Are you better?
Describing and Explaining Changes in Health Status in Persons with Upper-Extremity Musculoskeletal Disorders.

By

Dorcas Eleanor Beaton, BScOT, MSc

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy
Graduate Department of Medical Sciences
University of Toronto

© Copyright by Dorcas E Beaton, 2000
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.

0-612-49951-0

ABSTRACT

“Are you better?” is a fundamental question for both clinical practice and clinical research. Measuring changes in health status questionnaire scores is one way we attempt to answer the question in a standardized way, but how do we know if they do it well? The ability of a questionnaire to accurately detect change when it has occurred is called responsiveness. This is tested by examining change score in patients deemed to be improved by some reasonable indicator of true change. This thesis shows that the description of an instrument's responsiveness is dependent on the indicator used. Measures are not “responsive” but are “responsive to...”. A taxonomy is developed describing different types of change (i.e., important change) and different perspectives that can be taken (patient, clinician). Each category is a different kind of responsiveness.

One perspective, the patient’s, is examined in detail in a qualitative study. The results show that being “better” is highly individualized, and may have little to do with a change in health state. This perspective is operationalized in a cohort study evaluating if different categories of change (cells in the taxonomy) result in different descriptions of an instrument’s responsiveness. The results show a high degree of variability.

“Are you better?” is challenging to measure, but it is usually the question we are trying to answer when health status questionnaires are used in evaluative manners. Responsiveness studies need to be designed with attention to the indicator being used for change, which should reflect the target application.

-ii-
ACKNOWLEDGMENTS

There are many people to thank for their help with this thesis that is the product of a journey over the course of many years.

First of all my committee, for many great discussions and for reining me in when necessary. My supervisor, Claire Bombardier, a wealth of knowledge and a forward thinker, always finding places where the work could be applied. James Wright, for his availability, his attention to clarifying meaning, and his patience -- enduring many fruitful debates that really formed the taxonomy. Jeffrey Katz, whose experience and good nature (along with Holly Fossel's) made the cohort data collection move along smoothly, and made it so enjoyable to talk about its progress and its analysis. And Valerie Tarasuk, a wonderfully opinionated spirit who pushed me to defend almost each component of the qualitative theory as it developed. I would like to also thank the examiners for their time and the thought they put into the review of this thesis: Dr. Jack Williams, Dr. Jacek Kopec, Dr. Susan Jaglal, Dr. Linn Holness, and Dr. George Wells.

Second, my family. In particular my father, George Beaton, for his example of dedication to understanding the breadth of a subject. My mother, Mary Clarke Beaton, who always encouraged me to do my best. My sister Pat and brother-in-law Merle who gave me hours of encouragement long distance and phud to help along the way. The Kincardine crew: my brother Jim, sister-in-law Dawne and “the kids”: Patrick, Kathryn, Scott and Heather who helped keep me aware of the world outside of school (in particular of parades), and shared their home for writing weeks. And Granny, who always thought I had already done my doctorate.

Third, my supports: my friends, the dog-walkers, the Monday night gang, and the Wednesday night yoga class -- for their interest in my work, and in my leisure. And a few of my ‘rocks’ in particular: Janet, Lillian, Bill, Anna, Kate, Carolyn, Michael, Jo-Anne and Trudy. I thank you most sincerely, and I hope I’ll see you in Mahone Bay!

And also my supports at work. My friends, colleagues and mentors at the Institute for Work & Health; Dr Robin Richards, Elaine Harniman and the therapists from St. Michael’s Hospital; and Dr. Jack Williams.

Finally, the patient companion who was with me through it all, Wunder Kunde Hund, my wonderful smart dog. She kept me on a schedule and was a constant reminder for me to take breaks, to play ball, and to celebrate.

I would like to acknowledge the support of the Medical Research Council of Canada who supported me with a PhD Fellowship in Health Research; the Institute for Work & Health for their support; and the Toronto Star and Southern Ontario Newspaper Guild for their support of the qualitative study.
# Table of Contents

Abstract ........................................................................................................................................... -ii-

Acknowledgments ................................................................................................................................. -iii-

Table of Contents ................................................................................................................................. -iv-

List of Tables ......................................................................................................................................... -viii-

List of Figures ....................................................................................................................................... -ix-

List of Appendices ............................................................................................................................... -xii-

Chapter One: Introduction ........................................................................................................................ 1
  1.0 Introduction .................................................................................................................................... 1
  1.1 Interest in change in health status ................................................................................................. 3
  1.2 Describing and explaining changes in health status ........................................................................ 5
  1.3 Study population ............................................................................................................................ 8
  1.4 Purpose and objectives of the thesis .............................................................................................. 11
  1.5 Structure of the thesis. .................................................................................................................... 12

Chapter Two: A taxonomy for responsiveness ........................................................................................ 15
  2.0 Preface .......................................................................................................................................... 15
  2.1 Introduction .................................................................................................................................... 15
  2.2 Methods ......................................................................................................................................... 19
  2.3 Results. .......................................................................................................................................... 21
    2.3.1 Building a taxonomy of responsiveness .................................................................................. 21
    2.3.2 The Who Axis: Who is being looked at? Individual-level or group-level measurement ............. 22
    2.3.3 The When Axis: When is the comparison being made? .......................................................... 25
    2.3.4 The “What” Axis: What type of change is being quantified in a study of responsiveness ............ 27
      2.3.4.1 Minimum change potentially detectable by the instrument .............................................. 27
      2.3.4.2 Minimum change detectable given the measurement error of the instrument (“minimally detectable change”). ........................................................................................................................................... 28
      2.3.4.3 Change observed between measurements in a given population (“Observed change”) .......... 32
      2.3.4.4 Observed change measured by the instrument in a population deemed to have improved (“estimated change”) ........................................................................................................................................... 33
      2.3.4.5 Observed change measured by an instrument in a population........................................... 34
Appendices ................................................................. 153
Reference List ............................................................. 226
LIST OF TABLES

Table 2.1. Examples of definitions of responsiveness found in the literature. .................. 17

Table 2.2: Application of the classification system using published studies of responsiveness
from the literature. ................................................................. 38

Table 4.1 Attributes for indices of change ................................................................. 81

Table 4.2. Responses to the status indices ................................................................. 89

Table 4.3. Frequency of responses (out of a possible 30) to status measures of pain ........ 89

Table 4.4. Transitional indices response patterns. ....................................................... 90

Table 5.1. Types of change used as indicators of the occurrence of change .................. 103

Table 5.2. Sociodemographic data on the study participants ......................................... 113

Table 5.3. Clinical characteristics of study participants in the cohort study .................. 114

Table 5.4. Description of DASH scores using all available subjects at each data collection point
........................................................................................................ 116

Table 5.5. Percent of sample responding to each response category for overall rating of problem
........................................................................................................ 121

Table 5.6. Percent of sample responding to each response category at 12 weeks of follow up
........................................................................................................ 121

Table 5.7. Correlations (Spearman) between the different external markers of change ...... 123

Table 5.8 Summary of four-week responsiveness of DASH outcome measure scores ...... 127

Table 5.9. Summary of responsiveness of the DASH to different categories of change at 12-week
follow-up ......................................................................................... 128

Table D.1. Construct validity of the DASH. ................................................................. 190
Table D.2. Responsiveness of the DASH to clinical changes. .................. 192
Table D.3. Comparing the responsiveness between DASH and two joint-specific measures . . 193
Table D.4. Correlation between change in DASH score and change in self-ratings of pain.

function and severity of upper-limb problem measured in two ways ............... 194
Table D.5. Sensitivity and specificity of different levels of change .................. 196
Table E.1 Response patterns to five questions of importance ......................... 214
Table F.1. Standardized response means (SRM) across instruments .................. 220
LIST OF FIGURES

Figure 1.1 An example of the expression of change scores using summary responsiveness statistics ............................................................ 1

Figure 1.2 A theoretical model for responsiveness based on work of Testa(1996). ................ 4

Figure 2.1 Flow chart describing literature reviewed ........................................................ 20

Figure 2.2 The “Who” Axis. ................................................................................ 22

Figure 2.3 The “When” axis. Answering when in time the comparison is being made. ....... 25

Figure 2.4 The “What” axis, answering what type of change is being examined ............... 28

Figure 2.5 The taxonomy of responsiveness ................................................................. 37

Figure 3.1 Construction of being better. ..................................................................... 55

Figure 3.2 A description of the process of formulating a response to the question “Are you better?” ................................................................................ 61

Figure 5.1 Profiles of the baseline health status of sample using the Short Form-36 questionnaire .................................................................................................................. 114

Figure 5.2 Distribution of baseline DASH score ............................................................. 115

Figure 5.3. Bland-Altman limits of agreement ............................................................... 117

Figure 5.4 Number of subjects in each of the 10 point score ranges for the DASH at baseline and 12 week follow up point ................................................................. 117

Figure 5.5 DASH scores in individuals at three data collection times ......................... 119

Figure 5.6 Change scores in the DASH related to different kinds of change (“what” axis in taxonomy). ............................................................................................... 129

Figure D.1. Baseline distribution of DASH Outcome Measure Scores ................... 188

-x-
Figure D.2. Distribution of observed change score in the DASH ........................................... 191

Figure D.3 Standardized response means for three outcome measures and for three types of 12-week change ........................................................................................................ 194

Figure D.4 Receiver Operator Characteristic (ROC) Curve ...................................................... 195

Figure E.1. Distribution of importance ratings ................................................................. 212

Figure F.1 Graphic depiction of responsiveness of four instruments to different types of change .............................................................................................................................. 221
LIST OF APPENDICES

Appendix A: A non-exhaustive list of available upper extremity self-completed questionnaires. (n=30) ................................................................. 154

Appendix B: Items used in pilot testing of indices of responsiveness ....................... 155

Appendix C: Example of follow-up questionnaire package used in the cohort study. ................................................................. 165

Appendix D: Measuring the whole or the parts? Validity, reliability & responsiveness of the DASH Outcome Measure in different regions of the upper extremity .......... 179

Appendix E: Indices used as external markers for change in the cohort study .......... 203

Appendix F: Application of the taxonomy to other measures fielded in this study .... 215

Appendix G: Guidelines for questions to ask of a study of responsiveness. ............... 224

Appendix H: Qualitative interviews and documentation. ........................................ 225
Chapter One
INTRODUCTION

1.0 Introduction

Clinicians and clinical researchers, even health policy makers, spend a lot of time describing and explaining changes in health of patients. In fact, it is often their role to induce, evaluate and act on them. Change, or the lack of it, indicates whether a treatment has worked or whether it is preferable to another treatment approach; change also helps guide subsequent intervention.

Instruments, or outcome measures, are widely used to quantify health attributes. For example, goniometers measure joint motion, rulers measure length, and self-report questionnaires can measure health status. The responsiveness of such an instrument reflects its ability to detect change, and is considered to be the most important attribute of a questionnaire if the goal is to evaluate change in the target attribute over time (MacKenzie et al. 1986; Deyo et al. 1984; Meenan et al. 1984; Guyatt et al. 1987).

Verification of the responsiveness of a particular outcome measure involves the use of relevant statistics such as Cohen's effect size (Cohen, 1988). Clinicians and

Figure 1.1 An example of the expression of change scores using summary responsiveness statistics
researchers wishing to make use of the findings from the literature must be able to interpret data presented in those terms. Figure 1.1 shows two estimates of responsiveness showing the standardized response mean (SRM, a special version of Cohen’s effect size statistic for paired data and a commonly used summary statistic of responsiveness) across the horizontal for each of two estimates: A and B. Estimate B (SRM = 1.27) exceeds Cohen’s guidelines for a “large” effect size (0.80), whereas estimate A (SRM = 0.13) would be considered a “small” effect size (Cohen, 1988). The initial impression that B is more responsive might be correct if the results came from a head to head comparison of A and B in the same patients who were known to have improved. But what if A and B were estimates from the same instrument as indeed they were in this example. Such an example exemplifies the complexity of responsiveness (Hays et al., 1993) and raises the question of how indicators of the responsiveness of a measurement instrument can be so divergent.

This thesis will investigate that question by describing and explaining changes in health status (the target attribute) in one patient group—individuals with upper-limb musculoskeletal disorders. As used here, change refers to “any variation in the quantity or quality of an entity’s attributes” (Burke et al. 1990)(pg.4). Health status (or health-related quality of life) refers to the components of quality of life attributable to health (Ware, Jr., 1987; Bergner, 1989; Faden et al., 1992) where health is physical, social and emotional well being (World Health Organization, 1980; Ware, Jr., 1987).

Health status is measured with self-completed questionnaires. The description and explanation of how it changes is therefore essentially an exercise in understanding and interpreting the responsiveness of those questionnaires. In doing that here, it is hoped that light
will be shed on the divergent descriptions of the property illustrated in Figure 1.1. The present introduction is also intended to place this work within the broader context of health status measurement.

1.1 Interest in change in health status.

Measures of health status (or health-related quality of life) have become increasingly popular over the past 10-20 years (Fitzpatrick et al. 1992) as we move towards what Relman has called the era of accountability in health care (Relman, 1988; Guyatt et al. 1997; Rosser, 1993; Jette, 1995; Lepere et al. 1997; Testa et al. 1994; Wood-Dauphinee, 1999; Ware, 1993; Gill et al. 1994; Faden et al. 1992). Valid and reliable health status questionnaires (Lohr et al. 1996; Hays et al. 1993; McDowell et al. 1996) are among the few tools clinicians and researchers can use to quantify patient’s perceptions of their health in such a way that the data can be used in statistical models, or as a means of describing the health/illness of groups of people (Jenkinson, 1995).

Such applications operate under two assumptions: first that the questionnaires are valid measures of health, and second that they can accurately detect changes in health. Under these circumstances, they have the potential to allow the measurement of change in a patient’s perception of their health. Change from the patient’s perspective could then be used to determine treatment effectiveness (Ware, Jr. et al. 1981; Jenkinson, 1995), set boundaries for the occurrence of change and of no change (Guyatt et al. 1987; Feinstein, 1998; Brant et al. 1999), and monitor an individual’s or a group’s progress and response to treatment (Stratford et al. 1999; Ware, Jr. et al. 1981). Clinicians, researchers, program planners and health policy makers would also be able to take account of the patient’s perspective when making decisions.
Although interest in changes in health status has a long history (Rosser, 1993) a focus on the practice of measuring and interpreting it has been a more recent development (McHorney, 1999; Gill et al. 1994; Bergner, 1987; Patrick et al. 1990). Several challenges or barriers have had to be addressed along the way. At one point, the limiting factor was probably been the availability of questionnaires (Deyo et al. 1984; Meenan et al. 1984). However many different questionnaires are now available -- arguably too many (Pocock, 1991). For example, in the population of interest here (people with upper-limb musculoskeletal disorders) there are 30 different disease-specific questionnaires to choose from, 13 of them designed for measuring functional status in persons with shoulder disorders alone (see Appendix A). Studies show evidence of their validity and reliability and in many cases now, their responsiveness to certain clinical changes. The new challenge is therefore not the availability of a questionnaire, nor of the apparent evidence of its reliability, validity and responsiveness. As Guyatt states the new challenge facing us is the 'interpretability' of responsiveness statistics, and that remains elusive (Guyatt et al. 1994). We now have many questionnaires with numeric summaries of change, that as yet have little meaning for us (Fitzpatrick et al. 1992; Guyatt et al. 1997; Guyatt et al. 1994; Bergner, 1989; Gill et al. 1994; Liang, 1995), or are confusing. There is a great interest in measuring change, and many examples of how it is used for determining the statistical significance of treatment effects, or in documenting the changes in scores observed in individuals patients. However, neither of these attach meaning to the measured change in scores. In fact, the practical use we can make of these changes in scores, for example in health care decision making, is limited by our inability to interpret the numbers produced.
1.2 Describing and explaining changes in health status.

Testa provides a framework for the study of responsiveness and allows the present thesis to be set in the context of similar work (Testa et al. 1996). A modified version of Testa's model (focusing on health measurement rather than quality of life, and opening up the inner triangle in her original figure) is shown in figure 1.2. The left vertical axis indicates a measure of true health status (a multifaceted, latent trait that we will never be able to directly observe or measure) and the right vertical axis depicts health status instruments as a measure for this latent trait (Z). As with any instrument, and particularly one assessing latent traits, measurement includes error. Observed change in health status scores would be shown on this axis as a change from $Z_0$ to $Z_1$. Theoretically, and indeed practically, we believe or assume that observed change directly reflects true change. However, Testa (Testa et al. 1996) then describes the role of the "criterion" measure, the small intermediate measure between true and observed change. Because true change is unobservable, an observable proxy ($\Delta C$) is used as a proxy or an external marker for true change where there is some hypothesized relationship between the proxy and true change in the attribute. Studies of

Figure 1.2 A theoretical model for responsiveness based on work of Testa(1996). See text for detailed discussion of components.
responsiveness then look at the relationship between the external markers of change (ΔC) in figure 1.2 and change in the observed score (ΔZ). Indeed, because of the central role this marker plays in studies of responsiveness, it may be worthwhile considering it a "criterion" (Deyo et al. 1986; Testa et al. 1996; De Bruin et al. 1997). However, in this thesis the term external marker or proxy of true change will be used to avoid the assumption that this marker is able to capture true change in health status. There is no such "criterion" or "gold standard", and it is in fact the absence of such a "gold standard" that makes the interpretation of change scores so challenging (Lachs, 1993; Liang, 1995; Deyo et al. 1986; Testa et al. 1996; Deyo et al. 1992; MacKenzie et al. 1986). This model helps us to realize that the description of responsiveness is dependent on the ΔC built into the study.

Two approaches to research into responsiveness can be seen, though the division is somewhat artificial, as both depend on each other for their meaning or their descriptions. Approaches can be more statistical in nature, or more conceptual (Liang, 1995). The statistically-oriented research focuses on the outputs: the numbers produced by these questionnaires when data is collected over time (ΔZ, Z₀, Z₁). This important area of research would include work on the mathematics of quantifying change, such as measurement error (Nunnally, 1975; Cronbach et al. 1970; Healy, 1989; Vollmer, 1988; Streiner et al. 1991a; Nunnally et al. 1994), scaling assumptions (Nunnally et al. 1994; Stucki et al. 1996; Michell, 1986), modeling change (Norman et al. 1997; Testa et al. 1994; Rogosa et al. 1982) and work on the statistics used to quantify responsiveness (Wright et al. 1998; Murawski et al. 1998; Wyrwich et al. 1999a). In contrast, the conceptual work of responsiveness research would be that which focuses on the relationship between the component parts of responsiveness (figure 1.2), true change, the marker of change
and the observed change scores.

The present thesis focuses on the conceptual component, and it will seek to answer questions such as: What do patients mean when they say they have changed? How does one determine that change has occurred before measuring responsiveness? How does that relate to the concept of true change in health status? And how dependent is the description of responsiveness (numerically) on that marker of change ($\Delta C$) in Testa's model (Testa et al. 1996)?

This conceptual focus might be considered more "upstream" as it is an attempt to understand the relationships of the underlying components of a study of responsiveness and elucidate how the relationship between the components determines the resultant effect size statistics. Upstream work is challenging because of the complex, latent nature of self-perception of health status (Hyland, 1993; Leplege et al. 1997), and of changes in it (Allison et al. 1997; Bandura, 1997; Eisenberg, 1980; Leplege et al. 1997; Stiles, 1996). It may take several different methodological approaches to even partially describe the relationships. For instance in this thesis, the relationship between the external markers and true change was examined using qualitative methods. The relationship between the same markers and the numeric output (responsiveness statistics) was described using quantitative methods.

There are a wide variety of approaches that could be used to examine the concept of change, either focusing on the statistical estimations, or the conceptual meaning. It might be fair to conclude that at present our ability to describe changes in health numerically (Wood-Dauphinee, 1999) exceeds our ability to explain their meaning (McHorney, 1999; Liang, 1995; Nunnally, 1975; Greenfield et al. 1992). Unfortunately practical use of the perspective of the patient that we gain through these measures will only be achieved when the scores can be
interpreted (Testa et al. 1996; Greenfield et al. 1992; Liang, 1995).

The present thesis aims to try to understand the role of the external markers of change in the study of responsiveness. This will be both in terms of the relationship between these markers and the underlying construct of true change, and in terms of the effect the markers have on the numeric estimates of change and responsiveness of a given instrument. The implications of the findings on interpreting studies of responsiveness will also be described.

One particular patient population is focused on in order to restrict the investigation to the measurement of similar (somewhat consistent) clinical phenomena. The choice of people with upper-limb musculoskeletal disorders is explained below.

1.3 Study population

The responsiveness of health status measures is best studied in populations with a health burden that is: 1) detectable in terms of the target construct, health status; 2) attributable to the disorder being treated; and 3) likely to vary with treatment. That is, it is a state (variable) rather than a trait of the individual. Long-term sequella of injury or disease are less likely to change over time, and hence though detectable and attributable, would not be a disorder group well suited to the study of responsiveness.

Upper-extremity musculoskeletal disorders meet all three criteria outlined above. First, the burden is detectable. Using survey data gathered in the province of Ontario and nationally (Badley, 1992; Badley et al. 1994; Badley et al. 1992), it has been estimated that 20.6% of Canadian adults over 16 years of age reported long term health problems due to musculoskeletal disorders (Badley, 1992) and that it accounts for 10% of short term disability and 30% of long-
term disability. Musculoskeletal problems are the primary reason for persons in the working age group to consult a health care professional (Badley, 1994). The burden of musculoskeletal disorders, particularly in the upper limb, has also become an issue at workplaces. For example, in a recent workplace study at a large urban newspaper over 50% of the work force was reported to have had upper limb pain, and 20.5% had pain meeting the National Institute of Occupational Safety and Health (NIOSH) criteria for having a work-related musculoskeletal disorder (Polanyi, 1997). The burden personally, socially and economically is clear.

However, markers such as case status, disability claimants’ benefits and work absence only provide part of the picture of burden. Other work has shown that while only 15% of those NIOSH cases (Polanyi, 1997) had missed work due to their arm pain, 81% reported having difficulty in some aspect of their daily activities. Health status measures capture this aspect of the burden — the difficulties experienced from the perspective of the patient — that would otherwise be missed in broader measures. Studies have also shown that the burden of upper-limb musculoskeletal conditions is detectable using available instruments be they generic (Rozencwaif et al. 1998; Gartsman et al. 1998a; Gartsman et al. 1998b) or disease-specific measures (Katz et al. 1994; McConnell et al. 1999; Kirkley et al. 1998; Nvasarikar et al. 1999; Turchin et al. 1998). Change in the burden is also noted to be detectable, however much more so with the disease-specific measures than with the generic Short Form-36 or Short Form-12 item health measures (Beaton et al. 1997; Beaton et al. 1998; Kirkley et al. 1998; MacDermid et al. 1999a), a difference that appears to be more pronounced in the upper limb than in comparisons made between generic and disease specific measures in lower limb or lower back pain. Thus, to varying degrees, the burden and change in burden of upper limb musculoskeletal disorders is
detectable, meeting the first criterion.

The second criterion for an optimal study population, that they should have a burden that is attributable to the disorder under treatment, is also met. Because disorders of the musculoskeletal system are often isolated (for example, isolated pain in the shoulder or in the elbow) the burden (largely manifested in terms of function and symptoms, such as limitations in dressing or reaching) is often easily attributable to the disorder itself. The exception would of course be persons with a comorbidity that also interferes with physical function or causes pain independent of musculoskeletal-system impairment. However, it is fairly easy to obtain a sample of people with relatively isolated musculoskeletal disorders.

Finally, the burden associated with many of these disorders can be alleviated with treatment. In the cohort study described in Chapters Five and in Appendices D and E, a 12-week follow up after onset of treatment (operative or non-operative) sufficed to allow patients to describe at least an initial improvement in health status.

Upper-extremity disorders were selected over either lower extremity or low back pain for several reasons. First, because they are less well studied already and therefore present an opportunity to gain new knowledge. Second, they are unique in that the upper limb functions as a single kinetic chain positioning the hand for prehensile and non-prehensile tasks (Davis. 1999). Furthermore, the upper limb can also compensate for injured components along the chain, not only within the extremity but also across extremities. for example changing dominance might reduce difficulties associated with writing.

The final consideration in the decision to work within this study population was the opportunity to apply a new and unique instrument, the DASH Outcome Measure, released in
1996. The DASH (Disabilities of the Arm, Shoulder and Hand) is a self-completed whole extremity measure that, though only recently showing evidence of its responsiveness in formal testing (Kirkley et al. 1998; MacDermid et al. 1999a), was developed with great attention to its intended concept (disability and symptoms), purpose (discriminative and evaluative) and population (any or multiple disorders of the upper limb) (Hudak et al. 1996; McConnell et al. 1999). Repeated use of this measure will improve our understanding of a potentially useful measure developed with the conceptual underpinnings often lacking in available instruments (McHorney, 1999).

A population of persons with upper-extremity musculoskeletal disorders is thus well suited to the study of responsiveness. As described above, the burden of their disorder is detectable, attributable and likely amenable to treatment. Furthermore, they pose unique challenges in the measurement of health status and might be expected to benefit from efforts to describe and explain change.

1.4 Purpose and objectives of the thesis

The overall purpose of this thesis is to advance the understanding of responsiveness by describing and explaining change in persons with upper-extremity musculoskeletal disorders. Attention is specifically focused on the concept of change and how this concept impacts on the numeric description of a given instrument’s responsiveness. Within this framework this thesis has four specific objectives:

1. To develop a framework of responsiveness based on a consolidation of the literature;
2. To explain what change (being better) means from a patient’s perspective;
3. To develop methods to integrate the patient's perspective into studies of responsiveness; and
4. To describe the responsiveness of a single instrument to different types and perspectives on change.

1.5 Structure of the thesis.

The thesis will be structured in the following manner in order to address these objectives:

Chapter Two. A taxonomy of responsiveness. This chapter will review the conceptual literature on responsiveness, and describe the nature of the change that is commonly used to study responsiveness. The results will be presented in the form of a taxonomy of change within studies of responsiveness.

Chapter Three. Are you better? In this chapter we will explore what the proxy of change ($\Delta C$) would look like from a patient's perspective (Mitchell, 1995; Wright, 1996; Leplege et al. 1997). Often it is assumed that clinicians and researchers have an understanding of what patients mean when asked "are you better?" This chapter will describe a qualitative (grounded theory) study conducted to elucidate the meaning of change from the point of view of individual workers with upper-limb musculoskeletal disorders.

Chapter Four: Creation and pilot testing of external indicators of change from the patient's perspective. As noted above, studies of responsiveness are dependent on markers of change (indices, or hypothesized change). Chapter four will use the theory arising from the qualitative
study (Are you better?, Chapter three), to develop hypotheses about new markers of change from a patient’s perspective. These markers were created and tested for their relevance, distribution and comprehensiveness in a clinical setting (upper extremity clinic at a teaching hospital).

Chapter Five. Validating the taxonomy.

Chapter five will focus on describing the responsiveness of one measure, the DASH Outcome Measure while varying the type of change (ΔC) in the study. Data gathered on a cohort of 200 patients undergoing treatment for either a shoulder or a wrist/hand disorder will be used. Patients completed the DASH Outcome Measure as well as the indicators of change developed and tested in Chapter four, indicators that also describe different categories in the taxonomy of responsiveness. Questionnaire packages were completed pre-treatment and then at four and twelve weeks after treatment onset or surgery. Responsiveness across the different indicators of change will be described.

Chapter Six. Discussion. Conclusions from across the thesis, the implications for the study of responsiveness and the directions for further research will be described in this chapter.

Quantifying and interpreting changes in health status is a goal worth pursuing as it provides an avenue to have the patient’s perspective play a role in health-care decision making at individual, clinical and health-policy levels. However, it is a goal that is also extremely challenging because the focus is on complex traits (the meaning of change) that are influenced by factors within and outside the specific disorder under study.
Burke and Nesselroade (Burke et al. 1990) describe an approach to science that they apply to the challenge of analyzing and interpreting change:

“Science is concerned with both the description and explanation of phenomena,...attempts at both...may facilitate improvements in the other in an iterative, recursive process”. (Burke et al. 1990) pg. 6.

These realities indeed embody the nature of this thesis: an iterative process, moving from quantitative to qualitative, and back to quantitative methodologies to describe, explain and then describe change again. Each methodological approach is independent and valid on its own, yet can also inform the other (Buchanan, 1992; Pope et al. 1995; Creswell, 1994; Steckler et al. 1992). It is an iterative process that will dismantle and finally rebuild our understanding of responsiveness.
Chapter Two

A TAXONOMY FOR RESPONSIVENESS

2.0 Preface

In this chapter a literature review focusing on the definitions and concepts of responsiveness is completed with the results suggesting the need for a taxonomic framework that delineates different categories of responsiveness. The taxonomy is then developed and presented. This chapter has been submitted for review at the Journal of Clinical Epidemiology.

2.1 Introduction

Clinicians often use indexes, instruments or questionnaires to evaluate their patients over time, and they must select which one(s) to use. The most appropriate measure must be sensible, reliable, valid and (if used to evaluate change over time) it must also be “responsive” (Guyatt et al. 1987; Hays et al. 1993). Responsiveness is defined as the ability of an instrument to accurately detect change when it has occurred (De Bruin et al. 1997; Wright et al. 1998) and is usually quantified by a statistical or numeric score, such as an effect size statistic (Kazis et al. 1989; Beaton et al. 1997; Katz et al. 1992; Wright et al. 1998; Murawski et al. 1998) or a standardized response mean. The question arises, however, of whether such scores provide clinicians with enough information about the usefulness of an instrument in its intended application.

In many ways, the interpretation of statistics of responsiveness is analogous to the interpretation of p-values used in treatment trials, where too much emphasis is often placed on
the magnitude of its numeric estimates (i.e., $p = 0.05$), with too little attention paid to the nature and meaning of the change being quantified. Which patients are being compared? How long was the follow-up? Which treatments are involved? How many patients were involved in the trial? The answers to these questions provide the context of the trial and are essential for interpreting the results (the $p$-value). Similarly, with responsiveness, the magnitude of the effect size (or other) statistic alone is unlikely to provide enough information. To date, attention has focused primarily on finding measures with the largest responsiveness statistics or determining if a measure can produce a "large" effect size statistic ($>0.80$, according to Cohen) (Cohen, 1988) and therefore is "responsive". This assumes, however, that responsiveness is a static, quantifiable attribute of the questionnaire -- an assumption we feel has fueled the numerous discussions about interpretation of these statistics.

It is argued here that concentrating on the context of measurement (what the study is set up to measure), rather than the magnitude of the statistic is more likely to advance the already protracted debate (Donovan et al. 1993; Redelmeier et al. 1996a; Deyo et al. 1992; Norman, 1989; Nunnally, 1975; Murawski et al. 1998; Leplege et al. 1997; Norman et al. 1997) over the interpretability (Guyatt et al. 1994) of responsiveness statistics and that it is essential to look deeper than the statistic's numeric value. As Michell suggests, "protracted controversy suggests that the disagreement lies much deeper than the arguments hitherto presented imply" (Michell, 1986)(pg 398). We suggest that there are three core aspects of the responsiveness debate: first, the interpretation of the statistic (i.e., is the change relevant or important?) (Cronbach et al. 1970; Braitman, 1991; Fortin et al. 1995; Liang, 1995; Naylor et al. 1994; Donovan et al. 1993; Drummond et al. 1993; Deyo et al. 1992; Feinstein, 1998; McDowell et al.
1996; Redelmeier et al. 1996a; Wright, 1996); second, methodological issues (i.e., how should studies of responsiveness be designed? (Deyo et al. 1992; Beaton et al. 1997; Wright, 1996; Naylor et al. 1994; Norman et al. 1997) and how should the property be quantified and analyzed (Murawski et al. 1998; Wright et al. 1998; Beaton et al. 1997; Buchbinder et al. 1995a; Norman et al. 1997; Fortin et al. 1995; Norman et al. 1997; Testa, 1987; Testa et al. 1996)); and finally, the conceptual and definitional issues (i.e., what is being quantified (Stratford et al. 1996a; Testa et al. 1996; Norman et al. 1997)). Although the conceptual and definitional issues have received least attention, they may be the most fundamental factors and are probably critical to making sense of the data (interpretability).

The literature contains many definitions of responsiveness (Table 2.1), and the differences between them are instructive, though baffling. Most authors agree that responsiveness involves the ability of a measure to detect change but there are wide variations in opinion about nature of the change that is being detected. For example, in 1997 Guyatt (Guyatt et al. 1997) defined responsiveness as “the ability to detect change, specifically, important change, in the way patients are feeling, even if those changes are small” (Guyatt et al. 1997) thus focusing on individual feelings and fine discrimination. In 1994, Testa took a broader view, defining responsiveness as “the ability to detect meaningful treatment effects” (Testa et al. 1994), whereas Anderson, 1993, (Anderson et al. 1993)says it was the “ability to detect important changes in disease activity over time”. The latter definition emphasizes the importance of the change. Differences between the definitions are critical, as they each reflect
Table 2.1. Examples of definitions of responsiveness found in the literature.

<table>
<thead>
<tr>
<th>Author, year [reference]</th>
<th>Definition of responsiveness (direct quotes from articles)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeBruin, 1997 (De Bruin et al. 1997)</td>
<td>accurate detection of change when it has occurred.</td>
</tr>
<tr>
<td>Guyatt, 1987 (Guyatt et al. 1987)</td>
<td>ability to detect minimal clinically important differences</td>
</tr>
<tr>
<td>Guyatt, 1997 (Guyatt et al. 1997)</td>
<td>ability to detect change, specifically important changes in the way patients are feeling, even if those changes are small.</td>
</tr>
<tr>
<td>Irvine, 1996 (Irvine et al. 1996)</td>
<td>sensitivity to detect important changes in clinical status.</td>
</tr>
<tr>
<td>Hays, 1993 (Hays et al. 1993)</td>
<td>the ability of a measure to reflect underlying change.</td>
</tr>
<tr>
<td>Wright, 1998 (Wright et al. 1998)</td>
<td>the ability of an index to measure clinical change.</td>
</tr>
<tr>
<td>Liang, 1995 (Liang, 1995)</td>
<td>the ability of an instrument to measure a clinically meaningful or important change in a clinical state.</td>
</tr>
<tr>
<td>Murawski, 1998 (Murawski et al. 1998)</td>
<td>ability to detect change over time.</td>
</tr>
<tr>
<td>Pfennings, 1995 (Pfennings et al. 1995)</td>
<td>a property of an instrument for detecting clinically important changes in the status of the patient.</td>
</tr>
<tr>
<td>Siu, 1993 (Siu et al. 1993)</td>
<td>validity in longitudinal analysis focusing on within person changes.</td>
</tr>
<tr>
<td>Jenkinson, 1995 (Jenkinson, 1995)</td>
<td>ability to detect change.</td>
</tr>
<tr>
<td>Testa, 1996 (Testa et al. 1996)</td>
<td>a measure of the association between the change in the observed score..and the true value of the construct.</td>
</tr>
<tr>
<td>Testa, 1994 (Testa et al. 1994)</td>
<td>ability to detect meaningful treatment effects.</td>
</tr>
</tbody>
</table>

distinct types of change being quantified in a given study of responsiveness, and thus different concepts of responsiveness. We suggest that agreement must be reached about what responsiveness is in order to advance the debate about the interpretability of related statistics.

