Coefficients within the heterotrait-monomethod matrices were high between related constructs. For example, correlations between WOMAC subscales ranged from .58 to .75. Seventeen out of 28 correlations between SF-36 subscales were $> .30$ pre-surgery and all were $> .30$ post-surgery. Correlations between SF-36 subscales within the broader mental and physical health constructs were moderate. For example, correlations between the four SF-36 mental health subscales ranged from .27 to .50 pre-surgery and from .46 to .61 post-surgery. As well, correlations between subscales across mental and physical health were moderate. For example, correlations between SF-36 vitality and all other SF-36 subscales were $> .30$ (pre-surgery) and $> .40$ (post-surgery). Finally, the last criterion for discriminant validity, that the same pattern of correlations between traits should be shown in the heterotrait-monomethod and the heterotrait-heteromethod matrices, was difficult to assess because of the different variables measured by the three measures. However, correlations between three of the most comparable dimensions (pain, physical function, and mental health) were assessed for the EQ-5D and SF-36 (Table 19). Although correlations between EQ-5D items were lower than the other correlations, patterns were similar in the

three matrices with the highest correlations between pain and function and the lowest between mental health and physical function.

Table 19

Heterotrait-monomethod and Heterotrait-heteromethod Correlation Coefficients^a
for Three Comparable Dimensions

	EQ-5D			SF-36	
	MO	PD	AD	PF	BP
EQ-5D					
Mobility [MO]					
Pain/Discomfort [PD]	.45				
Anxiety/Depression [AD]	.29	.30			
SF-36					
Physical Functioning [PF]		-.46	-.34		
Bodily Pain [BP]	-.49		-.39	.60	
Mental Health [MH]	-.28	-.32		.35	.43

Note. Heterotrait-monomethod coefficients are bold.

^a Spearman correlation coefficients.

Responsiveness

Effect Size

Effect sizes for EQ-5D index scores were .76 (knee) and 1.16 (hip) compared with 1.14 to 1.88 (knee) and 1.83 to 2.68 (hip) for the WOMAC subscales (Table 20). EQ-5D VAS effect sizes were lower than for EQ-5D index

Table 20

Comparison of Effect Size by Type of Surgery

	Hip (n = 200)				Knee (n = 236)			
	Baseline Mean ^a	Baseline SD ^a	Change	ES	Baseline Mean ^a	Baseline SD ^a	Change	ES
EQ-5D								
Index	0.35	0.31	0.35	1.16	0.40	0.32	0.24	0.76
VAS	52.55	21.35	18.03	0.84	56.26	19.46	9.61	0.49
WOMAC								
Pain	43.40	15.84	42.53	2.68	43.20	17.92	33.62	1.88
Stiff	38.06	19.67	36.06	1.83	39.14	21.83	24.84	1.14
Function	38.80	15.41	38.89	2.52	42.32	17.67	28.87	1.63
SF-36 ^b								
PF	19.94	17.79	30.72	1.73	21.28	18.43	23.86	1.29
RP	8.75	21.08	37.88	1.80	12.08	25.63	25.00	.98
BP	26.01	14.91	37.67	2.53	30.72	18.22	22.67	1.24
GH	61.14	21.77	7.97	0.37	62.46	19.52	1.56	0.08
VT	39.78	18.98	18.53	0.98	41.79	21.34	11.12	0.52
SF	50.25	28.88	27.75	0.96	53.28	27.14	18.11	0.67
RE	50.00	45.04	26.67	0.59	56.78	43.77	10.59	0.24
MH	66.30	20.79	9.57	0.46	68.64	19.61	6.06	0.31
PCS	25.28	6.44	13.12	2.04	25.89	7.65	9.00	1.18
MCS	48.24	12.29	5.36	0.44	50.14	11.31	2.15	0.19

Note. n = 436.

^a Baseline is pre-surgery

^b PF physical functioning; RP role physical; BP bodily pain; GH general health; VT vitality; SF social functioning; RE role emotional; MH mental health; PCS physical component summary; MCS mental component summary.

scores, .49 for knee and .84 for hip. Effect sizes for the SF-36 subscales varied from .31 (mental health) to 1.29 (physical functioning) for knee surgery and from .46 (mental health) to 2.53 (bodily pain) for hip surgery. Effect sizes were larger for hip replacements than for knee replacements for all outcome measures.

Repeated Measures

Using a 2 X 2 (joint by time) ANCOVA with repeated measures on the second factor and age as a covariate, mean EQ-5D index scores showed a significant joint-by-time interaction with a greater improvement for hip replacement patients (Table 21 and Figure 14). EQ-5D VAS scores showed a similar pattern with a significant joint-by-time interaction. No significant age effect was found for either measure. All EQ-5D dimensions improved significantly from pre- to post-surgery using the Wilcoxon Signed Rank Test (Table 22).

Table 21

Repeated Measures Analysis of Covariance (Time by Joint) with Age as a Covariate

Source	EQ-5D Index		EQ-5D VAS	
	df	F	df	F
Between				
Age (A)	1	3.65	1	0.91
Joint (J)	1	0.39	1	0.15
Error	433		433	
Within				
Time (T)	1	12.94*	1	10.23*
T * A	1	0.44	1	1.14
T * J	1	12.59*	1	17.30*
Error	433		433	

* $p < .05$

Table 22

Wilcoxon Signed Ranks Test for EQ-5D Dimensions

EQ-5D Dimension	Z-Score
Mobility	-11.93*
Self-Care	-8.74*
Usual Activities	-11.97*
Pain/discomfort	-13.29*
Anxiety/depression	-6.03*

$p < .05$

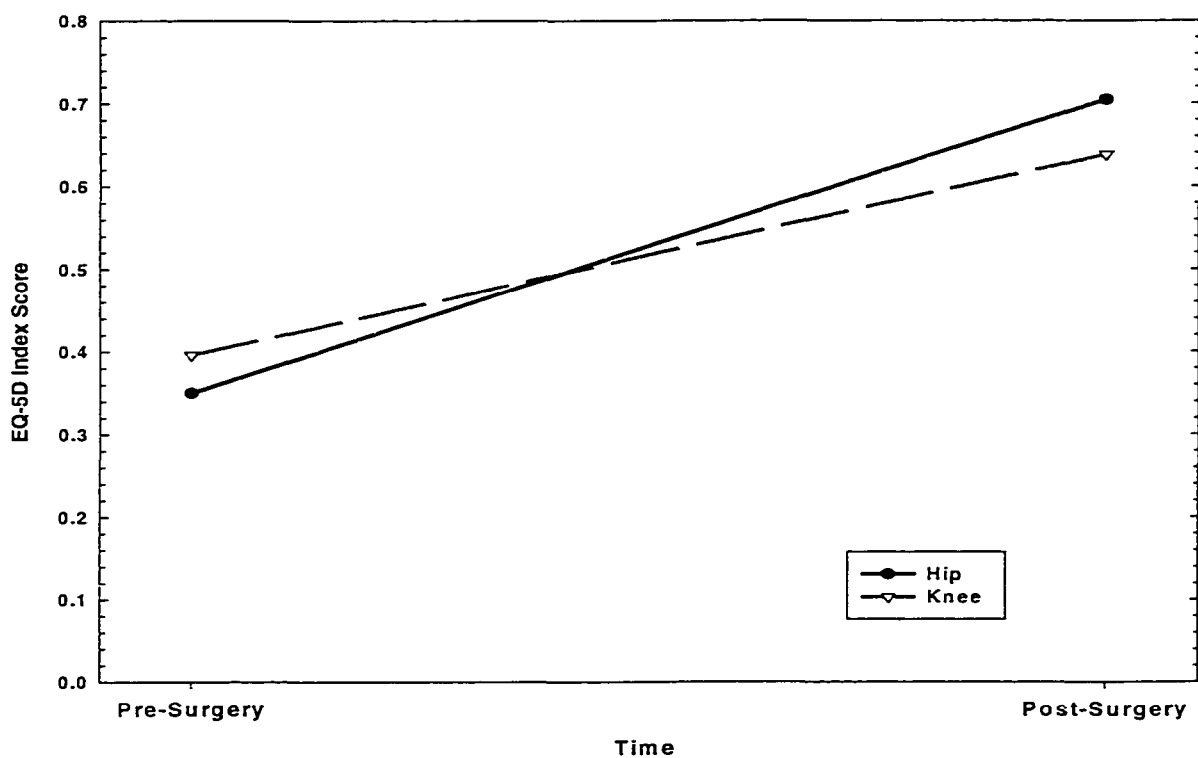


Figure 14. Joint by time interaction for EQ-5D index scores from pre- to post-surgery.

Comparison of Changes in EQ-5D Dimensions with SF-36 and WOMAC Change Scores using the SEM

The percentage of respondents who changed in each EQ-5D dimension is shown in Table 23. Improvements in EQ-5D dimensions and SF-36 subscales were compared with WOMAC change scores (using 2SEM) measuring similar constructs (Table 24). To facilitate the discussion and because the condition specific WOMAC was likely to be more responsive, it was used as the base of comparison. Of the respondents that improved in each EQ-5D dimension, few improved two levels : 0% in mobility, 0.2% in self-care, 0.4% in usual activities, 5.7% in pain/discomfort, and 1.1% in anxiety/depression. For this reason, improvements of one and two EQ-5D levels were combined for this analysis.

Table 23

Percent of Respondents Who Changed at Least One Level from Pre- to Post-surgery on each EQ-5D Dimension

EQ-5D Dimension	% Improved	% No change	% Worse
Mobility	35.3	63.8	0.9
Self-Care	32.3	61.7	5.9
Usual Activities	46.1	49.3	4.6
Pain/discomfort	50.5	47.0	2.5
Anxiety/Depression	25.2	66.1	8.7

Note. $n = 436$.

Where WOMAC and SF-36 subscales measured more than one EQ-5D dimension, EQ-5D changes were compared with similar item changes using the WOMAC and SF-36 items. For example, because the WOMAC function and SF-

36 PF subscales measured mobility, self-care, and usual activities, EQ-5D mobility was compared with WOMAC and SF-36 items measuring walking and climbing stairs.

As seen in Table 24, EQ-5D dimensions are consistently less responsive than SF-36 subscales and items in comparing change with the WOMAC. For example, 390 (85.6%) respondents reported an improvement on the WOMAC Function subscale. Of these, 39.1% improved one level on EQ-5D mobility compared with 74.3% on SF-36 physical functioning.

At the item level for self-care and mobility items, less than 41% of respondents changed at least one level on the EQ-5D as compared with the WOMAC. In contrast, over 60% of respondents changed at least one level on the SF-36 items measured on a 3 point scale. For mental health, 47.7% of those who improved on SF-36 mental health improved at least one level on EQ-5D anxiety/depression.

Table 24

Comparison of Percent Improved on EQ-5D and SF-36 with WOMAC for Comparable Constructs

WOMAC Subscales	Percent of total that Improved	Percent Improved of Those That Improved with WOMAC							SF-36 Items (Percent Improved 1 or 2 levels)			
		EQ-5D Items (Percent Improved 1 or 2 levels)				SF-36 Subscales (Percent Improved > 2SEM)			walk	stairs	Bath or dress	
		Mobility	Self-Care	Usual Activities	Pain Discomfort	Anxiety Depression	Bodily Pain	Physical Functioning	Role Physical			
Pain	82.1				55.9		77.7					
Stiffness	67.7				58.6							
Function	85.6	39.1		51.7			76.7	74.3	59.5			
WOMAC Items												
Walk	82.0	40.1								62.9		
Ascending	78.7	39.6									67.3	
Descending	74.5	40.4									69.3	
Socks on	72.0		37.5									65.4
Socks off	72.2		36.1									62.3
Bathing	68.8		36.8									63.8
Toileting	77.6		36.1									61.8
SF-36 Subscales												
Mental Health	25.0					47.7 ^a						

Note. n = 436; Criteria for Improvement for WOMAC and SF-36 Subscales is 2SEM;

Criteria for Improvement for EuroQol Items, WOMAC Items (5 point scale) and SF-36 Items (3-point scale) is an improvement of at least one level.

^a Reference is SF-36 Mental Health Subscale

QALYS

QALYS were calculated using the three sets of societally derived weights, a model with no N3 term, the sum, and the pre-surgery patient derived weights. Table 25 shows the differences in mean changes and effect sizes of the index scores and QALYS using the different weights. Although the mean change in index scores from pre- to post-surgery was the largest using the TTO weights, the effect size was only the fifth largest due to the large standard deviation pre-surgery. Similarly when QALYS gained were compared using different weights, the QALYS gained were the largest using the TTO weights, and higher by 1.5 than the QALYS gained using the 10 year VAS weights. However, the effect sizes for QALYS using TTO weights and 10 year VAS weights were similar. Although the effect size for QALYS is not usually calculated, it facilitated the comparison across different scales. Because the standard deviation is not taken into account in calculating QALYS gained, a weight that results in lower index scores will likely produce a higher change score. As can be seen from Table 25, the standard deviation of baseline index scores using the TTO is higher than with the other weights, thus resulting in a larger SD for the QALYS using TTO weights. A comparison of QALYS gained using TTO weights with and without the N3 term resulted in a difference of 1.7. Because the N3 term results in a lowering of poor states of health, the likelihood of a higher mean gain in QALYS is increased. Patient derived weights resulted in the lowest number of QALYS gained and the lowest QALY effect size (.49).

Table 25

Descriptive Statistics for EQ-5D Index Scores and QALYs Using Different Weights

Weights	Index Scores				QALYs			
	Baseline Mean	SD	Mean Change	Effect Size	Baseline Mean	SD	QALY gained	QALY Effect Size
TTO	0.38	0.31	0.29	0.93	6.12	6.12	5.14	0.84
VAS 10 year	0.45	0.20	0.21	1.07	7.48	4.58	3.64	0.80
VAS 1 year	0.51	0.18	0.19	1.07	8.45	4.68	3.32	0.71
TTO No N3	0.49	0.19	0.20	1.02	8.17	4.71	3.44	0.73
Sum	0.55	0.14	0.18	1.22	9.21	4.58	3.03	0.66
Patient ^a	0.53	0.09	0.12	0.82	8.88	4.19	2.05	0.49

Note. $n = 451$.

^aPatient weights are derived from pre-surgery data.

Summary of Results

The sample consisted of 540 respondents, 85.4% with a diagnosis of osteoarthritis, undergoing hip and knee replacement. Pre-surgery mean mental health scores reflected levels of mental health close to U.S. population means, while low mean WOMAC and SF-36 subscale scores reflected a substantive level of dysfunction in physical function and pain. The distributions of responses of EQ-5D items generally reflected the distribution of similar dimensions in the WOMAC and the SF-36. However, comparisons of the pre-surgery distributions of EQ-5D dimension scores with comparable WOMAC and SF-36 subscales and items showed some degree of ceiling effect in EQ-5D self-care. Although the majority of respondents scoring a level 1 on EQ-5D self-care reported little or no limitation on a similar SF-36 self-care item, well over half of these respondents experienced moderate to extreme difficulty with specific areas of self-care on the

WOMAC. Pre-surgery 96.3% of respondents reported a level 2 (some problems) on EQ-5D mobility. When their responses were compared with mobility items in the WOMAC and SF-36, respondents showed a wide range of function. As well, few respondents used level 3 for self-care and mobility even when dysfunction as measured by the WOMAC was severe. Differences appeared to be related to the number of levels, and the wording of the extreme levels.

Index scores using the York TTO weights ranged from -0.484 to 1. The distribution was bimodal with 20% of respondents scoring a value of 0 (dead) or <0 (worse than dead). Distributions using 1 year and 10 year VAS weights were similar, but compressed with only one (-0.021) negative score. Earlier, two questions were asked: 1. Do respondents with health states valued by society as 'dead' or 'worse than dead' rate their own health similarly?; and 2. Does the bimodal distribution reflect the distribution of the underlying construct or is it an artifact of the scoring system? The answer to the first question is no, there is great variation. Although correlations between EQ-5D index scores and self-rated health (VAS and SF-36) were moderate, for those individuals with index scores of 0 or less than 0, large discrepancies existed in respondents' perceptions of their own health and in community valuations of their health states. To the second question, the data support the argument that the bimodal distribution is an artifact of the scoring system. The bimodal distribution and substantive number of negative scores appear to be due to a combination of factors: the method of measuring preferences, the N3 term in the regression model, the high disutility weight for pain, and the relationship between health

dimensions. Although correlations of index scores using different weights were high, weights had an effect on both the range and mean of EQ-5D scores. Patient derived regression weights were generally smaller than those derived from societal preferences.

Using confirmatory factor analysis with combined EQ-5D and SF-36 items, two models were tested: a one factor and two factor model. Using the chi-square statistic, the RMR, and the RMSEA, both models resulted in a poor fit. The two factor model was also tested using only the SF-36 items. Although the SF-36 is a widely used HQRL tool with standardized scoring based on extensive psychometric evaluation, the two factor model was a poor fit. Therefore, interpretation of the factor structure of the EQ-5D based on combined SF-36 and EQ-5D items was not possible.

Essink-Bot et al.'s (1997) model using SF-36 subscales and EQ-5D items was tested with CFA and also resulted in a poor model fit. Finally, an exploratory factor analysis with EQ-5D items using a polychoric correlation matrix yielded one factor, while a confirmatory factor analysis resulted in a poor fit. It was concluded that, due to the lack of a stable two factor structure for the SF-36, and the brevity of the EQ-5D, with only one item per dimension, the appropriateness of factor analysis for the assessment of the factor structure of the EQ-5D was questionable.

Evidence was generally supportive of convergent validity. All convergent validity coefficients were statistically significant ($<.01$) and most were greater than .4. Evidence for discriminant validity was mixed depending on which of three

criteria was tested. Most hypothesized validity coefficients were higher than correlations between different traits using different methods. Exceptions were between variables that were conceptually related, such as pain and function. The second criterion for discriminant validity was not met. Many coefficients between different variables using the same method were moderate to high, including those between SF-36 subscales measuring physical and mental health. Finally, the third criterion of discriminant validity examined patterns of correlations between different traits using both the same method and different methods. Because of the variability in number of kinds of dimensions measured by the three instruments, only three dimensions from the EQ-5D and SF-36 subscales were used. Although EQ-5D correlations were lower, correlations followed a similar pattern in the heterotrait-monomethod and heterotrait-heteromethod matrices.

Evidence supported the ability of the EQ-5D to respond to clinically relevant change in patients undergoing hip and knee replacement surgery. Effect sizes were medium to large and larger for hip replacement than for knee replacement patients. The repeated measures ANCOVAs with age as a covariate showed no significant age effect and a significantly larger improvement in HRQL (EQ-5D index and VAS scores) for hip replacement patients than for knee replacement patients. Using the WOMAC as a reference for comparison for improvement in physical functioning and the SF-36 as a reference for mental health improvement, the EQ-5D dimensions were less responsive than either the WOMAC or the SF-36.

Because EQ-5D index scores were designed primarily for use in the calculation of QALYs, QALYs gained were compared using three sets of societal derived weights, patient derived weights, and TTO weights using a model with no N3 term. Results showed that QALYs gained using TTO weights were higher by 1.5 than those using 10 year VAS weights, and higher by 1.7 than those using TTO weights without the N3 term. QALYs gained were the lowest using patient derived weights.

CHAPTER VI

Construct Validity

The purpose of this chapter is to integrate the study findings with other empirical findings to evaluate the construct validity of the interpretation of EQ-5D scores. Both empirical evidence and conceptual analysis were used to assess construct validity. To assess the construct validity of the EQ-D, sources of evidence were examined in terms of the three components of validity (Loevinger, 1957) used by Messick (1989) to organize his discussion on construct validity. These are: substantive, structural, and external and the discussion will follow this order. Finally, the aspect of consequential validity will be discussed. The substantive component is the extent to which the content of the instrument can be accounted for in terms of the trait to be measured and the context of measurement (Loevinger, 1957). The structural component refers to the accuracy of the scoring model as a measure of the theoretical construct. The external component refers to the degree that the relationships with other measures are consistent with the construct theory. The consequential aspect includes an evaluation of the value implications of test interpretation and the social consequences of test use (Messick, 1989).

Substantive Component of Construct Validity

The process of test construction begins with defining a purpose for the test. It is that purpose which determines the area of content to which the items included in the instrument will be referenced. The items should be chosen from a pool of such items (Loevinger, 1957). According to Loevinger the final set of

items should be based both on theory and empirical findings. Key issues in the evaluation of the substantive component include the purpose of, and theoretical basis for, the construct, the nature and boundaries of the construct, and the relevance and representativeness of the items.

The EQ-5D Health State Descriptive System

Purpose and theoretical basis of the EQ-5D descriptive system. The process of measurement begins with a purpose. The EQ-5D was designed as a preference-based HRQL measure to enable comparisons of results across groups, conditions, and settings. Requirements were that it be simple to complete, usable in postal surveys, relatively undemanding, relevant to all respondents, capable of providing a single index score, and consistent with health states worse than dead (Kind, 1996). Although, ideally, theory should guide all aspects of the measurement process, there is a lack of published literature on the theoretical background of the EQ-5D.

The foundation for the work of the EuroQol Group was laid by Rosser and her collaborators (Williams, 1995). The Rosser Index (Rosser & Kind, 1978) was developed for use as a health indicator and output measure. Based on disability (states of illness) and subjective distress, its aim was to place valuations on these states and on death. The Rosser Index items measuring disability (mobility, self-care, usual activities, social and personal relationships) and distress (depression, anxiety, pain) formed the basis of the current EQ-5D. To identify the EuroQol dimensions, the EuroQol Group used their expertise, and reviewed both the literature and the content of the Rosser Index as well as other

health status measures (Kind et al., 1994) including: the Quality of Well Being Scale (QWB) (Patrick, Bush, & Chen, 1973), the Sickness Impact Profile (SIP) (Bergner, Bobbitt, Pollard, Martin, & Gilson, 1976), and the Nottingham Health Profile or NHP. A summary of the dimensions included in these HRQL measures is provided in Table 2. The QWB was based on a concept of health as functional status (social preferences for levels of health) and prognosis (Patrick et al., 1973). Optimal function was defined as conformity to the social norms of well-being, including the ability to perform daily activities usual for a person's age and social role. It also measured symptom/problem complexes as factors that cause deviations from well-being. The SIP was developed as a measure of perceived health status. Its purpose was to provide a measure of the outcomes of health care that could be used for evaluation, program planning, and policy formulation. It was based on a construct of 'dysfunction' and measures behavioral impacts of sickness in terms of dysfunction. The NHP was developed as a standardized tool for the survey of health problems. An initial pool of statements describing the typical effects of ill-health (social, psychological, behavioural, and physical) was collected from patient and community populations. The NHP measures negative aspects of health only and not positive feelings of well-being.

Background research in the development of the EQ-5D also included interviews on people's attitudes towards health, including a random sample of 200 community members (aged 18 and over), and two other groups each with a control group: 100 physically disabled young people living at home and 100 individuals who had been caring at home for physically disabled children

(Williams, 1995). Researchers found that notions of health as functional capacity, feelings, and general fitness predominated. Importance of items was also rated by respondents and used to determine the EQ-5D dimensions.

Although many HRQL instruments, including the EQ-5D, appear to be atheoretical, various theories have influenced the development of HRQL preference-based instruments, such as the EQ-5D. Two major groups of theories have influenced the content of HRQL measures: theories of positive well-being and functionalism (Patrick & Erickson, 1993). Theories of well-being are numerous and include models of coping, adaptation, stress, self-mastery, self-efficacy, and subjective well-being. Functionalism, founded by Durkheim, focuses on the social roles, such as work, and the activities of daily living needed to function independently in society (Patrick & Erickson, 1993). It appears to be the predominant theory underlying the EQ-5D, the SIP, the Rosser Index, the QWB, and the NHP. The four measures that were used as a basis in the development of the EQ-5D were all based on a concept of health status in terms of function or dysfunction. All five EQ-5D dimensions directly or indirectly measure function. Three measure the performance of physical function (self-care, mobility, and usual activities) while anxiety and depression are indicators of psychological functioning. Pain and discomfort are symptoms which affect functioning.

The nature and boundaries of the construct. There is no one agreed upon definition of HRQL, also referred to as health status. The term, HRQL, is differentiated from a much broader term, quality of life, which includes

dimensions not necessarily directly related to an individual's health, such as economic status, vocational status, standard of living, environment, and culture. However, the terms quality of life and HRQL are sometimes used interchangeably. Using concepts of functionalism in conceptualizing health and illness, Parsons (as cited in Patrick & Erickson, 1993, p. 61) defined health as "the state of optimum capacity for the effective performance of valued tasks." Patrick and Erikson (1993), well known in the area of health policy, define HRQL as "the value assigned to duration of life as modified by the impairment, functional states, perceptions, and social opportunities that are influenced by disease, injury, treatment, or policy" (p. 22). HRQL has also been conceptualized as well-being, life satisfaction, happiness, and need satisfaction (Cella, 1992; Oleson, 1990; Patrick & Erickson, 1993). It is difficult to find a conceptual basis for the EQ-5D. Rather than basing the EQ-5D on a conceptual model of HRQL, the dimensions appear to define the construct. Williams (1974), one of the original members of the EuroQol Group, conceptualized health as comprised of two dimensions, pain and restriction of activity. In one paper, Essink-bot, one of the founding members of the EuroQol Group, defined health status as "physical, psychological and social functioning" (Essink-Bot et al., 1995, p. 200).

Although the definitions of HRQL or health status differ, most HRQL measures are based on function and assess three broad areas of health: physical, psychological, and social (Patrick & Erickson, 1993; Schipper, Clinch, & Olweny, 1996; Osoba, 1991; Torrance, 1986; Ware, 1987; Wood-Dauphinee,

1992). These areas reflect the World Health Organization (WHO) definition of health as “a state of complete physical, mental, and social well being, and not merely the absence of disease or infirmity” (as cited in Patrick & Erickson, 1993, p. 19).