The debate over the interpretation of responsiveness statistics is not just of methodological or academic interest, but has direct implications for how we assess patients and
how we decide if treatments have truly made them better. Without a clear understanding of responsiveness statistics, a meaningless change could be misconstrued as clinically significant when it is merely \textit{statistically} significant (Redelmeier et al. 1996b; McDowell et al. 1996); alternatively, a small gain in mobility might be \textit{statistically} insignificant (Testa et al. 1994; Russell, 1998) but have a dramatic effect on the patient's well being. As Jenkinson (Jenkinson, 1995) suggests, "the results of health status measures could not simply be misleading, but actually harmful." (Jenkinson, 1995)

The purpose of the present paper is to propose a classification system that reflects the context of the measurement of responsiveness by suggesting that different categories or types of responsiveness can be defined in terms of the attributes of the particular change being quantified.

\subsection*{2.2 Methods}

The present classification system for studies of responsiveness was developed from a review of the literature. The initial review was based on articles from the personal files of the authors. In addition Murawski and Miederhoff (Murawski et al. 1998) (who published a review of the literature on responsiveness up to 1994) shared their list of 324 references and search strategy. Murawski (Murawski et al. 1998) searched several databases (psychology, sociology and medicine) for any articles indexed under headings such as quality of life, health status or health status indicators. They reviewed over 20,000 titles and abstracts to find those with longitudinal data collection. Three-hundred and twenty-four met this criteria.

A modified search strategy was used to identify first that we could identify the key articles Murawski had identified (Murawski et al. 1998), and then update the search with
additional articles in both Medline (after 1994) and CINAHL (see figure 2.1). It was necessary to restrict Murawski's approach to those articles dealing with measurement properties or statistics because the less restrictive approach resulted in too many abstracts to review. Our strategy was validated by ensuring that it caught key articles in the area of responsiveness known to the authors. Our approach was as follows: first, the text words (abstract or title) of "responsiveness" and "health" were searched (English only). This produced 238 articles (set A). Second, MeSH headings of "reproducibility of results" or "psychometrics" or "treatment outcomes" were used in conjunction with a MeSH heading of "questionnaires" or "health status indicators" resulting in 3,520 articles. This set was then restricted to 1024 articles reflecting English only publications which also had a text word of "health" (set B). Articles from both sets (A or B) were combined for a total of 1,210 citations (there was some overlap). Murawski's list of 324 articles were then added making 1,534 potentially useful citations. The restriction to the English only should be recognized as a potential limitation, particularly given international efforts in health measurement currently underway in Europe. However many of these groups publish in English journals including the

![Flow chart describing literature reviewed as basis for the taxonomy on change within studies of responsiveness.](image-url)
official journal of the International Society of Quality of Life Research. Quality of Life Research. This journal is indexed on medline.

Abstracts were reviewed for their relevance to the theory of responsiveness, and the reference lists of those retrieved were also reviewed. Grateful Med (on the Internet) was then used to make lateral searches on key references (these search for comparable articles to a key reference by using a similar combination of subject headings).

A classification system, or taxonomy, of studies of responsiveness based on these articles was derived on the basis of the following guidelines described by Buchbinder (Buchbinder et al. 1996a) and/or Feinstein (Feinstein.1968): have mutually exclusive yet exhaustive categories, use multiple axes if necessary, and make sure the classification system is simple to apply and easy to interpret. Ultimately, the goal was to have a system that would facilitate accurate communication (Katz et al. 1991) about the context of measuring responsiveness.

2.3 Results.

2.3.1 Building a taxonomy of responsiveness

Several articles concerning responsiveness have discussed different aspects of the nature of the change being studied and how they relate to interpretation of resultant statistics. Three groups of articles were identified: those that discussed individual-level versus group-level assessment of change (Testa,1987; Guyatt et al. 1994; Redelmeier et al. 1996a; Redelmeier et al. 1996b; McHorney,1997; Lydick et al. 1993; Wood-Dauphinee,1999), those that considered the contrast of between-person change versus within-person change (Norman,1989; Redelmeier et al. 1996c; Juniper et al. 1994; Bombardier et al. 1991; Jaeschke et al. 1989; Buchbinder et al. 1995a;
Redelmeier et al. 1996a); and those that addressed different types of change (i.e., the importance of the change, and from whose perspective) (Liang et al. 1990; Liang et al. 1985; Katz et al. 1992; Laupacis et al. 1993; Deyo et al. 1984; Norman et al. 1997; Stratford et al. 1996a; MacKenzie et al. 1986; Redelmeier et al. 1996c; Juniper et al. 1994; Jaeschke et al. 1989; Redelmeier et al. 1996b; Beaton et al. 1997; Naylor et al. 1994; Drummond et al. 1993; Wright, 1996; Fortin et al. 1995). These groupings became the defining axes of the system. All possible subheadings (based on examples from the literature) within each of the three axes were then defined, to form a 2x3x5-cell matrix that formed 30 cells each representing a category of change. Although not every cell had a supporting example from the literature, it was usually possible to hypothesize a study that could be designed to measure responsiveness to that type of change.

The three axes were labeled Who (Who is being looked at? An individual or a group?), When (When is the comparison being made? i.e., Over time or at one point in time?) and What (What kind or amount of change is being estimated?). There is no hierarchical relationship between the axes.

### 2.3.2 The Who Axis

The Who axis considers whether information from the study under consideration is provided at a level that will be interpretable for an individual-level or group-level measurement.

---

**The Who Axis: Individual-level versus Group-level of measurement.**

1. Individual level interpretation
2. Group level interpretation

**Figure 2.2** The "Who" Axis. Defines whether the study of responsiveness is looking at group of individual level of analysis and interpretation.
individual patient (a cut-off for change in individual patients) or at a group level (the average amount of change for a group) (see figure 2.2). Most summaries of responsiveness use statistics that describe the distribution of change scores such as mean change or standard deviations (Bombardier et al. 1991; Deyo et al. 1986; Buchbinder et al. 1995a; Wright et al. 1998).

However, as Williams and Naylor point out "probability and chance are not easily related to applications on the individual level" (Williams et al. 1992) (pg 1347). Therefore, change that has meaning for groups may not have meaning for individuals (Redelmeier et al. 1990).

The distinction between individual-level versus group-level interpretation of change scores is important for two reasons. First, studies have shown that a given change has different meaning whether you are looking at individuals or at groups. Research has shown that when experts are shown profiles of change for individuals and for groups, they are willing to accept a much smaller amount of change as being "important" for group comparisons than for comparisons of individuals (Goldsmith et al. 1993; Redelmeier et al. 1990; Paulus et al. 1991). For example, as Deyo pointed out, a 5 mm change in diastolic blood pressure, is likely to be considered insignificant at an individual level, but significant if it is the average change for a group of patients (Deyo et al. 1995). Because measurement involves estimating an attribute within a certain margin of error, single, individual measures have a wide margin of error. This may partially explain why we have less confidence in change scores for single patients (Paulus et al. 1991). In groups of patients, the increased sample size reduces the margin of error, and thus we can be more certain that the mean change is close to the true change for that group. This gives us more confidence in treating small changes as important in groups.

The second, related issue in identifying the difference in interpreting at an individual or at
a group level is the degree to which one can generalize between the two. Clinicians often need to apply group-level information to their individual patients. Using group-level estimates of change (such as average change or median change) as a guide to when an individual has improved is a challenge, not only because of the different standards that are at play, but also because any cut point will have some degree of misclassification error (Testa, 1987). Mean change is an estimate of a central tendency in a whole distribution of change scores, where each data point represents change (even important change) for an individual patient. Accepting the mean as representative of an indication of any single individual's improvement would mean several people, who may have considered themselves improved, would be erroneously considered unchanged (Feinstein, 1996; Jenkinson, 1995; Fletcher et al. 1996) because their numeric change score was below the cut point (the average). Improved measurement will provide narrower confidence intervals, and increase confidence in interpreting when a change score in an individual exceeds error (McHorney, 1997; Nunnally et al. 1994; Williams et al. 1992), but it will still not determine how much change should be considered important or indicative of improvement or deterioration for that patient. Redelmeier suggests using the lower bound of the confidence limit of the change scores as a threshold for improvement in order to be sensitive to include 95% of the individuals' change scores (Redelmeier et al. 1993). Goldsmith considered the 25th percentile in the distribution of change scores as the threshold for the same reason (Goldsmith et al. 1993). Both point to the possible problems with using a mean change score as representative of the most accurate cutpoint in the distribution of change scores. Similar dilemmas face the establishment of nutrient requirements for the general population. A cut-off at any given point on the distribution of individual requirements will misclassify some of the sample (Beaton, 1996) and
consequently set a requirement level that may result in a proportion of the population being either deficient or exposed to excessive amounts of a nutrient. The decision about where to set the recommended requirement level must be based on an estimation of the risks of under- or overestimating the true requirement level. The aim is to set a level that has the greatest meaning, while causing the least amount of harm. The same is true of attempts to determine a threshold for meaningful change from a distribution of change scores.

Clinicians routinely faced with the challenge of applying the results of a clinical trial or program evaluation to an individual patient find relatively few guidelines in the literature (McHorney, 1997) on how to do it (McHorney et al. 1995). An equally important and related challenge is how to interpret responsiveness at an individual level. This axis identifies that the results of a study of responsiveness can be provided in a manner that makes them useful at an individual level in the form of a threshold for important change for example, or in a manner useful for group-level interpretation, such as the mean change observed.

### 2.3.3 The When Axis: When is the comparison being made?

The When axis refers to the timing of the data collection (see figure 2.3). Information can be gathered over time in order to assess how a patient changes over time (within-person change (Testa et al. 1994; Testa et al. 1996)). Alternatively, questionnaire scores can also be compared using data gathered at one point in time but from different people (between-person change).
differences) (Redelmeier et al. 1996b; Redelmeier et al. 1996c; Redelmeier et al. 1993). A combination of so-called between and within change is used in randomized controlled trials where the between-group differences of within-person change is being analyzed.

Conceptually, within-person change and between-person differences have long been held to be very different, with only the former being considered relevant for discussions of responsiveness. Burke describes change as “any variation in the quantity or quality of an entity’s attributes” (Burke et al. 1990). From this perspective, it could be seen that both between- and within-person change could be estimated. Both are included in our taxonomy for two reasons: first, because the same statistics can be applied to each, giving “effect sizes” for each and thus their uniqueness will be clarified. Second, research in this area suggests that the results of within-person change and between-person differences yield similar results (having observed similar magnitudes of change using either approach) (Redelmeier et al. 1993; Redelmeier et al. 1996a; Redelmeier et al. 1996c). These investigators also suggest that estimates of effect size or minimal clinically important differences observed in the analysis of “between-person differences” could be used to guide the estimation of minimally clinically important differences in the “within-person change” situation (Redelmeier et al. 1996a; Redelmeier et al. 1996b; Redelmeier et al. 1993). We would suggest, however, that they are conceptually different types of change the magnitude of which could vary a great deal when comparing “between-person differences” and “within-person change” (Jacobs et al. 1996; Testa et al. 1996; Redelmeier et al. 1996a; Wright, 1996). Hemingway et al. (Hemingway et al. 1999), measured and compared both “between-person difference” scores and “within-person change” scores using a similar external marker to differentiate people who had changed (one year difference in age). They found that the
estimates were not the same, and that between-person differences systematically underestimated within-person change (Hemingway et al. 1999). This supports the concept that within-person change does not necessarily equate to between-person differences; hence the two are held as different in our taxonomy.

The final category in the second axis is a combination of the first two and is called the "hybrid" change: between-group differences in within-person change. For example, in clinical trials (Bombardier et al. 1991; Buchbinder et al. 1995a) the focus is on *between-person differences* (treated groups versus control groups) of *within-person change* over time (measuring improvement over time). The change is no longer what we have labeled an observable change in a single-arm cohort sense, but rather a relative change – the change in health in the treatment group relative to the control group. Thus, it is a combination of within-person and between-person types of change.

2.3.4 *The "What" Axis: What type of change is being quantified in a study of responsiveness.*

The third axis distinguishes the different kinds of change that may be quantified in a study of responsiveness (see figure 2.4): for example, change observed before and after treatment and change that was deemed to be clinically relevant or important (Cronbach et al. 1970; Wright et al. 1998; Lachs, 1993; Redelmeier et al. 1996b; Deyo et al. 1995; McDowell et al. 1996; Feinstein, 1998; Braitman, 1991; Russell, 1998; Testa et al. 1994). There are five subheadings along this axis.

2.3.4.1 *Minimum change potentially detectable by the instrument.*
The minimum change is the smallest amount of change that the instrument is capable of measuring on an individual level and is dependent on the construction of the instrument (i.e., the number of items and the number of response categories). For example, with the DASH questionnaire (Hudak et al. 1996), there are 30 items, with five response options (0-4) giving a possible score ranging from 0-120. A change of one response level on just one item (1/120 or 0.83%) would be the minimal amount of change possible. In another scale with 24 yes/no items the minimum change would be 1/24, or 4.2%. For a given individual, a change score less than this amount would be impossible. However, at a group level the average change could be lower than this minimum potentially detectable change. As McDowell and Jenkinson (McDowell et al. 1996) state, interpretation should be content-based (p241). An average change score that is less than the minimum possible for an individual patient should be seriously scrutinized to determine if it can be interpreted in a group-level analysis. This is the only type of change that is stable in magnitude across the other two axes because it is an attribute of the questionnaire, using items and scaling, and not the application.

2.3.4.2 MINIMUM CHANGE DETECTABLE GIVEN THE MEASUREMENT ERROR OF THE INSTRUMENT ("MINIMALLY DETECTABLE CHANGE").
The second type of change is defined by the error associated with the measurement. Measurement of change reflects true change plus error. When the error is great, wider confidence intervals apply to the observed change score — the “true” change could be anywhere within the given range. The upper limit of this confidence interval, when placed around a change score or zero (no change), helps to define another boundary of meaningful change. Theoretically, change scores greater than this upper limit would have less than a 5% chance of being change due to chance (error) alone (Christensen et al. 1986a) and could therefore be confidently considered as true improvement (or deterioration). Following this logic, the minimally detectable change (MDC) also constitutes a lower boundary of meaningful change.

Similar boundaries of change have been described in the literature. Radiographic changes are often based on their magnitude compared to the smallest detectable change (SDC) calculated in the same fashion (Sadler et al. 1992; Ravaud et al. 1999). Bland and Altman describe the concept of limits of agreement for alternate forms reliability, which sets 95% confidence intervals around the mean change in a stable group (Altman, 1991). Ravaud (Ravaud et al. 1999) compares formulae and resolves that the MDC or SDC for an individual is equivalent to the limits of agreement, except that the MDC is set around a change score of zero rather than the observed mean change in the stable group.

Application of this in health status measures usually is found under the rubric of the reliability change index (RCI) based on the work of Christensen (Christensen et al. 1986a). Ottenbacher (Ottenbacher et al. 1988; Stratford et al. 1997) and Stratford (Stratford et al. 1996b; Stratford et al. 1996c). The RCI is defined as the change score divided by the standard error of measurement (SEM) of the change. If the latter is not known, it can be calculated by using the
information on reliability as: standard error of measurement = standard deviation \times \sqrt{2 \times (1 - r_{xx})} where \( r_{xx} \) = reliability. There is some variability in the calculation of this statistic specifically over which reliability coefficient to put into this equation and whether the former should be multiplied by two as shown above. Christenen (Christensen et al. 1986b) and McHorney (McHorney et al. 1995) describe two different goals: first, to get an estimate of the error around a single observed score at one point in time (in which case Cronbach’s alpha would be used); and second to get an estimate of the precision of change scores (in which case the test-retest reliability coefficient would be used). Nunnally suggests that in very large samples (\( n \sim 300 \)) the two reliability coefficients become convergent (Nunnally et al. 1994), however often in health research we are not deriving these coefficients from such large samples. It would seem most sensible to use the related reliability statistic, either Cronbach’s alpha for cross sectional precision or the test-retest coefficient for longitudinal precision for the calculation.

The second point was disagreement around the need to make adjustments to the calculation of the RCI when looking at change in score over time. It has been suggested that an adjustment should be made to the determination of the SEM in a change situation because the reliability data is being obtained from two samples. Christensen (Christensen et al. 1986a) renamed the adjusted coefficient the SC index for significant change index. The adjustment corrects for the additional error involved in obtaining reliability estimates from two samples (test and retest) by multiplying the SEM by the square root of two (McHorney et al. 1995). Earlier work by Jacobson (Jacobson et al. 1984) and more recent work by Wyrwich (Wyrwich et al. 1999b; Wyrwich et al. 1999a) do not make this last adjustment, thus their estimates of minimally detectable change would be lower by a factor of \( \sqrt{2} \) (or 1.41). However, in 1991 Jacobson
adjusted his statistic to reflect this the multiplier to adjust for using two samples (Jacobson et al... 1991). Because our interest is in the precision around an estimate of change which means taking data collected at two separate times, we will use this adjusted formula. equivalent to Christensen’s significant change index (also labeled the RCI’) and accommodating the additional error encountered by bringing together two sets of data.

The reliability change index (or significant change index) would therefore be an expression of the change score in standard deviation units, much like a z score. Setting this equal to 1.96 (the value on a standard normal curve associated with a 95% confidence interval) and then solving for the change score would give the minimum change score considered significantly different to no change at all (at p=0.05) on that instrument in that application. Stratford (Stratford et al. 1996b; Stratford et al. 1996c) uses the term “minimally detectable change” (MDC) for this value. Accordingly, change above this level is considered greater than measurement error (for a given level of confidence, in this case 95%).

As a standard for a lower boundary for interpreting change, the MDC is usually quite large, especially at the 95% confidence level (De Bruin et al. 1997). MDC values can theoretically be created for any confidence level. Stratford shifted to reporting a 90% minimally detectable change and found this to be four (out of a possible maximum of 24) Roland Morris points, or roughly 16% (Stratford et al. 1996c). Wyrwich suggests that setting the RCI (unadjusted for two samples) to one (therefore a 67% confidence level) and solving for change (which will equal the SEM), and suggests that this might produce a figure closer to a minimal threshold for meaningful change (Wyrwich et al. 1999b; Wyrwich et al. 1999a). In this thesis we will use a subscript figure to suggest the level of confidence used in calculating the MDC
(therefore $MDC_{0.95}$, $MDC_{0.90}$ or $MDC_{0.63}$ would be used in the examples given).

Although some describe this measure as being independent of the sample from which it was derived (Wyrwich et al. 1999b), we felt that the minimally detectable change could vary with setting, populations, raters, etc., because reliability and variance can change in different samples. We suggest caution in linking this data-driven approach with clinical meaningfulness, or using it as a definitive cut-off point for a health status measure in any application, but put forth the concept of minimally detectable change as a data-driven approach to aid in the interpretation of change scores (Liang et al. 1985), and acknowledge the various approaches being used (varying the level of confidence, varying reliability coefficients) in its formulation.

2.3.4.3 Change observed between measurements in a given population ("Observed change")

Observed change is simply the change in health status that is measured by gathering data on two different occasions, such as before and after a treatment. Treatments of "known efficacy" are often used when testing a measure's responsiveness because change is expected. For example, because most patients improve in function after a total joint arthroplasty (Liang et al. 1985; Laupacis et al. 1993) or after the onset of uncomplicated, acute, low-back pain (Beaton et al. 1997) such patient populations have been used to test responsiveness. Observed change can be described for the individual person or for the group. Perhaps its most common and familiar use is in randomized controlled trials where it is measured and compared between treatment and control groups. It should be noted that observed change includes individuals who do get better as well as those who do not.
2.3.4.4 OBSERVED CHANGE MEASURED BY THE INSTRUMENT IN A POPULATION DEEMED TO HAVE IMPROVED ("ESTIMATED CHANGE")

With "estimated change," study subjects are stratified into those who improved and those that did not, according to some external standard. Responsiveness is then assessed (i.e., calculation of responsiveness statistics) using only data from those deemed to have improved, according to this external standard. Deyo coined the term "clinically estimated change" to describe a study of responsiveness in which he used a clinician’s rating of patients who had improved ("yes, this patient is better") to define the sample for the analysis (Deyo et al. 1984). MacKenzie used a similar approach in studying the patients’ ratings of whether or not they had improved (MacKenzie et al. 1986).

The determination of whether a change has occurred can come from different perspectives, for example that of the clinician versus that of the patient. The perspectives taken are used here to define four subcategories of estimated change. Patients can estimate their own change (by answering “yes I am much better” when asked). Alternatively, clinicians can estimate change in their patients by asking themselves “Is this patient better?” perhaps using a variety of criteria (such as test results, assessment findings, and general impression, etc.). Deyo used the clinician’s rating of the estimated change (Deyo et al. 1984). Payers may also be able to estimate when a change has occurred. For instance, change has occurred when a person returns to work, or when a wage-replacement claim is closed. It is suggested here that society can also be used to estimate the presence or absence of change in a person’s health. In this context, by society we mean those aspects outside of the patient-health services interaction, including institutions, communities, culture, government policy makers, workplace organizations,
etc. (Naylor et al., 1994; Testa, 1996). For example, at a societal level, patients can be considered to have improved from surgery when they have resumed their social or work roles (Drummond et al., 1993). Testa calibrated changes in measured quality of life with life events scales and thus suggested certain changes would be equivalent to the loss of a job, or a divorce (Testa et al., 1994). This could suggest relevance from society's perspective. It is most common to see the clinician's perspective being used, followed by the patients. Payer's and society's views are less often used within studies of responsiveness, though the ability of patient-based health status questionnaires to demonstrate change that is relevant to these players is worth pursuing. It is precisely through these types of connections that health policy changes could be framed around changes in patient's daily lives rather than in more traditional markers such as health care utilization. Estimated change is therefore divided into four subcategories, one for each of the four different perspectives that could be used in its determination.

2.3.4.5 Observed change measured by an instrument in a population deemed to have had an important improvement or deterioration ("important change").

*Important change* is estimated change that is seen to be valued or important and becomes a criterion for stratifying a sample prior to analysis. Subjects who undergo important change (usually separating important improvement from important deterioration) would be used in the analysis of this type of responsiveness. The distinction between estimated and important change is critical because many studies claim to be measuring the latter when, in fact, they are usually measuring the former (as defined above). Furthermore, it is often assumed that the magnitude and importance of change are correlated, but they are not the same. For instance, some research
suggests that the threshold for change to be described as "important" varies according to the severity of condition at the time of the baseline assessment (test-1) (Redelmeier et al. 1996b; Stucki et al. 1996).

Patients can be asked to discern not only if they had experienced change, but also if that change was important to them in their lives (Naylor et al. 1994). Such highly individualized appraisal could lead to very different experiences of change being considered important. For instance, an indication that pain is reduced is estimated change; it might only become important change when the reduction is sufficient for the patient to forget about his or her arm problem or do what they want or need to do (Beaton et al. 1999a). For someone else, reduction in pain could be experienced (estimated change), but nevertheless be entirely unimportant to an individual. Increased variance in change scores could be expected in this type of responsiveness, as the definition of "important change" is so highly individualized.

As was the case with estimated change, the perspective also has a bearing on when change is considered important. Clinicians or researchers may determine whether change is important by relating change scores to some "criterion": clinical findings such as lab work or observed findings, experience, or a consolidation of findings from past clinical trials (Naylor et al. 1994; van Walraven et al. 1999). Juniper (Juniper et al. 1994) and Jaeschke (Jaeschke et al. 1989) had patients estimate the magnitude of change on a 15-point ordinal scale (-7 to +7). The researchers then set the cut point of what was considered to be "minimally clinically important" change (i.e., transition score of +/-1-3 (Jaeschke, 1989) or +/- 2-3 (Juniper, 1994)). Although this work is considered seminal in the determination of minimally clinically important differences, we feel that it incorporates a mix of perspectives, in that the patient estimates the
change, but the investigators decide on its importance. The selection of a cut point has a great deal of influence on the estimation of the responsiveness; different cut points will result in smaller or greater amounts of change being considered the minimally clinically important difference (Wright, 1996). For the taxonomy presented here, it is suggested that a consistent rater determine both the existence and importance of the change in order to have a single perspective reflected in the external standard. Redelmeier, used similar techniques for estimating important between-patient differences, and recognized the potential conceptual inconsistency in having researchers set the cut point for importance (Redelmeier et al. 1996b).

Naylor and Llewellyn-Thomas support the role of broader, albeit differing, perspectives in determining the importance of change (Naylor et al. 1994). Payers such as insurance companies and workers' compensation boards and other facets of society (as defined above) can also determine the importance of change, often from a productivity or economic point of view. It would be anticipated that this would differ from an individual's or even a clinician's ratings of important change (Hemingway et al. 1999). The external standard might take the form of either a cut point set at an amount of change that the person has to experience in order to achieve a certain level -- population norms, a return to work, a perfect score) or a transitional stage (a certain magnitude of improvement, perhaps accounting for baseline scores). Thus, payers and society provide the final two perspectives on whether a given change would be deemed important.

2.3.5 Application of the classification system.

The proposed classification system describes many different categories of change that can
be quantified in studies of responsiveness. Each category is defined by the place it occupies on a triaxial matrix (see figure 2.5) in which the three axes define the Who, When, and What of the change being quantified within a study of responsiveness. Some examples might help to illustrate the practical application of the system. Table 2.2 briefly describes five different studies and outlines how they would be categorized.
Table 2.2: Application of the classification system using published studies of responsiveness from the literature. This table shows some examples of studies of responsiveness from the literature, and how they would be classified using the suggested classification system. Refer to text for description of each of the categories under the Who, When and What columns.

<table>
<thead>
<tr>
<th>Author</th>
<th>Brief Description</th>
<th>Who?</th>
<th>When?</th>
<th>What?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laupacis (Laupacis et al. 1993)</td>
<td>cohort of patient pre and post total hip replacement</td>
<td>1. group</td>
<td>1. within person</td>
<td>3. observed change in population</td>
</tr>
<tr>
<td>Redelmeier (Redelmeier et al. 1993)</td>
<td>group of patients comparing selves to others, and rating self as somewhat better than other.</td>
<td>1. group</td>
<td>2. between person (one point in time)</td>
<td>5. observed change in those estimated to have <em>important</em> improvement as defined by researcher.</td>
</tr>
<tr>
<td>Beaton (Beaton et al. 1998)</td>
<td>shoulder patients who said they were better after surgery</td>
<td>1. group</td>
<td>1. within person</td>
<td>4. observed change in those estimated to have improved (patient's perspective)</td>
</tr>
<tr>
<td>Buchbinder (Buchbinder et al. 1995a)</td>
<td>randomized controlled trial in rheumatoid arthritis</td>
<td>1. group</td>
<td>3. hybrid</td>
<td>3. observed change in a population</td>
</tr>
<tr>
<td>Wyrwich (Wyrwich et al. 1999b)</td>
<td>outpatients enrolled in trial of drug utilization in heart disease and rating improvement</td>
<td>1. individual</td>
<td>1. within</td>
<td>5. observed change in those estimated to have <em>important</em> improvement from the researchers perspective (using one SEM as criterion and McMaster MCID)</td>
</tr>
<tr>
<td>Stratford (Stratford et al. 1998)</td>
<td>patients with low back pain &lt; 6 weeks duration, undergoing physiotherapy. Follow up 3-6 weeks later. Average of patient and clinician rating used as indicator.</td>
<td>1. Individual</td>
<td>1. Within</td>
<td>5. Important: global scale used, defined improved / not improved. Most accurate change score to discriminate = important change to individual.</td>
</tr>
</tbody>
</table>

As summarized in table 2.2, Laupacis followed a group of patients before and after total hip replacement and provided overall summaries of change in health for the whole group whether or not each individual had actually improved (it would be probable that some did not improve). This is an example of group-level, over-time, observed change. In contrast Redelmeier had patients rate themselves as being better/worse/same compared with other patients in the same room, on the same day. The differences in health-status score between pairs of patients where it
was agreed that one was better than the other were used to estimate responsiveness and to determine an "important difference" (according to the researchers). This provides a rare example of group-level, between-person, important-change type of responsiveness study. In previous work done by the authors, patients undergoing shoulder surgery (rotator cuff debridement or repair, or total shoulder arthroplasty) were measured before and six months after surgery. At the second testing, patients were asked if they were better on a five-point scale. Those that responded "yes" (that they were somewhat or much better) were deemed to have improved, and their responses were analyzed for responsiveness (to estimated change) (Beaton et al. 1998). An excellent example of responsiveness in a hybrid group comparison is found in Buchbinder's paper on outcome measures in randomized controlled trials in rheumatoid arthritis care (Buchbinder et al. 1995b). Buchbinder describes the adaptation of the effect size statistics for use in a hybrid design and ranks the traditionally used outcomes by their responsiveness. Wyrwich (Wyrwich et al. 1999b) provides a rare example of analysis of responsiveness at an individual patient level using both the minimally clinically important difference approach of McMaster (Jaeschke et al. 1989) while proposing one standard error of measurement unit as reflecting an important difference (Wyrwich et al. 1999a). Not all of the analysis in this paper is at the individual level; however certain components are, such as describing the proportion of subjects exceeding one SEM criterion. The last row in table 2.2 describes the work of Stratford and colleagues (Stratford et al. 1998) in describing important change in patients with low back pain. The average of patient and clinician rating of improvement was used as the external marker, and if it exceeded 5/7 it was considered to be "important". It was therefore a combination of patient and clinician perspectives. Receiver operator characteristic curves were used to find the most
accurate cut point. Riddle led another paper from this group which also described important change, but used achievement of treatment goals as the dichotomous marker (Riddle et al. 1998a) of an important improvement.

Examples can also be found of studies that deliberately contrast different categories of responsiveness and emphasize the uniqueness of each approach. Hemingway (Hemingway et al. 1999) examines the same category of change (group-level, within-person observed change) as Laupacis, above but contrasts it with another category by changing the “When” component in the second group (group-level, between-person, observed change). In a study of injured workers, some of the present authors demonstrated the difference between observed change and estimated change holding the other two axes constant (group-level, within-person change) (Beaton et al. 1997). The differences found in these two studies show how different categories of responsiveness can lead to dissimilar estimates of change and effect size (in terms of magnitude) even within the same study situation.

2.4 Discussion

A review of the literature on responsiveness and a synthesis of the results suggested the need for a context-specific approach to its study. Context is defined in large part by the nature of the change being considered. We propose a classification system that reconciles many of the debates and discussions in the literature (Testa et al. 1996; Norman, 1989; Kirshner et al. 1985; McHorney, 1997; McDowell et al. 1996) by locating the nature of change (category of change) within a matrix defined by three axes: Who is the focus of the study? When were the two measures that are being compared gathered? and What is the nature of the change being
examined? These three axes define the essential attributes of all the different types of change that have been or may be studied under the rubric of "responsiveness." Because any measure may be tested for responsiveness to any of these categories of change, responsiveness is not a unified attribute "possessed" by an outcome measure, but, rather, a context-specific characteristic within a situation defined by the patient, the treatment under study and the category of change being assessed. We believe that responsiveness can only be attributed to that particular application (patient, treatment, category of responsiveness, outcome measure) and that a study of responsiveness therefore validates the application, not the instrument (Nunnally et al. 1994; Testa et al. 1996).

A definition of responsiveness should be congruent with its meaning. We suggest that the definition used by deBruin (the accurate detection of change when it has occurred (De Bruin et al. 1997)) is, in our view, probably the most appropriate because it balances the measurement of change (accurate detection) with the need to define how we know that change has occurred. This definition returns responsiveness to its etymological origins. The Canadian Oxford Dictionary states that the term is derived from the Latin respondere: to return to a pledge or promise (Barber, 1998), and thereby defining responsiveness as a reflexive term: a response to "x" where, in our situation, "x" is a category of change in a given setting. A limitation of many other definitions in the literature is that they include what we have called a category of change within their definitions of responsiveness (such as the ability to detect treatment effects, or important change even if it is small), yet do not recognize the potential lack of generalizability to other types of change. A new framework or conceptual model in which several types of responsiveness can be held as meaningful is called for. Testa (Testa et al. 1996) suggests that a
study of responsiveness compares fluctuations in a measure’s score against shifts in some
criterion (or proxy) for true change in the attribute (given a hypothesized causal link between the
proxy and the true change (Testa et al. 1996)). This comes closest to this context-specific
concept of responsiveness. We have elaborated on the Testa framework by suggesting the
different categories of change that could be considered proxies for true change in health status,
each of which would offer a valid assessment of a specific type of responsiveness. Thus, the
definition and the concept are complementary.

A classification system should have the goal of “facilitating accurate communication”
(Katz et al. 1991). The question of whether a triaxial classification system does indeed facilitate
the study of responsiveness remains open to debate. Nevertheless, it does define the type of
change being quantified in a study and (because it emphasizes the importance of the context) can
be expected to lead to less confusion between contrasts such as observed versus estimated or
important change. Results will be easiest to generalize to the same context – the same category
of change, as well as the same patient group and treatment.

Limitations to this research include the fact that it is theory driven. The proposed
classification system needs to be tested using practical examples of studies of responsiveness.
Some cells in the matrix, such as types or categories of change, might be missing, and others
rarely or never filled. Second, this work addresses only some of the challenges inherent in
measuring responsiveness. There are also statistical and mathematical debates over how to
quantify and model change (Nunnally, 1975; Testa et al. 1996; Norman, 1989; Pfennings et al.
1995; Cohen, 1990; Cronbach et al. 1970) as well as concern about the possibility of violating
assumptions in assuming an interval level of scaling in our measures (Michell, 1986; Nunnally et
al. 1975; Nunnally, 1975; Norman et al. 1997; Feinstein, 1998; Stevens, 1958; Stucki et al. 1996; Chambers et al. 1987). Some authors have raised issues such as how well respondents can be expected to recall previous health states when asked a transitional question such as “how are you now compared to...?” to determine estimated or important changes in an attribute (Ross, 1989).

Still others suggest that the whole framework for how a patient would respond to a given question on their health could shift over time (Llewellyn-Thomas et al. 1999; Schwartz, 1995).