Although it is agreed by most researchers that HRQL is multidimensional, the specific nature and number of dimensions differ. Dimensions that have been assessed include leisure, spirituality, somatic comfort, sexual functioning, cognition, symptoms, social or cultural disadvantage, resilience, general health perception, economic status, and vocational status (Patrick & Erickson, 1993; Schipper et al., 1996; Testa & Nackley, 1994). The EuroQol descriptive system was initially based on six dimensions: mobility, self care, main activity, pain, mood, and social relationships, each being measured with two or three levels which reflected the degree or extent of functioning along the dimension. For example, ‘mobility’ was measured in terms of three levels, while ‘social relationships’ was measured with two levels: 1. able to pursue family and leisure activities, and 2. unable to pursue family and leisure activities. Researchers later concluded that ‘social relationships’ played little part in determining health state valuations. No further explanation was given in the literature. Consequently, the number of dimensions was reduced to five with ‘social relationships’ being incorporated into the category, ‘usual activities’, which includes work, study, housework, family, or leisure activities as examples of usual activities. A seventh dimension, ‘energy/tired’, was also included on one survey, but that dimension has not been used in subsequent studies (Kind et al., 1994).

Although the dimensions for the EQ-5D have been agreed upon by the EuroQol Group, there is still some controversy as to whether additional dimensions should be included. In comparison with a competing HRQL preference-based measure, the Health Utilities Index (HUI) (Feeny, Furlong, Boyle, & Torrance, 1995), two dimensions, sensation and cognition, included in the HUI are not covered by the EQ-5D (Table 2). As well, the HUI covers fine motor function under mobility. Although it is recognized by EuroQol members that cognition is an important dimension not included in the EQ-5D, it has been argued that cognition is indirectly covered under usual activities (M.Buxton, personal communication, July 31, 1998). It could be argued that other symptoms, such as fatigue, nausea, weakness, dizziness, or shortness of breath, are as important to HRQL as is pain/discomfort. Conversely, it could be argued that all symptoms could fit under discomfort. However, brevity of the questionnaire was an important consideration of the EuroQol Group. The content of the EQ-5D broadly reflects that of competing preference-based HRQL instruments.

Content representativeness and relevance. Content representativeness is concerned about how well the items included in a test represent the construct or domain of reference. The domain of reference is the total body of information for which the construct is expected to account and about which inferences are to be drawn (Messick, 1989). According to Loevinger (1957) the areas of content should be represented in proportion to their life-importance. Clarity in describing the dimensions and boundaries of the domain are essential to assess the representativeness of the items to the construct, HRQL. Content relevance

refers to the relevance of the dimensions to the construct of interest and the relevance of each item to the dimension to which it is referenced.

Although the construct is multidimensional, each of the five dimensions should be conceptually unidimensional. However, with the EQ-5D, the dimension 'anxiety/depression' measures two concepts; although anxiety and depression are related, one could be anxious without being depressed. For example, in a recent study comparing responsiveness of the EQ-5D with a cancer specific instrument in breast cancer patients (Conner-Spady, Cumming, Nabholtz, Jacobs, & Stewart, 1999), patterns of anxiety and depression as measured by the *Functional Living Index-Cancer* differed over 4 time periods. The same problem exists with the dimension 'pain/discomfort'. This lack of unidimensionality may lead to lack of clarity in knowing what is being measured and could result in problems in interpreting results.

Unlike most instruments which contain many items for each dimension, the EQ-5D has only one item for each dimension. While the wording of the dimensions are broad (e.g., mobility), the items are not necessarily representative of the dimension. In addition, inconsistent wording of levels could lead to variations in the interpretation of items. For example, mobility is a term which could include the ability to drive or board a bus, climb stairs, or get into the bath. A person confined to a wheelchair could be considered to be mobile. However, two mobility levels restrict mobility to 'walking about', while the third is 'confined to bed'. Similarly, although self-care could include activities such as feeding, grooming, and elimination, two self-care levels are restricted to washing

and dressing, while the third level mentions neither (no problems with self-care). This inconsistency could lead to confusion about what is meant by the items and dimensions. Fox-Rushby (1996) asked EuroQol Group members to describe key terms or phrases within the EQ-5D. She found considerable variation among members in the meaning of key terms (for example, mobility and discomfort) in the EQ-5D questionnaire. Little is known about how patients interpret and respond to the EuroQol questions. Although no published studies have examined this issue, Selai (1994), a EuroQol Group member, reported on comments made by patients with epilepsy and Parkinson's Disease while completing the EQ-5D descriptive section and VAS self-rating task. The main problems with the descriptive system were the wide ranges within levels, the wording of 'usual activities' in patients with chronic problems, and the stigma of depression in the combined anxiety/depression dimension.

In the present study, although all of the EQ-5D dimensions are relevant to patients with osteoarthritis, four are particularly relevant: pain, mobility, usual activities, and self-care. Each of these four dimensions is also measured by the WOMAC, developed specifically for patients with osteoarthritis. Osteoarthritis is commonly associated with clinical symptoms of joint pain and disability. It is usually progressive and leads to worsening pain, particularly by joint movement, and increasing functional disability (Bombardier et al., 1995; Guccione, Felson, & Anderson, 1990). Difficulty with walking and stair climbing are two of the most frequent disabilities in patients with osteoarthritis of the hip or knee (Guccione et al., 1990; Laupacis et al., 1993). Other disabilities such as difficulty with putting

on shoes and socks, difficulty in standing for some time, difficulty in carrying out activities of daily living, and decreased socializing have also been frequently found (Laupacis et al., 1993). Forty-nine per cent of respondents used an aid to walking, reflecting a substantial clinical level of disability. Pre-surgery, 94.0% of respondents reported a problem in at least three EQ-5D dimensions, reflecting the high level of disability preoperatively, and the relevance of the dimensions in this population.

The item must be relevant to the dimension and cover the range of possible states within the dimension. One limitation of a three level scale is the broad range of disability covered by each level. Because the SF-36 physical functioning items were also measured on a 3-point scale, comparisons of EQ-5D self-care and mobility were expected to yield a similar pattern of distribution. For self-care, the EQ-5D distribution was fairly congruent with that of the 3-point SF-36 self-care item; that is, the majority of respondents who reported a 1 (no problem) on the EQ-5D also reported little or no limitation in self-care on the SF-36. This expectation was not met for mobility. Pre-surgery, 96.3% of respondents were at level 2 on EQ-5D mobility. In comparison with the 3-level SF-36 item 'walking one block', respondents were more evenly distributed with the SF-36 item than on EQ-5D mobility. Therefore, EQ-5D mobility did not reflect the range of underlying disability.

Two factors may have contributed to inconsistent patterns between measures: the time frame of the questions and the wording of the options. The EQ-5D asked respondents to indicate which statements best describe 'your own

health state today'; the WOMAC used 'currently experiencing' and most of the SF-36 questions varied from 'during the past 4 weeks' to 'does your health now limit you'. However, because the pain and functional limitations of osteoarthritis are typically chronic, the time frame was less likely to be a major factor.

Secondly, the interpretation for each measure most likely differed due to different wording. For example, the SF-36 questions for mobility and self-care ask the degree to which 'your health limits you', while the EuroQol items ask respondents to indicate whether they have problems. Respondents who had moderate to severe problems on various aspects of self-care (WOMAC), yet reported a 1 on EQ-5D self-care may have had different interpretations of the word 'self-care'. Little is known about the cognitive processes that occur when interpreting and responding to HRQL items. Unfortunately, in the present study, it was not possible to collect this data. Also, the extreme wording of level 3 in both self-care (unable to wash or dress myself) and mobility (confined to bed) may have restricted respondents from choosing a level 3 even when they had extreme problems in these dimensions. Confined to bed is not a realistic level for most osteoarthritis patients even with severe levels of disability. Wolfe and Hawley (1997) reported a similar finding in patients with rheumatic disorders and concluded that the EuroQol scaling was too narrow to capture changes in patients with rheumatoid arthritis. They compared the distribution of responses of the EQ-5D with that of the Health Assessment Questionnaire Disability Index (HAQ DI), a measure of activities of daily living with four levels: no problem, some problems, moderate problems, and severe problems. For the EQ-5D

'mobility' dimension 0.4% of the patients reported they were at level 3 (confined to bed) while for the HAQ DI 'mobility' 11.3% reported that they had a severe problem. Similarly for pain, 78.6% of the patients reported 'moderate pain or discomfort' on the EQ-5D while 36.4% reported 'moderate pain' and 31.2% reported 'some pain' on the HAQ DI.

In contrast, the level 3 wording for items measuring usual activities, pain/discomfort, and anxiety/depression did allow for respondents to use that level when problems were severe. Although the wording of level 3 (unable to perform) for usual activities is similar to that of self-care, the dimension is sufficiently broad for a wide variation in activities (work, study, housework, family, leisure) and an interpretation which is open to the respondent, compared with the specific wording of self-care (unable to wash or dress oneself). A person might be unable to perform the activities that one previously performed, yet still be able to function independently. Because EQ-5D usual activities measures a variety of activities, interpretation of comparisons with other instruments is difficult. However, the low mean scores of both the SF-36 physical and WOMAC subscales (Table 8) are consistent with the distribution of respondents in EQ-5D usual activities and pain/discomfort. Similarly, the distribution of responses on EQ-5D anxiety/depression reflected the moderately high mean SF-36 mental health scores.

In summary, although the EQ-5D is not based on a specific construct theory, the five dimensions cover the three broad dimensions of health included in most currently used HRQL tools: physical, psychological, and social

functioning. A lack of unidimensionality in the dimensions, the lack of criteria for describing the dimensions, and the inconsistent wording of the levels affects the clarity of the construct and the assessment of representativeness of the items to the dimension. Pain/discomfort and the functional dimensions are particularly relevant to this patient population. However, the range of levels in self-care and mobility were not adequate to cover the range of underlying disability.

Preference Scores

The main purpose of valuation is to establish trade-offs between quality and length of life (Nord, 1991b), for example, between living with the pain and disability of rheumatoid arthritis or receiving a medication which will relieve some of the pain but which has many potential side effects, such as liver damage, nausea, and allergic reactions. Many of our health care interventions involve similar trade-offs. The measurement of preferences forms the basis for the conversion of EQ-5D health states to index scores and their subsequent use in cost-utility analysis. The measurement of an underlying construct of preference for health states assumes that there is an underlying preference to measure. The previous section discussed the problem of determining what HRQL dimensions to measure; this section will briefly describe some of the theories underlying the measurement of preferences. What to measure and how to best measure health state preferences are as yet undetermined.

Preference-based HRQL measures are based on theories of welfare economics and utility theory. The notion of preference is rooted in welfare economics and is based on the economic idea of satisfaction, or utility

(differentiated from utility theory). One of the assumptions made about the motive of an individual is that the consumer desires to obtain a maximum of utility or satisfaction. The individual who attempts to obtain these maxima is said to act rationally. Utility theory provides a mathematical theory about how people act rationally to maximize utilities.

The standard gamble is based directly on the axioms of utility theory. Utility theory, developed by von Neumann and Morgenstern (1944), is a normative theory that describes how a rational person ought to make decisions when faced with uncertain outcomes. Three axioms of rational behavior govern utility theory:

1. Preferences for outcomes exist and are transitive (ordering axiom). If $A > B$ and $B > C$, then $A > C$.
2. Preferences for a risky prospect are independent of whether it has one stage or two (independence axiom).
3. There is a continuity of preferences. If there are three outcomes where $A > B$ and $B > C$, there is a probability p where the individual is indifferent between the certain outcome of B or a risky outcome of outcome A with a probability of p and outcome C with a probability of $1-p$ (Drummond, O'Brien, Stoddart, & Torrance, 1997; Patrick & Erickson, 1993).

Many economists assert that the SG is the gold standard against which other methods of measuring preferences should be compared (Gold et al., 1996). However, researchers have demonstrated that the assumptions of utility theory are not always met (Tversky & Kahneman, 1981) and that the SG is subject to

framing or contextual effects including risk preference (Tversky & Kahneman, 1981), duration of health states (Sutherland, Llewellyn-Thomas, Boyd, & Till, 1982) and other states with which the health state is compared (Patrick, Starks, Cain, Uhlmann, & Pearlman, 1994).

TTO is based not only on the assumptions of utility theory, but also on the assumption that the perception of the severity of illness is independent of the time spent in the state. However, findings have challenged this assumption (Dolan, 1996). It has also been assumed that individuals have a positive rate of time preference, prefer benefits today rather than in the future, and prefer to defer undesirable outcomes (Drummond et al., 1997; Pigou, 1960). However, this assumption has also been challenged. Dolan and Gudex (1995) varied duration and time preference in using TTO to value six EQ-5D health states. They found that individual's time preferences varied considerably and concluded that TTO was a function of framing rather than preferences. Whether these violations of assumptions could seriously compromise the validity of the SG and TTO methods in clinical practice is controversial. Finally, the VAS method of measuring preferences is an indirect way to establish these trade-offs between quality and length of life; respondents are asked to locate the health state on a VAS but are not asked to make any choices or decisions (Nord, 1991b). Because the primary purpose of the EQ-5D valuation is for use in cost-utility analysis, the validity of the index is based not only on its validity as a measure of HRQL but also its meeting of the assumptions of a utility and interpretability of the score as a utility.