Finally, our work did not address the other contextual issues that define and influence the interpretation of responsiveness statistics: the type of treatment experienced between measurement times (for example, joint arthroplasty will have a large and dramatic effect compared to some non-operative interventions for easing joint pain (Katz et al. 1992)), the patient mix, and the timing of the measurement (Greenfield et al. 1992). These matters were beyond the scope of this paper, which had the goal of classifying just one component of responsiveness: the type of change being quantified. Our work reflects a more conceptual approach to responsiveness which has been called for by others (Liang, 1995; Hyland, 1993; Marquis, 1998; Cronbach, 1975).

The work presented here has two main implications: first, it will change the way we (clinicians and researchers) talk about responsiveness; and second, it will influence how we will choose instruments to use in our work. In using this taxonomy, clinicians and researchers will begin to talk about responsiveness in a contextualized way. Instruments will be considered responsive to a type of change (treatment effects or important change from a patient’s perspective). Conceivably, though outside the scope of this paper to demonstrate, a given instrument may be responsive to one category of change, but not to another, suggesting a limit to
the generalizability of the findings, even when using the same questionnaire. Certain disagreements in the field of responsiveness (exemplified by papers such as Norman's (Norman et al. 1997) and the debate between Redelmeier and Wright (Redelmeier et al. 1996a; Wright.1996; Redelmeier et al. 1996c)) could be resolved with this new way of thinking. The taxonomy would not promote a right or wrong way for studying responsiveness, but would suggest that there are different approaches each valid for a specific application. For instance, Norman (Norman et al. 1997) raises the criticism that measures of what we have called “Who: group-level, When: within-person over time, What: estimated change” are biased indicators of change for clinical trials. Although we would concur, we would say it is because it is the incorrect type of change that is being looked at for the target application. A statistical result from a study that has used an external marker to estimate the occurrence of change does not reflect the type of change being tested in a clinical trial (where you would need hybrid comparison of group-level, observed change). It is suggested here that what Norman et al (Norman et al. 1997) argue from a mathematical point of view about the correlation between the external criterion and the change score, is for us an error of context: the wrong type of change to look at for that application. Our findings would also caution against many statements found in the concluding lines of responsiveness studies such as that the findings derived from observed change will aid in “choosing more sensitive outcome measures for a clinical trial” (p 537) (Anderson et al. 1993). This is to mix categories of change, and an instrument that works well in a particular application may or may not do so in a clinical trial. Likewise, we would be hesitant at this point in time to generalize from Redelmeier’s between-person estimates (Wells et al. 1993; Redelmeier et al. 1996b) to determining clinically important differences in a clinical trial (as the authors suggest).
Again, this generalizes to a situation where the instrument may or may not perform in the same manner.

The second implication is that taxonomy described here will change the way we select instruments for either clinical applications or research. Those choosing an instrument based on the criterion of responsiveness will now look not only for evidence of "responsiveness" but for evidence of a particular category of responsiveness that matches how they intend to measure change when using this instrument. In 1985 and again in 1992, Kirshner and Guyatt (Kirshner et al. 1985; Guyatt et al. 1992a) proposed a now well-accepted, though also openly challenged (Hyland, 1993; Williams et al. 1992; Guyatt et al. 1992b), taxonomy for the purposes of health status measures. The goal was to clarify and simplify the ways in which instruments are used (Guyatt et al. 1987; Guyatt et al. 1992a; Guyatt et al. 1992b). As Guyatt suggests, their work addressed only one area in health status measurement: the purpose of the instrument (Guyatt et al. 1992b). With a similar goal, we have developed another taxonomy, in this case, one for responsiveness. Like Kirshner and Guyatt (Kirshner et al. 1985), we hope that it will help the user select the questionnaire most likely to accurately detect the type of change being quantified. An appendix to this thesis (Appendix G) contains a form to be used when appraising published studies of responsiveness in terms of how much an article can inform the reader about the instrument's responsiveness in their particular targeted application. It helps define the context of the study guiding the user through the components of the proposed classification systems as well as clarifying that the patient group and treatment are similar. It is suggested that if the patients, treatment, and category of change are the same as those of the target situation, the study should provide useful and relevant information about the measure. If the context is not the same (not the
same type of change as the user wishes to estimate), then the reader is best served by moving on: even if the study is very impressive, it will not provide the information needed for the intended application.

In conclusion, the present paper puts forward a taxonomy for the change being quantified in studies of responsiveness. By defining responsiveness as a context-specific attribute (with the context being the type of change the study is designed to measure) we are more likely to be able to interpret the meaning of change scores and determine the comparability of estimates of responsiveness across different studies and different instruments. The goal is, like Guyatt's, to simplify "what appears to be chaos" (Guyatt et al. 1992b)(pg 1353) in the field of health measurement.
Chapter Three

ARE YOU BETTER? A QUALITATIVE STUDY OF THE MEANING OF RECOVERY

3.0 Preface

In the taxonomy of responsiveness (Chapter 2), it was identified that the determination of estimated and important change could come from differing perspectives: the patient, the clinician/researcher, the payer or society at large.

Health status instruments, such as those designed to describe the health of persons with upper-limb disorders, have the goal of capturing the patient’s perspective (Wood-Dauphinee, 1999; McHorney, 1999). It therefore seems logical that considerable attention be paid to the perspective of the patient when trying to describe and explain changes in these measures.

What does it mean to get better? What is the patient’s perspective on this question?

In much of the responsiveness literature the external marker of change is a question which asks something like: Overall, would you say that your arm problem is better?, the same?, worse? And these seem entirely valid (however imprecise), as they reflect the ultimate goal of health care, that is, to help the patient feel better and they also parallel the types of questions asked in actual clinical practice (how are you now? did that treatment help?). Ultimately, we would like to have a measure (i.e., a self-report questionnaire) that would be able to capture this shift to a “better” state as defined from the patient’s perspective. However, what is often observed is that in those who say that they are better, the change scores on the health status index are highly variable and, as Guyatt says, their interpretability remains elusive (Guyatt et al. 1994). Perhaps some of the variation and elusiveness is because first, we have not yet understood the
latent trait of "being better", and second, we are assuming or hoping that most of this shift is capable of being detected on the questionnaires we are using (i.e., health status measures) which may or may not be the case.

This chapter describes the meaning of being better from a "patient's", or an injured worker's perspective. Using a qualitative approach (grounded theory), the study set out to listen to that perspective and to develop a theory of what it means to be better from that perspective.

3.1 Introduction

Patients go to their clinicians with individual worries, concerns or complaints. Treatment is aimed at relieving what ails the patient. Thus, the question "are you better?" is a fundamental one for a practicing clinician (Feinstein, 1996). Following treatment, responses to the question "Are you better?" give critical information on the effect of that treatment and guides clinical decision making regarding the next step (alter care, terminate care etc.) (Feinstein, 1996). Few would debate the significance of this information in the management of a disorder at an individual clinician-patient level. If we knew more about what being better meant to patients, then we might better be able to help them.

Health care evaluation has recently turned to the use of standardized measures of health status and quality of life in order to provide some consistency to this meaning in an era of accountability (Relman, 1988). This has led to a search for the elusive "minimally clinically important difference" (Feinstein, 1998; Redelmeier et al. 1996a; Wright, 1996; Redelmeier et al. 1996c) in the scores of these measures which ideally would identify when an individual or a group is better. However, that fundamental question of "are you better?" is at risk of getting lost
in the numbers and statistical debates (Michell, 1986; Nesselroade, 1991). Recalling Lord's warning in 1957: "the numbers do not remember where they come from" (Lord, 1953), it behooves us to struggle with the underlying meanings. What does it mean for a person to say they are better? Only after answering that, can we determine if getting better is actually reflected in change in health status scores.

Perhaps the most pressing needs for understanding questions of "are you better" are in areas where there are fewer observable characteristics of the disorder to use to determine progress. In contrast to a healing fracture or laceration for instance (the healing of which can be evaluated more easily with xray or clinical exam), painful disorders of the soft tissue require the person to describe the impact, and the recovery from their perspective (Borkan et al. 1995).

Work-related musculoskeletal disorders of the upper extremity (WMSD) -- also referred to as repetitive strain injuries (RSI) or cumulative trauma disorders (CTD) -- are disorders of the muscle, tendon or nerve that are caused or aggravated by work (Armstrong et al. 1993; Hagberg et al. 1995). In the province of Ontario in 1992, these disorders represented up to 22,000 new lost-time claims within the workers compensation system in the province of Ontario and were second only to low back pain in terms of measurable burden (wage replacement benefits, lost days) within that system (Beaton, 1995). These disorders can include very specific conditions such as confirmed carpal tunnel syndrome, but the majority of WMSDs are pain-related syndromes that often elude specific diagnostic criteria (Buchbinder et al. 1996b; Beaton, 1995). As such, people with WMSD often present for care of truly "what ails them" without necessarily having clear, specific pathologies (Feinstein, 1967). This population would therefore be ideal for the exploration of the meaning of being better not only because it would help us to understand
their experience with this significant disorder (WMSD), but also because they have likely had to articulate their understanding of their symptoms and recovery in the past (to themselves, or to health care providers). They might therefore be more at ease with exploring this area in a qualitative study than people with more easily diagnosable disorders.

The purpose of this study was to develop a theory of the process and meaning of moving from an ill state to one of “being better” in a sample of people with work-related musculoskeletal disorders of the upper limb. Both the concept of this transition, as well as the factors influencing it will be explored.

3.2 Methods

3.2.1 Methodological approach.

Constructivist grounded theory (Charmaz, 1993; Charmaz, 1988; Charmaz, 1990; Charmaz, 1983) was selected as the qualitative approach for this study. It is a qualitative approach which focuses on building a new theory which is rooted, or grounded in the interviews of the participants but which also permits both factual and constructed realities to be part of the theory.

We used this approach for three reasons. First, our research question was one of developing theory around a process (becoming better). Grounded theory approaches are often considered to be best suited for evaluating the meaning of processes (Strauss et al. 1990; Creswell, 1998). Second, the methods used within grounded theory (i.e., constant comparisons between findings, analysis as data is being collected, and sampling based on theories emerging from the data) all seem well suited for this study in which we had no a priori certainty of who should be involved. Grounded theory would allow us to get a fuller understanding of the
concept because our theory was not yet developed. And third, the constructivist grounded theory approach developed by Charmaz (Charmaz, 1993; Charmaz, 1988) and adopted here incorporates a more abstract level of the 'meaning' of what is being said, and the construction of that meaning based on physical, psychological and social contexts (as shown in her work with chronic disease) (Charmaz, 1990). In contrast with some more traditional grounded theorists, Charmaz tries to "get at meaning and not at truth . . . a constructivist grounded theory may remain at a more intuitive, impressionist level than a realist, positivist approach" (Charmaz, 1993)(pg. 11).

3.2.2 Setting and Sampling strategy

This study took place within the context of a larger workplace-based study of work related musculoskeletal disorders of the upper limb (WMSD) or RSI (repetitive strain injury) (Polanyi et al. 1997) carried out at a large urban newspaper.

Twenty-four participants were recruited from respondents to an earlier workforce survey conducted in March 1996 (Polanyi et al. 1997) who had agreed to participate in additional research in the large project (56% of the 1007 subjects agreed). Each participant in the study had sufficient pain in their upper limb to meet one of two epidemiological definitions for WMSD/RSI (Hunting et al. 1994; Bernard et al. 1994) (Polanyi et al. 1997; Beaton et al. 2000). The least stringent of the survey based definition required a self-report of pain in the upper limb occurring at least 3 times in the past year, or lasting at least five days (Hunting, 1994). Of the 24 participants, 45% were female; 70% had sought some sort of care for their symptoms.

Open, purposive sampling was used for this study (Strauss et al. 1990). Initially cases of RSI were selected from a randomly sorted list of people willing to participate in this phase of the
research (and having met the case definition). This included both women and men, as well as both editorial and non-editorial workers. Two people volunteered for the study in the early stages and were also interviewed. As theory emerged from these experiences, more severe cases, people seeking care, and opinion leaders (based on feedback from the early participants) were selected to enhance theoretical sensitivity (Strauss et al. 1990). The sample size of 24 persons is generally considered adequate for a qualitative study design. The sample size was determined before beginning the study as the study in order to balance involvement in this study with broader aspects of the overall project at this workplace. This study was reviewed and approved by the research ethics board of McMaster University, Hamilton, Ontario.

3.2.3 Data collection and analysis

One hour audio taped interviews were conducted by one researcher (DB) in a private setting on work time. A loosely structured interview guide was used and modified as new themes arose (see Appendix H). Immediately following the interview notes were taken documenting key findings, links with other findings, new questions and emerging themes (Miles et al. 1994). The audiotapes were transcribed and imported into Nud*ist (Qualitative Solutions and Research, 1996) software. Transcripts were compared to the tape by one researcher (DB) for accuracy. Corrections were made as necessary.

Analysis proceeded along the methods described by Strauss (Strauss et al. 1990), and Charmaz (Charmaz, 1993; Charmaz, 1990; Charmaz, 1988). As is typical of grounded theory, the first stage, open coding, was done as the data was being collected (Strauss et al. 1990; Charmaz, 1988). In this way both the field notes and the open coding could influence the
subsequent interviews and the theoretical sensitivity of both the data collection and analytic processes. The first six interviews were coded by two researchers (DB, and one of three other researchers) in order to expand the breadth and depth of the coding, as well as ensure the transparency of meaning in the codes as documented in a coding manual. Another six interviews were coded by a pair of researchers (DB and one other researcher) using this manual, and expanding it as necessary. This dual coding helps to verify that the findings can be replicated by another researcher, and are not biased by reflecting the view of the principal coder rather than the voice of the participants. This is an important aspect of establishing the credibility of a grounded theory study (Lincoln et al. 1985a; Lincoln et al. 1985b). Discrepancies in interpretation or coding were resolved by discussion between coders and documented. Once the manual had been established, the remaining 12 transcripts were coded by one researcher only (DB) using the coding manual. The potential for a coder bias in the transcripts coded by only one researcher was minimized by the use of the manual and ongoing discussion of results with other members of the worksite study team, and a co-investigator. Finally, the earlier transcripts were re-coded using the final set of codes to pick up themes that had been missed.

Open codes were then joined together by "axial coding", or codes that link between the initial codes by asking three questions of the initial open codes: What led to this? What was the context in which it happened? What was the consequence? (Strauss et al. 1990). At this point the work of Charmaz (Charmaz, 1988; Charmaz, 1990) began to influence the analysis, as findings were not only assumed to have face value (i.e., the meaning of the words said), but also to have value in terms of the meaning at a less concrete level, examining the context and reason for the words said. This was done by analyzing the data with questions such as: Why would that person
experience that? Why is there this reaction to this question? Memos were kept as a technique to track these links between codes and their meaning. All were recorded in a study note book or on the Nud*ist software. The transcripts were revisited to verify that the new concepts were indeed grounded in the data.

The theory was built from this foundation. Logic maps (diagrams linking the concepts and codes) and memos were used to gather thoughts about how all these fit together (Creswell, 1998; Miles et al. 1994; Strauss et al. 1990; Strauss et al. 1995). Presentations of the emerging theory were made to the work place parties (union and management, but not the participants per se), as well as to the co-researchers. The emerging theory was revised to improve its clarity, though the essence of the content remained the same, and linked to the data from the participants. During the final stage of theory building, the literature (primarily chronic disease experience and medical sociology) was reviewed to determine links with other relevant findings. Finally, as the theory was consolidated, one investigator (DB) re-coded 12 transcripts using focused or selective coding (Strauss et al. 1990)-- codes describing the theory -- to ensure that the proposed theory was indeed grounded in the original interview data (see Appendix H for a guideline used in the focused coding).

3.3 Results

The core finding of this study is that the participant’s response to the question “are you better” is constructed in a very individualized manner, and as such an affirmative response - yes, I am better - may mean very different things to different participants/clients. Being better could reflect not only a resolution of the disorder, but also a state of adjustment to an ongoing disorder.
Even in the former, the criterion for what was considered a 'resolution', to be discussed in detail below, also differed amongst participants. In addition, factors influencing the perception of self as better were found. These included the personal experience of RSI (the disorder), their appraisal of that experience, and the mediating influence of three factors: the perceived legitimacy of their illness, the comparators individuals used to ground their own experience, and their coping style.

3.3.1 Construction of 'being better'

As introduced above, participants in this study tended to belong in one of three groups based on how they reflected what they meant when they said that they were better (figure 3.1). First, there were those who had a change in the state of their disorder (resolution of the disorder): second, those who adjusted their lives to avoid the pain or symptoms (readjustment) and thirdly, those who adapted to their symptoms and learned to work with their pain and limitations (redefinition). The second and third groups could be defined as people who did not necessarily experience a change in the manifestations of their disorder at the level of pathology or symptoms, but who had adjusted in some other way.

3.3.1.1 Resolution of the disorder.

Many participants described changes in their health states as indications that they were better. This is perhaps the most anticipated response from a biomedical perspective. However, this
change did not always reflect similar concepts, neither in terms of the type of change being described, nor in terms of the amount of change being described. In fact, sometimes it was change which got them to a certain threshold (i.e., "pain I can ignore") regardless of the magnitude of change needed to get there. Therefore, we defined both the type of change and the nature of that change. The latter was to capture whether it was change that reflected an amount (magnitude of change in an attribute), or the reaching of a threshold. Both the type of change, and its nature (change in attribute versus reaching a threshold) differed amongst our participants.

**Type of state change.** Participants described improvements in different ways. For some it was an improvement in pain or symptoms, for others an improvement in the ability to do different tasks with little to do with pain or symptoms. For instance, one person suggested that, given the variability in pain tolerance, the ability to function is really the key to defining self as better (note that the numeric coding [ ] after each quote identifies the specific lines and transcript used for the exemplar, they have no other meaning):

"Everybody has a different threshold [for pain] . . . some people can put up with a fair amount of pain and still function . . . So I mean, it [being better] is more the ability to do whatever you want to do" [691:1195-1205].

Some people depended on external markers - such as the number of times the therapist wanted to see them or pain medication prescriptions as indicators that they must be better (or worse): "Therapy did go quite well because she had me down to twice a week and at one point once a week"[352:198-200]. However, such a change in schedule may or may not reflect the clinician’s view of a patient’s progress. It would be easy to imagine that other factors could lead to a reduction or increase in visits (vacation, covering for other staff, etc.).

Participants were also unique in terms of defining what attribute of the change (i.e., of
pain) was most indicative of an improvement. Frequency of occurrence, intensity and duration of pain-free activity were all described as attributes defining transitions toward being better. Therefore, even within one type of change, there could be even more subtle differences in meaning. For example:

**Intensity:** “Every diminishment in pain, every step of diminishment gave me more motivation.” [355:745-748].

**Frequency:** “It’s a lot better. Like I said, it doesn’t happen as frequently.”[418:737-738]

**Duration of time until recurrence of pain:** “That would have said to me, it’s getting better. If [my RSI] was improving the pain maybe didn’t come back until Wednesday instead of having it on Monday. OK? I can’t think of any other way....I’m going for a longer period of time without feeling uncomfortable.” [1031:1413-1441].

**The nature of the change.** Across the different types of change, there were also different concepts in terms of the nature of the change which would lead to a definition of self as better or not. The first type would be what we commonly think of: a shift or transition in the level of the given state, for instance a change in pain level that would indicate to them that they were better. Many examples of this have been given above (for instance under intensity).

In other individuals the nature of the change was more of achieving a certain level - a threshold which defined being better. Better would occur when a certain meaningful threshold had been reached regardless of the magnitude of the change required to get them to that threshold. One participant provides an example of a threshold type of symptom relief: “I just think that I can type and I’m not aware of any parts of my body”[11:224-225]. In this case the threshold was found at that point when he could forget about the injured part -- a common theme. Often it was in the context of a specific marker activity: “Gee, I did all that ironing and my hand was okay”[991:959]. Others described being better in terms of reaching that threshold when the
pain had decreased enough that they could cope with it.

“It doesn’t get, as often, to the horrible way it used to get...where you would wake up and you would have pain...I consider pain problematic when I can’t ignore it....It’s amazing just how much [pain] you can live with constantly and just ignore...it isn’t gone....I’m coping.”[272:1108-1116: 2242-2250].

Resolution of symptoms was also considered an improvement when it got to a level where the person was able to use their resources to manage and control their pain.

“You feel that you really manage it, and that you’re in control and that you know if you make sure your chair is okay and manageable and you have the wrist rests....you know you are going to be able to manage it” [268:455-460].

This theme is closely linked to readjustment (to be discussed later) but in this case it is the resolution to that point of being able to control the disorder that was the focus rather than the adjustment to an ongoing disorder (readjustment theme).

The threshold can also be understood to be very high. When asked “are you better” some people think of having to be all better (cured) as shown in this response:

“that’s a good question ...because it’s a deceptive question isn’t it?....I will say Yes, but then that’s not it...I still have it [RSI]”[355:956-965].

Finally, regardless of the concept or nature of the resolution of the disorder, participants were able to evaluate the change in terms of whether or not it was important to them.

“You feel good about yourself when you can do that, things you used to do all the time....[Int: so that is important?]...Knowing that I could do it again...important, well you know. I can miss them [activities] if I had too, but important in the way that knowing that I could do it again”[252:965-996].

Changes could happen that were not important, examples given included small or short-lived changes in symptoms.

The resolution of the disorder, as defined by both the type of change and the nature of that
change (transition versus a threshold), was by far the most common underlying construct for ‘being better’ in this study, though many different perspectives were found within this theme (type of change, nature of change, importance). It is also the most common frame of reference for the health professional, and hence might be what we are expecting clients to be articulating when we ask ‘are you better’? However, some participants in this study did not frame being better in terms of a change in their disease state. These people had other constructs underlying their definitions of being better, which except the existence of an ongoing disease state.

3.3.1.2. Readjustment of life to accommodate the disorder.

A group of people described adjustments to their daily activities or lifestyles that have helped them to work around, or avoid aggravating the disorder that is still present. We have called this readjustment. One woman who still has ongoing symptoms but has made adjustments to her lifestyle to accommodate the disorder, is able to describe herself as better because of what she is doing (or not doing) to manage the symptoms. This quote highlights the sometimes delicate balance and sacrifices this readjustment of life requires.

“It [the RSI] is problematic, but it’s manageable. So maybe it is not problematic? ...I can manage semi-effectively...it’s well. I guess you could say it’s a problem now if you look at the fact that I’m devoting all my energy into work and getting better. I can’t imagine having to pick that child up at the end of the day. I want kids but I can’t imagine having to do that. You know...having a life outside of journalism. That would be problematic. I’m manageable now.....putting so much energy into [therapy] and then working. That just sucks up my whole life. If I had to do anything outside of that...serious problem.”[119:967-984].

Many other examples were found including adjusting work stations, using braces, altering techniques used at work, abandoning or carefully selecting leisure activities. All are adjustments being made to life activities in order to work around the disorder.

“So I was doing very little typing and able to keep things under control. Keeping things under
control is different from being cured" [106:395-398].

3.3.1.3. REDEFINING MEANING OF SELF, HEALTH, AND QUALITY.

A number of people described yet another construct underlying the meaning of being better. These people had adapted to the existence of the disorder in their lives. It differs from readjustment in that these people are not changing their activities to accommodate and avoid pain (for example) but rather have adapted to the pain itself as part of their lives.

"[The pain] is not getting better, it is a constant. It’s there every day you know...[and later in the interview]...I just won’t give in to it, I try to keep on doing...I want my life to continue as it is. I live with it, that’s what I do." [1015:360-362; 686-694]

"What is the opening line of that book called A Road Less Traveled? The first line is Life is Hard. Everybody has pain. Everybody has pain in their heads and their bodies...I don’t think just because somebody is hurting they can not work, don’t work, shouldn’t work....maybe at some level they should." [344:789-797]

These quotes, and others reflect the adaptation to having a painful condition. Participants in this group would forcefully state “I’m not better, I’m coping....it won’t go away.” [272:2250-2252] as an initial reaction to the question “so, are you better?” Adaptation to the disorder, much like adaptation described in other qualitative studies of chronic diseases (Charmaz, 1983; Reid et al. 1991; Shaul, 1995; Sprangers, 1996) should not be equated to a cure, or to a state that should be an acceptable goal for those seeking to manage the disorder, particularly in the case of work-related disorders. It was clear that these people were moving on with life despite, not without, their pain. Their pain shifted from the dominant part of the life they shared in the interview to a less prominent part. Shaul captures this in describing mastery: “not control over the disease, but mastery in living with it, a redefinition of what constitutes quality of life”(pg. 295) (Shaul, 1995).
Others have described similar adjustments made by persons with long term illnesses (spinal cord injury, cancer, AIDS, arthritis (Allison et al. 1997; Shaul1995; Rapkin et al. 1994; Reid et al. 1991; Ross,1989)). Allison described these shifts in perspective, and says the quality of one’s life should be considered a “dynamic construct” (Allison et al. 1997), suggesting like Rapkin, that definitions, outlooks, goals etc. may change in order to maintain “a sense of well-being” in life (Rapkin et al. 1994). This study is finding the same results in a disorder with a less well understood course.

“Nobody likes to feel that they’re the victim, and I don’t like that feeling myself. So I, myself, wanted to be in control. I didn’t want to suffer pain all of the time, and I wasn’t wanting people to pity me that I had pain” [396:1147-1177].

In summary, three different health states were described by the participants when they spoke of being better. Being better can reflect a shift in some aspect of the disease state (resolution), or a complex readjustment or redefinition of their lives to accommodate an ongoing disease state. Though the response to the question “Are you better?” might be a simple “yes”, our participants explained that the underlying latent construct is not only complex for a given individual, but highly variable across individuals. There is also some suggestion that it may differ within individuals over time (redefinition for example).

Along with these three different states (resolution, readjustment, redefinition), participants also provided insights into how they construct their response to “are you better?”.

As shown in figure 3.2, participants when asked if they were better thought about their experience of the disorder, for instance in terms of the impairments (i.e., pain, stiffness), the effect it had on their capacity to do specific activities (i.e., doing the laundry, writing) and finally the effect it had on their ability to participate in their roles, in particular their work and family
roles. The participants then moved through a phase of appraisal of that experience (see Figure 3.2 again) where they decided if they were indeed "sick" or if their pain was a "problem". Both

![Diagram of appraisal process](image)

**Figure 3.2** A description of the process of formulating a response to the question "Are you better?" and the resultant states that could be described in an affirmative answer. Participants, when asked "Are you better?" reviewed their own experience and their appraisal of it and then formulated a response. See text for detail.

these processes in turn influenced how they decided if they were better (resultant states as described above). Figure 3.2 also shows that there were influential factors that could be barriers or facilitators to the process of declaring oneself as "better" or indeed, "worse". Participants went through this process, often describing what it was like at its worse and then describing ways
in which they appraised that experience as a way of explaining why they were answering "yes, I am better". Figure 3.2 therefore captures the process participants used in developing an answer to the question Are you better?. Each component will now be described.

3.3.2 Personal appraisal of the disorder.

Three clusters of experience emerged from the interviews. For the purposes of this paper, we will limit the discussion to a brief introduction of each and then how it relates to the process of being better. The clusters that arose were easily labeled using the same titles as the ICIDH-2 (International Classification of Impairment, Disability and Handicap, second edition) classification system (World Health Organization, 1999). This framework is also helpful in that it does not dictate a hierarchical arrangement between these clusters thus allowing an individual's experience to be in any or multiple levels.

Body functions. Participants clearly described the experience of symptoms such as pain and numbness at varying locations in their arms, as well as varying in duration and intensity. However, other symptoms were also experienced, including the emotional distress attributed to the burden of pain, and limitations in capacity and participation (Kleinman et al. 1992; Eccleston et al. 1997; Reid et al. 1991):

"I get bitchy and I cry......like last night I was sitting on the couch...and lean on me the wrong way, and then I just break out crying......I want to cut my arm off. I wish I could just do that, you know. I'm right handed so I can't do that."[1015:701-716].

Others described episodes of depression and periods of feeling: "so confused and so unhappy and so scared and so worried"[352:708-710]. In order to be better, elements of the emotional burden must be overcome along with the symptoms or physical limitations. thus they
form an important part of the disorder experience. As Eisenberg suggests, all aspect of what is affecting the ill person will likely need to be addressed in an effective treatment (Eisenberg, 1980). One participant described the approach she would use to help enable someone to overcome the emotional impact (fear and anxiety in this case) of having RSI:

"I know this upsets you, you can control it, it is never too late. you can make it better, you are in control of your fate, you know, those are the things that give people some feeling of power over their circumstance. But fear is not one of them. Fear only produces helplessness and flight and escapism." [171:953-958]

The need to overcome or adapt to the emotional distress of the painful disorder is an integral part of the process of becoming better (Eisenberg, 1980; Charmaz, 1983).

*Activity or capacity to do activities in their lives.* The experience of having RSI was often described in terms of function, or the ability to do different activities. Participants described what was difficult: gardening, grasping, washing clothes, typing, holding pens. Often this was the area of major concern for them, surpassing concerns about specific symptoms. For example: "I can't type as fast as I think"[11:290] or "I hated not being able to open a jar"[171:169]. In fact, when asked what it was like at its worse, participants often shifted and described the burden in terms of activity. As an example, at its worse, they were unable to pick up groceries, hold their kids, shower and so on - "just the little things you take for granted"[1015:442]. In terms of being better, this emphasis was also found. Being 100% meant "being able to do things outside of work"[106:339].

*Participation in roles.* Many described difficulty participating in their roles, be it leisure, work or self care. However, two themes with particular application to getting better warrant special
attention. The first was the level of commitment to their work -- a theme labeled work ethic -- and with that, the impact of not being able to work because of their pain. The second was the need to balance roles of parenthood (predominantly mothers), and the disorder.

Participants in this study were either passionate about their own work, or about working at this newspaper. For the reporters and editors, who had competed to get to where they are, the ability to write and to write for this paper was very important in their lives but was also threatened by their RSI:

"I think it's the nature of journalism. We're very competitive...it attracts competitive sorts of folk...one of the hardest things, I think, when you come down with RSI is to understand that your horizons are diminished."[119].

"Not typing [is the solution to the pain]. Yes, absolutely. And yet I love to type. I like all the mechanics of my work, I love pens, paper. I have a pen fetish...I like many different kinds of pens, but I won't write with a ball point...fountain pens are my favourite...so I like all the equipment, I love it! I'm infatuated with it! I love my computer! I love the internet! I like all that stuff, so, you know, if it gets to the point where it hurts me to type, that will be horrible".[171:515-527]

Along with this passion for writing and for their work, came deadlines and a work ethic that often meant they would not stop even if they were experiencing pain: "short of a heart attack at 9:30 [pm] you get your work done"[80:1053]. Similar ethics and pressures were found in other departments, though often motivated by a need to meet quotas and deadlines for advertisements, reducing burden on co-workers, or being a representative of the paper to the public/customer. For instance one person answering the phones in circulation, can define a bad day for RSI by how late the papers are in getting out.

"Today would be a particularly bad day....the papers were three hours late today...I think there were just press problems. So when the paper is late that increases the volume of calls and you just try to get through them as quickly as you can...it's pretty stressful...It's just if the papers aren't late it's not as tense"[352:230-276].

Another describes the area manager's way of making those bad days a little less tense:
"The manager of the department brought in...donuts or something. Just to, sort of say I know it's a bad day and thank you...and that helps. That they're acknowledging that day is worse than the norm...So that kind of makes you smile. Me anyway. So the next customer that yells at you isn't quite so bad."[693:1092-1109]

The participants in this study almost universally described pressure in their work environment, but it remained where they wanted to work. This poses a particular challenge to getting better in that one source of their pain (working under deadlines and doing a lot of keyboarding) was also very much what they want to do. In a sense, their body is obstructing their ability to work – a tension Gadow refers to as the breakdown of the self-body unity (Gadow, 1980). Their passion is also associated with their pain.

Others shared the difficulties they were having with their parenting roles, and often framed being better in terms of when they could do things like pick up their children again. Mothers described their shift in roles from one concerned about the needs of others to one who needs to focus on themselves. One woman describes having to say her children: "listen. I can't help. You'll have to do it yourself"[419]. Being better would mean either a shift back to the original roles, or retaining this new role with caring for self as a priority - what Reid (Reid et al. 1991) describes as a challenging and often permanent shift for the whole family. The same participant later said: "you just learn to adjust...You learn to accept things. That you're not like before."[419:329-330]. Furthermore, being better may not be achieved until the impact on home is alleviated:

"It is a problem. As long as it's interfering with...not just my work, but outside work, it's a problem....I think the hardest part was not being able to lift up my kids...that kind of stuff." [106:470-475,494-499]

Like with other disorders - chronic pain for example – the full impact of RSI must be considered, and rarely will this be limited to only the symptoms and biological findings. Our participants
recounted both the psychological and social impact of this painful disorder as well (Kleinman et al. 1992).

3.3.3 Evaluation of experience.

People gather their experiences together and make a decision about their health state. Is this normal or abnormal (Eisenberg, 1980) for them? Medical sociologists speak of this appraisal and transition (Reid et al. 1991; Eisenberg, 1980; Chrisman, 1977; Mechanic, 1972; Good et al. 1980) and describe it as particularly difficult in disorders that are symptom based (like RSI) in contrast to those having a clear underlying accepted pathology (i.e., cancer, arthritis) (Reid et al. 1991; Brown, 1995; Chrisman, 1977; Good et al. 1980). This phase of self-appraisal comes before an encounter with a health provider, and hence is a personal and social phenomenon (Chrisman, 1977; Reid et al. 1991; Brown, 1995; Eisenberg, 1980).

No I'm not better because I am not sick.

Some participants in the study had symptoms, but did not feel they had “RSI”. In fact they did not describe their pain as a problem at all, despite meeting a perhaps overly sensitive epidemiological case definition for having RSI. (Hunting et al. 1994; Beaton et al. 2000). Rarely had their symptoms interfered with their capacity to do things, or ability to work. In general, they experienced some symptoms, and the symptoms resolved. In figure 3.2, their course is described in the upper arrow, briefly touching on the experience of resolution of symptoms, to result in a resolved state, but one that is not considered “better” because they had not appraised themselves as sick. They merely had the resolution of inconsequential, non-problematic pain.