Not only is the most valid method of measurement controversial, but there are many variations within methods. These include: wording, number, severity, and duration of health states valued; modelling of preferences; time preference; instructions; anchor states; and whether negative states are valued. In studies using the EQ-5D, there are variations in instructions and little rationale is given for the reasons behind the choice of instruction. For example, in measuring TTO weights, Krabbe, Essink-Bot, and Bonsel (1997) replaced 'being dead' with 'worst imaginable health state' and instructed respondents that after 10 years, the health state would return to its present form. In measuring VAS weights, Dolan and colleagues used '10 years followed by death' in one study (Gudex et al., 1996) and '10 years followed by don't know what is going to happen' (Dolan, 1996) in another. In contrast, the standard EQ-5D questionnaire for measuring preferences asks respondents to value health state descriptions 'for a person like yourself' on a VAS. The duration of each state is to be one year, with an unknown future.

Few studies have examined the meaning of the EuroQol valuation process to respondents. When nurses were asked how they approached the valuation task, they varied from using their own health state against which to judge the health states to prioritizing variables such as anxiety, depression, pain, and degree of independence from others (O'Hanlon et al., 1994). Nord (1991b) asked physicians and bioengineers to value a set of health states on a VAS and then to describe why they chose the particular numbers. Answers ranged from 'percentages of the best imaginable state' to 'no particular meaning' (Nord,

1991b). Kind found that when instructions included the idea of allocation of limited resources, scores were higher for the more severe EuroQol states (Kind et al., 1994). In a study examining the validity of the six dimension EuroQol as a measure of social value, Nord et al. (1993) compared EuroQol valuations with those obtained by a person trade off method, in which respondents were asked to choose between two treatment alternatives with limited resources. Because EuroQol valuations were lower than the person trade off valuations, particularly at the lower end of the scale, Nord et al. concluded that the EuroQol underestimated the social value put on health states. However, this conclusion assumes that the person trade off method is a more valid method of measuring preferences. Clearly, for some consistency in interpretation of the valuation process, instructions need to be more explicit and consistent with the intended interpretation.

Although preferences were not directly measured in this study, seven different weights were used to calculate EQ-5D index scores. Results demonstrated the effect of different weights on the distribution and subsequent interpretation of EQ-5D index scores. Index scores using TTO derived weights produced a wider range and more negative index scores than did VAS derived weights. The TTO method of measuring preferences was initially based on the assumption that death is the worst possible outcome (Rosser & Kind, 1978). However, researchers have found that many respondents value some health states as worse than death. Unfortunately, few studies have addressed states worse than death and yet, as seen in this study, such valuations are relatively

common. Some researchers have measured negative weights (Dolan, 1997) and some have set negative health states to zero if the respondent valued them worse than death (O' Hanlon et al., 1994). Consequently, values for 'states worse than death' are largely dependent on if they are measured and how they are measured.

One assumption underlying the measurement of preferences by the public is that people have some understanding of the health states they are asked to value. Moderate correlations (0.45 to 0.60) between self-rated health (EQ-5D VAS) and EQ-5D index scores give modest support to this assumption; however, it seems to fail under certain conditions. Thus, in spite of the fact that respondents with scores valued at 0 or < 0 had a high level of pain and dysfunction, 74.1% rated their health from fair to excellent. This finding is supported by other studies that have found a level of adaptation in people coping with disability (Albrecht & Devlieger, 1999). Therefore, a problem with the system is that it fails to take in to consideration adaptation to a particular state. This wide discrepancy between self-rated health and societal determined valuations challenges the assumption that the public has some understanding of the hypothetical health states they value. In trying to understand this discrepancy, one must examine the method of measurement of TTO weights. The EQ-5D TTO weights were derived from asking people to value health states in which the respondent would remain for 10 years. The weight is then applied to a health state measured as 'your health state today.' Is the hypothetical situation realistic? Are people in states of extreme pain or discomfort for 10 years or is the

pain more realistically intermittent? Does extreme pain mean the same thing to someone who hasn't experienced it and to someone who has?

In summary, the purpose of the measurement of preferences is to establish trade-offs between the quality and length of life. The construct of preference is based on theories of welfare economics, while the measurement of preferences is based on utility theory. Three methods commonly used to measure preferences are the VAS, the TTO, and the SG. Each method uses different techniques and instructions to arrive at a preference score which can be used in the calculation of QALYS. The validity, not only of EQ-5D index scores, but preference scores in general, is questionable because evidence increasingly supports the argument that preference scores are highly dependent on the measurement technique, do not necessarily meet the assumptions of the technique, and are subject to framing effects. Although this study did not directly assess the substantive component of validity, study findings showed weak support for the assumption that the public have some understanding of the health states they are asked to value, particularly for poor health states. Although there was moderate correlation between self-rated VAS scores and societal derived index scores, there was little congruence between scores valued at 0 or less than 0 and self-rated health. Interpretation of index scores is difficult due to the discrepancy between self-rated HRQL and societal based valuations. The difficulty in interpretation may be due to the techniques used to measure preferences.

Structural Component

Michell (1990) defines measurement as “a procedure for identifying values of quantitative variables through their numerical relationships to other values” (p. 63). A measurement model or scoring model describes the method of scaling by which item responses are combined to form scores. The structural component of validity assesses both the fidelity between the scoring model and the construct theory and the degree of interitem structure (Messick, 1989). The scoring model should be guided by the purpose and reflect the underlying construct theory, so that scores represent the underlying quantitative attribute of the construct being measured. As well, the degree of homogeneity of items should reflect the degree of homogeneity of the underlying construct. Because the EQ-5D was primarily designed to produce a single index value for any given health state, this assessment will focus on the fidelity of the scoring model used to produce the index score. Two aspects of the structural component were addressed: 1. the congruence of the scoring model with the construct theory; and 2. the assumption of interval scaling.

To adequately assess the structural component, a sound construct theory should be the basis for measurement. Two constructs, HRQL and preference, underlie the EQ-5D scoring models. As discussed under the substantive component, the EQ-5D is not based on a well-defined theory, but on an accumulation of evidence on what people view as essential components of HRQL. Because there is little published on the theoretical underpinnings of the

EQ-5D, an assessment of the congruence of the scoring model with the construct theory is difficult.

The scoring model for the EQ-5D is complex. A rating scale, with an assumption of unidimensionality and interval scaling, is used for the VAS. The health state dimensions are measured on an ordinal scale, while the index score is a combination of the measurement of HRQL and the measurement of value judgements. This combination results in uncertainty as to what evidence is needed to assess the structural component.

The scoring model underlying the EQ-5D index score is based on classical test theory and is a variation of the summative model (Scott, 1968); that is, the score is obtained by adding the weighted scores from each item to produce a disutility score, which is then subtracted from one to produce the single index score. The assumption underlying a summative scale is that each item is a linear or at least a monotonic function of the same attribute (Scott, 1968). However, conceptually, this assumption does not hold for the items in the EQ-5D, particularly for pain, which is not a function of HRQL, but rather a symptom which affects aspects of HRQL. While the EQ-5D descriptive system is based on a multidimensional model of HRQL, both the EuroQol VAS and the index score are based on a unidimensional model. That is, the phenomena are mapped onto a numerical continuum according to variations in their characteristics (Maguire, Hattie, & Haig, 1994). What is mapped along the continuum is not a variation in an amount of HRQL that an individual possesses. Rather, the value that judges place on some combination of HRQL attributes is what is mapped. Kaplan

(1964) refers to this as a configurational method, where the measure is a combination of the behavior of the subjects and judges. The consequence of a multidimensional construct summarized in a single score is that relations between the construct, HRQL, and other constructs are difficult to interpret. For example, researchers have tried to predict future health outcomes, such as mortality, with HRQL but knowing which attributes of HRQL are causally related to mortality cannot be interpreted with a multiattribute measure such as the EQ-5D. As well, these two models may not be congruent with each other.

The EQ-5D is in a relatively early stage of development and the rules for the measurement of preferences have not been standardized. Because the scoring model is guided by the proposed application of the scores, two requirements of scaling were that the index score be interval and that it be anchored by 0 (dead) and 1 (perfect health) for its use as a QALY. The assumption underlying the measurement of preferences is that there is an underlying quantitative dimension of value judgements for various health states that can be ordered on a continuum and that these value judgements are measured on an equal-interval scale. Theoretically, responses using the TTO and VAS methods should be measuring the same underlying construct and should be correlated. Although there is no method to test the assumption of interval properties when there is no physical continuum with which to compare responses, Torgerson (1958) suggests that if the same rationale underlies two scales, the ratio of scale values for three or more stimuli should be invariant, within sampling error, for the two scales. That is, the plots of values should be

linear. Although this test was to be applied to scales measured by the same methods, it was used in this study as a test to assess the property of interval scaling. Using the scatterplots of index scores using TTO and VAS (10yr) weights, the scatterplots appeared to be linear, supporting the assumption of interval scaling. However, because index scores are not only dependent on the weights but also on the N3 model, this test does not address the question of whether either scale is interval.

The distribution of scores using patient derived weights resulted in a symmetric distribution of index scores. In contrast, the bimodal distribution of index scores (using both TTO and VAS weights) with a gap containing scores for improbable health states raises the question of whether the bimodal distribution is an artifact of the N3 model or a reflection of the underlying preferences for a continuum of health states. The bimodal distribution may reflect the large qualitative gaps between levels 2 and 3 and the subsequent low preferences that the public gives to extreme health states that they might imagine to be intolerable (for example, confined to bed for 10 years).

Finally, a serious problem exists with the scoring model regarding negative health states. Most studies only address and measure health states preferred to death (Patrick et al., 1994). As a result, negative index scores only exist if negative preferences are measured. When health states 'worse than death' are measured using the TTO method, the procedures must be altered. Thus the scaling rules for these health states differ from those for health states valued 'better than dead': values below 0 are measured on a different scale from

those above 0, resulting in scores that have been measured using different instructions. As well, the lower boundary is dependent on the duration used, resulting in a lack of a theoretical lower boundary to negative health states. For measuring EQ-5D TTO values, Dolan (1997) used a maximum duration of 10 years with a resulting lower bound of -39. The negative scores were then transformed to produce a lower bound of -1. Thus, because the measurement of states worse than death is dependent on whether they are measured and how they are measured, negative scores are largely uninterpretable (Patrick et al., 1994).

In summary, there are a number of unresolved problems with the structural component of the EQ-5D. Although the TTO weights are currently recommended by the EuroQol Group, the scoring system has not yet been standardized. The evidence from this study does not support the fidelity of the construct theory with the scoring model. Two main problems exist: the lack of a construct theory, and how to measure and model the preferences so that the scores represent an underlying unidimensional phenomena. Until the problem of the measurement of health states 'worse than death' is resolved, EQ-5D index scores will not be suitable for use as QALYs.

External Component

The external component assesses the degree to which the EQ-5D's relationships with other measures is consistent with underlying theory. In this study, two aspects of the external component were assessed: convergent and discriminant validity evidence and responsiveness.

Convergent and Discriminant Validity Evidence

Convergent validity evidence was supported by the significant correlations between subscales measuring similar traits as hypothesized. These findings were similar to those of Chetter et al. (1997), who compared the SF-36 with the EQ-5D in patients with lower limb ischaemia. In this study the evidence for discriminant validity was mixed. Convergent validity coefficients were generally higher than the correlations between variables having neither trait nor method in common. However, there were also moderate to high correlations between related variables measuring different constructs, such as pain and function. As well, many correlations between different constructs using the same method of measurement were high, within the SF-36 and the WOMAC heterotrait-monomethod matrices. For example, vitality was correlated moderately with all other SF-36 subscales post-surgery. This makes substantive sense when items within subscales are examined. Vitality measures energy levels and tiredness which are symptoms of both mental and physical health. Finally, using only three dimensions of the EQ-5D and SF-36, patterns of correlations between different constructs within the heterotrait-monomethod and heterotrait-heteromethod matrices supported discriminant validity. However, this was a weak test due to the limited number of dimensions used for assessment.

The MTMM approach was intended to assess distinct theoretically distinct constructs using independent methods (Campbell & Fiske, 1959; Messick, 1989). In the situation where one is comparing subscales of different HRQL measures, this ideal situation is rarely found. According to Campbell and Fiske high

heterotrait-heteromethod values can be due to correlation between methods or to correlation between traits. When either of these situations occurs, evaluation of validity, although more difficult, can still take place and must include an assessment of the factors which could affect the magnitude of the correlations between different traits (Campbell & Fiske, 1959).

Dependence between traits in these HRQL measures was due partly to the multidimensional nature of some of the items. The EQ-5D item usual activities measures work, study, housework, family, and leisure activities. Concepts overlap when subscales of different measures are compared. For example, WOMAC functioning measures mobility, self care, and function. As well, the EQ-5D items mobility, self-care, usual activities, and pain/discomfort are conceptually related due to their common underlying physical traits. For example, pain and problems with mobility will directly affect self-care and usual activities. Finally, where there is a common underlying process, such as the experience of joint pain, both the physical and psychological dimensions of HRQL will likely be affected, thus possibly explaining some of the covariance between variables measuring different traits.

Although the MTMM approach is useful, it has its limitations in a study such as this where the study was not designed specifically for this approach. When measures differ as to the constructs measured, and subscales contain overlapping and conceptually mixed items, interpretation of results using this method is problematic.

Responsiveness

Two methods were used to assess the responsiveness of the EQ-5D index scores and EQ-5D VAS scores: effect size and a repeated measures ANCOVA (time-by-joint) with age as a covariate. The EQ-5D effect sizes were moderate (knee) to large (hip) for both index and VAS scores, with larger effect sizes for the index scores. As hypothesized, all effect sizes were larger for hip than for knee replacement patients. Effect sizes were smaller for the EQ-5D index scores and VAS scores than for the WOMAC subscales and SF-36 PCS, but larger than the SF-36 MCS. Considering that the EQ-5D measures both physical and emotional dimensions, the moderate to large effect sizes are supportive of an ability to respond to change in this patient population. Few studies have reported EQ-5D effect sizes, but in those that have, effect sizes have been small for both EQ-5D index and EQ-5D VAS scores compared with large effect sizes for condition specific measures (Jenkinson et al., 1997; Jenkinson et al., 1998). The mean EQ-5D index change scores of .24 (knee) and .35 (hip) were similar to those found by James et al. (1996) in patients with TJA, but comparisons are difficult because different weights were used to calculate index scores.

The results of the repeated measures ANCOVAs support the hypothesized change in HRQL due to TJA. The greater improvement in HRQL for hip replacement patients is consistent with other findings (Rissanen et al., 1997). These findings are supported by other studies which have shown improvement in physical functioning as well as sleep and rest, emotional

behavior, social interaction, and recreation following hip replacement (Laupacis et al., 1993). There is evidence that age may be related to physical functioning and degree of improvement following TJA (Jacobsson et al., 1991). However, in this study, age had no effect on improvement in HRQL. With the Wilcoxon's rank test, all EQ-5D dimensions improved over time following surgery.

Using the WOMAC as a 'gold standard' with which to assess responsiveness in physical functioning and pain, the EQ-5D was the least responsive HRQL tool (Table 24). For mental health the EQ-5D dimension anxiety/depression was less responsive than the SF-36 mental health subscale. To be responsive, a measure must have items that are relevant to the population, are amenable to health interventions, have little floor or ceiling effect, and have an adequate range of responses (Fitzpatrick et al., 1992; Fletcher et al., 1992). As discussed previously, the EQ-5D items were found to be relevant to this patient population and there was little ceiling effect. However, when change scores were compared with the WOMAC and the SF-36, the EQ-5D dimensions were less responsive. For two items, self-care and mobility, less than 41% of respondents who improved on the WOMAC improved on the EQ-5D, compared with over 60% improving on similar 3-level SF-36 items. Although the change in mobility was statistically significant, 65% of respondents did not change on this dimension. This may have been due in part to the wording of level 3 where less than 1% of respondents reported a pre-surgery rating of 3 despite severe to extreme problems. These findings suggest that diminished responsiveness of the EQ-5D is not only due to the number of levels but the wording of the options.

To summarize, two sources of evidence were used to evaluate the external validity of the EQ-5D. Although convergent validity evidence was supported, discriminant validity evidence was weak, partly due to the multidimensional nature of the dimensions 'usual activities' and anxiety/depression, as well as the overlapping concepts when different measures are compared.

This study supported the ability of the EQ-5D to respond to clinically relevant change in patients with hip and knee replacement surgery. Although responsiveness of the EQ-5D was less than the WOMAC and the SF-36, the relevance of the dimensions support its usefulness as a preference-based HRQL tool in this population. The lesser responsiveness of the EQ-5D in comparison to condition-specific tools, while a recognized characteristic of generic HRQL tools, could most likely be improved with a change in the wording of the extreme levels for EQ-5D self-care and mobility, or by an additional level. Whether the EQ-5D would be responsive to smaller clinical changes in patients with osteoarthritis is yet to be determined.

The Consequential Aspect of Validity

'The issue is no longer whether to take values into account, but how'

(Messick, 1989, p.58)

Validity inquiry in the measurement of HRQL has thus far in the literature been limited to various traditional kinds of validity evidence, such as content, criterion, and construct related evidence. Messick (1989) theorizes that validity inquiry should include an evaluation not only of this evidence but also of the

value implications of test interpretation and the potential and actual social consequences of test use. He refers to this as the consequential aspect of validity. An example of social consequences is the attempt by the Oregon Health Services Commission in 1991 to use QALYs to ration health care. This was the first large scale attempt to apply cost-effectiveness analysis to set priorities for approximately 1600 medical services. Services were to be covered in order of their appearance until the budget was exhausted. However, the preliminary rankings had unexpected consequences. For example, surgical treatments for life-saving events such as appendicitis were placed below tooth capping (Hadorn, 1991). Treatments for headaches and thumbsucking were ranked higher than treatments for cystic fibrosis and AIDS. These rankings were revised because they did not reflect public values, and yet a methodical and apparently rational procedure to create the rankings had been employed. This attempted application of cost-effectiveness analysis generated considerable debate and critique of the measurement process to understand what went wrong.

This section deals with issues surrounding the EQ-5D index score and its subsequent use to calculate QALYs. An EQ-5D index score is the result of a process of measurement, embedded in theory based on philosophy and societal values. A valid interpretation of scores can only take place within the context of these factors. Although researchers have addressed the ethical consequences and the shortcomings of the QALY (Gafni, 1989), no studies have examined the potential consequences of preference-based HRQL measures within a validity framework. The purpose of this discussion is to examine this aspect of validity in

the measurement of EQ-5D index scores and their potential use as QALYs.

Following a brief overview of Messick's conception of the consequential aspect of validity, the values and philosophical theory underlying the measurement and interpretation of EQ-5D index scores in their proposed use as QALYs are examined. Next, the issues in the social debate on the measurement of preference-based measures and their role as QALYs are reviewed. Finally, the consequential aspect of validity of the EQ-5D is assessed by examining how study findings can be used to anticipate potential adverse consequences of the applied use of the EQ-5D.

Messick's Conception of the Consequential Aspect of Construct Validity

Messick (1995) describes two interrelated and overlapping parts of the consequential aspect of construct validity: "the value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use, especially in regard to sources of invalidity related to issues of bias, fairness and distributive justice" (p. 745). Values are standards or principles of worth (Kaplan, 1964). The value implications of score interpretation include the constructs, the theories in which the construct is embedded, and the broader ideologies influencing the construct, such as the nature of humanity, society, and science. The issue is not that science should be value free but that the scientist should be aware of the impact of values on the questions we ask, the labels we attach to constructs, our scoring procedures, the content of the measure, and the meanings we attach to scores. What kinds of evidence are needed? Messick argues that validity itself is inherently a value judgment, which

links evidence and rational argument in the evaluation of inferences based on scores. "What serves as evidence is a result of a process of interpretation-facts do not speak for themselves" (Kaplan, 1964, p. 375). Values can be examined by open debate, by putting forth counter-hypotheses, by empirical findings, and by logical analysis (Messick, 1989).

The second part of the consequential aspect of validity addresses the functional worth of scores by evaluating the potential and actual consequences of test use. A significant question is: are the social consequences supportive of the intended testing purposes and consistent with other social values? (Messick, 1989). Critics argue that this aspect is better located as debate within the broader social milieu and not within the technical confines of psychometrics (Maguire et al., 1994), while Messick views social consequences as a form of evidence which reflects on the validity of inferences (p. 21). The latter point of view was adopted in the present study. To judge whether a measure serves its intended purpose requires an evaluation of the intended and unintended consequences of test use. Negative social consequences are often associated with bias or unfairness (Messick, 1995). In the context of Messick's argument, bias is "a prejudgment, a conclusion arrived at prior to the evidence and maintained independently of the evidence" (Kaplan, 1964, p. 375). Other sources of bias can derive from test construction, such as lack of relevance and representativeness of item content. A key concern of negative social consequences is when they derive from a source of test invalidity. For example, the unintended social consequences of the Oregon Plan were, in part, linked to

measurement error, thus reflecting on the validity of aspects of the measurement process (Eddy, 1991; Nord, 1993b; Tengs, 1996). As well, the process did not take into consideration the value people place on lifesaving technologies (Eddy, 1991). By anticipating these consequences, potential biases can be openly debated before they affect social policy.

What are some ways to approach the assessment of the social consequences of validity? Messick (1989) suggests the use of counterproposals to provide a context of debate. Potential consequences can be anticipated and weighed against other proposals or against the potential social consequences of not testing at all. Adverse consequences of test use can be examined as to whether they stem from test invalidity. Results in the present study, such as the effect of different weights on the calculation of QALYs, were used to assess the potential and actual consequences of EQ-5D use in this population and across populations.

Values and Philosophical Theory Underlying the Measurement of the EQ-5D

Measurement begins with a purpose. The primary purpose behind the development of preference-based HRQL measures, such as the EQ-5D, was to provide tools with which to answer the question, 'How do we provide equitable health care with limited societal resources?' Economists have proposed various methods of economic analyses to answer this question, one of which is cost-utility analysis based on providing the greatest effectiveness for a given cost. The EQ-5D was designed as a measure of HRQL which, when combined with duration of health, could be used as a measure of effectiveness. Although other

measures of effectiveness have been developed, the QALY is currently the accepted measure (Gold et al., 1996).

Underlying this purpose are principles of distributive justice which are normative principles designed to allocate societal goods in limited supply relative to demand. There are various philosophical theories of distributive justice, related to conceptions of the human good and to different ideas of man's dependence on society to realize the good (Taylor, 1985). Cost-effectiveness analysis is based on the idea that society wishes to maximize the total amount of goodness that can be produced with the available resources (Eddy, 1991; Weinstein & Stason, 1977). The theory of distributive justice underlying this idea is utilitarianism (Cubbon, 1991; Hadorn, 1991). The moral good to the utilitarian is human happiness or satisfaction and the utilitarian position is to maximize general happiness (Mill, 1991). The utilitarian principle for distributing economic benefits is to distribute them so as to maximize preference satisfaction.

The principles of utilitarianism are not without criticism. Critics argue that utilitarianism when applied to society may result in some people suffering or being sacrificed so that others may gain (Kymlicka, 1989; Rawls, 1971). Utilitarians believe that it is rational for society to sacrifice an individual's happiness to increase someone else's happiness if it maximizes overall social welfare (Kymlicka, 1989, p. 24). Critics argue that while sacrificing present happiness for later happiness to increase an individual's overall happiness may be rational for an individual, it is not rational for society. That is, one should not be asked to sacrifice their present happiness to increase someone else's future

happiness. Rawls (1971), in his criticism of utilitarianism, says that there should be limits on the sacrifices that can rightfully be asked in the name of the overall good. Finally, Taylor argues that reducing moral good to a single dimension does not take into consideration the diversity of goods that must go into normative thinking (Taylor, 1985).

One of the difficulties of implementing policy guided by cost-utility analysis is that the underlying theory of utilitarianism may be incompatible with other societal views of distributive justice. However, cost-utility analysis was designed to inform decision-making, not to replace it (Canadian Coordinating Office for Health Technology Assessment [CCOHTA], 1997). Proponents argue that the QALY approach, in spite of its limitations, is a reasonable approach to collective priority-setting (Williams, 1996), while others maintain that there is no proof that QALYs are better than doing nothing, and in fact that they might be worse than doing nothing (Gafni & Birch, 1993). How best to determine health priorities in the allocation of scarce resources is unresolved.

The QALY Debate

QALYS (in the context of cost-utility analysis) can be used in two ways. One is to determine which competing therapy might be used to treat a particular condition. The other is to determine which group of patients or which conditions should be given priority in the allocation of health care resources (Harris, 1987). It is the second use that is the most controversial. Typically, alternate programs or services are ranked from the lowest value to the highest and selected from the top until available resources are exhausted (Weinstein & Stason, 1977).

Attempts to use QALYs in this way have been largely unsuccessful. Although the QALY is the accepted measure of benefit in cost-utility analysis, there are a number of serious criticisms to the use of QALYs. Even so, in spite of the criticisms, Canadian guidelines for the economic evaluation of pharmaceuticals recommends the QALY as a measure of benefit in cost-utility analysis for comparability across programs of interventions involving pharmaceutical products (CCOHTA, 1997).

The QALY is based on the idea that a rational person would prefer a shorter healthier life to a longer period of survival in a state of severe discomfort and disability (Harris, 1987). A priority health care activity is one in which the cost-per-QALY is low (Harris, 1987). The QALY approach is based on an egalitarian position and contains a number of assumptions (CCOHTA, 1997):

1. All QALYs are regarded as of equal value to everybody, regardless of age, comorbidity, or other circumstances of the individual.
2. It is equally desirable to provide a small gain to many people or a large gain to a few, as long as the QALY totals are equal.
3. The preferences that individuals have for paths of changing health states can be reasonably estimated by adding up the time-weighted preferences that the individual has for the components of that path.
4. The relative weights for health states are independent of the duration of the health states.