"How do I know if it's better? See, it's not a question of better or worse. I don't think it flares up enough time for me to take concern. ...I don't necessarily say that my pain is getting better. I
don't think it's there to begin with."[866:555-562]

Although this may seem a concern, it is far from the first time it has been described. In fact, Chrisman (Chrisman, 1977) details all the factors that could influence an individual making the decision that their symptoms might reflect a disorder of concern. Reid, in her study of RSI, clearly described the period she called "symptom evaluation" that was the period where people knew they had symptoms, but they did not consider them problematic (Reid et al. 1991), attributing them to natural aging or activity outside of work. For these people, symptoms had not exceeded what they considered to be the range of normal for them, what is to be expected - a decision made by comparing self to their current understanding of being well (Good et al. 1980; Brown, 1995; Eisenberg, 1980; Mechanic, 1972; Allison et al. 1997). Few participants in our study described themselves in that state, but the majority of the others who believed they did have a problem remember being in a state of uncertainty, or denial before defining/realizing their symptoms were 'a problem'. This suggests that of those who go on to have more acknowledged problems, this failure to address symptoms as RSI in the early stages was almost universally experienced. These people often expressed that, in hindsight, they wished that they had paid more attention to those early signs and symptoms.

*Yes, I am sick.*

Persons who decided through either a conscious appraisal or an experience of illness that made it quite clear to them that they were ill, experienced one or more of the three states of being better. Within this group were those whose experience was either mild or brief ("yes I am sick, but not that bad") , and they seemed to pass directly through the resolution of disorder state, with no hint
of either adjustment or redefinition as it was not necessary: “Like a mild ache....it’s not so bad that I can’t use it, but I know it would feel better if it didn’t [ache]” [693:507-509]. Others showed a dominance of one type, but may also have had to do some readjustment or redefinition.

3.3.4 Influencing factors.

Three factors were found that would influence the appraisal of the experience, and the declaration of being better, particularly the states of readjustment and redefinition. They are the perceived legitimacy of the disorder, the comparator(s) used in evaluating their symptoms, and the coping style of the individual.

3.3.4.1 PERCEIVED LEGITIMACY OF THEIR DISORDER.

Reid has described the long difficult ‘pilgrimage of pain’ for women seeking a legitimate diagnosis of RSI in Australia. RSI in our use of the term is a cluster of soft tissue disorders many of which elude specific medical diagnosis (pathology). It is therefore up to the sufferer to define themselves first as ill (and then pursue a diagnosis of RSI), and then as being better in this same cloud of uncertainty (does it ever go away?). These self-defining illnesses are often open to disbelief and stigma (Kleinman et al. 1992; Eisenberg, 1980; Brown, 1995; Reid et al. 1991; Eccleston et al. 1997). Examples of stigma against RSI at our workplace were found:

“But every time I’d go like this [rubs arm] and they [co-workers] could see -- Oh there goes xxx again, he’s building up to it [taking time off work]...They laughed at me”[252: 386.598-602]

As Eccleston (Eccleston et al. 1997) suggests, “pain is most often constructed as a symptom in need of corroborating evidence” (pg 700) (Eccleston et al. 1997).

Our participants described their own pilgrimages towards getting a legitimate diagnosis:
"You'll hear this story a million times over. When you initially get RSI as a reporter... RSI is very tough, because there are no physical scars. So you're there... there's an initial "Is she really sick?" kind of thing. Secondly you can't get a proper diagnosis. Your doctor, your family doctor isn't going to believe you, because he doesn't even know what RSI is. So you're stuck between the medical establishment not understanding and the workplace not understanding. Then you go to workers' comp which is a whole other nightmare, because they don't understand RSI, especially back then."[119:1154-1201]

Perceived legitimacy is reaffirmed as an issue by the situation in which it was overcome. In several participants it became a non-issue once they had received recognition as an excellent worker, either inside or outside the workplace. RSI was often considered a way out of doing work: "people [with RSI] always get derided as slackers"[171:1162]. In their minds their disorder became legitimate when they (with their RSI) were recognized for excellence in their work, an indicator therefore that they were not trying to get out of work by claiming to have RSI: "My stock had gone up [after getting recognized for work]. I think that was really important to me because now I don't feel like a liability, I feel like an asset"[106:592-606].

This regaining of self respect despite the ongoing pain was key for this person in terms of his/her sense of being better: "You are doing a good job even with RSI" [106:1175]. RSI is no longer the source of their self-definition, and getting better doesn't just mean wanting to work again.

3.3.4.2 Comparators.

In order to define the meaning of one's pain, people make social comparisons (Gibbons, 1999; Eisenberg, 1980; Chrisman, 1977). Gibbons has described both the positive and negative influences these can have on sense of well being (for instance, "well at least it isn't as bad as 'x'" or "oh, look where I am headed"). Participants made comparisons internally, perhaps to other more serious health concerns that they may have adjusted to (Gadow, 1980) in their lives, or externally to other persons with the disorder. In our participants the internal comparisons tended
to lead to a better sense of well being.

"I mean the old thing "why me?". Well, why not you. It's going to affect 1/1000 people why shouldn't it be you?...I think in some cases you just press on and get used to it or things like that...I've had some setbacks and heartbreaks, but what the hell! As I say, it's part of life's rich teeming pageant, what do you expect?" [351:1024-1046]

And from a participant with permanent physical impairment:

"People's coping skills...depending on what else is going on in their lives...That was my first reaction -- nothing can happen to my arm, I can't let it...I know what it's like to lose something you know" [268:140,573,704]

Without that internal comparator, other participants described their experience in relation to other people with RSI. The stories of the "other" helped to frame their experience, their treatment choices and their expectations. These others were often co-workers, however family, or friends of the family were also used. These stories often took the form of "horror" stories, portraying a negative view of the experience.

"I knew a girl and she said that her mother is practically crippled in her one arm from RSI. And I'm like, no does it get that bad?" [668:693-698]

These comparisons had influence in ways very similar to those summarized by Gibbons (Gibbons,1999). In the internal comparison the person felt a distinction between their RSI and this other condition and described a "not as bad" or coping attitude. In the external comparison group, the downward comparison target was perhaps where these people could be and hence the results were more negative, creating a negative expectation.

3.3.4.3 COPING STYLE/ SELF EFFICACY.

In the quotes given above, as well as in every other interview, individuals described their coping styles. By coping we mean a set of beliefs about being able to manage the RSI problem: either through a belief that they can effect control over their symptoms (Bandura's self efficacy)
(Bandura, 1997), or that they can adjust to the challenge of having RSI (Reid et al. 1991; Gadow, 1980). In our study those participants who had either had symptoms for a long time or believed RSI would never go away ("but it doesn't go away" [1031]) often articulated their coping pattern. Some had difficulty seeing any way of coping: "I can't bear to think about going on like this" [352:1000], with little sense that they could effectively manage their pain or situation. Others had a more active coping style, and feeling more effective in their role they described more involvement in the problem solving and management of their symptoms. This latter group were also the ones more likely to describe themselves as better in terms of redefinition or readjustment end points, and described their attitude as being key in coping with RSI: "I think it has something to do with attitude. I won't let it defeat me" [419:350]. However, given the cross sectional nature of this study, it is impossible to determine the cause and effect nature of this relationship because we are unable to track whether this was an evolving attitude, or a preexisting one. Positive coping could be a result of redefinition, rather than a cause of it.

The literature in self-efficacy, coping or locus of control - all indicators of a persons feeling that they can (or cannot) do what is necessary to manage their pain - suggests a mediating role for coping style in the management of painful disorders (DeGood et al. 1992; Turner et al. 1987; Bandura, 1997), the most studied being chronic pain. Furthermore, some suggest that it can be a learned skill (Bandura, 1997).

The description of legitimacy, comparators and coping style are presented as potential influencing factors, not alternatives to the process of being better. They influence perceptions, and create barriers or catalysts to being better (Chrisman, 1977; Brown, 1995; Eisenberg, 1980; Bandura, 1997). They can make the process of perceiving oneself as better easier, or more
difficult, because of the way they shape the perspective/attitude/fear one has about having RSI. However, they are not essential to the process, as it is also possible that people can get to the same end point without these influences.

3.4 Discussion.

The main finding in this study is that individuals construct their response to questions like "are you better?" in unique, highly individualized manners, both in terms of the type of change that they considered to be indicative of improvement, and the importance of that change to them. Our work would therefore suggest a link between Feinstein’s (Feinstein, 1996; Feinstein, 1967) fundamental questions: “What’s wrong?” and “Are you better?”

In our participants, the experience with the disorder influenced the experience of getting better. Borkan’s work with low back pain had similar findings: “the “illness behaviours” can only be understood within the framework of past experience, personality, family and cultural standards and current interpersonal interactions (Borkan et al. 1995). These factors act as “filters for the physiologic changes involved in pain” (p 985) (Borkan et al. 1995). Participants also helped us to understand that there are different states that could all be considered “better”, they do not all reflect a reduction in the disease process but can also be an adjustment to self-definition or the environment.

3.4.1 Limitations

There were limitations to this study. The sample size was small - only 24 persons - which may affect whether or not we achieved a full understanding of the meaning of becoming better (theoretical saturation (Strauss et al. 1990)). Also we only had one interview with each subject
as each interview was conducted on work time, and the subjects were involved in other parts of the overall study as well. Re-interviewing the early participants may have helped to verify the theory arising over the course of the study with the initial participants. Another limitation could arise from the fact that we interviewed these participants on work time, and in the context of a study of work-related arm pain, in an office on site. We may therefore have heard and developed a more work-oriented theory. Though still valid and, in our opinion, strengthened by the richness of this setting, the resultant theory may differ from what might have arisen if the interviews were conducted in the participants homes or in a community health setting.

Finally, we sampled people with WMSD because it had less observable findings than something like a fracture or laceration, and hence people may have been accustomed to articulating and exploring the meaning and course of their problem. Although this was a good choice, there is a need to test the theory in another population (such as easily definable, well accepted disorders) to make sure the results are transferable.

3.4.2 Implications.

Along with describing the complex and highly individualized nature of responses to “are you better?”, this study has three important implications: first for clinicians, second for researchers doing similar worksite based studies of WMSD, and third for the outcome measurement movement.

3.4.2.1 IMPLICATIONS FOR CLINICIANS.

For clinicians, such as doctors or therapists, this study points to the need to understand the full impact of the disorder on the individual, perhaps taking a more client-centred stance to
understand what is happening in the person's life (Townsend et al. 1997). One frustration for clinicians is when gains can be observed in clinical measures (motion, strength) and yet the clients are not perceiving those gains as indicators of improvement - they are not saying that they are better. By understanding the complexities of constructing a response to “are you better” clinicians will be able to appreciate that the question may have very different meaning for the client, than it does for the clinicians. Working with the client to understand how the disease/disorder is affecting their life, and what their experience with it has been, the clinician could broaden their concept of what needs to be achieved to help the client feel better. Much like Feinstein’s concept of the clinician as treating not only the pathology, but what ails the patient (Feinstein, 1967). As such, apparent schisms between what the clinician feels is important improvement and what the clients says is not good enough could be redefined. Certain instruments might help to elicit the client’s perspective such as the Canadian Occupational Performance Measure (Law et al. 1990; Pollock et al. 1990), the Patient-specific index (Wright et al. 1997), or the PET (Patient elicitation technique) (Buchbinder et al. 1995a). Goal attainment scaling, if applied to an individual’s rather than a program’s goals could also be a useful tool (Rockwood et al. 1993). Any of these will facilitate understanding the client’s experience of the disorder and its impact, and facilitate discussions of improvements.

The clinician must also take note of the role of their cues in the client’s perception of improvement. In our study, the most overt sign was when the number of appointments given at therapy or the number of pills prescribed became an indicator of “being better”. Cues such as these could reflect processes far removed from clinical progress: accommodating a colleague’s absence may reduce available appointments one week and increase them the next. Far from a cue
that the patient is better and then worse, this could be a practical issue. Brown (Brown, 1995) might consider this an 'attributional error' emphasizing the clinician's role in being aware of this and other examples of shaping appropriate understandings and expectations. Clinicians should be sensitive to the potential for both negative and positive repercussions of these perceptions. As Good (1980) says, health care providers could take on the role of mirroring back a more realistic view of the disorder, providing honest, realistic hope (which our participants found helpful).

"Clinical models...are "models for" reality, influencing perception and behavior. shaping the course of illness and the evaluation of medical outcomes, and producing the realities they posit" (Good et al. 1980)(pg. 178). Thus, the way we work with clients, and the messages that gives should be considered seriously.

3.4.2.2 IMPLICATIONS FOR WORKSITE STUDIES OF UPPER-LIMB MUSCULOSKELETAL DISORDERS.

The experiences described by the participants will be relevant to others doing research in this area. Greater insight into the burden of these disorders arose, complementing the work of Reid (Reid et al. 1991)and others (Tarasuk et al. 1994)in this area. In addition, workplace researchers and care givers should be aware of the impact of these disorders beyond the workplace, and hence, the need for workers to recover their roles outside of work as well as their work roles before they may consider themselves "better". It should also be noted that the people in this study often held on to their desire and ability to work, but were fatigued or unable to carry out what was perceived to be more discretionary activities outside of work. The strong work ethic in this sample of workers should be recognized, whether it be a characteristic of people in this workforce (which we believe to be the case) or whether it be due to selection bias.
3.4.2.3 IMPLICATIONS FOR RESEARCHERS INTERESTED IN THE PATIENT’S PERSPECTIVE IN OUTCOMES RESEARCH.

The primacy of the patient’s perspective on their health and the benefit of care is reflected in the plethora of self completed questionnaires on health status, physical function and symptoms developed over the past two decades. However, in doing stochastic tests of statistical significance to show the benefit of treatments, researchers are assuming that differences in the scores from these measures will have equal meaning across individual clients in a study. The findings of our study would suggest otherwise. Two people could have the same amount of change in a certain area, but the meaning of that change could be quite different.

Similarly, studies testing the responsiveness (ability to accurately detect change when it has occurred (Beaton et al. 1999b; De Bruin et al. 1997; Wright, 1996; Wright et al. 1998)) of an outcome measure seek to determine the minimally clinically important difference — that threshold at which a change in health becomes an important one. In doing so they sometimes use an external criterion for important change (ie. a question quantifying how much better the person is since last time), against which the questionnaire scores are compared. Again, these approaches assume that a change in health as reflected in a change score will have the same meaning, and importance across people. Our work would suggest that such uniformity is unlikely. Perhaps the best criterion for change which is important would again be to use questionnaires such as the patient specific index (Wright et al. 1997) or the COPM (Law et al. 1990), which allow clients to select or to generate their own items. This would lead to different content between individuals for the criterion of the occurrence of change, but increased validity at the level of the individual.

Indeed, Guyatt and Cook (Guyatt et al. 1994) assert that responsiveness requires the
ability to make sense of observed change scores. adding that this "interpretability remains elusive for many of the questionnaires in common use" (pg 631). Our study would suggest that this will not improve until more work is done on how the presence or absence of clinical change should be ascertained — the criterion for studies of responsiveness (Testa et al. 1996). We would assert that the interpretability of the question "are you better" also remains elusive in terms of how, or whether, it can be structured in a way that allows us to continue with the currently accepted format for studies of responsiveness (group-level questions on change). Perhaps it will be more interpretable (though more complicated to carry out) if we allow an individualized construction of what 'being better' would mean for the individual clients against which the questionnaires are compared.

Our results suggest increased complexity in understanding what it means to be better for people with upper-limb pain. We have discussed the implications for both clinicians, and for researchers especially in the area of responsiveness. Our work is not making the quest for understanding this concept of being better more complex without reason. Rather we are asserting that we can no longer afford to look at 'being better' as used in clinical practice or as the criterion marker for change in studies of responsiveness in a simple manner. The complexity of this concept must be recognized given that a clinician's or researcher's interpretation of responses to 'Are you better?' may have the very real potential of influencing the care of our clients and ultimately their expected quality of life.

ACKNOWLEDGMENTS

The authors wish to thank participants in this study for sharing their experiences; Dr. Donald
Cole, Ms. Sue Ferrier and Mr. Michael Polanyi for assistance in the initial coding; and the RSI Watch steering group at the newspaper along with the Upper Extremity Working Group at the Institute for Work & Health for their feedback and support of the study. This study was supported by the Toronto Star, the Southern Ontario Newspaper Guild and the Institute for Work & Health. Ms Beaton is supported by a PhD fellowship (health research) from the Medical Research Council of Canada and by the Institute for Work & Health. Dr Wright is the R.B. Salter Chair of Surgical Research and a Medical Research Council of Canada Scientist. Dr Katz is supported through a grant from the NIH #AR36308 and the US National Arthritis Foundation.
Chapter Four

CREATION AND PILOT TESTING OF EXTERNAL INDICATORS OF CHANGE FROM THE PATIENT’S PERSPECTIVE.

4.1 Introduction

Studies of instrument responsiveness require some sort of external standard of change against which the change in the instrument is compared (De Bruin et al. 1997; Deyo et al. 1986; Testa et al. 1996). There are different kinds of change that could be used as external standards and we have described these in Chapter two: for example, observed change in patients undergoing a specific treatment or change in those who say they have improved (either an estimated improvement, or an important one). Within the latter, different perspectives can come to bear on what is an improvement (or an important improvement). The patient’s perspective is most congruent with the goal of self-perceived health status measurement (Guyatt et al. 1997). In Chapter three the focus was on trying to describe what patients would think of as a change, or an important change, from their perspective. The findings suggested that change was highly individualized, and concluded by suggesting that patient-specific indices (Law et al. 1990; Wright et al. 1997) may be the most accurate way of identifying the occurrence of change in this attribute (self-perceived health). They may in fact be the “Rolls Royce” (Guyatt et al. 1986) approach to creating markers of change. However, these patient-specific markers of change may not be feasible to apply in a study of responsiveness for two reasons. First because they are time consuming to complete, and second because the resultant scale will be quite different across patients (each has their own items), lacking the uniformity one might like to have for a group-
level analysis (McDowell et al. 1996). Although the core finding from the qualitative work was that being better was highly individualized, there seemed to be some agreement in the concepts/attributes that arose in the qualitative work in terms of the types of change that were indicative of an improvement including: pain, dimensions of pain, ability to function, ability to cope. These could be thought of as the most shared component of the experience of being better across participants in what would otherwise be a unique experience (based on the influence of the individuals experience of the disorder and their appraisal of it). We therefore felt it was important to create standardized indices of change that were more likely to reflect changes from a patient’s perspective. These indices would not replace patient-specific indicators, nor an individual’s full description of their experience of being better (as described in Chapter three) but would improve upon existing (usually researcher-created) indicators by grounding them in the themes and language of the qualitative study participants (Chapter three). Thus they could be an improvement on existing approaches in terms of being able to reflect the patient’s perspective.

In this chapter new external markers for change are developed that try to capture themes arising from the qualitative work on

<table>
<thead>
<tr>
<th>Attributes:</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Overall rating of problem</td>
</tr>
<tr>
<td>- Symptoms</td>
</tr>
<tr>
<td>severity of symptoms</td>
</tr>
<tr>
<td>frequency of symptoms</td>
</tr>
<tr>
<td>night pain/symptoms</td>
</tr>
<tr>
<td>ability to manage symptoms</td>
</tr>
<tr>
<td>absence of pain/symptoms</td>
</tr>
<tr>
<td>- Physical Function</td>
</tr>
<tr>
<td>ability to do daily activities</td>
</tr>
<tr>
<td>ability to work</td>
</tr>
<tr>
<td>do all you want/need</td>
</tr>
<tr>
<td>do normal activities &amp; forget about arm</td>
</tr>
<tr>
<td>- Emotional Function</td>
</tr>
<tr>
<td>- Social Function</td>
</tr>
<tr>
<td>- Ability to cope with the problem</td>
</tr>
</tbody>
</table>

Table 4.1 Attributes for indices of change
the meaning of being better (Chapter three). This chapter will describe the methods used to create these markers of change and the pilot testing of these indices in patients with upper limb musculoskeletal disorders recruited from a clinical setting. The purpose of the pilot testing is to evaluate if the indices created from the qualitative work are relevant (important attributes) to a clinical population, likely to detect deficits and changes in the attributes, and are comprehensive (that is, few additional attributes are suggested by the pilot testing sample).

4.2 Methods.

4.2.1 Developing markers of change

Two decisions had to be made regarding the creation of the potential indices that were to reflect change from a patient’s perspective: the attributes to be included and the conceptual approach used in the formatting of the indices. Each will be described.

Attributes. The attributes are the concepts being measured within each of the indices. Table 4.1 shows a list of the concepts extracted from the qualitative theory (Chapter three). In total there were 13 attributes that were possible indicators of improvement from a patient’s perspective. Generating these attributes was not the main purpose of the qualitative work on the meaning of being better. It is only in seeing a pattern of responses in that study that were judged as likely being able to fit in a clinical context that the generalization took place. The clinical context is that of persons with very similar upper limb musculoskeletal disorders. Other aspects of the qualitative theory were likely dissimilar in a workplace versus a clinical setting, such as the role of legitimacy or support of co-workers, and hence were not felt to be generalizable (Lincoln et al. 1985b; Lincoln et al. 1985a; Kvale, 1996). The findings from the qualitative findings were also considered to be theoretical constructs, rather than absolute truths in the new
context (Lincoln et al. 1985a) and it was from this perspective that the findings were brought into a clinical context, and a quantitative framework.

*Format.* Decisions were then made about the formatting of each index. Both the wording, and the scaling were addressed. The qualitative interviews were used to create the wording used for each index. Often the wording was a direct quote or a paraphrase of the text from the interviews - often recognizable from the quotes in Chapter three. One exception to that is the index of ability to function in daily activities. For this index, we returned to the developmental work for the DASH (Disabilities of the arm, shoulder and hand) outcome measure and selected the field testing items which patients rated as most important to include in a measure of upper limb function and symptoms (Marx et al. 1999; McConnell et al. 1999). The top ranked items were included as examples in the index about daily activities.

The wording was then finalized by using two approaches: transitional indices, or serially measured descriptions of the person’s current state. Transitional indices ask the person to rate the change in the concept between two points in time, for instance “comparing where you are now to before your surgery, how would you rate the change in your pain?”. This is a common (Beaton et al. 1997; Beaton et al. 1998; Deyo et al. 1986; Jaeschke et al. 1989; MacKenzie et al. 1986; Redelmeier et al. 1993), though criticized (Norman et al. 1997) approach to identify the occurrence of change in studies of responsiveness. Serially measured descriptions of current states ask for ratings of the attribute (pain, for example) at the present time, and then asks the same question at a second point in time (how is you pain today?). The arithmetic differences in the two serial measures is taken as an indication of change. The serial measures approach parallel data collection techniques we tend to use in outcomes research (difference in the functional status
using data collected before and after treatment). The transitional approach is often what we use in clinical practice ("how are you now compared to last week? or compared to when we started that new exercise?"). Recent research suggests that transitional versus serial indices produce different results (Fischer et al. 1999; Herrmann, 1995; Ross, 1989; Sprangers et al. 1999), but does not suggest which is the "best" or most valid method (Fischer et al. 1999). Both approaches were used in this study.

The final issue was scaling. Numeric rating scales were used for each of the indices. The majority of the state measures used seven-point scales (Streiner et al. 1991a). Streiner and Norman suggest five to seven point scales offer enough range to discriminate between levels of difficulty, while not providing too many options which would add to noise (Streiner et al. 1991b). The pain markers used 10 point (1-10) scales and this was chosen because of familiarity people have with 10 as an anchor for pain measures. Six items were scaled with dichotomous responses because they reflected concepts framed in terms of achievement of thresholds (framing that arose from the qualitative work). These included ability to cope, being free of pain, being able to manage pain, working in usual manner (normal duties, not modified), doing all the activities they need to do and finally being able to forget about their arm problem during daily activities. Each was formatted as a yes/no response with additional subcategories available to those who answered no in order to describe the degree to which they were missing this threshold.

4.2.2 Pilot testing.

A pilot study was conducted to test the items in a clinical population. As described earlier the
objectives of the pilot testing were to assess: first, the relevance of the items; second the likelihood that the items would describe the experience of difficulties and changes in a clinical population; and third, the comprehensiveness of the battery. The first objective was assessed by asking patients to rank the importance of each item in terms of reflecting "being better" or worse. The second objective was assessed by examining the distribution of responses (minimal floor or ceiling effect, responses not expected to all be in one category) along with the importance ratings. And the last objective would be evaluated by responses to an optional question where they could add additional indicators of change.

A convenience sample of thirty patients with upper-extremity musculoskeletal disorders was recruited from the waiting room of a surgical outpatient clinic at St. Michael's Hospital. Those attending this clinic would typically be coming for the follow-up after surgery, or coming for an initial consultation with one of two surgeons specializing in the management of upper-limb musculoskeletal conditions. The number thirty was chosen as it is generally considered the transition from a small sample to a sample that may begin to reflect characteristics of a broader, similar population (i.e., enough patients to start getting data that is approaching a normal distribution). It was felt that this sample could give us an indication of how well they responded to the markers as indicators of their own experience. Any unusual findings would indicate the need to pursue this further in a larger study or in more detail. The nature of this study did not require that they were at a specific stage, so participation was offered to patients who were willing to participate and were able to complete the written questionnaire package in English.

Each patient completed the package over the course of their visit, usually in the waiting
room. It asked them to respond to each of the indices created for the pilot (13 state measures designed according to the list in table 2.2, and 6 transitional measures). A full description of the indices and the source of their wording is provided in Appendix B. In addition they were able to comment on what they thought of each item, and how important this item was to include if we wanted to ask about them being better. The importance was rated on a four-point scale from very important (1) to not important at all (4). At the end of the state measures, they were asked to identify which index was, in their opinion, the best in terms of asking about an important part of their upper limb problem. Similarly, at the end of the transitional questions they were asked to choose the transitional scale that they felt was the best to ask people if their upper limb problem was changing. It should be noted that the transitional indices will reflect very different intervals of change for the different patients. The actual numeric score is therefore of limited interpretability except that we would expect to see variation in this mixed population. If the entire sample said “no change” it could be assumed that the item is not changing for this population and is therefore irrelevant.

Two open-ended questions were also included at the end of the package; one asking them to describe changes in their condition that were indicators to them that they were improving, and the second asking respondents to describe indicators that they would see as signs of improvement from their current status (defined as ‘today’, the day they filled out the questionnaire package). At the end of the questionnaire additional room was provided for any other comments.

A brief description of the region involved, gender, and age was obtained. Patients remained anonymous, and no identifying information was gathered. Data on the exact diagnosis was not felt to be needed for the goals of this study, and therefore was not collected obviating the
need to access medical records on these patients. This study was approved by the research ethics board at St Michael’s Hospital.

4.3 Results.

A sample of the indices fielded in this study is included in Appendix C of this thesis. There were thirty participants in the study, 14 males and 16 females with an average age of 43 (standard deviation of 15.6). Nineteen subjects had their shoulder area involved, eight their wrist/hand region, and two their elbow. One person did not indicate the region involved.

The tables (4.2 and 4.3) below show the results for the status indices. The number of people checking each response level are shown, as are the responses to the question “would this be an important way to ask about being better?” (1 = very important to 4 = not important at all). In the seven and 10 point scales the distributions were wide, with no index having a large ceiling or floor effect (although the mode tended to be at one extreme or the other). Floor and ceiling effects reflect the number of respondents who score at the extremes of the response scale, and can indicate a lack of sensitivity of the scale to either more or less of the attribute at that extreme. For instance, 14 patients indicated that difficulty working was 7/7 (very difficult). This could indicate a floor effect on this item. These results were checked against other markers and found to be reasonable as many subjects were not able to work because of their arm problem. Thirteen of the patients scored one (ceiling) on the social function index, and also indicated the more ambivalent response to the importance rankings (see table 4.2). This marker was therefore not likely to be a good indicator of change because it lacked importance, and it appeared that it was not affected in this sample of patients with arm disorders.

The dichotomously scored items (numbers 8-13 on table 4.2), in particular the first two
(8. Can you do all that you would like to do?, and 9. Can you forget about your upper limb in activities?) were distinctly different from the 7-10 point scales in terms of their distributions. In each case only one person indicated “yes” to these questions, in contrast to more people who indicated ratings on the lower end of the scale for ability to work, and do daily activities on the seven point scale. The dominance of “no” responses to the dichotomous markers continued in numbers 10 and 11, and then shifted when the attribute became one of coping or managing, rather than being “free” of restrictions as required in items 8-10. Sixteen patients said they were coping with their problem, but only two were free of pain. Similar to coping, 23/30 said that they could manage their pain - but some said comments like “I have a high tolerance for pain” or that they were taking pain killers to “manage” the pain - delineating the difference from being free of pain and being in control of the pain. Patients were reflecting that coping with or managing their problem does not mean they can do everything they want or need to do or that they are free of pain, a strong distinction also found as a theme in the qualitative work as well.
Table 4.2. Responses to the status indices (describing how the person is that day). The number of subjects (out of 30) endorsing that response option is shown, as well as the importance rating subjects gave to that concept as a potential index for change from their perspective. Number reflects the position in the pilot testing questionnaire (Appendix B).

<table>
<thead>
<tr>
<th>Concept</th>
<th>Rating</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>non/</td>
<td>mild</td>
</tr>
<tr>
<td>1. Rating overall</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>4. Ability to do ADL</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>5. Ability to work</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>6. Emotional health</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>7. Social health</td>
<td>13</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>8. Do all you want</td>
<td>1</td>
<td>29</td>
</tr>
<tr>
<td>9. Forget about arm</td>
<td>1</td>
<td>29</td>
</tr>
<tr>
<td>10. Free of pain</td>
<td>2</td>
<td>28</td>
</tr>
<tr>
<td>11. Work Normally</td>
<td>4</td>
<td>26*</td>
</tr>
<tr>
<td></td>
<td>*3 for reasons other than upper limb</td>
<td></td>
</tr>
<tr>
<td>12. Cope with problem</td>
<td>16</td>
<td>14</td>
</tr>
<tr>
<td>13. Manage pain</td>
<td>23</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 4.3. Frequency of responses (out of a possible 30) to status measures of pain along the 10 point scale with 1 described as “no pain”, and 10 described as “pain as bad as it can be.”

<table>
<thead>
<tr>
<th>Concept</th>
<th>Rating</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>no pain</td>
<td>1</td>
</tr>
<tr>
<td>2. Pain Intensity</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>3. Pain at night</td>
<td>4</td>
<td>3</td>
</tr>
</tbody>
</table>
The subjects were asked to rate which item they felt was the most important indicator of change from this group (if it was asked repeatedly over time). Can you cope with your problem (cope with problem, #12), and are you working in your usual way (work normally, #11) received the highest number of votes. It is also noted that ability to do daily activities was rated the highest in importance (22/30 rated it as very important) but it was not selected as the one most important attribute by the most subjects.

The transitional indices were examined in a similar manner, again with the number responding to each of the categories listed, and then the importance score for the index (Table 4.4).

Table 4.4. Transitional indices response patterns. One was defined as much worse, six as no change and 11 as a lot better. There were no other verbal descriptors on the scale.

<table>
<thead>
<tr>
<th>Concept</th>
<th>Rating 1 (much worse)</th>
<th>Rating 2</th>
<th>Rating 3</th>
<th>Rating 4</th>
<th>Rating 5</th>
<th>Rating 6 (same)</th>
<th>Rating 7</th>
<th>Rating 8</th>
<th>Rating 9</th>
<th>Rating 10</th>
<th>Rating 11 (much better)</th>
<th>Importance 1 (very)</th>
<th>Importance 2</th>
<th>Importance 3</th>
<th>Importance 4 (not)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Overall</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>8</td>
<td>0</td>
<td>5</td>
<td>25</td>
<td>4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>B. Pain severity</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td>8</td>
<td>2</td>
<td>4</td>
<td>20</td>
<td>7</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>C. Pain frequency</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td>8</td>
<td>3</td>
<td>3</td>
<td>21</td>
<td>6</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>D. Function</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>8</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>18</td>
<td>9</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>E. Work</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>8</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>21</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>F. Other (see text)</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>11</td>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

The transitional scales again had a wide distribution and few at the ceiling or floor. The patients in the study were at a variety of stages in their recovery, so there is no "correct" or valid response. If an index was irrelevant or meaningless in a clinical setting, we would expect a cluster of responses in one category ("no change" for instance) and importance ratings to be low.
Neither was found. Subjects also chose their preferred transitional scale, and 10 voted for the change in ability to function (Scale D in Table 4.4) as the most important indicator of improvement or deterioration. This was followed by the overall rating of the problem (Scale A, 7 votes) and the ability to do usual work (Scale E, 6 votes).

Sixteen (of thirty) people made use of the option to add an item — listed in the Table 4.4 above under ‘other’ — to identify some aspect of change that had not been included originally. The content varied widely: crocheting, mobility, future potential, reaching overhead, sleeping, social activities/sport, strength, and your everyday life. Four others identified concepts already covered in other indices: improvement in the overall problem, and change in pain level. Therefore only 12 people identified items that we potentially missed. Of those only sleep and social function (in a transitional format as opposed to the state indicator in item seven Table 4.2) was identified by more than one person, and in each case it was only by two patients.

The broad open-ended questions were the final component of this pilot study. The subjects in this study responded in a similar manner to those in the qualitative study when asked in an open-ended manner about the change that they felt indicated that they were getting better (or worse). Degree or pain, ability to do daily activities or chores, and ability to sleep and forget about my arms were listed as were return to recreational and social activities. When describing what had changed, the predominant theme was function and lifestyle, with only a few mentioning impairment (Verbrugge et al. 1994) level markers (range, stability). Even when mentioned, these were often put in a functional context. For example “less pain, more able to do things”, or “my shoulder is more stable, allows me to do normal activities.”

When people responded to what would need to change from today to indicate a change.
they tended to respond at an impairment level: "less pain" "increased strength", "mobility".
"getting at least 90% usage back in my wrist", "full range of motion, strength to return to work."

4.4 Discussion

The purpose of this study was to create and then to pilot test new indices of change in a clinical population (patients with upper-extremity musculoskeletal disorders). The indices were created based on part of the theory generated in a qualitative study of the meaning of being better. Pilot testing in this study, along with the open ended questions fielded as well suggest that the concepts were relevant and meaningful to this clinical population. There was no significant, or unjustifiable floor or ceiling effect on those scales with 7-11 point ranges. There was a large proportion of the sample not achieving the first four of the threshold markers (being free of pain for instance), followed by a dramatic shift towards reaching the threshold when the wording changed to coping or managing their problem.. Given our study sample, the waiting room at a tertiary-level clinic, we would not anticipate that patients would be "free of pain" or "able to do everything they need or want to do" – hence the responses seem valid. The dichotomized scores reflect a high level of function and, as indices of change, may well identify those with improvements at a high level of functioning. The ability to cope reflects the findings of the qualitative study, that coping did not mean an absence of limitations or symptoms.