Most of these assumptions have been challenged by rational argument and/or empirical findings. Although QALYs are assumed to be equal, the value

of life for each person is not necessarily treated as equal (Harris, 1987). For example, Hadorn (1991) argues that the cost-effectiveness approach is in conflict with the 'Rule of Rescue', a 'people's perceived duty to save endangered life whenever possible' (p. 2218). Harris (1987; 1991) cautions that QALYs are potentially ageist, racist, and sexist if used to determine which groups of patients to treat. Because QALYs take into account life expectancy, QALYs will favor the young. He argues that the QALY is potentially biased towards groups of people that will gain the maximum QALYs, such as people that have diseases that are relatively cheap to treat, or against groups that have conditions that are not QALY efficient. For example, the treatment for AIDS is expensive and if not QALY efficient, the QALY approach could systematically be biased against treating groups that have a higher incidence of AIDS.

Researchers have criticized the theoretical basis of QALYs, arguing that there is no theoretical or empirical basis for the assumption that the value of a health state for a given period of time is the 'value score' that the individual ascribes to that state multiplied by the time spent in that state (Gafni & Birch, 1993; Harris, 1987). Gafni and Birch (1993) argue that duration of a health state should be taken into account when measuring the value of that state and that the QALY is biased against interventions that are aimed at improving quality of life for short durations or interventions that are aimed at reducing minor side-effects over a very long period of time.

Finally, are we asking the right questions when we measure preferences and is the measure appropriate for the intended use? Eddy (1991) argues that

the questions that people are asked don't correspond to the use to which the answers are put. Questions are asked from the viewpoint of the individual, not society, and do not necessarily correspond to the intended use. For example, the EQ-5D asks people to value how good or bad a health state would be for them by imagining they are in a health state for a specified duration. However, the intended use of the valuations is to value health states of others so health outcomes can be compared and health services prioritized. It may be that given more information about the intended use of the valuation process, respondents would value the health states differently. Other methods of determining health benefits may be more appropriate, such as asking people more directly how they would allocate services (Eddy, 1991; Nord, 1993a). For example, Nord (1993a) interviewed a small sample of Norwegian citizens to measure their ethical preferences in prioritizing health care. He found that people emphasized equality in the value of life and the entitlement to treatment, rather than a better quality of life outcome. Some researchers have suggested a combination of patients and society's values may be more appropriate. As well, other methods have been developed for economic analyses, such as the person trade off and willingness to pay, but each has its own shortcomings (Drummond et al., 1997).

How can Validity Inquiry Assess the Potential Consequences of EQ-5D Use?

Potential consequences of the use of preference-based HRQL measures can be assessed through logical analysis as well as empirical findings. Apart from the ethical implications of the QALY approach, how can we consider the consequential aspect of validity in validity inquiry? The purpose of this section is

to use the study findings to examine some of the potential consequences of the interpretation and use of EQ-5D scores.

The EQ-5D index measure is the end result of a number of steps in the measurement process and is affected by a number of factors; the content and number of levels of the health state descriptive system, the methods used to derive the weights, and the scoring model for aggregating the weights into a single index score. The health state descriptive system determines the states that individuals are being asked to value. All of the EQ-5D items are relevant to patients with osteoarthritis, and each was responsive to TJA. Effect sizes of the EQ-5D index scores were moderate to large, compared with other studies where items may not have been as relevant for the particular population (Jenkinson et al., 1997; Jenkinson et al., 1998). If used to compare patient outcomes in different populations, the EQ-5D could potentially be positively biased towards interventions such as TJA, where the EQ-5D dimensions closely match those found in the WOMAC, a condition specific measure.

The broad levels of the EQ-5D include a large range of disability. This can affect preference measures, as the state imagined when assigning a value can have an equally broad range. For example, in the Oregon Plan, levels of health states descriptions were broad, thus 'trouble speaking' ranged from a mild lisp to mutism (Eddy, 1991). Similarly in the EQ-5D, respondents recording a 2 on mobility ranged from using no aids to walking to using a walker. The same broad interpretations would be likely to occur when the hypothetical health states are valued by the community. The consequence of spreading values over a broad

range in the EQ-5D could be more beneficial to some groups of patients. The extreme wording of the levels for mobility limits responsiveness in this dimension. For patients with limitations in mobility but with no pain, change may not be easily detected by the EQ-5D.

Items measuring more than one concept could also lead to a variety of interpretations in the valuation process. For example, usual activities measures work, study, housework, family, and leisure activities. Although the developers argue that its advantage is that it measures whatever is important to the individual, wide discrepancies in interpretation would likely exist, both with the valuation process and the self-rating of health states to which the weights are applied. Little is known about how the health states are imagined and interpreted and about the cognitive processes that occur during the valuation process.

The weights to be used for the EQ-5D are not standardized. Different countries involved in the development of the EQ-5D have each collected their own weights, using TTO and/or VAS methods. What are the potential consequences of the 'N3 model' and the method of determining weights? As can be seen from Table 25, QALYs gained using different weights yield quite different results with the TTO weights producing the most QALYs gained. The inclusion of the N3 term contributes to this large gain by spreading the scores. Although the effect size of index scores using the TTO weights is less than with VAS weights, the measure of change that is used (i.e., QALYS gained) is the largest. Because 'QALYs gained' do not take into account the standard deviation of scores, any method that lowers the QALYs of poor health states pre-intervention will

potentially increase QALYs gained. When used as a denominator in cost-utility analysis, the difference (QALY's gained) could have a substantive effect on the cost-utility ratio. Although cost-effectiveness studies usually carry out sensitivity analysis, no study has addressed this issue with the EQ-5D. In theory, when QALYs are used to compare the outcomes of health care interventions across programs, the same method of measurement should be used. In practice, there is currently a large variation in methods and HRQL measures. If QALYs were compared across programs using different methods, certain groups could potentially gain more QALYs with the EQ-5D N3 model.

As well as the weights applied to each dimension, the index score is also weighted by the number of items measuring each concept. For example, there are 3 items measuring function, one item measuring anxiety/depression, and one item measuring symptoms (pain/discomfort). The index score of a person with clinical depression but able to function adequately (i.e., 11113), would be 0.41 while the lowest score attained in this study of patients with osteoarthritis was -0.48. Because joint pain affects mobility, usual activities, and self-care, as well as mental health, scores for patients with osteoarthritis are likely to be lower. Consequently, the possible change score for mental health problems may be lower than for someone suffering from joint pain. No published studies have assessed responsiveness of the EQ-5D in patients with mental disorders.

Who should value the health states is controversial. Results from this study show a wide discrepancy in the value of health states by the community and the individuals experiencing those health states. This could be partly due to

a number of factors: the way the question is asked, the regression model (N3 term), the effect of duration, or by a lack of understanding of the health state to be valued. Variations in the way questions are asked has been shown to affect valuations. What is the right question to ask to determine preferences? Duration has been shown to have an effect on valuations. Yet the duration of health states for different methods of collecting valuations has varied. As well, little is known about how people's adaptation to illness would affect their valuation. The differences in values placed on health states by an uninformed public and the people experiencing the health states could have potential adverse consequences for those people with chronic conditions who may have adapted to their health state, but to society seem 'worse than death.'

Finally, the scoring model should be based on theory. Is there any theoretical rationale for the addition of dimensions? The cumulative model assumes a underlying unidimensional construct. Symptoms are not effect indicators of HRQL but causal indicators. That is, pain affects physical and psychological functioning, but pain is not necessarily an indicator of HRQL. The EQ-5D may not be sensitive to health technologies that treat patients with other symptoms, such as nausea or fatigue. Because pain is the only specific symptom measured, technologies that treat pain would likely result in more QALYS gained than would other treatments. Without theory to guide the development of the EQ-5D as a measure of HRQL, interpretation of findings is difficult. Without a theoretical basis of the interrelationship between HRQL dimensions the

consequences of applying the scores in real situations will be difficult to predict and difficult to interpret.

In summary, Messick proposes that the consequential aspect of validity should be part of validity inquiry. This includes the value implications of test interpretation and the potential and actual social consequences of test use. Measurement of HRQL takes place within a context of theory, philosophy, and societal values. The development of preference-based HRQL measures grew out of a societal need to allocate limited health resources. Principles of distributive justice guide how to distribute these resources in a fair and equitable manner. Cost-utility analysis is a method of determining health priorities through the ratio of cost per QALY. This is based on a theory of utilitarianism, which may not be congruent with societal principles of distributive justice.

Anticipation of adverse social consequences should guide instrument development and testing before HRQL tools are used to guide health policy. Potentially adverse consequences of EQ-5D index scores derive from the content and wording of levels, the method of determining weights, the method for aggregating weights, and the lack of a theoretical model. Until these issues have been resolved, empirical findings should be limited to assessing the reliability and validity of inferences in various populations and informing future research.

Measurement is a means to an end. It is the bridge between purpose and action. Flaws in the measurement process are not always evident until a measurement tool is used in a real situation. Not to consider value implications

and social consequences as part of validity inquiry is to miss an essential part of the validity puzzle.

Summary

Study findings were integrated with the literature and conceptual analysis to assess the construct validity of EQ-5D scores. Four aspects of construct validity were examined: substantive, structural, external and consequential. Although the EQ-5D is not based directly on theory, a number of theories underlie the measurement of HRQL: functionalism, welfare economics, utility theory, classical test theory, and utilitarianism. The five dimensions of HRQL measured by the EQ-5D reflect the three broad areas measured by competing instruments: physical, social, and psychological functioning. Evidence supporting construct validity included dimensions relevant to the patient population, an ability to respond to clinical change in an expected direction, and convergent validity. Areas of concern included the lack of a theoretical basis, clarity and wording of items, the incongruence with the interpretation of negative scores and patient values, the inconsistency of the scoring model with the theoretical model, and the effect of different weights on the calculation of QALYs.

CHAPTER VII

Conclusions, Study Limitations, and Future Research

Conclusions

Validation is a “continuous process that starts sometimes with a construct in search of proper measurement and sometimes with an existing test in search of proper meaning” (Messick, 1980, p. 1023). Assessing the construct validity of the EuroQol is a little of both. Results from this study support the responsiveness of the EQ-5D as an outcome measure in patients undergoing hip or knee replacement. As well, support for convergent validity is moderate, while discriminant validity evidence was difficult to interpret due to the overlapping of constructs and the theoretical relationships between different constructs. Exploratory data analysis in combination with a conceptual analysis raised a number of issues that threaten the validity of inferences of EQ-5D scores. A lack of a conceptual model for the EQ-5D health state descriptive system, lack of clarity in the EQ-5D items, and wording of extreme levels in mobility and self-care potentially affect the interpretability of EQ-5D profile and index scores. Incongruity between the scoring model and the dimensionality, the discrepancy between the index scores and the patient’s self-reported health, and the presence of a bimodal distribution threaten a valid interpretation of index scores. Finally, potentially adverse consequences of EQ-5D scores are affected by the content and wording of items, the methods of determining weights, the scoring model, and the lack of a theoretical model.

Few studies have examined the construct validity of the inferences drawn from the EQ-5D in depth. Although the EuroQol Group includes researchers with a variety of backgrounds including psychometrics, the EQ-5D was developed out of a health economics perspective. As a result it has not undergone the rigorous psychometric tests of reliability and validity conducted in the development of other HRQL tools such as the SF-36. In the few studies which have addressed validity issues of the EQ-5D, most have examined relationships with other instruments through methods such as correlational analysis. The issue of validity has barely been addressed in the EQ-5D literature and may be summed by the following excerpt from a paper written by EuroQol Group members.

Even supposing that a full set of values (or even utilities) for all the EuroQol states had been obtained, from a large representative sample, there will be those who return to the issue of validity. Much of this ritual seems to arise from the legacy of physical measurement in the world of laboratory science...Insofar as the measurement of weight can be treated as an analogue for human judgment, then the process of validating health state valuations might take a similar path. However, there are so many clear differences that the researcher might be forgiven for declining to embark on the process of investigating validity at all (Kind et al., 1994, p. 241).

This perspective on validity may be due, in part, to a conception of validity as correlation with a criterion, or gold standard. Due to increasing amounts of recent evidence that questions aspects of validity, issues such as the wording of mobility and self-care are currently being examined by members of the EuroQol

group (personal communication, EuroQol Scientific Meeting, November 6-7, 1999). In spite of the weak validity evidence, an increasing number of studies are being published using the EQ-5D in cost-effectiveness analysis. Although the results may not yet be used in policy decisions, the eagerness to use these tools in cost-utility analysis while there are still unresolved measurement issues is a concern. However, the measurement of HRQL arises out an important gap in the traditional outcome measures of mortality and morbidity, particularly with the increase in health technologies and people living with chronic illnesses.