The open-ended responses confirmed that there are a diversity of markers that can indicate improvement. Furthermore, there appeared to be a shift in what was considered when the subjects were looking back on the markers they had experienced, versus looking forward to those they anticipated experiencing. This study was not the correct design to explore this further, however it might be considered in future research.
The pilot work confirmed that the questions fielded had the potential to be valid markers of change from the patient’s perspective. The importance of each of the items was rated highly. Only social functioning was considered less important to include, but even so, only seven said it was not very important. Emotional functioning was next, with three saying it was not very important. All others had two or less endorsements in that response category. The majority of the respondents rated most of the items as “very important” (see tables 4.2-4.4).

Based on the results of this pilot study, only the ‘other’ transitional scale will be dropped, and the rest carried forward into a study of responsiveness to different types of change. The ‘other’ item is being dropped due to the high number of missing items in a relatively brief questionnaire, as well as the overlap between the attributes arising in the “other” category and those already found in the other indices in many instances. Both serial state (numeric rating scales and dichotomous threshold types of state descriptors) and transitional indices will be used in the cohort study. In four cases measuring the same attribute (overall rating of disorder, pain intensity, ability to function, ability to work) allowing ongoing comparison of these two conceptually different approaches. The final indices can be found in the back of the questionnaire package used in the cohort study (see Appendix C). The responses to each of the indices over time (in the final cohort study) is summarized in another appendix (Appendix E).

The pilot study showed that many different indices have importance in describing the course of upper limb disorders. A “responsive questionnaire” should therefore be able to accurately detect the changes indicated on these hypothesized indicators of true change from a patient’s perspective.
Chapter Five

VALIDATING THE TAXONOMY:

EVIDENCE FOR THE CONTEXT-SPECIFIC NATURE OF RESPONSIVENESS.

5.1 Introduction

In Chapter 2, the unique types or categories of responsiveness were described as cells within a triaxial taxonomy of responsiveness. Each cell represents a valid form of change ($\Delta C$ from figure 1.2) that could be built into a study of responsiveness. This taxonomy defined the change along three axes: who is the focus, when are the groups being compared (over time, or at one point in time) and what type of change is being measured. Within two categories along the "what" axis, estimated and important change, the perspective that can come to bear on whether or not this change has occurred was also described. Based on this work it was suggested that patients, clinician/researchers, payers, or society as a whole could each have valid perspectives on estimating change, or defining its importance. In Chapter three, our focus continued with the patient's perspective, as this is likely to be most congruent with the concept of self-perceived change in health status ("I am better"). This perspective was explored (Chapter three), and indices of change developed and tested reflecting the occurrence of change that is relevant from the patient's perspective (Chapter four).

The current chapter focuses on the description of responsiveness of a single questionnaire across different cells in the taxonomy using longitudinal data from the same patient sample. The
purpose is to evaluate whether the different categories of responsiveness will provide different description of the same instrument's responsiveness, in the same patients. The instrument used in the study is the DASH Outcome Measure (Hudak et al. 1996; McConnell et al. 1999). If the taxonomy is useful (discriminating between truly unique categories of responsiveness), the responsiveness of an instrument will vary depending on the category of responsiveness being examined if the clinical circumstances are the same (patient type, treatment, setting). Therefore, if the responsiveness of the DASH varies depending on the category of change built into the analysis of responsiveness, then responsiveness is context-specific, and it must be defined in terms of the type of change being measured in the study. If there is no difference in the description of responsiveness across categories, then the findings would not reject the null hypothesis: that an instrument's responsiveness is the same regardless of the context of change. hence responsiveness would be considered more of a fixed attribute of the instrument and not of the application. Such a finding (the null) would be consistent with the way responsiveness is currently being described in the literature.

5.2 Methods.

5.2.1 Design

A convenience sample of patients waiting for treatment of their upper-limb conditions at one of two teaching centres (St Michael’s Hospital in Toronto and Brigham and Women’s Hospital in Boston) were invited to participate in this prospective cohort study. These sites were selected because the author (DB - St Michael’s Hospital) and one of the committee members (JNK - Brigham and Women’s Hospital) worked out of these facilities and could coordinate data collection. The study did not affect their care, but rather observed their progress over a three-
month period by measuring aspects of their health status (through questionnaires only) before and after treatment.

Four groups of patients, two with proximal disorders and two with distal were targeted. The proximal disorders included those patients with glenohumeral arthritis who were undergoing joint replacement, and those with soft tissue disorders around the shoulder (predominantly rotator cuff tendinitis). Distally, patients undergoing carpal tunnel release, or receiving treatment for a tendon disorder in their wrist or hand (predominantly trigger finger and tendinitis) were the focus. Acute injuries such as tendon lacerations or fractures were not included because of the inability to obtain pre-treatment measures of disability.

No guidelines were available for calculating sample sizes for studies of responsiveness. A traditional paired sample size calculation was used, and was structured to be able to detect the differences in the magnitude of change that would be found if two different types of responsiveness were compared in the same patient sample. Type I error rate was set at 0.05 (α = 0.05) and Type II at 0.10 (β). Traditionally settings for the type II error rate would suggest a power (1-β) of 0.80. In this study it was raised to 0.90 in order to have high power in order to avoid missing a true difference if it had occurred (Rosner, 1990; Norman et al. 1994; Guyatt et al. 1987). The amount of change (delta) we wished to be able to detect was based on the difference in change scores observed in those who said that they had a large improvement in their health, and those who said that they were only somewhat better, simulating the kind of change we would like to make sure we could differentiate when looking at different descriptions of change using the DASH. This was done using data from a work site study (Polanyi et al. 1997; Beaton et al. 2000). The group that said they were “much better” had mean change scores of 5.79/100 (standard deviation (sd) = 11.2) and those that were “somewhat better” had mean change scores
of 2.42/100 (sd = 12.3). The difference between them was 3.37/100, which was the delta used. The variance came from the same study, however from the stable subgroup (those who said they were the same: mean change 0.89, standard deviation = 9.9), as suggested by Guyatt (Guyatt et al. 1987). These data were placed into a paired sample size calculation formula as described by Guyatt (Guyatt et al. 1987), and Rossner (Rosner, 1990):

\[ n = \left( Z_{1-\alpha/2} + Z_{1-\beta} \right)^2 \sigma^2 / \Delta^2 \]

Substitution results in a sample size of 113 persons. A 20% correction was added to allow for missing or unusable data which raised the requirement to 142 patients (where the target of \( n = 113 \) was assumed to represent 80% of the data). We anticipated up to 20% would not complete the follow-up package, hence recruitment was targeted at 178 patients. When that target was approached, we continued to recruit until it was apparent that we would have adequate numbers at follow-up, and that it would represent 80% of the baseline sample.

Patients at the St. Michael’s Hospital site were approached in person by study personnel (either at a surgeon’s clinic, or in a pre-admission clinic). The study was explained to them, and they were asked to sign a consent form if they wished to participate. At that time they were given the baseline package to complete and return, as well as a second package to be completed 3-5 days later (in some instances this was not done as the operative date fell within that 3-5 day window). Two follow-up packages were mailed out to the subjects with a stamped return envelope. This was done at four and 12 weeks post treatment. Up to two reminder packages and phone calls were made to encourage response (Dillman, 1978).

In Boston, study personnel identified patients by reviewing ICD9-CM (International Classification of Disease, 9th edition, Clinical modification) (Anonymous, 1993) diagnostic codes in billing data for patients receiving or awaiting operative or non-operative care for the target
disorders. These potential subjects were sent detailed letters explaining the study and inviting them to participate. The baseline questionnaire package and a stamped return envelope were included with the letter. Participation (returning the baseline questionnaire package) was considered consent. Subjects from Boston were sent their second package by mail at 12 weeks. Again phone calls and reminder packages (up to two) were sent as necessary.

The study was approved by the Research Ethics Board at both sites.

5.2.2 Rationale for the timing of data collection

As described in the previous section, questionnaires were gathered at three follow-up times in Toronto (three to five day, four weeks and twelve weeks) and one in Boston (twelve weeks). There was a specific rationale for each time point.

The three to five day follow up in the Toronto subjects was designed to capture people at a stable phase of their disorder and allow the assessment of test-retest reliability if they were indeed unchanged in the interim period of time.

At four weeks many patients (particularly the operative cases) would probably have experienced a temporary decline in their health-related quality of life (due to post-operative restrictions and morbidity). Other patients may have made small gains in function in that amount of time (particularly non-operative treatments). Four-week data was therefore analyzed in the Toronto subjects for earlier and smaller changes in health, as well as providing an opportunity to test the DASH’s responsiveness to deterioration in health status.

At 12 weeks the patients undergoing treatment for one of the four targeted conditions would probably have experienced some improvement in their function. Further gains beyond the 12-week point would be anticipated in many, if not most, cases. This 12-week point provided
the opportunity to assess responsiveness to smaller changes in health (Greenfield et al. 1992), a task requiring more measurement precisions than capturing the larger changes associated with a fuller recovery at six months or one year.

5.2.3 Measures

Data was gathered using a self-completed questionnaire package. The following sections describe the contents.

5.2.3.1 DASH OUTCOME MEASURE

The DASH Outcome Measure (Disabilities of the Arm, Shoulder and Hand) was selected as the principal measure for this analysis. The development of the DASH has been documented elsewhere (Hudak et al. 1996; Marx et al. 1999; McConnell et al. 1999). The DASH is a 30 item questionnaire, evaluating symptoms and physical function (at the level of disability (Jette. 1994; Verbrugge et al. 1994)), in any upper limb disorder (McConnell et al. 1999; Davis et al. 1999). As such it is a whole-extremity questionnaire, applicable in both the shoulder and the wrist/hand patients in this study.

There are five response options for each of the 30 DASH items. Scoring is done by summing up the circled responses and subtracting 30 (in order to anchor the score with a base of zero - a correction made necessary because the response scale is 1-5, and needs to be changed to a 0-4 equivalent). This figure is then divided by 1.2 to get a DASH function/symptom score out of a possible 100. A higher score on the DASH reflects more disability. Missing items (up to 3 items, or 10% of items) are replaced by the mean value of the responses to the other items before summing. If more than three items are missing, the overall score cannot be calculated.
Preliminary work on the validity (against constructs of function and pain) and reliability (Cronbach’s alpha = 0.96 (McConnell et al. 1999), test-retest reliability = 0.92 (Turchin et al. 1998)) has been carried out by those involved in the development (Hudak et al. 1996; McConnell et al. 1999; Marx et al. 1999; Davis et al. 1999). Others outside this group have also done work on the DASH by comparing it to other more specific measures (Kirkley et al. 1998; MacDermid et al. 1999a). These two groups also demonstrated that the DASH was responsive to change observed over the course of physiotherapy treatment for shoulder instability (Kirkley et al. 1998) and to change over the course of recovery from wrist fractures (MacDermid et al. 1999a), though slightly less so than their own disorder and joint-specific instruments. These studies provide initial evidence of the validity, reliability and responsiveness of the DASH scores.

5.2.3.2 OTHER QUESTIONNAIRES

Other questionnaires were also included in the package, although they will not be the focus of the analysis presented in this chapter (see Appendix F for a presentation of the results for these other questionnaires). Joint-specific measures were included, one for each target region: i) the shoulder pain and disability index or SPADI (Beaton et al. 1998; Roach et al. 1991; William et al. 1995). The SPADI has thirteen items each scored on a visual analog scales, eight items form the disability subscale (our focus) and five form the pain index. ii) The Brigham carpal tunnel questionnaire (Brigham) was used as a wrist/hand outcome measure (Katz et al. 1995; Levine et al. 1993; Amadio et al. 1996a; Amadio et al. 1996b). It has been tested extensively in the carpal tunnel population but its use has extended beyond that into groups of Colles fracture patients for example (Amadio et al. 1996b). The Brigham also results in two
scores: functional limitations and symptom severity, and our analysis focused on the functional limitations scale. iii) The well known Short-Form 36-item (SF-36) generic health status questionnaire was also used in this case as an indicator of overall health (Ware, Jr. et al. 1994; Ware, Jr. et al. 1992; Ware, Jr. et al. 1993; McHorney et al. 1993; McHorney et al. 1994). The SF-36 has been used extensively in upper limb patients, although some concern has been raised about its sensitivity to relevant changes in upper limb conditions — a problem not found with the SF-36 in lower limb or low back patients as the content in the physical function domain relies heavily on mobility items (Beaton et al. 1998; Beaton et al. 1997; Gartsman et al. 1998b; Gartsman et al. 1998a). In this study we used it to describe the baseline health of the sample, and to allow further evaluation of its responsiveness to change in upper limb patients. The acute version of the SF-36 was used. This version asks respondents to consider their health/abilities over a one week period rather than a one month period in the original version. iv) Finally we included the EQ-5D, a five-item (three-response options each) questionnaire that describes the overall health state (profile of responses to the five items) and assigns value weights to each described state (derived from a general population survey in Britain) (EuroQol Group, 1990; Brazier et al. 1993). The battery of questionnaires therefore spanned the different types: health valuations, generic, region-specific (DASH), and joint-specific (applied to both the targeted, and non-targeted joints). The focus of the analysis presented in this chapter will be on the DASH. The comparable results for the other measures will be shown in Appendix F.

5.2.3.3 Categories of responsiveness being tested.

According to the taxonomy presented in Chapter two, the change being quantified in a study of responsiveness can be described as a cell in a triaxial matrix. The three axes are defined
as follows: When? (when is the comparison being done -- over time or at one point in time). Who? (who is being considered -- individuals or groups?) and What? (what type of change is being quantified?). In our analysis, we will hold two of the three axes (the when and who) constant by only considering within-person change over time (when axis), and only considering group-level analysis (who axis). With these held constant, the type of change (what axis) will be varied, and the description of responsiveness given for the DASH across these cells in the taxonomy. Table 5.1 shows the categories (and sub-categories) for the What axis of the responsiveness taxonomy as well as indicating the types of change that were designed into the cohort study for this analysis.
Table 5.1. Types of change used as indicators of the occurrence of change (ΔC, figure 1.2) in responsiveness analysis in this chapter (from the "what" axis of the taxonomy of responsiveness, see figure 2.5). “Who” (group level) and “when” (within-person change) axis were held constant.

<table>
<thead>
<tr>
<th>Type of change (what axis)</th>
<th>✓ = included in analysis</th>
<th>Ways in which the occurrence of this type of change was ascertained for this analysis.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimal potentially detectable change</td>
<td>✓</td>
<td>Structure of the questionnaire (scaling and number of items)</td>
</tr>
<tr>
<td>Minimum change detectable beyond error</td>
<td>✓</td>
<td>Calculation of the change associated with a reliability change index (RCI') of 1.96 or 1.64 (95% and 90% confidence levels respectively) (Christensen et al. 1986a) (McHorney et al. 1995) (Ottenbacher et al. 1988) (Stratford et al. 1996b) (Stratford et al. 1999).</td>
</tr>
<tr>
<td>Observed change in a population</td>
<td>✓</td>
<td>Subgroups where change “expected”: 1. All subjects in cohort 2. Proximal vs. distal disorders 3. Specific diagnostic/treatment groups 4. Surgical vs. non-surgical management</td>
</tr>
<tr>
<td>Observed change in those estimated to have improved (estimated change) from the perspective of the.....</td>
<td>✓</td>
<td>1. Differences (follow-up - baseline) in state ratings of overall problem, pain, function, ability to work, ability to cope, ability to forget about their arm and do what they wanted to do. 2. Transitional scales asking about same attributes: rating of problem, pain, function, work ability * see chapter four and appendices B and C for the format of the actual indices used.</td>
</tr>
<tr>
<td>Observed change in those deemed to have an important change (important change) from the perspective of the.....</td>
<td>✓</td>
<td>Combining transitional markers (above) of change PLUS the importance of that change (either very or extremely important ratings considered = &quot;important&quot; change): rating of problem, pain, function and work ability.</td>
</tr>
</tbody>
</table>

Observed changes in populations reflect the group change in all those being followed in a cohort, or in a subsample of the cohort, regardless of whether a given individual felt that they had improved. In this study we anticipated improvements in most patients by 12 weeks, and a mixed response at four weeks (some better some worse). We will consider the entire cohort, but also examine responsiveness in subgroups defined by region involved (proximal versus distal), diagnostic group (one of the four), and treatment type (operative or non-operative care).
As is depicted in the table (table 5.1), the focus of the estimated and important changes was based on the patient's perspective (Najman et al. 1981; Wright, 1996). The findings of the qualitative work (Chapter three) served as the basis for developing theoretical indices of change that were subsequently tested for relevance in a clinical settings (Chapter four). The positive results in this testing (patient ratings of importance of the attributes and response patterns) then enabled us to use these as the external markers of change in order to capture these attributes as reflecting change from the patient's perspective. They therefore served as Testa's "criterion" of change (figure 1.2) (Testa et al. 1996), a proxy for true change in health. For the analyses presented in this chapter we focused on six attributes: overall rating, pain intensity, ability to function, ability to work, ability to cope, and ability to forget about their arm.

The actual indices for these attributes were constructed in two ways reflecting the two different approaches now considered in the literature (Fischer et al. 1999). First, the arithmetic difference between serially measured ratings of current state (i.e., current rating of attribute at follow-up minus current rating at baseline) was calculated and the resultant score considered indicative of the direction and magnitude of change. The rating of problem, function and work were done on a seven point scale, pain was collected on a 10 point scale, and the last two were gathered on yes/no (1/2) scales. Negative differences reflected less disability, hence an improvement in health (follow-up reflected less difficulty then baseline, or a shift from no to yes in the dichotomous rating). On the scales using a 7 or 10 point response, a difference of zero +/- one was considered to be no change. The second approach was the use of a transitional scale to rate change in those same attributes (overall rating of the problem, pain intensity, ability to function, ability to work) since beginning treatment (either four or 12 weeks prior to the date of completion). Ability to cope, or forget about their arm were not gathered in a transitional format.
The score was rated on an eleven point scale (0-10) with three verbal anchors: a lot worse (0), no change (5), much better (10). A score of four, five or six was considered no change. The decisions around the boundaries of no change were based on the consensus of the committee before the analysis was begun.

Important change was assessed by asking the importance of the change reported on each of the transitional indices. A four point scale was used. Those (with transitional scale ratings > six) stating that the change they experienced was "very important" or "extremely important" to them in their lives, were considered to have an important improvement.

5.2.4 Analysis

Each completed questionnaire package was reviewed by research staff to check for data integrity identifying such things as missing items and duplicate responses. Responses were highlighted to facilitate data entry. Data at each site was then entered into a customized database at each site. These were converted into SAS data sets (SAS Statistical Analysis Systems, version 6.12. Cary, NC ) and the two data sets merged. All analysis was done in SAS.

5.2.4.1 Sample Description.

Baseline demographics for the whole cohort were analyzed descriptively. This was repeated separately for subjects from each of the two data collection sites. The variables described included age, gender, education, and clinical variables (comorbidity, pain medication use, duration of symptoms etc.). The SF-36 generic health status measure (Ware, Jr. et al. 1994; Beaton et al. 1997) was also used to describe baseline overall health in the cohort. Means and medians for each dimension of the SF-36 were calculated for the entire cohort and plotted against
US general population mean values (Ware, Jr. et al. 1993). No equivalent data for Canada was found.

5.2.4.2 Description of the DASH in this population.

DASH data from baseline, 3-5 day test-retest (on subset of the sample) and each follow-up point were analyzed to provide a description of distribution of scores, internal consistency of the item scores, the pattern of change over time and reproducibility over time (test-retest reliability). Frequency distributions in the whole sample, as well as proximal and distal samples were created as well as descriptive summaries of central tendency and spread. The percent of the sample found at the floor (0) and ceiling (100) at each of the three testing times was calculated, as was Cronbach's alpha (Nunnally et al. 1994; McHorney et al. 1995).

The distribution of change over time was described first using frequency distributions (histograms for each time interval) and second by plotting the pattern of change for a subset of patients, those patients with four week data (n=93). The use of this subgroup alone was chosen because the visual display of the data for all 172 individuals was too difficult to read and because this subgroup had the advantage of 4 week data allowing the visualization of potentially non-linear patterns (as per the hypothesis of diminished function at four weeks in some patients) that could not be seen with only two data points (Rogosa et al. 1982).

Test-retest reliability was analyzed using data from those who had two measures before treatment began (3-5 days apart) and said that their arm problem was “unchanged” on a transitional 11 point scale (how is your problem now compared to when you last did this questionnaire package about 3-5 days ago?) (n=56 said “no change” +/- one category). Bland Altman limits of agreement (Altman,1991) were described and plotted on a graph with individual
change score on the vertical, and the mean of that same individual’s test and retest scores across the horizontal. This provides a view of the distribution of individuals’ change scores across the range of observed questionnaire scores. Spearman correlation and intraclass correlation (ICC 2,1) (Shrout and Fleiss 2,1 coefficient) (Shrout et al. 1979) coefficients were calculated to describe the reproducibility of scores between the baseline and retest assessment in this stable group. Confidence intervals (95%) were constructed for both the Spearman (Altman. 1991) and the ICC (2,1) (Shrout et al. 1979) coefficients.

A detailed description of the other tests of validity and reliability of the DASH is described in the appendix (see Appendix D).

5.2.4.3 DESCRIPTION OF RESPONSES ON THE EXTERNAL MARKERS OF CHANGE.

Table 5.1 describes the categories of change (observed, estimated and important change) as well as the specific attributes (right hand column) in each of the indices that were fielded and analyzed in this study. In total 23 different indices, each reflecting one way to construct change for a study of responsiveness are presented, spanning three different categories (observed, estimated and important within-person, group level change). Individuals were asked to respond to each of these indices (see questionnaire package in Appendix C). The frequency distributions depicting the subjects’ responses to each of the indicators of change (estimated and important) at four week (n=93) and 12-week (n=172) follow-up were examined for floor and ceiling effects, variations in response pattern, and their tendency to follow the anticipated course in these patients (some deteriorating at 4 weeks, most improving at 12 weeks).

5.2.4.4 STATISTICAL SUMMARIES OF RESPONSIVENESS TO OBSERVED, ESTIMATED AND IMPORTANT
Responsiveness of the DASH was described for each of the categories of change outlined in Table 5.1 above.

Minimal change potentially detectable on the instrument was described based on the item number and scaling of the DASH.

Minimal change detectable beyond error was described by constructing a reliability change index (RCI') (Jacobson et al. 1984; Stratford et al. 1996c; Stratford et al. 1996b; Stratford et al. 1999; Ottenbacher et al. 1988; Christensen et al. 1986a) and setting it equal to the critical value of z for a given confidence level. The formula for the RCI' (a modification of Jacobson's original RCI to account for sampling being done from two samples (Ottenbacher et al. 1988; Christensen et al. 1986a; Jacobson et al. 1984)(Jacobson et al., 1991) is as follows:

$$ RCI' = \left( \frac{X_1 - X_0}{\sqrt{2 \cdot \sigma_i \sqrt{1 - r_{xx}}} } \right) $$

where $\sigma_i \sqrt{1 - r_{xx}}$ = standard error of measurement, and where $r_{xx}$ = reliability of the measure, and $\sigma_i$ = standard deviation of observed scores.

When the RCI' is therefore set to 1.96 and the equation is solved for the change score, an estimate is obtained of where change can be confidently considered to be greater than noise (error). This is what Stratford has called the MDC. Minimally detectable change (Stratford et al. 1999; Stratford et al. 1996b). It is very close in concept to the smallest detectable change used in radiology or biochemical assays (Ravaud et al. 1999; Sadler et al. 1992). It should be noted that in there is inconsistency in terms of which reliability coefficient should be used (test-retest or Cronbach's alpha). Both Jacobson and Wyrwich use of Cronbach's alpha coefficient (Wyrwich et al. 1999b; Wyrwich et al. 1999a; Jacobson et al. 1984) even when speaking of precision around change scores which is when the test-retest coefficient is usually favoured (McHorney et
al., 1995; Stratford et al. 1999). In 1991 Jacobson (Jacobson et al. 1991) published using the amended version of Christensen (Christensen, 1986), and continuing to use the test-retest coefficient. It should also be noted that it is possible to determine the MDC associated with any confidence level. Both the MDC associated with the 90% and 95% confidence levels were calculated, and for clarity they are described using subscripts: $MDC_{90}$ and $MDC_{95}$ respectively. Wytrich (Wytrich et al. 1999b; Wytrich et al. 1999a) would add the $MDC_{67}$ where the Z statistics equals one.

The remaining categories of change require summary statistics which are performed on the whole sample being considered in observed change, or those who are deemed to have improved/deteriorated in the case of either estimated or important change. Mean change scores and the corresponding standardized response means (SRM) (Wright et al. 1998; Liang et al. 1990) (mean change divided by standard deviation of the change scores) were used as the parametric summary of responsiveness. The SRM was selected as the responsiveness statistic because of its face validity as an indicator of change. Mean change is divided by the standard deviation of that change providing a unit free summary of the distribution of change scores in that sample. Other indicators are available but are either dependent on the baseline distribution (effect size), sample size (paired t-statistic) or upon information that we do not have (Guyatt’s responsiveness statistic (Guyatt et al. 1987) - we do not have the minimally clinically important difference for the numerator).

Other approaches to summarizing responsiveness were also used. It is well documented that the magnitude of the change score can be associated with the baseline score (Judd et al. 1981; Nunnally et al. 1994; Rogosa et al. 1982; Stucki et al. 1996; Cronbach et al. 1970; Vollmer, 1988). One component of this is regression to the mean (Cronbach et al. 1970;
Vollmer, 1988; Burke et al. 1990), where people with extremely high, or extremely low baseline scores will tend to have scores closer to the mean on repeated testing merely due to the distribution of the measurement error. In addition, there could be additional associations between baseline and change that go beyond regression to the mean. These could include different capacity to respond to treatment depending on baseline status. These real associations are important to understand (Rogosa et al. 1982) and, when combined with the problem of regression to the mean, plague the interpretation of studies with non-random assignment to treatment groups (Judd et al. 1981). One suggested (Norman, 1989; Nunnally et al. 1994), though contested (Yanez et al. 1998) approach to handle the regression to the mean is to eliminate the component of change that is linearly predictable by the baseline score. The disagreement lies in that removing the effect of baseline sometimes over-corrects (Rogosa, 1982; Streiner et al., 1991) for regression to the mean and is uncertain if only the error component of the effect can be taken out. In this study residuals from the regression of time two measures (either four week or 12-week follow-up scores) on the baseline (time one) provide what is called a residualized change score (predicted follow-up score minus observed follow-up score for each person). The correlation between baseline and the residualized change is zero. Residualized change scores were calculated for each individual using the entire sample available at each follow-up time. They were therefore calculated based on the four-week data, and on the 12-week data using a simple linear regression of follow-up on baseline. Mean residualized change and standardized response means were calculated for each category of responsiveness using the residualized change (mean residualized change divided by standard deviation of the residualized change). It should be noted that the responsiveness statistic for the entire cohort (observed change) at four and twelve weeks (the sample on whom the residuals were calculated) were expected and
observed to be zero because, by definition, the mean of the residuals (mean residualized change) is zero.

Finally, some would criticize the use of parametric statistics in many if not all of our analyses either because of non-normality in smaller subsets of data, or because the data from a questionnaire might be considered non-interval level. A non-parametric approach was also used to summarize responsiveness: the median change score and the responsiveness statistic reported by van der Heijden (van der Heijden et al. 1996) who called the statistic the Calibrated Responsiveness Ratio (CRR). The CRR is calculated by dividing the median score in the patients who are considered improved (by whatever external marker) by the interquartile range of change scores in stable subjects (van der Heijden et al. 1996). We returned to the data on the 86 persons who had a retest three to five days after baseline for the analysis of test-retest reliability. Fifty-six indicated that they had not changed between assessments. The interquartile range for DASH change scores in this stable subgroup was 5.48. As suggested by van der Heijden (van der Heijden et al. 1996), this was used as the denominator in the CRR. It should be noted that Fletcher used the interquartile range from the sample deemed to have improved for the denominator (same sample as used for the median in the numerator) (Fletcher et al. 1992). Taylor (Taylor et al. 1999) used both approaches providing statistics using the interquartile range of the change, and of a stable group in two different statistics. The approach (CRR) of van der Heijden was used in the analysis of the current data.

5.3 Results:

5.3.1 Sample description.

Two hundred persons were enrolled and completed the baseline portion of this study. Ninety-
three people responded to the four week questionnaire (one study site only, 86% response). At
twelve weeks, data from both sites was combined. One hundred and seventy-two completed the
12-week follow-up questionnaire (86% follow up rate). The description of the samples differed
between sampling sites. However, because we are making within-person analyses and not
between-site, the differences should not affect the results. The description of the sample is
shown in Table 5.2. The first column reflects the findings in the whole cohort, the second and
third columns the Boston (n=91) and Toronto (n=109) subsamples respectively. The mean age
and marital status was similar across sites. The average age was 42, and the majority of people
were married. There was a fairly even split between men and women in the whole sample, but
most of the males came from Toronto. The sites differed in terms of level of education. In
Boston, 67% had graduated from university or college, while in Toronto only 34% had done so.
Correspondingly, a larger proportion of the Toronto sample indicated that their highest
educational level was ‘completed some high school’ (25%) compared to the Boston sample
(3%). Level of employment differed with more people being off work due to their upper-limb
problem in Toronto (19%) versus Boston (2%). Likewise the Toronto group had a higher
proportion of persons on workers’ compensation. The differences described here would have to
lead to a bias between Boston and Toronto in terms of how women, or persons with less
education (as examples), responded to the markers versus their response to the outcome
measures. We could not hypothesize a situation where this would be the case, but in principal
decided that analysis would only be presented for the cohort as a whole. Table 5.3 summarizes
some of the clinical findings for the whole cohort as well as for each site.
Table 5.2. Sociodemographic data on the study participants. The first column represents the whole cohort, the second and third that portion of the cohort coming from each of the sites. The differences between the Boston and Toronto cohorts are noted, though not likely to affect analysis of the questionnaires in the whole cohort (see text). Analysis is not done across study sites.

<table>
<thead>
<tr>
<th></th>
<th>Whole Cohort (n=200)</th>
<th>Boston (n=91)</th>
<th>Toronto (n=109)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Follow-up rate</td>
<td>86% (n=172)</td>
<td>86% (n=78)</td>
<td>86% (n=94)</td>
</tr>
<tr>
<td>(% completing baseline and 12 week)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age: Mean</td>
<td>53.6</td>
<td>54.4</td>
<td>52.9</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>86</td>
<td>29</td>
<td>57</td>
</tr>
<tr>
<td>Female</td>
<td>113</td>
<td>62</td>
<td>51</td>
</tr>
<tr>
<td>Missing</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Marital status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Married/Living</td>
<td>146</td>
<td>65</td>
<td>81</td>
</tr>
<tr>
<td>Divorced/Separated</td>
<td>17</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>Widowed</td>
<td>14</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Single</td>
<td>21</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>Schooling</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; grade 8</td>
<td>10</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>Some high school</td>
<td>28 (14%)</td>
<td>3 (3%)</td>
<td>25 (23%)</td>
</tr>
<tr>
<td>High school</td>
<td>51</td>
<td>11</td>
<td>20</td>
</tr>
<tr>
<td>Some college/university</td>
<td>29</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>Graduated college/ univ.</td>
<td>98 (50%)</td>
<td>61 (67%)</td>
<td>37 (34%)</td>
</tr>
<tr>
<td>Work status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full time</td>
<td>78</td>
<td>39</td>
<td>39</td>
</tr>
<tr>
<td>Part time</td>
<td>20</td>
<td>14</td>
<td>6</td>
</tr>
<tr>
<td>Disabled - due to us</td>
<td>23 (12%)</td>
<td>2 (2%)</td>
<td>21 (19%)</td>
</tr>
<tr>
<td>Disabled - other</td>
<td>8</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Homemaker</td>
<td>15</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>Retired</td>
<td>49</td>
<td>22</td>
<td>27</td>
</tr>
<tr>
<td>Student</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>WCB</td>
<td>. = 7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No, not on WCB</td>
<td>156</td>
<td>82</td>
<td>74</td>
</tr>
<tr>
<td>Yes, but not yet</td>
<td>14</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>Yes, receiving it</td>
<td>14</td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td>Yes, but no longer</td>
<td>9</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Lawyer for UE?</td>
<td>(% yes)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6%</td>
<td>6 (7%)</td>
<td>7 (6%)</td>
</tr>
</tbody>
</table>
Table 5.3. Clinical characteristics of study participants in the cohort study. Data presented for entire cohort, and stratified by data collection site.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Whole Sample (n= 200)</th>
<th>Boston (n=91)</th>
<th>Toronto (n= 109)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region affected</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shoulder</td>
<td>138</td>
<td>61</td>
<td>77</td>
</tr>
<tr>
<td>Wrist/hand</td>
<td>62</td>
<td>30</td>
<td>32</td>
</tr>
<tr>
<td>Duration of symptoms</td>
<td></td>
<td>178</td>
<td>206</td>
</tr>
<tr>
<td>(mean weeks)</td>
<td>193 (sd 374)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medication use</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aspirin, NSAID</td>
<td>117</td>
<td>60</td>
<td>57</td>
</tr>
<tr>
<td>Tylenol (OTC)</td>
<td>90</td>
<td>39</td>
<td>51</td>
</tr>
<tr>
<td>Narcotics</td>
<td>59</td>
<td>20</td>
<td>39</td>
</tr>
<tr>
<td>Other</td>
<td>45</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>Not taking meds.</td>
<td>17</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>Comorbidity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypertension</td>
<td>52</td>
<td>23</td>
<td>29</td>
</tr>
<tr>
<td>Asthma</td>
<td>23</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>Diabetes</td>
<td>13</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>Ulcers</td>
<td>29</td>
<td>10</td>
<td>19</td>
</tr>
<tr>
<td>Depression</td>
<td>51 (26%)</td>
<td>21 (23%)</td>
<td>30 (27%)</td>
</tr>
<tr>
<td>Cancer</td>
<td>16</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Arthritis</td>
<td>83</td>
<td>41</td>
<td>42</td>
</tr>
<tr>
<td>Low back pain</td>
<td>124 (62%)</td>
<td>59 (65%)</td>
<td>65 (60%)</td>
</tr>
<tr>
<td>Number of comorbid conditions:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>29</td>
<td>11</td>
<td>18</td>
</tr>
<tr>
<td>One</td>
<td>48</td>
<td>23</td>
<td>25</td>
</tr>
<tr>
<td>Two</td>
<td>57</td>
<td>25</td>
<td>32</td>
</tr>
<tr>
<td>Three</td>
<td>29</td>
<td>18</td>
<td>11</td>
</tr>
<tr>
<td>Four</td>
<td>25</td>
<td>9</td>
<td>16</td>
</tr>
<tr>
<td>&gt;5</td>
<td>12</td>
<td>5</td>
<td>7</td>
</tr>
</tbody>
</table>
The majority of the shoulder patients were accrued from Toronto, especially those with osteoarthritis (undergoing shoulder replacement). The mean duration of symptoms was 193 weeks, over three years, not surprising given that both accrual sites offer secondary or tertiary levels of care. The majority of the sample were taking some sort of medication to manage symptoms preoperatively including 59 persons of the 200 who stated they were taking narcotics for their pain. Of interest is the high proportion of the samples with a history of depression (23 and 27%) and low back pain (65 and 60%).