Limitations of the Study

A weakness of this research project is that the data set used was not designed for the purpose of assessing the construct validity of the EQ-5D. Therefore, some questions of validity were not possible to assess with the available data. Secondly, some of the limitations of the EQ-5D may be related to limitations of the instruments with which it was compared and an imperfect match between comparable dimensions. However, a strength of the study is that it examined the EQ-5D in a real clinical situation rather than a laboratory situation where results may not be generalizable. As well, some of the findings led to questions which may not have otherwise been raised. In a tool that was designed to compare results across different patient groups, and different countries, it is essential to examine the test responses in the various populations where it could be used. Results from this study have added to the body of knowledge of the validity of EQ-5D scores in patients undergoing total hip and

knee replacements and have raised some measurement issues which have not been addressed in the literature.

Future Research

Although HRQL tools are in a relatively early stage of development, the measurement of HRQL has raised the awareness among health professionals that the patient's perspective of HRQL must be taken into account. Further areas for research include theoretical work on the conceptualization of HRQL and the causal network between dimensions. Also, when we measure preferences what are we really measuring and what is the best way to measure them? More research using qualitative techniques, such as protocol analysis, would give us more insight into the cognitive processes that people use in determining preferences thereby enabling us to determine the differences between methods and which method is best. Also, the gap between patient and societal weights is a cause for concern. Qualitative techniques could be used to capture patient's and societal interpretations of the health state dimensions. If there is a wide discrepancy in interpretation, then applying societal derived weights to patient health states would be invalid, as they are not measuring the same things. The EQ-5D health dimensions are broad, leading to possible different interpretations. However, very little is known about the cognitive processes of interpreting the health state and then valuing the health state. If the EQ-5D is to be used in cost-effectiveness analyses to compare across populations, then it must be tested in all populations in which it would be used. As an outcome measure it must be sensitive enough to detect important clinical outcomes. More research

needs to be done to test responsiveness in various populations, and to examine the factors that affect responsiveness, such as the relevance of items, and the distributions of responses. The validity of using the EQ-5D to compare across interventions such as mental health and physical interventions or chronic and acute conditions should be revisited and debated.

REFERENCES

Albrecht, G. L., & Devlieger, P. J. (1999). The disability paradox: high quality of life against all odds. Social Science and Medicine, 48, 977-988.

Badia, X., Diaz-Prieto, A., Rue, M., & Patrick, D. L. (1996). Measuring health and health state preferences among critically ill patients. Intensive Care Medicine, 22(12), 1379-84.

Badia, X., Fernandez, E., & Sequra, A. (1995). Influence of socio-demographic and health status variables on evaluation of health states in a Spanish population. European Journal of Public Health, 5(2), 87-93.

Bellamy, N. (1995). WOMAC user's guide. London, Ontario.

Bellamy, N., Buchanan, W. W., Goldsmith, C. H., Campbell, J., & Stitt, L. W. (1988). 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. Journal of Rheumatology, 15(12), 1833-1840.

Bennett, K. J., & Torrance, G. W. (1996). Measuring health state preferences and utilities: Rating scale, time trade-off, and standard gamble techniques. In B. Spilker (Ed.), Quality of life and pharmacoeconomics in clinical trials (2nd ed., pp. 253-265). Philadelphia: Lippincott-Raven.

Bergner, M., Bobbitt, R. A., Pollard, W. E., Martin, D. P., & Gilson, B. S. (1976). The Sickness Impact Profile: Validation of a health status measure. Medical Care, 14(1), 57-67.

Bollen, K. A. (1989). Structural equations with latent variables. New York: John Wiley & Sons.

Bombardier, C., Melfi, C. A., Paul, J., Green, R., Hawker, G., Wright, J., & Coyte, P. (1995). Comparison of a generic and a disease-specific measure of pain and physical function after knee replacement surgery. Medical Care, *33*(4), AS131-AS144.

Brazier, J., Jones, N., & Kind, P. (1993). Testing the validity of the Euroqol and comparing it with the SF-36 health survey questionnaire. Quality of Life Research, *2*(3), 169-80.

Brazier, J. E., Walters, S. J., Nicholl, J. P., & Kohler, B. (1996). Using the SF-36 and Euroqol on an elderly population. Quality of Life Research, *5*(2), 195-204.

Brooks, R. (1996). EuroQol: the current state of play. Health Policy, *37*(1), 53-72.

Brooks, R. G., Jendteg, S., Lindgren, B., Persson, U., & Bjork, S. (1991). EuroQol: health-related quality of life measurement. Results of the Swedish questionnaire exercise. Health Policy, *18*(1), 37-48.

Busschbach, J. J., Horikx, P. E., van den Bosch, J. M., Brutel de la Riviere, A., & de Charro, F. T. (1994). Measuring the quality of life before and after bilateral lung transplantation in patients with cystic fibrosis. Chest, *105*(3), 911-7.

Byrne, B. M. (1989). A Primer of LISREL Basic Applications and Programming for Confirmatory Factor Analytic Models. New York: Springer-Verlag.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56(2), 81-105.

Canadian Coordinating Office for Health Technology Assessment (CCOHTA). (1997). Guidelines for economic evaluation of pharmaceuticals: Canada. (2nd ed.). Ottawa: Canadian Coordinating Office for Health Technology Assessment.

Cella, D. F. (1992). Quality of life: The concept. Journal of Palliative Care, 8(3), 8-13.

Chetter, I. C., Spark, J. I., Dolan, P., Scott, D. J. A., & Kester, R. C. (1997). Quality of life analysis in patients with lower limb ischaemia: Suggestions for European standardisation. European Journal of Vascular & Endovascular Surgery, 13(6), 597-604.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. (2nd ed.). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Conner-Spady, B., Cumming, C., Nabholtz, J-M., Jacobs, P., & Stewart, D. (1999). Responsiveness of the EuroQol in breast cancer patients undergoing high dose chemotherapy. Manuscript submitted for publication.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Orlando: Harcourt Brace Jovanovich College Publishers.

Cubbon, J. (1991). The principle of QALY maximisation as the basis for allocating health care resources. Journal of Medical Ethics, 17, 181-184.

de Haan, R., Aaronson, N., Limburg, M., Hower, R. L., & van Crevel, H. (1993). Measuring quality of life in stroke. Stroke, 24(2), 320-7.

Dolan, P. (1996). Modelling valuations for health states: the effect of duration. Health Policy, 38(3), 189-203.

Dolan, P. (1997). Modeling valuations for EuroQol health states. Medical Care, 35(11), 1095-108.

Dolan, P., & Gudex, C. (1995). Time preference, duration and health state valuations. Health Economics, 4(4), 289-99.

Dolan, P., Gudex, C., Kind, P., & Williams, A. (1996a). The time trade-off method: results from a general population study. Health Economics, 5, 141-154.

Dolan, P., Gudex, C., Kind, P., & Williams, A. (1996b). Valuing health states: A comparison of methods. Journal of Health Economics, 15(2), 209-231.

Dolan, P., & Kind, P. (1996). Inconsistency and health state valuations. Social Science & Medicine, 42(4), 609-15.

Drummond, M. F., O'Brien, B., Stoddart, G. L., & Torrance, G. W. (1997). Methods for the economic evaluation of health care programmes. (Second ed.). Oxford: Oxford University Press.

Eddy, D. M. (1991). Oregon's methods Did cost-effectiveness analysis fail? Journal of the American Medical Association, 266(15), 2135-2141.

Essink-Bot, M. L., Bonsel, G. J., & van der Maas, P. J. (1990). Valuation of health states by the general public: Feasibility of a standardized measurement procedure. Social Science in Medicine, 31(11), 1201-1206.

Essink-Bot, M. L., Krabbe, P. F., Bonsel, G. J., & Aaronson, N. K. (1997). An empirical comparison of four generic health status measures. The Nottingham Health Profile, the Medical Outcomes Study 36-item Short-Form Health Survey,

the COOP/WONCA charts, and the EuroQol instrument. Medical Care, 35(5), 522-37.

Essink-Bot, M. L., Stouthard, M. E. A., & Bonsel, G. J. (1993). Generalizability of valuations on health states collected with the EuroQol questionnaire. Health Economics, 2, 237-246.

Essink-Bot, M. L., van Royen, L., Krabbe, P., Bonsel, G. J., & Rutten, F. F. (1995). The impact of migraine on health status. Headache, 35(4), 200-6.

The EuroQol Group. (1990). EuroQol—a new facility for the measurement of health-related quality of life. Health Policy, 16(3), 199-208.

Feeny, D., Furlong, W., Boyle, M., & Torrance, G. W. (1995). Multi-attribute health status classification systems: Health Utilities Index. Pharmacoeconomics, 7(6), 490-502.

Fitzpatrick, R., Fletcher, A., Gore, S., Jones, D., Spiegelhalter, D., & Cox, D. (1992). Quality of life measures in health care. I: Applications and issues in assessment. BMJ, 305(6861), 1074-7.

Fletcher, A., Gore, S., Jones, D., Fitzpatrick, R., Spiegelhalter, D., & Cox, D. (1992). Quality of life measures in health care. II: Design, analysis, and interpretation. BMJ, 305(6862), 1145-8.

Fox-Rushby, J. (1996). First steps to assessing semantic equivalence of the EQ-5D: Results of a questionnaire survey to members of the EuroQol Group. Paper presented at the EuroQol Plenary Meeting, Oslo.

Gafni, A. (1989). The quality of QALYs (quality-adjusted-life-years): Do QALYs measure what they at least intend to measure? Health Policy, 13(1), 81-83.

Gafni, A., & Birch, S. (1993). Searching for a common currency: critical appraisal of the scientific basis underlying European harmonization of the measurement of health related quality of life (EuroQol). Health Policy, 23(3), 219-28.

Gierl, M. J., & Rogers, W. T. (1996). A confirmatory factor analysis of the test anxiety inventory using Canadian high school students. Educational and Psychological Measurement, 56(2), 315-324.

Gold, M. R., Patrick, D. L., Torrance, G. W., Fryback, D. G., Hadorn, D. C., Kamlet, M. S., Daniels, N., & Weinstein, M. C. (1996). Identifying and valuing outcomes. In M. R. Gold, J. E. Seigel, L. B. Russell, & M. C. Weinstein (Eds.), Cost-effectiveness in health and medicine (pp. 82-134). New York: Oxford University Press.

Gorsuch, R. L.. (1983). Factor Analysis (2nd ed.). Hillsdale: Lawrence Erlbaum.

Guccione, A. A., Felson, D. T., & Anderson, J. J. (1990). Defining arthritis and measuring functional status in elders: Methodological issues in the study of disease and physical disability. American Journal of Public Health, 80(8), 945-949.

Gudex, C., Dolan, P., Kind, P., & Williams, A. (1996). Health state valuations from the general public using the visual analogue scale. Quality of Life Research, 5(6), 521-31.

Gudex, C. M., Hawthorne, M. R., Butler, A. G., & Duffey, P. O. F. (1997). Measuring patient benefit from botulinum toxin in the treatment of dystonia. Feasibility of cost-utility analysis. Pharmacoeconomics, 12(6), 675-684.

Gulliksen, H. (1950). Theory of Mental Tests. New York: John Wiley & Sons, Inc.

Hadorn, D. C. (1991). Setting health care priorities in Oregon. Journal of the American Medical Association, 265, 2218-2225.

Harris, J. (1987). QALYfying the value of life. Journal of Medical Ethics, 13, 117-123.

Harris, J. (1991). Unprincipled QALYs: a response to Cubbon. Journal of Medical Ethics, 17, 185-188.

Hollingworth, W., Mackenzie, R., Todd, C. J., & Dixon, A. K. (1995). Measuring changes in quality of life following magnetic resonance imaging of the knee: SF-36, EuroQol or Rosser index? Quality of Life Research, 4(4), 325-34.

Hunt, S. M., McEwen, J., & McKenna, S. P. (1985). Measuring health status: A new tool for clinicians and epidemiologists. Journal of the Royal College of General Practitioners, 35, 185-188.

Hurst, N. P., Jobanputra, P., Hunter, M., Lambert, M., Lochhead, A., & Brown, H. (1994). Validity of Euroqol--a generic health status instrument--in patients with

rheumatoid arthritis. Economic and Health Outcomes Research Group. British Journal of Rheumatology, 33(7), 655-62.

Hurst, N. P., Kind, P., Ruta, D., Hunter, M., & Stubbings, A. (1997). Measuring health-related quality of life in rheumatoid arthritis: validity, responsiveness and reliability of EuroQol (EQ-5D). British Journal of Rheumatology, 36(5), 551-9.

Jacobsson, S. A., Rehnberg, C., & Djerf, K. (1991). Risks, benefits and economic consequences of total hip arthroplasty in an aged population. Scandinavian Journal of Social Medicine, 19(1), 72-78.

James, M., St Leger, S., & Rowsell, K. V. (1996). Prioritising elective care: a cost utility analysis of orthopaedics in the north west of England. Journal of Epidemiology & Community Health, 50(2), 182-9.

Jenkinson, C., Gray, A., Doll, H., Lawrence, K., Keoghane, S., & Layte, R. (1997). Evaluation of index and profile measures of health status in a randomized controlled trial. Comparison of the Medical Outcomes Study 36-Item Short Form Health Survey, EuroQol, and disease specific measures. Medical Care, 35(11), 1109-18.