The general health of the cohort, as described by the SF-36 (short form - 36 item) questionnaire, is shown in Figure 5.1 where the sample’s mean and median scores in each of the SF-36 (short-form, 36 items) health status measure dimensions as well as the physical component score (PCS) and mental component score (MCS).

**Figure 5.1** Profiles of the baseline health status of sample using the Short Form-36 questionnaire. Profile given for sample mean, sample median and average population score (American data) (Ware et al., 1993)

Figure 5.2 Distribution of baseline DASH score in the whole sample (top), the shoulder patients (bottom right) and the wrist/hand patients (bottom right). Far left and right bars represent floor and ceiling effects on scale.
are plotted along with data from the general population of the United States (Ware, Jr. et al. 1992; Ware, Jr. et al. 1993).

5.3.2 Performance of the DASH in this population

The distribution of the baseline scores on the DASH appear to be normally distributed (see Figure 5.2) with mean of 43.9 and median of 44.6. Only one person was at the "ceiling" perfect health (score of 0), no one is at the floor (100 = maximum disability on scale).

The distribution for the shoulder patients and the wrist hand patients is shown in the same figure, and demonstrates the less severe disability in the wrist/hand patients described by the DASH scores. Table 5.4 shows the percent of people at the floor (poorest described health state, score of 100 on the DASH) and the ceiling (best DASH score = 0), along with the Cronbach's alpha value for the data at that point in time and provides a description of the distribution of the DASH at each data collection point (McHorney et al. 1995; Nunnally et al. 1994).

Table 5.4 Description of DASH scores using all available subjects at each data collection point.

<table>
<thead>
<tr>
<th></th>
<th>% at floor = poorest score = DASH score of 100</th>
<th>% at ceiling = best score = DASH score of 0</th>
<th>Cronbach's alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (n=200)</td>
<td>0.0%</td>
<td>0.5%</td>
<td>0.97</td>
</tr>
<tr>
<td>Four week (n=93)</td>
<td>0.0%</td>
<td>3%</td>
<td>0.98</td>
</tr>
<tr>
<td>12 week (n=172)</td>
<td>0.0%</td>
<td>4%</td>
<td>0.97</td>
</tr>
</tbody>
</table>

Figure 5.3 provides a description of the reliability of the DASH in the subset of patients who completed a questionnaire package twice before treatment (n=86, 56 of whom said they were the same) using the Bland-Altman limits of agreement. This provides an understanding of the change on the DASH observed in those individuals who have not changed, which is useful to
understand before embarking on an analysis of change. The standard deviation is relatively high (6.08), however as the figure demonstrates that could be due to an outlying change score at about the midrange of the average scores. The remaining individuals remained fairly close to the mean, or around zero. The Spearman ranked correlation between baseline and retest was 0.95 (95% confidence intervals: 0.92 - 0.97). The intraclass correlation coefficient (2.1) for test-retest reliability was 0.96 (95% confidence intervals: 0.93-0.98). The fuller description of the psychometric properties of the DASH can be found in Appendix D.

The pattern of change score on the DASH are depicted in two ways. First the distribution of the scores at baseline and 12-weeks is shown in Figure 5.4. A shift towards lower scores (less disability) is seen. The average change score was -13.3 (standard deviation 17.1). Second, figure 5.5 uses a subset of the data (the 93 subjects with
four week data as well as baseline and twelve week) and depicts the pattern of scores over these three points in time. The figure shows the pattern of change for each diagnostic group. Only the rotator cuff group (most of whom had surgery) had the pattern we anticipated might occur: that being a worsening at 4 weeks and then recovery at 12 weeks. Carpal tunnel patients and total shoulder replacements were showing improvement in DASH scores by four weeks.
Figure 5.5 DASH scores in individuals over time using a subset of the sample (n=93) who had additional data collected at four weeks. DASH1 = baseline assessment, DASH4 = four week follow-up, DASH12 = 12 week follow-up. The lines reflect the pattern of disability (as measured by the DASH) over time.
5.3.3 Response patterns to the different indicators of change.

5.3.3.1 Populations for observed change:

Different hypotheses were used to define the groups in which observed change could be quantified. Responsiveness to the change observed in a defined population (observed change) was quantified in the whole cohort and the subgroups defined by region involved, diagnostic group, and operative versus non-operative management.

At four weeks there were 93 people with complete data. This represented 85% of the 109 subjects accrued in Toronto (the Boston subjects, n = 91 at baseline, had no four week data collection). The remaining 15% did not complete their four week package. Sixty-seven of the 93 subjects with complete data had shoulder problems, 26 had problems in their wrist/hand. There were 15 patients who had undergone a total shoulder arthroplasty, 52 with rotator cuff tendinopathy, nine with carpal tunnel syndrome, and 17 with wrist hand tendinitis. Of the four week sample 82% had operative care, and the rest were managed non-operatively.

At twelve weeks of follow-up the samples used for the observed change shifted as both the Toronto and the Boston cohorts had this data (n=172 with twelve week follow-up data). The majority (n = 119, 69%) had shoulder conditions (21 total shoulder arthroplasty, 92 rotator cuff tendinopathy, 6 other soft tissue conditions in the shoulder). There were 27 subjects with carpal tunnel syndrome, and 26 with tendinitis in the wrist/hand. Of the 12-week sample 65% had undergone operative treatment.

These different groups therefore formed the nine samples described under "observed change" in table 5.8 and 5.9.

5.3.3.2 Responses to different indicators of change.
Detailed descriptions of both the indices and the frequency distribution of the responses at 4 and 12-week follow-up are included in an appendix (see Appendix E). Table 5.5 shows a sample of the response pattern for one of the many indices at four weeks. The same is replicated in Table 5.6 for the 12 week follow-up data. The rating of overall problem is shown in both formats: transition scales in the upper half of the table, and serial measures on the bottom. The proportion falling into each is shown. At four weeks a larger proportion reported that they were worse (13.7% on the transitional scale) than at twelve weeks (4.2%), as we expected. In general, more subjects appeared to be improved using the transitional scale in comparison with the serial measures.

Table 5.5. Percent of sample responding to each response category for overall rating of problem (shown for 4 week data using the transitional format (top) and serial measures (bottom). The four week sample consisted of 93 persons.

<table>
<thead>
<tr>
<th></th>
<th>Worse (%)</th>
<th>No change</th>
<th>Improved</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transition</td>
<td>0 1 2 3</td>
<td>4 5 6</td>
<td>7 8 9 10</td>
</tr>
<tr>
<td>Serial measures</td>
<td>0 0 4.7 5.9</td>
<td>16.5 24.7 21.2</td>
<td>12.9 5.9 3.5 3.5 1.2</td>
</tr>
</tbody>
</table>

Table 5.6. Percent of sample responding to each response category at 12 weeks of follow up in the transitional (top) and serial measures (bottom) for the rating of change in their problem. The sample included 172 persons with follow up data.

<table>
<thead>
<tr>
<th></th>
<th>Worse (%)</th>
<th>No change</th>
<th>Improved</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transitional</td>
<td>0 1 2 3</td>
<td>4 5 6</td>
<td>7 8 9 10</td>
</tr>
<tr>
<td>Serial measure</td>
<td>0 0 2.4 3</td>
<td>7 25 20</td>
<td>14 14 9 4.8 1.2</td>
</tr>
</tbody>
</table>
Tables 5.5 and 5.6 reflect the proportion of the sample that would be considered improved in the analysis of estimated change as presented in Tables 5.8 and 5.9. Furthermore, in the case of four week data only (table 5.5, and table 5.8) table 5.5 shows the proportion considered to have worse health according to these proxy indicators. This would be an example of the people falling into the left hand side of table 5.8. Analysis of responsiveness to deterioration was not carried out at twelve weeks as we had the a priori expectation of improvement at that point in time, and as table 5.6 shows, this expectation was met with a very small proportion of the sample reporting deterioration at twelve weeks.

As mentioned earlier, four attributes were measured using both the transitional and serial state measures (overall problem, pain intensity, ability to function, and ability to work). Correlations (Spearman rank) between the two approaches are shown for each attribute on the following table (Table 5.7, using twelve-week data). Correlations between each index and the change in DASH scores (baseline to twelve-week change scores) were also calculated (again using Spearman ranked correlation coefficients). The correlations between methods are low ($r \sim 0.4$) considering they are supposedly measuring the same attribute (pain, function etc.) in the same patients. Correlations with the DASH change scores were highest with the difference in serial state measurements rather than the transitional scales, suggesting that the similarity in the method of data collection produces higher correlations with change scores in health status instruments.
Table 5.7. Correlations (Spearman) between the different external markers of change. T = transitional (T-pain), D = difference in serial state measures (D-pain). Bolded figures along the diagonal represent the correlation between approaches (T versus D) for the same attribute. Correlations with change in DASH score is also shown.

<table>
<thead>
<tr>
<th></th>
<th>T-rating</th>
<th>T-pain</th>
<th>T-function</th>
<th>T-work ability</th>
<th>Change in DASH</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-rating</td>
<td>0.42</td>
<td>0.43</td>
<td>0.38</td>
<td>0.30</td>
<td>0.60</td>
</tr>
<tr>
<td>D-pain</td>
<td>0.41</td>
<td>0.41</td>
<td>0.34</td>
<td>0.32</td>
<td>0.62</td>
</tr>
<tr>
<td>D-function</td>
<td>0.42</td>
<td>0.42</td>
<td>0.39</td>
<td>0.37</td>
<td>0.63</td>
</tr>
<tr>
<td>D-work ability</td>
<td>0.31</td>
<td>0.35</td>
<td>0.41</td>
<td>0.40</td>
<td>0.48</td>
</tr>
<tr>
<td>Change in DASH</td>
<td>0.43</td>
<td>0.37</td>
<td>0.39</td>
<td>0.39</td>
<td></td>
</tr>
</tbody>
</table>

5.3.4 Responsiveness of the DASH outcome measure to different categories of change.

5.3.4.1 Minimal change detectable on the DASH.

The DASH questionnaire has 30 items, with a five point response scale for each. The minimal change (equivalent to a one increment change on one item) is 0.83/100.

5.3.4.2 Minimally detectable change above error.

The standard error of measurement of the DASH, using the test-retest reliability described above (and in more detail in Appendix D), is 4.6 DASH points (/100). Wywich would suggest that this is indicative of a lower bound of meaningful, or likely relevant change (Wywich et al. 1999b; Wywich et al. 1999a), although they would have used Cronbach’s alpha in the calculation. In our data, Cronbach’s alpha and the test-retest reliability were equivalent (0.97, 0.96 respectively). As described in the analysis section, minimal detectable change values were calculated at the 90 and 95% confidence intervals. The results were MCD$_{90}$ = 10.7 DASH points (/100) and MDC$_{95}$ = 12.75/100. We can therefore have confidence that change above these
levels would not likely be observed without actual change was occurring.

5.3.4.3 Observed change, estimated change and important change.

The final three categories of change from the taxonomy are summarized in tables 5.8 and 5.9. In the left hand column of these tables the different categories of change, and the approaches used in this study to estimate the occurrence of that type of change are listed. For instance under the category of observed change (from the taxonomy), there are the samples described above: all patients (expected to improve in the twelve week window because of the treatments chosen: treatments of likely efficacy), the sample divided by region (analysis of the shoulder patients separate from the wrist/hand patients) etc. Under estimated change there are two approaches described earlier where the attributes of change in overall problem, pain intensity, function and work ability are assessed using either the difference between two independent state ratings, or the transitional approach where the patients recalled of the amount of change they had experienced in the attribute since baseline. Two additional threshold concepts: the ability to cope, and the ability to forget about their injury and do what they want to do were also measured. These concepts arose in the qualitative work, and were described again in Chapter four. The threshold concepts were only done using arithmetic differences of serial state measures. Important change was deemed to have occurred if that subject had an estimated change on the transitional scale and stated that this experience of change was important to them in their lives (either very important or extremely important). Again the same four attributes were measured.

The columns of tables 5.8 and 5.9 provide the statistical summaries of responsiveness using the three statistical approaches described earlier: mean change score and SRM; mean residualized change and $\text{SRM}_{\text{res}}$; and finally median change and CRR. Four week and twelve
week are presented in separate tables. At four weeks, where it was hypothesized a priori that some people would have a decline in health (post-operative course after surgery) the table is split separating that portion of the sample that described themselves as worse at four weeks according to the indices of change, and those who described themselves as improved according to the same indices. The numbers who indicated that they were either worse or improved is shown under the “n” column. The category for observed change was not split into improved or worsened because there is no external marker to make that determination when examining observed change in responsiveness, the results can only be presented for the whole group (arbitrarily placed on the left hand side of the table). At twelve weeks only those reporting improvement are shown as the numbers who reported worsening were, as was expected, very small (less than 10 in many cases).

The results in tables 5.8 and 5.9 demonstrate variation in the description of the responsiveness of the DASH outcome measure in three ways: first, across the statistical approaches used (parametric mean, residualized mean and non-parametric median); second across timing of data collections where the same type/category of change resulted in different magnitudes of responsiveness (SRM’s for important improvement in problem = 0.34 at four weeks and 1.15 at 12 weeks). Finally, variation in the description of responsiveness varied across the different categories of change (from the taxonomy) studied. This was found at the categorical level (category along the What axis of the taxonomy: observed, estimated and important change) as well as at a finer level, the specific attributes being measures, or the approach (difference versus transitional) used to estimate its occurrence. Figure 5.6 provides a summary of the findings by showing the 12-week change score related to each category of the what axis of the taxonomy, while the when axis was within-person change and the who axis was group-level.
Improvement in function (using a transitional scale) was used as an example for estimated and important change.
Table 5.8 Summary of four-week responsiveness of DASH outcome measure scores observed over time according to different markers for type of change. Negative change scores indicate less disability and therefore an ‘improvement’. SRM and CRR statistics are presented in absolute values. N = number of subjects (of 93 with follow-up) having change according to that marker (or hypothesized to have that change in the case of observed change).

<table>
<thead>
<tr>
<th>Category of change</th>
<th>Estimates of responsiveness (pre-treatment to 4 weeks) in those hypothesized to, or reporting decline since treatment.</th>
<th>Estimates of change and responsiveness (pre-treatment to 4 weeks follow-up) in those reporting “improvement”</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean change &amp; (SRM)</td>
<td>Residualized change mean &amp; (SRM&lt;sub&gt;c&lt;/sub&gt;)</td>
</tr>
<tr>
<td>Observed</td>
<td>n</td>
<td></td>
</tr>
<tr>
<td>- all patients</td>
<td>93</td>
<td>+1.36 (0.07)</td>
</tr>
<tr>
<td>- all shoulders</td>
<td>67</td>
<td>+7.44 (0.40)</td>
</tr>
<tr>
<td>- all wrist/hand</td>
<td>26</td>
<td>-13.8 (0.90)</td>
</tr>
<tr>
<td>- all TSA</td>
<td>15</td>
<td>-3.43 (0.16)</td>
</tr>
<tr>
<td>- all RCT</td>
<td>52</td>
<td>+10.8 (0.64)</td>
</tr>
<tr>
<td>- all CTS</td>
<td>9</td>
<td>-15.9 (1.41)</td>
</tr>
<tr>
<td>- all tendinitis (w/h)</td>
<td>17</td>
<td>-12.9 (0.76)</td>
</tr>
<tr>
<td>- all operative</td>
<td>76</td>
<td>+5.00 (0.26)</td>
</tr>
<tr>
<td>- all non-operative</td>
<td>17</td>
<td>-13.0 (0.76)</td>
</tr>
</tbody>
</table>

Estimates of change: Difference in states
- rating of problem
- function
- pain
- ability to work
- able to cope now
- able to forget arm

Estimates of change: Transitional approach
- change in problem
- change in pain
- change in function
- change in work

Important change
- change in problem
- change in pain
- change in function
- change in work

nt = not tested in this analysis. SRM = standardized response mean = mean change score/standard deviation of the change. SRM<sub>c</sub> = SRM calculated using residualized gains, based on regression of follow-up on baseline scores. CRR = Calibrated Responsiveness Ratio = median change / interquartile range of change scores in stable patients. Note: the data for observed change is arbitrarily placed on the left side of this table, but is not stratified into improved or worsened patients as it reflects change observed regardless of status.
Table 5.9. Summary of responsiveness of the DASH to different categories of change at 12-week follow-up. N = number of subject (172 with follow-up data) identified as having "improved" by this marker of change (or hypothesized direction of change in the case of "Observed change").

<table>
<thead>
<tr>
<th>Type of change</th>
<th>Estimates of responsiveness</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0-12 weeks (based on n=172)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>Mean baseline</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Observed</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>172</td>
<td>44.5</td>
</tr>
<tr>
<td>Shoulders</td>
<td>119</td>
<td>49.3</td>
</tr>
<tr>
<td>Wrist/Hand</td>
<td>53</td>
<td>33.8</td>
</tr>
<tr>
<td>TSA</td>
<td>21</td>
<td>58.0</td>
</tr>
<tr>
<td>RCT</td>
<td>92</td>
<td>47.6</td>
</tr>
<tr>
<td>CTS</td>
<td>27</td>
<td>39.6</td>
</tr>
<tr>
<td>Operative</td>
<td>26</td>
<td>28.3</td>
</tr>
<tr>
<td>Non-operative</td>
<td>112</td>
<td>49.8</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>34.6</td>
</tr>
<tr>
<td><strong>Estimated change</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Difference in states</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rating of problem</td>
<td>71</td>
<td>50.4</td>
</tr>
<tr>
<td>Pain</td>
<td>84</td>
<td>50.6</td>
</tr>
<tr>
<td>Function</td>
<td>55</td>
<td>52.9</td>
</tr>
<tr>
<td>Ability to work</td>
<td>23</td>
<td>42.3</td>
</tr>
<tr>
<td>Able to cope now</td>
<td>49</td>
<td>48.0</td>
</tr>
<tr>
<td>Able to forget arm</td>
<td>26</td>
<td>38.9</td>
</tr>
<tr>
<td><strong>Estimated change</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Transitional approach</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Change in problem</td>
<td>121</td>
<td>42.9</td>
</tr>
<tr>
<td>Change in pain</td>
<td>123</td>
<td>42.8</td>
</tr>
<tr>
<td>Change in function</td>
<td>86</td>
<td>40.6</td>
</tr>
<tr>
<td>Change in work</td>
<td>94</td>
<td>43.3</td>
</tr>
<tr>
<td><strong>Important change</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Change in problem</td>
<td>109</td>
<td>43.5</td>
</tr>
<tr>
<td>Change in pain</td>
<td>101</td>
<td>43.2</td>
</tr>
<tr>
<td>Change in function</td>
<td>79</td>
<td>41.9</td>
</tr>
<tr>
<td>Change in work</td>
<td>84</td>
<td>43.8</td>
</tr>
</tbody>
</table>

SRM = standardized response mean = mean change score/standard deviation of the change. Reported in absolute values.

SRM<sub>res</sub> = SRM calculated using residualized gains (from regression of follow-up on baseline scores).

CRR = Calibrated Responsiveness Ratio = median change / interquartile range of change scores in stable patients. Reported in absolute values.
5.4 Discussion.

5.4.1 Principal findings.

This study has provided evidence that the responsiveness of a single questionnaire, in the same patient group, will vary in its description depending on the category of change being examined. Furthermore it varies enough that these estimates of change would lead to substantially different interpretations: for instance in sample size calculations (Guyatt et al. 1987) and other applications of responsiveness (such as in determining the cut-point for an “improved” score, and using that to estimate the number of patients that should be considered “improved” as a result of a given treatment).

The category of change refers to the individual cells in the tri-axial taxonomy of responsiveness, and in this study only the “what” axis, the type of change, was varied. Similar findings were found with other instruments (see Appendix F) tested in the same manner. Further differences could be found if the other two axes were altered, however this was not done in the current study. A study combining Redelmeiers (Redelmeier et al., 1993) cross sectional approach with a longitudinal cohort (as recommended by Wright (Wright, 1996) would allow for
examining the effect of the difference between within-person and between-person discrimination. An example of this would be the work of Hemingway (Hemingway et al., 1997) who used a year of aging rather than self-rating of change in health as a proxy for deterioration in health. As described in Chapter Two, this group described that between-person change underestimated within-person change.

These findings suggest that responsiveness is context-specific and that the context of measurement will impact on the magnitude of change being quantified in a study of responsiveness. This context is defined by the type of change (as defined by a cell in the taxonomy), as well as other factors such as the patient group, the treatment they are undergoing (some will be expected to result in larger or smaller changes in the target construct such as function), and the timing of the follow-up. This last point can be demonstrated by observing the differences in the responsiveness of the DASH at 4 weeks and at 12 weeks (SRM of 0.07 and 0.78 observed in whole cohort at four and twelve weeks respectively). The DASH did not become “more responsive” at 12 weeks, rather the subjects likely experienced more recovery and the DASH was responding to this larger experience of change. Things like this are very obvious, though often overlooked when effect sizes or standardized response means are extracted from articles and used as evidence of the questionnaire’s responsiveness in another setting.

The variations observed for the responsiveness of the DASH would span from what Cohen (Cohen, 1988) called a small effect size (i.e., SRM = 0.16 in the total shoulder arthroplasty group at four weeks), to a large effect size (i.e., SRM = 0.98 in the same group at 12 weeks, or SRM = 1.27 in those with an important change in function at 12 weeks). Cohen’s guidelines are often misinterpreted as being indicators of a “responsive” or “non-responsive” instrument. Their intention seems to be to assist in the interpretation and the implication of having a large or small
observed effect size in terms of statistical power and experimental design. Our findings, by demonstrating such a wide range of effect sizes in the same instrument, reinforce the fact that magnitude of the responsiveness statistic alone does not indicate that an instrument is, or is not “responsive”.

The DASH instrument was able to detect negative changes in health as well as positive. The four week responsiveness demonstrated this with consistently positive change scores (more disability) in those who indicated that they were worse, and negative (less disability) in those who were better. At four weeks, Patients treated operatively had increased disability, in contrast to those treated non-operatively as was anticipated (see table 5.8, figure 5.5) and as was described by the DASH. Of interest, however, were that two of the operative groups had improvements by four weeks. The total shoulder arthroplasty and the carpal tunnel syndrome (operative release) patients seemed to be improving even by four weeks. This was a fairly consistent pattern when individual trajectories were examined (see figure 5.5). The rotator cuff patients did what was expected and had slightly worse scores at four weeks. The patients with rotator cuff surgery have several post-operative restriction in place to protect the surgery in addition to having a painful procedure. Therefore this was the expected trajectory. However, the contrast between the patients undergoing total shoulder arthroplasty versus rotator cuff tendinopathy treatment was not expected, and demonstrates the different trajectories that can occur with different treatments, and that was detectable on a disability/symptom scale like the DASH. It also supports the improved understanding of recovery obtained by having more than two repeat measures of disability.

5.4.2 Summary statistics for responsiveness
In this study three summary statistics for responsiveness were used: the standardized response mean (SRM) (Liang et al. 1990), the same statistic substituting in the residualized gain rather than the raw change (SRM<sub>reg</sub>) (thus reducing the association between baseline and change scores), and a non-parametric responsiveness statistic the CRR (calibrated responsiveness ratio) (van der Heijden et al. 1996).

The residualized gain score analysis addressed the problematic issue of correlation between baseline and change scores. At twelve weeks there was a Pearson correlation between baseline and change of -0.39. This suggests a relationship between the two. The regression eliminated that correlation (that which is explained in a linear relationship), however as a tool for expressing responsiveness the use of residualized gains had limitations. First, by definition, the mean of the residuals is zero. Therefore when the responsiveness for the observed change in the entire cohort (the groups we used to do the regression) was calculated the statistics were equal to zero. Second, it was not clear which group should be used for the regression in a study such as this. The whole samples were used (all four week and then all 12 week data regressed on to the baseline) however, it could be argued that each subsample defined as “improved” according to a certain external markers should be used in the regression equation. This was not done because every statistic would then equal zero. Finally, the pattern of responses on the residualized change approach was not always logical: the mean residualized change for the carpal tunnel patients at 12 weeks was +0.12 which seemed small in comparison with the raw change of -1.16. However this example serves to remind us that the concept being measured is no longer the amount of change, but rather how much the post-test measure varied from the predicted post-test score. Hence, the concept of change has shifted when applying residualized change scores. The carpal tunnel cases had follow-up scores that were closer to the predicted value for patients at that
follow-up point. The magnitude is now reflecting greater deviance from the predicted post test score rather than greater absolute change. The residualized change approach offers some advantages, particularly in adjusting for regression to the mean. However, it can also lead to over adjustment for true differences in response attributable to baseline scores that really should be considered in interpreting observed change scores (Yanez et al., 1999). The results should be interpreted carefully.

The calibrated responsiveness ratio offers the non-parametric version of the standardized response mean (van der Heijden et al. 1996). In many cases the median change was fairly close to the mean change score. The choice of denominator, the interquartile range in stable patients (van der Heijden et al. 1996), parallels Guyatt’s use of the error term from a group of stable patients (Guyatt et al. 1987). In our analysis, the stability of the denominator offered little to the interpretation because it was the same across all analyses. Similar criticisms to Guyatt’s statistic are applicable here. The denominator comes from an entirely independent sample, introducing another source of bias (Norman, 1989). The interquartile range is very dependent on the study sample, and in the case of using test-retest reliability data it is often based on a small number of subjects (our sample of 86 persons, is large for a test-retest study). As mentioned earlier in the text, Fletcher suggests the use of the interquartile range of the changed subjects (Fletcher et al. 1992) when describing a non-parametric version of the standardized response mean rather than the IQR of stable patients. Finally, the magnitude of the CRR was often found to be high in comparison with the other two statistics, and as it is a newer approach there are few guidelines for interpreting this statistic. Taylor (Taylor et al. 1999) uses Cohen’s guidelines, however these were not designed for non-parametric estimates of responsiveness and should be interpreted with caution. The usefulness of the CRR statistic is its attractiveness as a non-parametric approach,
and the information gained by observing how close the mean and median change values in our samples were.

5.4.3 **Boundaries on change and no change.**

This study also set some parameters for identifying when change has *not* occurred. Feinstein suggests that along with the attention paid to the significance of delta (change), the ability to interpret change might also be advanced by trying to determine what the lower bound of relevant change would be (or the upper bound of no change) (Feinstein.1998). In the present work, there are three possibilities. First, change would need to be greater than 0.83 because that is the size of a one increment step on the DASH, and as such should be considered an anchor for interpretability (it is difficult to interpret a change smaller than this amount). Second, we have two estimates of change around noise: 1) The standard error of measurement (SEM, using test-retest reliability coefficients) which Wyrwich would suggest equates to a lower bound of relevant change (Wyrwich et al. 1999a; Wyrwich et al. 1999b). This was 4.6 out of a possible 100 on the DASH; and 2) The minimally detectable change was also discussed, and was 12.75/100 on the DASH (95% confidence, using test-retest data, MDC_{0.05}; MDC_{0.00} = 10.7). However, such cut points have the danger of becoming perceived as "gold standards", when in fact they are variable depending on the sample being used to determine the components of the SEM statistic, specifically the adjustments to the variance and the reliability (Burnand et al. 1990; Sadler et al. 1992; Ravaud et al. 1999).

Furthermore there is some debate as to whether internal consistency or test-retest reliability should be used (Wyrwich et al. 1999a). Wyrwich favours internal consistency (Cronbach's alpha) and an unadjusted baseline variability (Wyrwich et al. 1999b; Wyrwich et al.
Nunnally states that with very high sample sizes (n~ 300), there will be little difference in the reliability coefficients (Nunnally et al. 1994). However, in clinical research, sample sizes rarely reach that level, in which case decisions need to be made deliberately about which coefficient is used. In our case we chose the test-retest coefficient in order to reflect longitudinal precision (around change scores), rather than estimates of the repeatability of one score (where Cronbach’s alpha would be the preferred coefficient) (McHorney et al. 1995). We also used the modified (more stringent) formula for the reliability change index to calculate the MDC supporting the fact that the data is being collected from two samples (test and retest) (Christensen et al. 1986a; Jacobson et al. 1984; Stratford et al. 1996b; Stratford et al. 1996c; Stratford et al. 1999). Although some claim that these figures are independent of the testing situation, we do not believe this to be so. In this study, for the DASH we reported that these values for MDC₉₀ and MDC₉₅ were 10.70 and 12.75 DASH points respectively. However, in another study (Turchin et al. 1998) the test-retest reliability of the DASH was 0.92. The RCI’ (95%) worked out to be approximately 15/100 points. Although this is close to what is suggested from the results of our study, it is not the same, and the 2.25 point difference equates to three incremental changes on the DASH (0.83/100 each). In addition work by Stratford has also described the need to have Conditional SEM’s which reflect the fact that SEM varies with the magnitude of the observed score (Stratford et al. 1996b). This means that the minimally detectable change varies depending on the baseline score, being larger in the midrange of the scores (Stratford et al. 1996b). Further work needs to be done to determine how best to use these figures, and how stable they are in different settings as it seems they would vary considerably with the variance and distribution of the observed values in the study sample.
5.4.4. Implications of the findings.

The implications of the findings of this study are similar to those suggested in Chapter 2, except that now we have numeric evidence that the different categories of responsiveness will lead to differences (and clinically relevant differences) in the description of responsiveness of a single instrument. This work suggests that responsiveness needs to be described in a context-specific manner (i.e., the DASH is responsive to the change observed in carpal tunnel release patients at both four and twelve weeks after surgery). Furthermore, it suggests that when looking for a responsive instrument to apply in a particular clinical setting or treatment trial, it is important to look for evidence of the same type of responsiveness.

The study findings also suggest that clear decisions need to be made regarding the theoretical approach taken to measuring change. The three approaches used here: parametric estimated, non-parametric estimates, and residualized change estimates, each provided different impressions of responsiveness, and indeed reflected different concepts (residualized change) in the study of change and responsiveness. Additional work is needed to understand the appropriate mathematical approach to the summary of responsiveness.

5.4.5 Limitations of the study.

This study had limitations. The most obvious being the number of analyses being conducted on the data. However, statistical comparisons across types of responsiveness were purposefully avoided, as were the creation of confidence intervals around the responsiveness statistics. The analysis was a description of the change that is quantified with the different approaches to responsiveness. Based on the amount of difference found between categories of responsiveness, the sample size, and the number of comparisons being made, statistically
significant differences would likely have been found at an alpha (type I error) level of 0.05. However, adjustments to correct for the multiple comparisons (Bonferroni adjustments to alpha for example (Fleiss, 1986; Kleinbaum et al. 1988)) would have run the risk of labeling clinically relevant changes in responsiveness “insignificant” from a statistical perspective because of the substantial reduction in alpha such an adjustment would entail. The decision was made not to do statistical comparisons for that reason (Kleinbaum et al. 1988). There is little doubt as to the potential clinical relevance of the difference in responsiveness described in this study.

The two study sites provided samples with different sociodemographic features: males were more likely to come from the Toronto cohort, as were shoulder patients. Highly educated patients were more likely found in the Boston cohort. This could also reflect the less direct method of accrual in the Boston site where they were mailed information on the study rather than talking directly to research staff. Lacking any belief that this would differentially affect the correlation between proxy and change in the health status questionnaire we combined the data and did not differentiate between data from the Boston versus Toronto site. Some undetected bias could have existed as an interaction of site and another factor such as gender, or region. We did not analyze for this and recognize this as a potential limitation.

Another limitation is the dependence of a great deal of the analysis on the single-item global measures of change and function. Norman et al (Norman et al. 1997) criticized approaches to responsiveness that use single-item, transitional indices of change for many reasons, one of which was the unreliability of single item indices. This one criticism is a valid point and is a limitation of this work. The lack of discrimination and lack of reliability of single item indices is precisely the reason there was a move towards multi-item indices (Spector, 1992; Nunnally et al. 1994) in the form of standardized health status measures. Yet responsiveness
researchers return to these single item scales as external markers, almost criterion, of change in studies of responsiveness (Beaton et al. 1997; Deyo et al. 1986) because they help us bridge the gap between a numeric estimate and its clinical meaning (Fortin et al. 1995; Testa et al. 1996; Stratford et al. 1996b; Lydick et al. 1993). In the present thesis, several attempts were made to improve the single item indices: using concepts arising from the qualitative study; framing the question in the patients’ own words (again from the qualitative work, and from patients’ rating of the importance of different tasks in an earlier study (Marx et al. 1999)); and finally expanding the number of response options (now between seven and eleven options depending on the item) in order to increase the discriminative ability. Each of these strategies should improve the external marker in terms of its ability to reflect true change from the patient’s perspective. Thus, allowing us to get a better idea of the accuracy of the change scores.

Finally, the responsiveness in this chapter has focused on the markers of change (either the theoretically expected change, or the use of actual questions), and then quantified the associated change in the DASH Outcome Measure. Our deliberate focus was on the construct of change (different ways of describing the occurrence of change) and how that effects the description of responsiveness using various summary statistics. The construct of change does not however, complete our understanding of responsiveness. We have defined responsiveness as the accurate detection of change when it has occurred (De Bruin et al. 1997), and subsequently focused on the determination of the occurrence of change in greater detail. Thus there is an ongoing need to look at the issue of accuracy of detection of change in health using instruments like the DASH (sensitivity, specificity of various change scores). Certain methods are more suited to addressing accuracy including the receiver operator characteristic curve (ROC) approach used by Deyo (Deyo et al. 1991; Deyo et al. 1986). Although these approaches are, by
definition, very dependent on a valid dichotomous indicator of improvement (Deyo et al. 1991) which returns the debate to the need to have good indicators of improvement. The results from the ROC type of approach allow more understanding of the accuracy (true positive and true negative rates) of various change scores relative to the external marker (a proxy for true change in health from the patients perspective), and for change in that questionnaire overall (area under the curve). A natural extension of the work in the chapter would be to attempt to look for the accuracy of change scores, a recognized limitation (or beyond the scope) of the current work.

5.4.6 Areas for future research.

There are also many opportunities for ongoing work. The data suggested different trajectories for the different treatments under study. One approach to interpreting change could therefore be shifting the focus from the magnitude of the change to more the rate of change. This would equate to analyzing the shape of the curves (scores plotted over time for individuals) presented for the different diagnostic groups earlier. Analysis could proceed along trying to describe the pattern of the change, or the rate of change (slope). Rogosa has suggested this in describing the growth curve approach to describing change (Rogosa et al. 1982). A great deal of attention has been given to this area in the book of Collins and Horn (Collins et al. 1991).

Perhaps responsiveness should also be shifting in focus from absolute change to rate of change, or shape of change in order to advance its interpretability.

Measuring change has been a longstanding problem in education, psychology and more recently in health. Work like this chapter in many ways clarifies change by separating it into distinct boxes, but at the same time it adds to the confusion because it suggests the need to look for a specific type of responsiveness, and not just label an instrument as responsive. However,
the temptation to make things seem simple when they are not leads to two problems: first, the simpler approach might be wrong; and second, the simpler approach sometimes reflects the status quo (maybe that is why it seems simpler), and does not push research to the edges of its paradigm – the place where new ideas and understandings are formed and advanced. This work therefore pushes our understanding of responsiveness in new directions and opens new possibilities for research for instance into the application of the taxonomy in different clinical populations.
6.1 Conclusions

The present thesis focuses on conceptual aspects of responsiveness; that is, it is an attempt to understand the meaning and implications of the construct of change built into the study of responsiveness (Testa et al. 1996). In describing and explaining change in health status, the work moves through iterations of quantitative to qualitative and back to quantitative perspectives. This approach is directly and indirectly recommended in the literature (Donovan et al. 1993; Burke et al. 1990; Liang, 1995; Leplege et al. 1997), particularly when complex concepts such as change in quality of life and health status are involved.

The principal conclusion of the thesis is that responsiveness is a context-specific attribute, and that context is largely defined by the nature of the change integrated into the study (as defined by the cell it occupies in the taxonomy of responsiveness). Furthermore, it was found that measurement of change from a patient’s perspective is particularly challenging because the meaning of change (or of being better) is highly individualized and dependent on an individual’s experience of the disorder concerned. Indeed, measuring change from the patient’s perspective may be more difficult than hitherto suggested (Donovan et al. 1993); as Jacobs et al put it: “the measurement of change in patients’ health status is an important but troublesome aim of clinical research” (Jacobs et al. 1996)(pg 163).

The question arises of whether the four objectives stated in the Introduction have been met.
**Objective one:** To develop a framework of responsiveness based on a consolidation of the literature.

The review of the literature led to the creation of a triaxial matrix in which each cell depicts a different type or category of change capable of being incorporated into a study of responsiveness. This became a taxonomy of responsiveness. The first of the three axes, the “Who” axis, differentiates whether individuals or groups of people are being considered. The second, or “When”, axis looks at when the change being examined occurred. Was it between people at one point in time or within individuals over time, or a combination of the two as in the analysis of a randomized controlled trial (between group differences of within person change)? Finally, the “What” axis distinguishes different types of responsiveness based on the type of change being considered. For example, some studies look at changes in a population before and after a certain treatment, whereas others look at change in a group deemed to have improved in some way (estimated change). Two categories along the “what” axis (estimated and important change) were recognized as being divisible into subcategories reflecting the different perspectives that can be used when they are determined, whether that of the patient, clinician/researcher, payer, or society. The location of a study of responsiveness along the three axes described defines the category of responsiveness being addressed.

Along with the context, as defined by the type (or category) of change, factors outside the taxonomy can also influence the description of any of the categories of responsiveness. They may include patient mix, timing of data collection, patient characteristics, measurement characteristics of the instrument, and the treatment being applied (some interventions have a marked effect on function, but with others it is more subtle) (Greenfield et al. 1992; Katz et al. 1992; Redelmeier et al. 1993).
The taxonomy developed in order to address this objective separates responsiveness into several different types and will, if accepted, alter how the results of a study of responsiveness are described. An instrument will now be described as responsive to a given category of change. The taxonomy also distinguishes between estimates of change from differing perspectives (patient, clinician, payer, and society) and suggests that they could lead to very different numeric results.

**Objective two: to explain what change (being better) means from the patient's point of view.**

One approach arising from the taxonomy would be to allow patients to determine if they have changed, or if they have experienced an important change (either improvement or deterioration). Broad global questions such as “are you better?” would be one approach that could be used, but responses carry too little, or perhaps too broad a range of information to be useful as proxies for true change in disease-specific health states in studies of responsiveness.

The present thesis aimed to elucidate what patients mean when they say they are better by conducting a qualitative study in 24 people with upper limb musculoskeletal disorders. The results demonstrated that being better is a highly individualized experience that does not always reflect a change in the underlying disorder. For example, some participants reported having adjusted their lives around their disorders, and others adapted their definitions of “good” health. Both occurred in the context of ongoing difficulties in terms of symptoms and functional status, with no actual improvement in disease state.

The implication for the assessment of responsiveness (only one component of the overall implications of this study), is that there is a need to establish a link between our target (true change in health status from a patient’s perspective) and the proxy markers used in assessing responsiveness. Broad measures do not necessarily achieve such a link and may capture many extraneous factors, thus obscuring the measurement and interpretation of change as measured by
the difference in health status scores. One way to avoid this would be to construct very specific indices intended to reflect the change thought to relate to alterations in true health status. Alternatively, we could attempt to develop proxies based on patient-generated indices, thereby allowing individuals to express their experience of change.

**Objective three: to develop methods of integrating the patient's perspective into studies of responsiveness.**

In addressing this goal, very specific indices of change were developed that (based on the qualitative findings) were expected to reflect, or serve as an indicator of, "true" change in health status from a patient's perspective. Using components of the theory of being better developed in the qualitative study, markers were developed that were thought likely to coincide with changes in physical functional status and symptoms. Indicators of social and emotional health were also developed and fielded in the study, as these important components of overall health status are not generally considered markers of musculoskeletal disorders. However, they were not included as proxies of true disease-specific change in subsequent analysis because there was no hypothesized rationale to include them as disease-specific indicators of change in the disorder.

Indices were created using wording taken directly from the qualitative study if at all possible and then tested in 30 patients attending a tertiary level clinic for treatment of an upper-limb musculoskeletal disorder. Their relevance and importance for capturing change from these patients' perspectives was confirmed. Open ended questions revealed no missing constructs. The indices of change developed were thus shown to reflect the patient's experience of change in health status. Although they do not capture the complexity of "being better" as expounded in the qualitative study, that was not the aim. The goal of the development of the indices was rather to
find improved proxies for true change in health status from a patient’s perspective in order to validate the change that would be detectable on disease-specific measures (Deyo et al. 1984; Deyo et al. 1986; Revicki et al. 1995; Liang, 1995; Fortin et al. 1995).

In the present patient group, it was anticipated that the change experienced would be resolution of the disorder due to treatment. The qualitative work demonstrated that this is not always the only route to being better, nor the only effect of a treatment. However, the indices we developed would likely only be sensitive to disease resolution (with the exception of the overall rating of their problem). Other indicators would be needed if the goal of treatment were to facilitate adaptation to a disorder or help the patient redefine the meaning of his or her life after an injury. In those situations, the markers of resolution developed here would be weak proxies of the target “true” change. Qualitative theory suggests that patients in our cohort may well have redefined their lives to some extent, or adapted to ongoing symptoms, but they were selected as a group likely to experience resolution of their disorder as the primary component of “being better”.

Although gaining insight into the patient’s perspective on change is necessarily complex, this thesis may have made some advances in researchers’ attempts to, as Feinstein said, return the patient “to the center [sic] of the clinical universe” (Feinstein, 1996)(pg. 616). The present work may allow the interpretation of change in the context of responsiveness to come closer to answering the central clinical question: do change scores on a questionnaire answer the question “Are you better?”

**Objective four:** to describe the responsiveness of a single instrument to different types of and perspectives on, change.

This objective sought to gather evidence of the degree to which the property of responsiveness is
dependent on the marker of change modeled into its evaluation. Two components of the thesis addressed this objective. First, the psychometric properties of a disease-specific instrument, the DASH (Disabilities of the Arm, Shoulder and Hand) Outcome Measure, were assessed in order to verify it as a reasonable measure of function and symptoms in upper-extremity patients with musculoskeletal disorders. (It was found to meet the necessary criteria, and exceeded many of the established psychometric standards.) Second, the results obtained in trying to achieve the objectives above were combined in order to establish whether the different categories of responsiveness in the taxonomy did indeed produce different descriptions. Holding the “Who” axis and the “When” axis constant (only group-level, within-person change was considered), we varied the “What” axis (type of change) was varied and the resultant descriptions of the responsiveness of the DASH compared (the results for other measures are described in an appendix). Both estimated change and important change were looked at from the perspective of the patient alone (using the results of objectives two and three above).

It can be concluded that this prospective cohort study in which 200 patients with upper-limb musculoskeletal disorders were followed over a 12-week period after the initiation of treatment provided evidence that the description of responsiveness varies depending on the category of change structured into the investigation (the proxy for true change).

6.2 Major implications of these findings

Specific findings and their implications have already been described at the end of each chapter. Here, we present the overall implications that emerge from a consideration of the thesis as a whole.
6.2.1 Describing responsiveness

A review of the existing literature and the findings of the cohort study presented here both suggest that responsiveness is not a property of an instrument but a context-specific attribute dependent on the patient group, the treatment, and the category of change being addressed. This has two important implications. First, instruments must now be described as being responsive to a given category of change in a given patient and clinical setting (in much the same way as findings from a clinical trial must be contextualized before they can be generalized to other situations). Second, numeric estimates of responsiveness should not be taken out of context. Thus, studies conducted on different categories of responsiveness should not be used for comparative purposes of any kind. Indeed, the magnitude of a statistic of responsiveness alone tells us little about an instrument, and certainly not enough to support an assertion that it is more or less responsive than another tool tested in a different situation. Ideally, the choice of one instrument over another should therefore be based on concurrent comparisons of different scales in the same study, rare though they are (Beaton et al. 1997; Beaton et al. 1998; Bombardier et al. 1991; Buchbinder et al. 1995a; Katz et al. 1992; Kirkley et al. 1998; Liang et al. 1985; Liang et al. 1990; MacDermid et al. 1999a). Different studies are unlikely to provide information about relative responsiveness of different questionnaires unless they happened to be of virtually identical design except for the measure used.

6.2.2 Gaining a patient’s perspective

The present work demonstrates the complexity of incorporating the patient’s perspective into a study of responsiveness. However, it is important to try if our goal is to treat the patient and what ails him or her (Feinstein, 1967).
One important implication of attempts to gain the patients' perspective is exemplified by the suggestion in the qualitative study that being better does not always mean resolution of disease. If the goal of health care interventions is to make the person feel better, the question arises of how much effort to put into decreasing the presence of the disease, or slowing its progress, and how much to put into improving patient-related factors such as ability to cope or to adapt to an ongoing disease state. The answer depends on the natural history or clinical course of the disorder concerned, and on the effectiveness of available interventions. When a patient has a chronic disease or a disorder thought unlikely to change much, it may be appropriate to focus on coping and adaptation rather than changing health status. However, the results presented here suggest that there is no clear-cut division. Although the study population used for the qualitative work had what are considered treatable disorders, they still reported adapting to ongoing disease and redefining the meaning of their health to include it. Such findings should serve as a reminder that focusing on the patient's perspective brings us back to asking questions about what Feinstein called dis-ease (all that ails the patient), rather than disease (as a purely pathological entity) (Feinstein, 1996). As he suggests: "their answers are the soul of humanism in patient care. They are also the essence of its science" (Feinstein, 1996)(pg. 617).

6.3 Areas for further research

Broad though the scope of the present thesis is, it is clear that countless opportunities to expand our understanding of responsiveness remain to be taken. Several have been mentioned within the summaries of each chapter, and those thought to be most critical are reviewed below.

6.3.1 Longitudinal qualitative work
The qualitative work presented here reveals at least two areas in need of further attention. First, there is some suggestion that over the course of recovery different attributes may become more or less important as indicators of improvement or deterioration. Second, participants described the ability to redefine health or adapt to a disorder as an indicator of being better. Both points could be considered suggestive of response shift, a phenomenon defined as a change in internal standards, values or conceptualization of health that may occur when patients undergo a change in health state (Schwartz et al. 1999a; Sprangers.1996; Llewellyn-Thomas et al. 1999; Sprangers et al. 1999). The effect is not restricted to health and has been described in both organizational behaviour and psychology literature (Armenakis et al. 1979; Armenakis.1988; Howard et al. 1979a; Howard et al. 1979b; Golembiewski et al. 1976). The implications of response shift for the measurement of changes in health are quite serious, as its presence would violate a fundamental assumption of the measurement of change: the stability of the underlying construct over time (Rogosa et al. 1982). When the construct itself shifts, the measurement of change is virtually meaningless as the difference score is being calculated across two totally unique concepts - equivalent perhaps to subtracting a subject’s height from his weight (Rogosa et al. 1982).

Both the response shift and alterations over time in indicators of change require further study, particularly in the context of research into the measurement of change. Sprangers and Schwartz have provided a theoretical framework for the response shift phenomenon that could enable testable hypotheses to be generated (Sprangers et al. 1999). But perhaps of primary importance would be the exploration of the meaning of being better, and the dynamics of that meaning in a longitudinal qualitative study. The methodological approach would no longer be grounded theory as we would be using the theory developed in this thesis work to explore deeper
meanings. It may therefore need shift to a phenomenological perspective (Creswell, 1998), an approach more suited for describing the meaning and depth of a particular theory. A qualitative approach would keep the patient as the primary source of understanding of whether or not our health status questionnaires can answer the question: “are you better?”.

6.3.2 Better indicators of change, patient-specific indices

Another potentially fruitful avenue of research would be an attempt to go one step further in obtaining the patient’s point of view on change. The present effort makes use of specific markers arising from the theory developed in the qualitative study. A worthwhile alternative might be to use patient-specific questionnaires such as the Canadian Occupational Performance Measure (Law et al. 1990), or the patient-specific index (Wright et al. 1997), as indicators of improvement. Difficulties in items that the patient identified as important could be tracked over time, and any improvement could be assumed to reflect improved health status and be used in Testa’s model as a proxy for true change ($\Delta C$) from a patient’s perspective (Testa et al. 1996). When used as health status measures themselves, these patient-specific indices are problematic in that varying content between patients limits their interpretability in any type of grouped analysis. But they could still be used as very valid, individualized indicators of change (change/unchanged). The only potential disadvantage of having a marker for change based on a highly individualized index is that the change scores described in the responsiveness analysis might be unpredictable. High variability in the content of the patient-specific index might lead to very different amounts and types of change (observed in the health status scores) in those who consider themselves “improved”. The variance would therefore be great, and lower the responsiveness statistic the distribution could be wide, and the responsiveness would be obscured
by the high variability. On the other hand, such variability could in fact be the most valid summary of change in a patient’s perspective.

6.3.3 Mathematical issues not addressed in this thesis

From the outset, the present thesis has been described as conceptual (Liang, 1995) or as focusing on the construct of change used in the modification of Testa’s model of responsiveness (Testa et al. 1996). This focus will minimize the discussion of the statistical management of change scores, an equally challenging task. Cronbach and Furby, among others, have questioned the very use of change scores, favouring instead randomized designs in which post-scores are examined only if possible (Cronbach et al. 1970). Difficult issues relating to statistical aspects of responsiveness may include the effect of using different summary statistics on the relative responsiveness of different measures (Wright et al. 1998) (raising the question of which is the best statistic). From a more mathematical point of view, there are problems concerning the quality of data (and its appropriateness for parametric analysis) (Michell, 1986; Nunnally et al. 1994; Stevens, 1958), how to work around error in measurement (Healy, 1989; Judd et al. 1981; Yanez et al. 1998; Nunnally et al. 1994; Cronbach et al. 1970; Burke et al. 1990), and how to deal with correlations between baseline and change scores (such as are found in the phenomenon of regression to the mean) (Nunnally et al. 1994; Judd et al. 1981; Rogosa et al. 1982; Nunnally, 1975; Norman, 1989). The detailed discussion this issue requires and deserves is beyond the present scope; however it is an important component of the measurement of change that needs to be recognized and attended to in future research on responsiveness.

The work presented here renders the interpretation of change scores and statistics of responsiveness more difficult, or at least more complex, but on the other hand it could also be
said to make change simpler to understand. Returning to the example in figure 1.1 (see Introduction), it is now easy see how one instrument could have produced such divergent yet valid descriptions of responsiveness. One bar (Bar A) represents responsiveness of the DASH outcome measure to observed change in a group of patients pre- and four-weeks following total shoulder replacement (standardized response mean = 0.13). The other bar (Bar B) represents the responsiveness of DASH to change in those patients deemed to have had an *important* change in function at 12 weeks, regardless of the specific treatment (standardized response mean = 1.27). Thus, different categories of responsiveness are capable of producing distinctly different descriptions of the responsiveness of the same questionnaire.

Guyatt and Cook warn that "the ability to make sense of the results [of an analysis of change], a property we have called interpretability, remains elusive for many questionnaires in common use" (Guyatt et al. 1994)(pg. 631). The present authors hope to have moved responsiveness one step closer to interpretability by introducing, and discussing the implications of, a taxonomy of responsiveness that identifies different types. The work has focused on the way the construct of change is defined and then built into studies of responsiveness. Deciding which proxy for true change to use in a study of responsiveness (or which study has the right proxy for your needs), depends on the category of responsiveness being considered, and the implications of that (what is important change?) in terms of the target construct - be it a measure of quality of life, health status, length, or range of motion.
Appendix A:

A NON-EXHAUSTIVE LIST OF AVAILABLE UPPER EXTREMITY SELF-COMPLETED QUESTIONNAIRES. (n=30)

<table>
<thead>
<tr>
<th>Whole extremity (n=6)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DASH: Disabilities of the Arm, Shoulder and Hand (Hudak et al. 1996) (McConnell et al. 1999)</td>
<td></td>
</tr>
<tr>
<td>TESS: Toronto Extremity Salvage Score (Davis et al. 1996)</td>
<td></td>
</tr>
<tr>
<td>NULI: Neck and upper limb index (Stock et al. 1996)</td>
<td></td>
</tr>
<tr>
<td>St. Mike's: St. Michael's Hospital Patient Self Evaluation (Beaton et al. 1998)</td>
<td></td>
</tr>
<tr>
<td>Upper Extremity Function Scale (Pransky et al. 1997)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Neck (n=4)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Neck disability index (Vernon et al. 1991) (Stratford et al. 1999; Riddle et al. 1998b)</td>
<td></td>
</tr>
<tr>
<td>Northwick Park Questionnaire (Leak et al. 1994)</td>
<td></td>
</tr>
<tr>
<td>Copenhagen Neck Functional Disability Scale (Jordan et al. 1998)</td>
<td></td>
</tr>
<tr>
<td>Neck Pain and Disability Questionnaires (Wheeler et al. 1999)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Shoulder (n=13)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>SSRS: Subjective Shoulder Rating Score (Beaton et al. 1998)</td>
<td></td>
</tr>
<tr>
<td>SST: Simple Shoulder Test (Matsen et al. 1995)</td>
<td></td>
</tr>
<tr>
<td>SPADI: Shoulder Pain and Disability Index (Roach et al. 1991)</td>
<td></td>
</tr>
<tr>
<td>SSI: Shoulder Severity Index (Beaton et al. 1998)</td>
<td></td>
</tr>
<tr>
<td>WOSI: Western Ontario Shoulder Instability Index (Kirkley et al. 1998)</td>
<td></td>
</tr>
<tr>
<td>Shoulder disability questionnaire (van der Heijden et al. 1996) (van der Windt et al. 1998)</td>
<td></td>
</tr>
<tr>
<td>Disability questionnaire (Croft et al. 1994)</td>
<td></td>
</tr>
<tr>
<td>Shoulder-Arm Disability Questionnaire (Bostrom et al. 1995)</td>
<td></td>
</tr>
<tr>
<td>Shoulder rating scale (L'Insalata et al. 1997)</td>
<td></td>
</tr>
<tr>
<td>Shoulder questionnaire (Dawson et al. 1996)</td>
<td></td>
</tr>
<tr>
<td>WUSPI: Wheelchair User's Shoulder Pain Index (Curtis et al. 1995)</td>
<td></td>
</tr>
<tr>
<td>Shoulder Pain Score (Winter et al. 1996)</td>
<td></td>
</tr>
<tr>
<td>ASES (American Shoulder &amp; Elbow Surgeons) patient self-evaluation (Richards et al. 1994)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Elbow (n=2)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient-rated forearm evaluation questionnaire (MacDermid et al. 1999b)</td>
<td></td>
</tr>
<tr>
<td>ASES patient self-evaluation (elbow) (King, 1999)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Wrist/hand (n=5)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Brigham carpal tunnel questionnaire (Levine et al. 1993) (Bessette et al. 1998; Katz et al. 1994)</td>
<td></td>
</tr>
<tr>
<td>PRWE: Patient-rated wrist evaluation (MacDermid, 1996) (MacDermid et al. 1999a; MacDermid et al. 1998a)</td>
<td></td>
</tr>
<tr>
<td>Anderson-McGill hand function questionnaire (MacDermid et al. 1999b)</td>
<td></td>
</tr>
<tr>
<td>Algofunctional index for osteoarthritis of the hand (*physician completed, but questionnaire) (Dreiser et al. 1995)</td>
<td></td>
</tr>
<tr>
<td>Michigan Hand Outcomes Questionnaire (Chung et al. 1998) (Chung et al. 1999)</td>
<td></td>
</tr>
</tbody>
</table>
Appendix B: Items Used in Pilot Testing of Indices of Responsiveness.

Sources for the status and transitional indices of improvement proposed for the DASH cohort.

This appendix summarizes the derivation of the external markers of change fielded in the pilot testing (described in Chapter four) and in the cohort study.

A: Status questions. (Asked at baseline and follow up, change = mathematical difference in responses)

1. How would you rate your arm/shoulder/hand problem over the past week? (check one box)

   1  2  3  4  5  6  7
   very mild
   neither severe, nor mild
ever very severe

Source: Overall severity rating of problem, consistent with very broad markers of change often used in the responsiveness literature.

2. How would you rate the intensity or severity of your arm/shoulder/hand pain or discomfort over the past week? (check one box)

   1  2  3  4  5  6  7  8  9  10
   no pain
   pain as bad as it can be

Source: Intensity of symptoms is important to patients. Will be used as a status measure of important change (time 2 - time 1) and will also appear as a transitional index. The wording and attribute
arose from the qualitative work, for instance: "every diminishment in pain, every step of diminishment gave me more motivation"[355:745--748].

3. How would you rate your arm/shoulder/hand pain or discomfort at night? (check one box)

<p>| | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>no</td>
<td>pain as bad</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pain</td>
<td>as it can be</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Discomfort at night taken from the Brigham symptom scale, included here for its relevance in carpal tunnel syndrome, and for rotator cuff tendinopathy. Note, this was only asked as a state measures, there is no parallel in the transitional approach.

4. How would you rate your ability to do your usual daily activities (such as opening jars, writing, doing household chores) over the past week? (check one box)

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>no</td>
<td>very</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>problem</td>
<td>difficult</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Ability to do daily activities was cited as important to participants of the qualitative study. "Everybody has a different threshold [for pain]....some people can put up with a fair amount of pain and still function...So I mean, it's more the ability to do whatever you want to do"[691:1195-1205]. "It took a good two weeks before I could turn that jar, or pick up a hammer again or squeeze a drill...even writing the bills for the first couple of weeks was painful enough"[252:1336-1341].

This scale will provide a construct for comparison (construct of physical dysfunction) as
well as a status index for change (compared to similar question at follow up). Items used as examples are those identified by patients as the most important and difficult of 70 upper extremity items. This was done during the development of the DASH outcome measure (Marx et al. 1999) (McConnell et al. 1999)

5. How would you rate your ability to do your paid work over the past week? (check one box)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no problem</td>
<td>very difficult</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Construct for validation of the DASH and the work module. Picking up the work or role disability that was found to be important in the qualitative study.

6. How would you rate your emotional health over the past week (sense of well being, happiness, coping with stress)? (check one box)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no problem</td>
<td>very bad</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Construct validity question, expecting that the DASH will not be as sensitive to this as the SF--36 mental health or mental component score. It would be interesting to use as a status index of change because we would hypothesize that the DASH or other measures of physical function should correlate less well with indices like this than one of symptoms and ADL.

7. How would you rate your ability to socialize with your friends or family or other support people over the past week? (check one box)

|   |   |   |   |   |   |   |   |
Source: Qualitative work suggested that this was part of the burden of musculoskeletal problem: the impact on family, and in particular when it involved child care. Given the focus of the DASH, we would not expect it to pick up change in social function as much as we would change in physical function, however this should be monitored over time as a marker of the impact of the disorder at a social level.

8. Can you do all the activities you would like to do?
   - Yes, I can.
   - No --- if no, are you:
     - restricted in your activities a little bit?
     - restricted in your activities a fair amount
     - restricted in your activities a lot

Source: This dichotomized question of function came up in the qualitative work. Wording was taken from the text of a participant. “It’s more the ability to do whatever you want to do”[252:1336-1341]. Traditionally this type of criterion has been used as a gold standard for a ROC analysis.

9. Can you do what you want to do and forget about your arms and your shoulder/arm/hand problem?
   - Yes, I can forget about my problem and do what I want to do
   - No ----> if no, are you having to think about your problem for
     - only a few activities (eg. only specific tasks)
     - about half your activities
all activities

Source: Forgetting about pain was another theme of important change arising from the qualitative work. This is a difficult one to word as it is hard to capture when people have actually forgotten something. “Gee, I did all that ironing and my hand was okay!” [991:950-959]. “Forget that your arms are there, you know. The way I work anyway, I type really fast and if there’s anything bothering me, it screws up, you know, the whole process” [11:237-239].

10. Are you free of pain or discomfort in your shoulder/arm/hand?

☐ Yes
☐ No

Source: Freedom from pain was identified as an absolute indicator of being better. It often came up in the qualitative work to describe how they were not yet completely better, because they were not yet free from pain.

11. Are you working in your usual way (normal duties, not modified)

☐ Yes
☐ No ---> if no, are you
       ☐ having only a bit of difficulty with work
       ☐ having a fair amount of difficulty with work
       ☐ having a lot of difficulty with work
       ☐ unable to work
       ☐ I am not working for reasons other than my arm problem

Source: Doing work in the usual, and sustainable way arose in the qualitative work.
12. Is your arm/shoulder/hand problem at a level now where you can ignore or cope with it and carry on with whatever it is you have to do in your daily life?

☐ yes, I can cope with it at the level it is at now.
☐ no, it is not at a level where I can cope with it.

Source: Important change did not require the eradication of pain, but rather the reduction of it to a certain threshold of being tolerable. As one participant suggest: “I don’t consider any pain a problem....when it would get to the point where it would be so distressing that I couldn’t put it out of my mind and carry on with what I was doing, that was the problem.”[272:1124-1136]. “The most important thing was it wasn’t bothering me anymore, you know, to the extent that it was”[80:373-375]

13. Do you feel you can manage or control your pain?

☐ yes, I can.
☐ no, I can not manage my pain.

Source: Being able to manage pain was one of the themes arising from the qualitative work. “You feel you really manage it and that you’re in control and that.... you know you are going to be able to manage it. [268:455-460]. This is a target state, feeling in control. It is described in work in the acceptance of chronic disease (Shaul,1995) and also in the self-efficacy work of Lorig and others (Bandura,1997) (Eisenberg,1980) (Lorig et al. 1989) (Lorig et al. 1993).

B: Transition questions (answered at follow up only, recalling time between follow up and baseline).
The transition scales reflect a subset of the attributes described above that were reframed into a transitional manner and administered at each follow-up.

Each of the transition scales will be followed by a four point scale of importance. The goal here is to capture the importance of the experience the subject has had. It is anticipated that people will frame this in terms of the magnitude of change they experienced. However, such instructions were not given to the subjects as we wished them to consider the concept of importance, which can be whatever part of the change experience they wish to call important.

Was this experience of change important to you in your life? (check one)

- [ ] not important at all
- [ ] a little important
- [ ] fairly important
- [ ] very important

A. Think about your arm/shoulder/hand problem now compared to when you started your treatment (or before your surgery). How would you rate the change in your upper extremity condition/problem? (check one)

- [ ] 1 Much worse
- [ ] 2
- [ ] 3
- [ ] 4
- [ ] 5
- [ ] 6 No change
- [ ] 7
- [ ] 8
- [ ] 9
- [ ] 10
- [ ] 11 A lot better

B. Again, compared to before your treatment/surgery, how would you rate the change in the intensity or severity of your arm/shoulder/hand pain or symptoms? (check one)
C. Compared to before your treatment/surgery, how would you rate the change in the frequency (how often) of your pain or symptoms? (check one)

   1  2  3  4  5  6  7  8  9  10  11
   Much worse  No change  A lot better

Source: Change in frequency of pain is the only attribute that was consistently framed in a transitional manner in the qualitative work (a decrease in the frequency of my pain). It is the only item that is an addition to the attributes capture in the state measures just described.

D. Compared to before your treatment/surgery, how would you rate the change in your ability to do the things you want or need to do (such as opening jars, writing, preparing meals, doing heavy chores, carrying bags, lifting children)? (check one)

   1  2  3  4  5  6  7  8  9  10  11
   Much worse  No change  A lot better

E. Compared to before your treatment/surgery, how would you rate the change in your ability to do your usual work (including homemaking if that is your primary role)? (check one)
F. We may have missed asking you about a change that you felt was important. Here is a chance to add that idea in. Add your important change into this questions, and then respond to it:

Compared to before your treatment/surgery, how would you rate the change in

________________________ (insert the change you think was important)? (check one)

1 2 3 4 5 6 7 8 9 10 11
Much worse No change A lot better

C: Open ended questions

At the end of the pilot testing of the questionnaire, we wished to offer the respondents an opportunity to add additional items to the list, or describe other changes they had experienced.

The following questions were placed at the end of the package.

“...A lot of these questions are trying to ask you about changes in your condition, and whether or not you felt they were important indicators of how you are now, as compared to before you started treatment. Can you describe changes in your arm/shoulder/hand condition, or changes in the impact it has had on your life, which were important to you in terms of knowing that you
were getting better, or worse?

“How will you decide that your condition is getting better than it is today? What changes will be important improvements from today?”
Appendix C: Example of follow-up questionnaire package used in the cohort study.

The following is the content of the follow-up questionnaire package for shoulder patients. In addition to these questionnaires, the Short-Form 36 item (acute version) was inserted after the SPADI scale.

The Brigham Questionnaire

1. On a typical day during the PAST WEEK, have shoulder, wrist or hand symptoms (pain, numbness, weakness, tingling) caused you to have any difficulty doing the activities listed below? Please circle one number for each activity.

<table>
<thead>
<tr>
<th>Activity</th>
<th>No difficulty</th>
<th>Mild difficulty</th>
<th>Moderate difficulty</th>
<th>Severe difficulty</th>
<th>Cannot do at all</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Writing</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>b. Buttoning clothes</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>c. Holding a book while reading</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>d. Gripping the phone handle</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>e. Opening jars</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>f. Household chores</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>g. Carrying grocery bags</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>h. Bathing &amp; dressing</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>i. Driving</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>j. Sleeping</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>k. Vigorous activities (heavy lifting, strenuous sports)</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

The following questions refer to your symptoms on a typical 24 hour period during the PAST WEEK.

2. How severe is the shoulder, wrist or hand pain you have at NIGHT?
   1. I do not have shoulder, wrist or hand pain at night
   2. Mild pain
   3. Moderate pain
   4. Severe pain
   5. Very severe pain

3. How often did shoulder, wrist or hand pain at NIGHT wake you up during a typical night in the PAST WEEK?
1. Never
2. Once
3. Two or three times
4. Four to five times
5. More than five times

4. **Do you typically have pain in your shoulder, wrist or hand during the DAYTIME?**
   1. I never have pain during the day
   2. I have mild pain during the day
   3. I have moderate pain during the day
   4. I have severe pain during the day
   5. I have very severe pain during the day

5. **How often do you have shoulder, wrist or hand pain during the DAYTIME?**
   1. Never
   2. Once
   3. Two or three times
   4. Four to five times
   5. More than five times

6. **How long, on average, does an episode of pain last during the DAYTIME?**
   1. I never get pain during the day
   2. 10 minutes or less
   3. 10-60 minutes
   4. Greater than 60 minutes
   5. The pain is constant throughout the day

7. **Do you have numbness (loss of sensation) in your hand?**
   1. No
   2. I have mild numbness
   3. I have moderate numbness
   4. I have severe numbness
   5. I have very severe numbness

8. **Do you have weakness in your shoulder, wrist or hand?**
   1. No weakness
   2. I have mild weakness
   3. I have moderate weakness
   4. I have severe weakness
   5. I have very severe weakness
9. Do you have tingling sensations in your hand?
   1. No tingling
   2. I have mild tingling
   3. I have moderate tingling
   4. I have severe tingling
   5. I have very severe tingling

10. How severe is numbness (loss of sensation) or tingling AT NIGHT?
   1. I have no numbness or tingling at night
   2. Mild
   3. Moderate
   4. Severe
   5. Very severe

11. How often did the numbness or tingling wake you up during a typical night during the PAST WEEK?
   1. Never
   2. Once
   3. Two or three times
   4. Four to five times
   5. More than five times

12. Do you have difficulty with grasping and using small objects such as keys or pens?
   1. No difficulty
   2. Mild difficulty
   3. Moderate difficulty
   4. Severe difficulty
   5. Very severe difficulty
The DASH Questionnaire.
This questionnaire asks about your symptoms as well as your ability to perform certain activities. Please answer every question, based on your condition in the last week, by circling the appropriate number. If you did not have the opportunity to perform that activity in the past week, please make your best estimate on which response would be the most accurate. It doesn't matter which arm or hand you use to perform the activity; please answer based on your ability regardless of how you perform the task.

<table>
<thead>
<tr>
<th>Activity</th>
<th>No difficulty</th>
<th>Mild difficulty</th>
<th>Moderate difficulty</th>
<th>Severe difficulty</th>
<th>Unable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Open a tight or new jar.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>2. Write.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>3. Turn a key.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>4. Prepare a meal.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5. Push open a heavy door.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>6. Place an object on a shelf above your head.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>7. Do heavy household chores (e.g., wash walls, wash floors).</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>8. Garden or do yard work.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>9. Make a bed.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>10. Carry a shopping bag or briefcase.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>11. Carry a heavy object (over 10 lbs.).</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>12. Change a lightbulb overhead.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>13. Wash or blow dry your hair.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>14. Wash your back.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>15. Put on a pullover sweater.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>16. Use a knife to cut food.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>17. Recreational activities which require little effort (e.g., card playing, knitting etc.).</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>
18. Recreational activities in which you take some force or impact through your arm, shoulder or hand (e.g., golf, hammering, tennis, etc.).