Jenkinson, C., Stradling, J., & Petersen, S. (1998). How should we evaluate health status? A comparison of three methods in patients presenting with obstructive sleep apnoea. Quality of Life Research, 7(2), 95-100.

Johnson, J. A., & Coons, S. J. (1998). Comparison of the EQ-5D and SF-12 in an adult US sample. Quality of Life Research, 7(2), 155-166.

Johnson, J. A., Coons, S. J., Ergo, A., & Szava-Kovats, G. (1998). Valuation of EuroQOL (EQ-5D) health states in an adult US sample. Pharmacoeconomics, 13(4), 421-433.

Joreskog, K., & Sorbom, D. (1996). LISREL 8: User's Reference Guide. Chicago: Scientific Software International.

Kaplan, A. (1964). The Conduct of Inquiry. San Francisco: Chandler Publishing Company.

Kind, P. (1996). The EuroQol Instrument: An index of health-related quality of life. In B. Spilker (Ed.), Quality of life and pharmacoeconomics (pp. 191-201). Philadelphia: Lippincott-Raven.

Kind, P., Dolan, P., Gudex, C., & Williams, A. (1998). Variations in population health status: Results from a United Kingdom national questionnaire survey. BMJ, 316(7133), 736-741.

Kind, P., Gudex, C., Dolan, P., & Williams, A. (1994). Practical and methodological issues in the development of the EuroQol: The York experience. In G. L. Albrecht & R. Fitzpatrick (Eds.), Advances in medical sociology (Vol. 5, pp. 219-253). Greenwich, CT: J. A. I. Press.

Krabbe, P. F., Essink-Bot, M. L., & Bonsel, G. J. (1997). The comparability and reliability of five health-state valuation methods. Social Science & Medicine, 45(11), 1641-52.

Kymlicka, W. (1989). Liberalism community and culture. Oxford: Clarendon Press.

Laupacis, A., Bourne, R., Rorabeck, C., Feeny, D., Wong, C., Tugwell, P., Leslie, K., & Bullas, R. (1993). The effect of elective total hip replacement on health-related quality of life. Journal of Bone & Joint Surgery American, 75(11), 1619-1626.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. Psychological Reports: Monograph supplement, 78-119.

MacDonagh, R. P., Cliff, A. M., Speakman, M. J., PJ, O. B., Ewings, P., & Gudex, C. (1997). The use of generic measures of health-related quality of life in the assessment of outcome from transurethral resection of the prostate. British Journal of Urology, 79(3), 401-8.

Maguire, T., Hattie, J., & Haig, B. (1994). Construct validity and achievement assessment. The Alberta Journal of Educational Research, XL(2), 109-126.

McHorney, C. A., Ware, J. E., Jr., & Raczek, A. E. (1993). The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. Medical Care, 31(3), 247-63.

Messick, S. (1980). Test validity and the ethics of assessment. American Psychologist, 35(11), 1012-1027.

Messick, S. (1989). Validity. In R. Linn (Ed.), Educational Measurement (3rd ed., pp. 13-103). New York: Macmillan Publishing Company.

Messick, S. (1995). Validity of psychological assessment. American Psychologist, 80(9), 741-749.

Michell, J. (1990). An Introduction to the Logic of Psychological Measurement. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Mill, J. S. (1991). On liberty and other essays. Oxford: Oxford University Press.

Nelson, E. C., Wasson, J. H., Johnson, D. J., & Hays, R. D. (1996). Dartmouth COOP Functional Health Assessment Charts: Brief measures for clinical practice. In B. Spilker (Ed.), Quality of life and pharmacoeconomics in clinical trials (pp. 161-168). Philadelphia: Lippincott-Raven.

Nord, E. (1991a). EuroQol: health-related quality of life measurement. Valuations of health states by the general public in Norway. Health Policy, 18(1), 25-36.

Nord, E. (1991b). The validity of a visual analogue scale in determining social utility weights for health states. International Journal of Health Planning & Management, 6(3), 234-42.

Nord, E. (1993a). The relevance of health state after treatment in prioritising between different patients. Journal of Medical Ethics, 19, 37-42.

Nord, E. (1993b). Unjustified use of the Quality of Well-Being scale in priority setting in Oregon. Health Policy, 24, 45-53.

Nord, E., Richardson, J., & Macarounas-Kirchmann, K. (1993). Social evaluation of health care versus personal evaluation of health states. Evidence on the validity of four health-state scaling instruments using Norwegian and Australian surveys. International Journal of Technology Assessment in Health Care, 9(4), 463-78.

Norum, J., Angelsen, V., Wist, E., & Olsen, J. A. (1996). Treatment costs in Hodgkin's disease: a cost-utility analysis. European Journal of Cancer, 32A(9), 1510-7.

O'Hanlon, M., Fox-Rushby, J., & Buxton, M. J. (1994). A qualitative and quantitative comparison of the EuroQol and time trade-off techniques. International Journal of Health Sciences, 5(3), 85-97.

Oleson, M. (1990). Subjectively perceived quality of life. IMAGE: Journal of Nursing Scholarship, 22(3).

Osoba, D. (1991). Effect of cancer on quality of life. Boca Raton, FL: CRC Press.

Patrick, D. L., Bush, J. W., & Chen, M., M. (1973). Methods for measuring levels of well-being for a health status index. Health Services Research, 8, 228-245.

Patrick, D. L., & Erickson, P. (1993). Health status and health policy. New York: Oxford University Press.

Patrick, D. L., Starks, H. E., Cain, K. C., Uhlmann, R. F., & Pearlman, R. A. (1994). Measuring preferences for health states worse than death. Medical Decision Making, 14(1), 9-18.

Pigou, A. C. (1960). The economics of welfare. (4th ed.). London: Macmillan and Company Ltd.

Ramey, D. R., Fries, J. F., & Singh, G. (1996). The Health Assessment Questionnaire 1995 - Status and review, Quality of life and pharmacoeconomics (pp. 227-237). Philadelphia: Lippincott-Raven.

Rawls, J. (1971). A theory of justice. Cambridge, Massachusetts: Harvard University Press.

Rissanen, P., Aro, S., Sintonen, H., Asikainen, K., Slati, P., & Paavolainen, P. (1997). Costs and cost-effectiveness in hip and knee replacements. International Journal of Technology Assessment in Health Care, 13(4), 575-588.

Rosser, R., & Kind, P. (1978). A scale of valuations of states of illness: is there a social consensus? International Journal of Epidemiology, 7(4), 347-58.

Rothwell, P. M., McDowell, Z., Wong, C. K., & Dorman, P. J. (1997). Doctors and patients don't agree: cross sectional study of patients' and doctors' perceptions and assessments of disability in multiple sclerosis. BMJ, 314(7094), 1580-3.

Schipper, H., Clinch, J. J., & Olweny, C. L. M. (1996). Quality of life studies: Definitions and conceptual issues. In B. Spilker (Ed.), Quality of life and pharmacoeconomics in clinical trials (2nd ed., pp. 11-23). Philadelphia: Lippincott-Raven.

Scott, W. A. (1968). Attitude measurement. (2nd edition ed.). (Vol. II). Reading: Addison-Wesley Publishing Company.

Sculpher, M. J., Dwyer, N., Byford, S., & Stirrat, G. M. (1996). Randomised trial comparing hysterectomy and transcervical endometrial resection: effect on health related quality of life and costs two years after surgery. British Journal of Obstetrics & Gynaecology, 103(2), 142-9.

Selai, C. (1994). The use of the EuroQol descriptive system (pages 2 & 3) with patients at the Institute of Neurology, London. Paper presented at the EuroQol Plenary Meeting, London.

Selai, C., & Rosser, R. (1995). Eliciting EuroQol descriptive data and utility scale values from inpatients. A feasibility study. Pharmacoeconomics, 8(2), 147-58.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. Educational Measurement: Issues and Practice. 16(2), 5-8, 13, 24.

Statistics Canada.(1990-1992). Life Tables, Canada and Provinces, (Cat. No. 84-537). Ottawa: Author.

Streiner, D. L., & Norman, G. R. (1995). Health measurement scales A practical guide to their development and use. (2nd ed.). Oxford: Oxford University Press.

Sutherland, H. J., Llewellyn-Thomas, H., Boyd, N. F., & Till, J. E. (1982). Attitudes toward quality of survival. Medical Decision Making, 2(3), 299-309.

Taylor, C. (1985). Philosophy and the Human Sciences: Philosophical Papers 2. (1st ed.). Cambridge: Cambridge University Press.

Tengs, T. O. (1996). An evaluation of Oregon's medicaid rationing algorithms. Health Economics, 5, 171-181.

Testa, M. A., & Nackley, J. F. (1994). Methods for quality-of-life studies. Annual Review of Public Health, 15, 535-59.

Torgerson, W. S. (1958). Theory and methods of scaling. New York: John Wiley & Sons, Inc.

Torrance, G. W. (1986). Measurement of health state utilities for economic appraisal: A review. Journal of Health Economics, 5(1), 1-30.

Torrance, G. W. (1996). Designing and conducting cost-utility analyses. In B. Spilker (Ed.), Quality of life and pharmacoeconomics in clinical trials (pp. 1105-1111). Philadelphia: Lippincott-Raven.

Torrance, G. W., & Feeny, D. (1989). Utilities and quality-adjusted life years. International Journal of Technology Assessment in Health Care, 5(4), 559-75.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. Science, 211, 453-458.

Uyl-de Groot, C. A., Hagenbeek, A., Verdonck, L. F., Lowenberg, B., & Rutten, F. F. H. (1995). Cost-effectiveness of ABMT in comparison with CHOP chemotherapy in patients with intermediate- and high-grade malignant non-Hodgkin's lymphoma (NHL). Bone Marrow Transplantation, 16, 463-470.

Uyl-de Groot, C. A., Vellenga, E., de Vries, E. G. E., Lowenberg, B., Stoter, G. J., & Rutten, F. F. H. (1997). Treatment costs and quality of life with granulocyte-macrophage colony-stimulating factor in patients with antineoplastic therapy-related febrile neutropenia: Results of a randomised placebo-controlled trial. Pharmacoeconomics, 12(3), 351-360.

van Agt, H. M., Essink-Bot, M. L., Krabbe, P. F., & Bonsel, G. J. (1994). Test-retest reliability of health state valuations collected with the EuroQol questionnaire. Social Science & Medicine, 39(11), 1537-44.

Vellenga, E., Uyl-de Groot, C. A., de Wit, R., Keizer, H. J., Lowenberg, B., ten Haaft, M. A., de Witte, T. J. M., Verhagen, C. A. H., Stoter, G. J., Rutten, F. F. H.,

Mulder, N. H., Smid, W. M., & de Vries, E. G. E. (1996). Randomized placebo-controlled trial of granulocyte-macrophage colony-stimulating factor in patients with chemotherapy-related febrile neutropenia. Journal of Clinical Oncology, 14(2), 619-627.

von Neumann, J., & Morgenstern, O. (1944). Theory of games and economic behavior. New York: John Wiley & Sons.

Ware, J. E. (1987). Standards for validating health measures: Definition and content. Journal of Chronic Diseases, 40(6), 473-480.

Ware, J. E., Kosinski, M., & Keller, S. D. (1994). SF-36 physical and mental health summary scales: A user's manual. Boston, MA: Health Assessment Lab, New England Medical Center.

Ware, J. E., Kosinski, M., & Keller, S. D. (1996). A 12-Item Short-Form Health Survey (SF-12): Construction of scales and preliminary tests of reliability and validity. Medical Care, 34(3), 220-233.

Ware, J. E., & Sherbourne, C. D. (1992). The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Medical Care, 30(6), 473-83.

Weinstein, M. C., Siegel, J. E., Gold, M. R., Kamlet, M. S., & Russell, L. B. (1996). Recommendations of the Panel on Cost-effectiveness in Health and Medicine. Jama, 276(15), 1253-8.

Weinstein, M. C., & Stason, W. B. (1977). Foundations of cost-effectiveness analysis for health and medical practices. The New England Journal of Medicine, 296(13), 716-721.

- Williams, A. (1974). Measuring the effectiveness of health care systems. British Journal of Preventative and Social Medicine, 28, 196-202.
- Williams, A. (1993). The importance of quality of life in policy decisions. In S. Walker, R. & R. M. Rosser (Eds.), Quality of life assessment: Key issues in the 1990s (pp. 427-439). Lancaster, UK: Kluwer Academic Publishers.
- Williams, A. (1995). The measurement and valuation of health: A chronicle. England: Centre for Health Economics, University of York.
- Williams, A. (1996). QALYs and ethics: A health economist's perspective. Social Science in Medicine, 43(12), 1795-1804.
- Wolfe, F., & Hawley, D. J. (1997). Measurement of the quality of life in rheumatic disorders using the EuroQol. British Journal of Rheumatology, 36(7), 786-93.
- Wood-Dauphinee, S. (1992). Quality of life as a rehabilitation outcome: are we missing the boat? Canadian Journal of Rehabilitation, 6(1), 3-12.
- Zigmond, A. S., & Snaith, R. P. (1983). The Hospital Anxiety and Depression Scale. Acta Psychiatrica Scandinavica, 67, 361-370.