19. Recreational activities in which you move your arm freely (e.g., playing frisbee, badminton, etc.).

20. Manage transportation needs (getting from one place to another).

21. Sexual activities.

<table>
<thead>
<tr>
<th></th>
<th>Not at all</th>
<th>Slightly</th>
<th>Moderately</th>
<th>Quite a bit</th>
<th>Extremely</th>
</tr>
</thead>
<tbody>
<tr>
<td>22. During the past week, to what extent has your arm, shoulder, or hand problem interfered with your normal social activities with family, friends, neighbours or groups?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Not at all limited</td>
<td>Slightly limited</td>
<td>Moderately limited</td>
<td>Very limited</td>
<td>Unable</td>
</tr>
<tr>
<td>23. During the past week, were you limited in your work or other daily activities as a result of your arm, shoulder or hand problem?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>None</td>
<td>Mild</td>
<td>Moderate</td>
<td>Severe</td>
<td>Extreme</td>
</tr>
</tbody>
</table>

Please rate the severity of the following symptoms in the last week

<table>
<thead>
<tr>
<th></th>
<th>No difficulty</th>
<th>Mild difficulty</th>
<th>Moderate difficulty</th>
<th>Severe difficulty</th>
<th>So much difficulty that I can't sleep</th>
</tr>
</thead>
<tbody>
<tr>
<td>24. Arm, shoulder or hand pain.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25. Arm, shoulder or hand pain when you perform any specific activity.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26. Tingling (pins and needles) in your arm, shoulder or hand.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27. Weakness in your arm, shoulder or hand.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28. Stiffness in your arm, shoulder or hand.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
29. During the past week, how much difficulty have you had sleeping because of the pain in your arm, shoulder or hand?

<table>
<thead>
<tr>
<th>Strongly disagree</th>
<th>Disagree</th>
<th>Neither agree nor disagree</th>
<th>Agree</th>
<th>Strongly agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

30. I feel less capable, less confident or less useful because of my arm, shoulder or hand problem.

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
</table>

**Work Module**: The following questions ask about the impact of your arm, shoulder or hand problem on your ability to work (including homemaking if that is your main work role).

Please indicate what your job/work is:

- I do not work. (You may skip this section)

Please circle the number that best describes your physical ability in the past week. Did you have any difficulty:

<table>
<thead>
<tr>
<th>No difficulty</th>
<th>Mild difficulty</th>
<th>Moderate difficulty</th>
<th>Severe difficulty</th>
<th>Unable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

1. Using your usual technique for your work?
2. Doing your usual work because of arm, shoulder or hand pain?
3. Doing your work as well as you would like?
4. Spending your usual amount of time doing your work?

© IWH & AAOS & COMSS

**The SPADI Questionnaire.**

This questionnaire is answered by placing an X along the line to show where you would be between end points. For example, if you had a lot of pain in the situation described, you might place your mark the right on the scale.

<table>
<thead>
<tr>
<th>No pain</th>
<th>X</th>
<th>Worst pain imaginable</th>
</tr>
</thead>
</table>
**Pain Scale.** In the past week, how severe has your pain been... (Mark an X on each line).

1. At its worse.....

| No pain | Worst pain imaginable |

2. When lying on the involved side....

| No pain | Worst pain imaginable |

3. Reaching for something on a high shelf.....

| No pain | Worst pain imaginable |

4. Touching the back of your neck.....

| No pain | Worst pain imaginable |

5. Pushing with the involved arm.....

| No pain | Worst pain imaginable |

**Disability Scale.**
How much difficulty did you have......

1. Washing your hair.....

| No difficulty | So difficult it required help |

2. Washing your back.....

| No difficulty | So difficult it required help |

3. Putting on an undershirt or pullover sweater....

| No difficulty | So difficult it required help |
4. Putting on a shirt that buttons down the front......

5. Putting on your pants....

6. Placing an object on a high shelf......

7. Carrying a heavy object of 10 pounds.....

8. Removing something from your back pocket.....

SF-36 Acute Version inserted here.

EQ-5D Questionnaire inserted here.

These questions ask a bit more about the impact of your shoulder problem......

During the last month, for how many days did you cut down on the things you usually do because of your shoulder problem?

_____ _____ days
During the last month, for how many days did your shoulder problem keep you from going to work or school or from doing housework?

____ ____ days

How many visits have you made to physicians because of your shoulder problem in the last six months?

____ ____ visits

In the last six months, have you seen an occupational or physical therapist because of your shoulder problem?

1. No
2. Yes, if yes how many visits have you made to the therapist? ____ ____ visits

These questions ask you to think about how your shoulder problem from a broader perspective. Please answer each question by checking the appropriate box.

1. How would you rate your shoulder problem over the last week? (Check one)

☐ ☐ ☐ ☐ ☐ ☐ ☐
1 2 3 4 5 6 7
Very mild Neither severe nor mild Very severe

2. How would you rate the intensity or severity of your shoulder pain or discomfort over the last week? (Check one box)

☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
1 2 3 4 5 6 7 8 9 10
No Pain as bad as it can be
pain

3. How would you rate your shoulder pain or discomfort at night over the last week? (Check one box)

☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
1 2 3 4 5 6 7 8 9 10
No Pain as bad as it can be
pain

4. How would you rate your ability to do your usual daily activities (such as opening jars,
writing, doing 
household chores) over the last week? (Check one)

☐ ☐ ☐ ☐ ☐ ☐ ☐ 
1 2 3 4 5 6 7
Not Very
Difficult difficult

5. How would you rate your ability to do your paid work over the last week? (Check one)

☐ ☐ ☐ ☐ ☐ ☐ ☐ 
1 2 3 4 5 6 7
Not Very
Difficult difficult

6. How would you rate your emotional health (sense of well being, happiness, coping with stress) over the last week? (Check one)

☐ ☐ ☐ ☐ ☐ ☐ ☐ 
1 2 3 4 5 6 7
No Very
problem bad

7. How would you rate your ability to socialize with your friends or family or other support people over the last week? (Check one)

☐ ☐ ☐ ☐ ☐ ☐ ☐ 
1 2 3 4 5 6 7
Not Very
Difficult difficult

8. Can you do all the activities you would like to do? (Check one)

☐ Yes, I can forget about my problem and do what I want to do
☐ No. If no, are you having to think about your problem for

☐ only a few activities (e.g. only specific tasks)
☐ about half your activities
☐ all activities

9. Can you do what you want to do and forget about your shoulder problem? (Check one)

☐ Yes, I can forget about my problem and do what I want to do
☐ No. If no, are you having to think about your problem for

☐ only a few activities (e.g. only specific tasks)
☐ about half your activities
☐ all activities

10. Are you free of pain or discomfort in your shoulder? (Check one)

☐ Yes
☐ No
11. Are you working in your usual way (normal duties, not modified)? (Check one)
   - [ ] Yes
   - [ ] No. If no, are you
     - [ ] having only a bit of difficulty
     - [ ] having a fair amount of difficulty with work
     - [ ] having a lot of difficulty with work
     - [ ] unable to work
     - [ ] I am not working for reasons other than my shoulder problem

12. Is your shoulder problem at a level now where you can ignore or cope with it and do whatever it is you have to do in your daily life? (Check one)
   - [ ] Yes, I can cope with it at the level it is at now.
   - [ ] No, it is not at a level where I can cope with it.

13. Do you feel you can manage or control your pain? (Check one)
   - [ ] Yes, I can manage my pain.
   - [ ] No, I cannot manage my pain

The next questions ask about how satisfied you are with the outcome of your treatment.

1. How satisfied are you with the results of the treatment for your shoulder syndrome?
   1. Very satisfied
   2. Somewhat satisfied
   3. Somewhat dissatisfied
   4. Very dissatisfied

<table>
<thead>
<tr>
<th>How satisfied are you with the results of treatment for your shoulder problem for......</th>
<th>Very satisfied</th>
<th>Somewhat satisfied</th>
<th>Somewhat dissatisfied</th>
<th>Very dissatisfied</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Circle one number on each line)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. relieving your pain?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>3. improving self-care activities such as bathing and dressing?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4. improving your ability to do home or yard work?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>5. improving your ability to do recreational activities?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>
These next questions ask you to think about how you are now **in comparison** with when you began this treatment, or had your surgery.

A. Think about your shoulder problem now compared to when you started your treatment (or before your surgery). How would you rate the change in your shoulder problem? (Check one)

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Much worse</td>
<td>No change</td>
<td>A lot better</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

How important was this experience of change for you in your life?

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not very important</td>
<td>Fairly important</td>
<td>Very important</td>
<td>Extremely important</td>
<td></td>
</tr>
</tbody>
</table>

B. Again, compared to before your treatment/surgery, how would you rate the change in the intensity or severity of your shoulder pain or symptoms? (Check one)

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Much worse</td>
<td>No change</td>
<td>A lot better</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

How important was this experience of change for you in your life?

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not very important</td>
<td>Fairly important</td>
<td>Very important</td>
<td>Extremely important</td>
<td></td>
</tr>
</tbody>
</table>

C. Compared to before your treatment/surgery, how would you rate the change in the frequency (how often) of your pain or symptoms? (Check one)

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Much worse</td>
<td>No change</td>
<td>A lot better</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

How important was this experience of change for you in your life?

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not very important</td>
<td>Fairly important</td>
<td>Very important</td>
<td>Extremely important</td>
<td></td>
</tr>
</tbody>
</table>
D. Compared to before your treatment/surgery, how would you rate the change in your ability to do the things you want or need to do (such as opening jars, writing, preparing meals, doing heavy chores, carrying bags, lifting children)? (Check one)

- Much worse
- No change
- A lot better

How important was this experience of change for you in your life?

- Not very important
- Fairly important
- Very important
- Extremely important

E. Compared to before your treatment/surgery, how would you rate the change in your ability to do your usual work (including homemaking if that is your primary role)? (Check one)

- Much worse
- No change
- A lot better

How important was this experience of change for you in your life?

- Not very important
- Fairly important
- Very important
- Extremely important

F. We may have missed asking you about changes in your shoulder problem that you felt were important indicators of whether you were getting better (or worse). What other change or changes in your shoulder problem were important to you?

Think about how you are today. During the next several weeks, what changes in your shoulder problem will tell you that you have had important improvement from the way you are today?

You are finished the questionnaire package. Thank you for your time....

Please mail this back in the envelope provided.
Appendix D: Measuring the whole or the parts? Validity, reliability & responsiveness of the DASH Outcome Measure in different regions of the upper extremity.

Preface

In order to measure and interpret change, there needs to be some level of confidence in the “yardstick” or the measurement instrument. This appendix will focus on examining the reliability, validity, and responsiveness (in a very broad manner) of the measurement device used in Chapter five to examine the implications of the taxonomy (Chapter two). This instrument is the DASH (Disabilities of the Arm, Shoulder and Hand) Outcome Measure, a measure containing items on disability and symptoms designed for persons with any upper limb disorder.

This chapter is provisionally accepted for publication in the Journal of Hand Therapy. Sections of the manuscript, in particularly the sampling methods, and the baseline description of the features of the cohort sample, are deleted from the appendix, because they also appear in Chapter five.

Introduction

The measurement of disability (Verbrugge et al. 1994; Jette, 1994; Cole et al. 1994) or capacity to function (World Health Organization, 1999) is critical to a comprehensive assessment of outcome following an injury in the upper limb. The fluid motion of a swimmer, artist, or
musician attests to the coordinated kinetic chain along the extremity which allows for such expression and function (Davis et al. 1999; Brand et al. 1993). However, measuring disability in patients with upper-limb disorders poses practical challenges. For example, distinct questionnaires have been developed for the different regions of the upper limb (Roach et al. 1991; Beaton et al. 1998; Stock et al. 1996; MacDermid et al. 1998b; MacDermid, 1996), or for various disorders in the limb (Levine et al. 1993; Davis et al. 1996) (Kirkley et al. 1998). Given that many patients could have multiple disorders or multiple regions affected, the choice between available measures is difficult. The DASH (Disabilities of the Arm, Shoulder and Hand) Outcome Measure (the DASH) provides one possible solution. It is a questionnaire designed to be used for any or multiple disorders in the upper limb offering the possibility of using a single questionnaire to measure disability for any of the upper-limb regions (Davis et al. 1999; McConnell et al. 1999; Hudak et al. 1996). The intent of the DASH is that it could be used no matter what region(s) is, or are, affected.

The development of the DASH has been documented elsewhere (Hudak et al. 1996; Marx et al. 1999; McConnell et al. 1999). The DASH is a 30 item questionnaire, evaluating symptoms and physical function (at the level of disability (Jette, 1994; Verbrugge et al. 1987; Verbrugge et al. 1994)), with 5 response options for each item. Scoring is done by summing up the circled responses, subtracting 30 (in order to anchor the score with a base of zero - a correction made necessary because the response scale is 1-5, and needs to be changed to a 0-4 equivalent). This figure is then divided by 1.2 to get a DASH function/symptom score out of a possible 100. A higher score on the DASH reflects more disability. Missing items (up to 3 items, or 10% of items) are replaced by the mean value of the responses to the other items before summing. If more than three items are missing, the overall score cannot be calculated.
(McConnell et al. 1999). Preliminary work on the validity (against constructs of function and pain) and reliability (Cronbach’s alpha = 0.96 (McConnell et al. 1999), test-retest reliability = 0.92 (Turchin et al. 1998)) has been carried out by those involved in the development (Hudak et al. 1996; McConnell et al. 1999; Marx et al. 1999; Davis et al. 1999) as well as indirectly by others who used it to compare with another instrument (Kirkley et al. 1998; MacDermid et al. 1999a). Kirkley (Kirkley et al. 1998) and MacDermid (MacDermid et al. 1999a) also demonstrated that the DASH was responsive, though slightly less so than more joint- or disorder-specific measures -- a wrist-specific and a shoulder instability-specific instrument in these studies respectively. These studies provide initial evidence of the validity and reliability of the DASH scores, however, additional work was needed to compare the DASH in patients with disorders in different upper-limb regions. Of particular interest was how the DASH would do in evaluating change over time (its intended role) in patients with disorders in different parts of the extremity, a role which requires evidence of construct validity, test-retest reliability and responsiveness (Guyatt et al. 1992a; Hays et al. 1993; DeVellis, 1996; Guyatt et al. 1997; Lohr et al. 1996). Only three of the above mentioned articles that used the DASH collected data over time for their study (Kirkley et al. 1998; Turchin et al. 1998; MacDermid et al. 1999a) but none focused on the DASH per se, and none provided information on all three attributes.

The purpose of this study was to evaluate the validity, test-retest reliability and responsiveness to change of the DASH Outcome Measure in a longitudinal study of patients with various upper-limb disorders.

**Methods.**

**Design**

A convenience sample of patients waiting for treatment of their upper-limb conditions at
one of two teaching centres (St Michael’s Hospital in Toronto and Brigham and Women’s Hospital in Boston) were invited to participate in this prospective cohort study. The study did not affect their care, but rather observed their progress over a three-month period by obtaining descriptions of their health (through questionnaires only) before and after treatment.

The sample size calculations, recruitment process and sample description are outlined in detail in Chapter five.

In this appendix we will report the results from the baseline data, the 3-5 day retest data, and the 12 week follow up data. Chapter five also describes four week follow-up data that was collected at only one site, Toronto.

*Measures and Analysis:*

Each completed questionnaire package was reviewed by research staff to check for such things as missing items and duplicate responses, and then entered into a customized database at each site. These were converted into SAS data sets (SAS Statistical Analysis Systems, version 6.12) and merged. All analysis was done in SAS.

**Construct Validity.**

Several different comparisons were done to evaluate the construct validity of the DASH scores using recommended methods (Nunnally et al. 1994; Lohr et al. 1996).

We hypothesized that the DASH scores would be sensitive to the range of disability in our sample. This was verified by looking at the distributions of baseline scores (whole sample, proximal, distal, operative, non-operative) and specifically looking for floor or ceiling effects (patients with scores at either extreme of the scale) which would indicate a lack of sensitivity to
the disability experience in this sample. Floor and ceiling effects would also lead to difficulties in trying to measure change (for instance if everyone scores at the best possible score – a ceiling effect – there is no place to move to on the scale if they improve) (Bindman et al. 1990).

We also felt that the DASH scores should be lower (less disability) in the following groups: those working full duty rather than not; those able to cope and do what they want rather than not. These contrasts were tested with an unpaired Students t-test, at a 0.05 level of error.

We posited that the DASH should also correlate at least moderately (Pearson correlations of > 0.5) with: visual analog scales of function, pain, and ability to work as well as with established joint-specific measures (Shoulder Pain and Disability Index (SPADI) for shoulder patients (Heald et al. 1997; Roach et al. 1991; William et al. 1995), and the Brigham Questionnaire (Brigham) (Katz et al. 1995; Levine et al. 1993) for wrist and hand patients). Furthermore, if the joint-specific measures are indeed specific to a particular joint, we should see lower correlations between the DASH and Brigham in the shoulder patients, as well as the DASH and SPADI in the wrist patients. The disability or function scores of the SPADI and Brigham were the focus of this analysis.

TEST-RETEST RELIABILITY.

Test-retest reliability was analyzed using data from those who had two measures before treatment began (3-5 days apart) and said that their arm problem was “unchanged” on a transitional 11 point scale (how is your problem now compared to when you last did this questionnaire package, about 3-5 days ago?) (n=56 said “no change”, score of 5 +/- 1). Mean change scores, and associated paired t-statistic (and p-value) were calculated. Correlation coefficients were done using both Pearson (parametric, normally distributed data) and
Spearman (non-parametric, using ranks) methods. These correlations indicate if scores for a given patient are high at time one and are also high at follow up, but not whether the scores were identical. Intraclass correlation coefficients (ICC) offer an estimate of how closely the numeric scores for each patient were to each other (called concordance), and are therefore considered a stronger statistic for describing reliability (Deyo et al. 1991; Cole et al. 1994; Shrout et al. 1979). Specifically, we used a Shrout and Fleiss (2,1) model derived from a two way analysis of variance (Shrout et al. 1979). By adopting this particular model, we are saying that the testing framework in this particular study (3-5 days apart, pre-treatment) is assumed to be only one of many possible ways the test-retest reliability could have been assessed. We considered a coefficient of 0.90 - 0.95 a minimum standard for reliability based on the guidelines of Lohr et al. (Lohr et al. 1996) and others (Nunnally et al. 1994; McHorney et al. 1995) for the ability to interpret questionnaire scores in individual patients.

The final estimate of reliability, the minimally detectable change (MDC) (Christensen et al. 1986a; Jacobson et al. 1984; Stratford et al. 1996b; Stratford et al. 1996c; Wyrwich et al. 1999b), was calculated using the reliability coefficient to estimate the standard error of measurement for the difference. The standard error was therefore: standard deviation(baseline) times the square root of (1 - $R_{xx}$) where $R_{xx}$ is the test-retest reliability (others use internal consistency (Wyrwich et al. 1999b)). Christensen (Christensen et al. 1986a) suggests multiplying the (1-$R_{xx}$) by two, in order to adjust for the fact that when looking at a change score two samples are being used (time one and time two) each with measurement error. The standard deviation is therefore increased to account for this. Wyrwich (Wyrwich et al. 1999b) and Jacobson (Jacobson et al. 1984) do not make this adjustment (hence the results of the adjusted approach will result in MDC's 1.42 times greater than Wyrwich (Wyrwich et al. 1999b)).
The MDC is calculated by taking this estimation of the standard error of measurement of the difference score and multiplying it by 1.96 (the point on the z distribution relating to a 95% confidence interval) and the variance at baseline (MDC\textsubscript{95} = 1.96 \times \sigma_\text{base} \times \sqrt{2 \times (1 - R_{\text{ICC}})}). It should be noted that an MDC could also be created for a 90% confidence interval, substituting 1.64 for the 1.96 in the formula. For this reason we have chosen to label the MDC with a subscript “95” to indicate which level of confidence we are using. The MDC\textsubscript{95} yields a threshold -- a minimum change score -- that allows you to be 95% confident that when you observe a change score in your patient which is greater than this value, it is likely to be a real change, rather than just measurement error alone (for that instrument in that population). The MDC provides a unique opportunity to translate the test-retest reliability coefficient (ICC) into units of change in the instrument.

RESPONSIVENESS TO CHANGE.

The ability to detect change when it has occurred (De Bruin et al. 1997; Wright et al. 1998), or responsiveness, is often incorrectly felt to be a property of an instrument (e.g. the XXX questionnaire is “responsive”). However, like construct validity, we are validating the application -- the test situation -- not the instrument in and of itself. Evaluation of responsiveness requires that some sort of change has occurred (and that it can be verified in some way) and then the questionnaire’s scores are tested against that change. It is possible that a given questionnaire could be responsive or sensitive to a particular type of change, but not to another type (Beaton et al. 1999b; Wright, 1996). Responsiveness needs to be described in relation to the relevant type of change. In our study we looked at three types of change known to have occurred between baseline and 12 weeks after treatment onset. First, we assumed that
patients would likely begin to show improvement with each of the treatments in the study (i.e.,
total shoulder replacement, carpal tunnel release), and therefore we compared pre- and 12-week
post-treatment scores in the whole group. We recognize that this does not reflect full recovery,
but patients likely will have had small improvements from the pre-treatment state by this time.
Second, we looked at people who had said that their upper-limb problem was better. We
determined this by their response to the 11-point scale asking “compared to before your
treatment/surgery how much has your arm (either shoulder or wrist/hand) problem changed?”
Those indicating >6 on the scale (where 5 = no change and 10 = much better) were estimated to
have improved. Different cut points could have been used, however we chose this point based
on an a priori consensus of four of the authors (DB, CB, JGW, JNK). Similarly, an indication
of >6 improvement in ability to function in daily activities was used as an external standard that
change had occurred, and responsiveness analysis carried out on that subsample who indicated
>6/10 on the scale.

In all cases responsiveness was summarized using the following statistics: change
scores, effect size (mean change divided by the standard deviation of baseline scores) and the
standardized response mean (mean change divided by the standard deviation of change scores
(SRM)). A comparison was then carried out contrasting the responsiveness (using the SRM) of
the DASH, Brigham and SPADI in each of the sub-cohorts: Shoulder patients, and wrist/hand
patients. Thus we compared the responsiveness of a shoulder questionnaire to improvements in
wrist patients and vice versa. It was hypothesized that the DASH should have comparable
responsiveness to the joint-specific, but the joint-specific will be highest in patients with that
joint involved.

We also described responsiveness by correlating change scores on the DASH with
change in pain intensity, function and severity of the problem. Two approaches were used to measure these attributes. First, numeric rating scales were used to gather these ratings at each testing time (we called these “status measures”) and the differences between the status measures at 12 weeks and at baseline (we called these “differences in status measures”) were correlated with the difference in DASH scores. Second, we also asked patients (at 12 weeks) to rate the amount these same attributes had changed since baseline (pre-treatment). We called this a transition approach. It would be hypothesized that both approaches should lead to at least moderate correlations with changes in the DASH scores. We expected slightly lower correlations because of the fact that the comparison measures were made up of only one or two items and hence are prone to more measurement error which will lower the correlation coefficient (attenuation of correlation (Nunnally et al. 1994)). As well change scores are derived from two samples (pre and post) and this also adds to the error in the measure (as discussed above), and therefore to the attenuation of the correlation. We therefore considered \( r \approx 0.4 \) indicative of a moderate correlation.

We also constructed receiver operator characteristic curves as described by others (Deyo et al. 1986; Stratford et al. 1996a). These curves demonstrate how accurately different change scores on the questionnaire distinguish those who are better from those who are not (as defined by some other criterion). In our study we used an affirmative answer to the question “can you cope with your problem and do what you would like to do” at follow-up, given an inability to cope at baseline, as the criterion of improvement. Although other questions could have been used, we selected this question on the basis of a qualitative study in which this was described as a threshold type of indicator of being “better” (Beaton et al. 1999a). Change scores of -1, -5, -7,
-10, -15 and -20 were considered*. The true positive rate (% of people with a change score of at least that amount who were also now able to cope) and false positive rate (or one minus specificity: the % with a change score of at least that amount but who had not shifted from being unable, to being able to cope) calculated and plotted on a graph -- a receiver operator characteristic curve (Fletcher et al. 1988). The area under the curve represents the responsiveness; the larger the area, the more responsive the instrument because the different change scores are able to discriminate improved from non-improved patients (as defined by our gold standard). In a ROC curve for one instrument, the point highest to the upper left might be considered the change score most able to discriminate between those who have shifted to coping with their condition and those who have not.

Results:

Sample.

Two hundred persons were enrolled and completed the baseline portion of this study. One hundred and seventy-two completed the 12-week follow-up.

*The reasons these change scores were selected for the cut-offs are as follows: one is closest to the smallest change detectable on the DASH (0.83) (Beaton et al. 1999a). A change of five is equivalent to one standard error of measurement which Wytrich (Wytrich et al. 1999b) suggests is close to a minimally clinically important difference, 7 (or 7% change) is the cut point suggested by Redelmeier's work (Redelmeier et al. 1993). 10 is selected for convenience only, 15 is a cut off which corresponds to the minimally detectable change if derived from the work of Turchin (Turchin et al. 1998), but also corresponds to Redelmeier's suggestion that the criterion for an important change is 0.5 points per item in a questionnaire (therefore 30 x 0.5) (Redelmeier et al. 1996c). Finally we selected a cut point of 20 to represent an extreme change score.
questionnaire (86% follow up rate). The samples are described in detail in Chapter five.

*Construct validity*

The distribution of the baseline scores on the DASH appear to be normally distributed (see Figure D.1) with mean of 43.9 and median of 44.6. Only one person was at the “ceiling” perfect health (score of 0), no one is at the floor (100 = maximum disability on scale). The distribution for the shoulder patients and the wrist/hand patients is shown in the same figure, and demonstrates the less severe disability in the wrist/hand patients described by the DASH scores.

Evidence of discriminative validity was found. Those currently working and able to continue to do so with their upper limb condition had significantly lower disability than those not able to work (26.8 vs. 50.7, \( t = -7.51, p < 0.0001 \)) (this analysis only contrasted these two subgroups, and did not analyze the responses from those who were retired, or not working for reasons other than their upper limb condition). Statistically significant differences in DASH scores were also found between those who were able to do all they want to do as opposed to not being able to do so (23.6 versus 47.1, \( t = -5.81, p < 0.0001 \)). Similar discrimination was found within the shoulder and wrist/hand patients when analyzed separately. Thus the difference was in the anticipated direction (people unable to do what they wanted, and those unable to work had more disability, higher DASH scores), and the difference was statistically significant.

Convergent construct validity of the DASH was demonstrated by finding correlations in the expected direction and of the expected magnitude with other measures of upper-limb function and symptoms. Table D.1 summarizes the results. In the whole cohort, all correlations exceeded 0.70 (Pearson). Correlations were highest with the measure of function, as well as the
function scores on the Brigham and the SPADI. Correlations between the DASH and these other joint-specific instruments were found even in the opposite joint.

### Table D.1. Construct validity of the DASH. Correlations between the DASH and other measures of upper extremity function. Pearson correlations and Spearman correlations (in square brackets [ ]).

<table>
<thead>
<tr>
<th></th>
<th>Whole cohort (n=200)</th>
<th>Shoulder (n=138)</th>
<th>Wrist/hand</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall rating of problem</td>
<td>0.71 [0.69]</td>
<td>0.69 [0.68]</td>
<td>0.68 [0.74]</td>
</tr>
<tr>
<td>Pain Severity</td>
<td>0.72 [0.72]</td>
<td>0.73 [0.71]</td>
<td>0.67 [0.71]</td>
</tr>
<tr>
<td>Ability to function</td>
<td>0.79 [0.79]</td>
<td>0.84 [0.85]</td>
<td>0.75 [0.78]</td>
</tr>
<tr>
<td>Ability to work</td>
<td>0.76 [0.77]</td>
<td>0.76 [0.76]</td>
<td>0.69 [0.74]</td>
</tr>
<tr>
<td>SPADI Pain</td>
<td>0.82 [0.82]</td>
<td>0.79 [0.76]</td>
<td>0.84 [0.83]</td>
</tr>
<tr>
<td>SPADI Function</td>
<td>0.88 [0.87]</td>
<td>0.85 [0.83]</td>
<td>0.92 [0.92]</td>
</tr>
<tr>
<td>Brigham Symptoms</td>
<td>0.71 [0.70]</td>
<td>0.73 [0.71]</td>
<td>0.70 [0.73]</td>
</tr>
<tr>
<td>Brigham Function</td>
<td>0.89 [0.89]</td>
<td>0.90 [0.90]</td>
<td>0.92 [0.48]</td>
</tr>
</tbody>
</table>

**Test Retest reliability**

Fifty-six of the 86 people completing the test-retest reliability package (3-5 days after baseline) indicated they had no change in their problem (no change +/- 1 response category on transitional scale). The mean change score in this group was -0.15 (median = 0), with a standard deviation of 6.08. The difference was not significant (Paired t statistic = -0.176, p-value = 0.86). The correlation between baseline and retest was 0.96 (Pearson correlation) and 0.95 (Spearman ranked correlation). The intraclass correlation coefficient (ICC) (2,1) on this sample (n=56) was 0.96 (95% confidence intervals 0.93-0.98) (Shrout et al. 1979) for the DASH, indicating excellent agreement (McHorney et al. 1995; Nunnally et al. 1994; Lohr et al. 1996; Staquet et al. 1998). The standard error of measurement is 4.6 DASH points which led
to a minimal detectable change (MDC_{95}) was 12.75 (standard deviation at baseline = 23.02) on a 100 point scale. A 90% MDC (MDC_{90}) would be 10.7/100.

**Responsiveness**

The DASH questionnaire was able to demonstrate change in all situations where change was presumed to have occurred: before and after treatment (SRM = 0.74 - 0.80); and in those patients who either said that their problem was better overall, or that their ability to function had improved (SRM = 0.92 - 1.40)(see Table D.2). Standards for a “good” or “large” responsiveness statistics have little meaning because they are dependent on the type of change being examined. The distributions for the change scores are shown in Figure D.2. The large histogram shows the change for all patients before and after treatment (whether they got better or in some cases worse), as well as for the subgroup who said their problem was better, and those who said their function was better.
### Table D.2. Responsiveness of the DASH to clinical changes.

The following table shows the mean (standard deviation) for the baseline score, follow up, change score, and then the effect size (mean change divided by the standard deviation of baseline) and standardized response mean (mean change divided by the standard deviation of change). First for the entire cohort, then for the shoulder and wrist/hand patients separately. Data is presented only on those with both baseline and 12 week follow up data available (n=172).

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Follow up</th>
<th>Change</th>
<th>Effect size</th>
<th>Standardized Response Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All patients (n=172 completing baseline and follow-up)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observed change</td>
<td>44.47</td>
<td>30.89</td>
<td>-13.28</td>
<td>0.59</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>[22.68]</td>
<td>[22.75]</td>
<td>[16.93]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Those rating problem as</td>
<td>42.93</td>
<td>24.86</td>
<td>-17.27</td>
<td>0.75</td>
<td>1.06</td>
</tr>
<tr>
<td>better (&gt;6/10)</td>
<td>[22.94]</td>
<td>[20.19]</td>
<td>[16.35]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Those rating function as</td>
<td>40.65</td>
<td>20.19</td>
<td>-19.72</td>
<td>0.84</td>
<td>1.20</td>
</tr>
<tr>
<td>better (&gt;6/10)</td>
<td>[23.43]</td>
<td>[19.24]</td>
<td>[16.45]</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Shoulder patients</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observed change</td>
<td>48.32</td>
<td>35.32</td>
<td>-13.41</td>
<td>0.64</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>[20.99]</td>
<td>[21.28]</td>
<td>[16.63]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Those rating problem as</td>
<td>48.98</td>
<td>30.02</td>
<td>-17.71</td>
<td>0.85</td>
<td>1.13</td>
</tr>
<tr>
<td>better (&gt;6/10)</td>
<td>[20.86]</td>
<td>[19.46]</td>
<td>[15.71]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Those rating function as</td>
<td>49.13</td>
<td>24.25</td>
<td>-23.69</td>
<td>1.08</td>
<td>1.44</td>
</tr>
<tr>
<td>better (&gt;6/10)</td>
<td>[21.93]</td>
<td>[17.55]</td>
<td>[16.44]</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Wrist/Hand patients</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observed change</td>
<td>33.80</td>
<td>20.45</td>
<td>-12.99</td>
<td>0.57</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>[22.77]</td>
<td>[22.90]</td>
<td>[17.53]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Those rating problem as</td>
<td>30.52</td>
<td>13.82</td>
<td>-16.32</td>
<td>0.73</td>
<td>0.92</td>
</tr>
<tr>
<td>better (&gt;6/10)</td>
<td>[22.24]</td>
<td>[17.24]</td>
<td>[17.83]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Those rating function as</td>
<td>27.38</td>
<td>13.51</td>
<td>-13.23</td>
<td>0.68</td>
<td>0.91</td>
</tr>
<tr>
<td>better (&gt;6/10)</td>
<td>[19.40]</td>
<td>[20.31]</td>
<td>[14.49]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The DASH was found to have comparable or slightly better responsiveness than the joint-specific measures (see Table D.3 and Figure D.3 which summarize these comparisons using the SRM statistic). The DASH remained comparable (only lower in the case of wrist/hand patients by 0.04-0.08) or better than the disease-specific measures when the
comparisons were made at the subgroup level (by injured region).

Table D.3. Comparing the responsiveness between DASH and two joint-specific measures (Shoulder Pain and Disability Index (function score) and the Brigham (function score)). Standardized response means (mean change score divided by the standard deviation of the difference) is used as the summary statistic.

<table>
<thead>
<tr>
<th></th>
<th>DASH (same as Table D.2)</th>
<th>SPADI function score</th>
<th>Brigham functional limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All patients</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observed change</td>
<td>0.78</td>
<td>0.62</td>
<td>0.64</td>
</tr>
<tr>
<td>Those rating their problem as better (&gt;6/10)</td>
<td>1.06</td>
<td>0.84</td>
<td>0.86</td>
</tr>
<tr>
<td>Those rating their function as better (&gt;6/10)</td>
<td>1.20</td>
<td>0.86</td>
<td>1.07</td>
</tr>
<tr>
<td><strong>Shoulder patients</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observed change</td>
<td>0.81</td>
<td>0.71</td>
<td>0.59</td>
</tr>
<tr>
<td>Those rating their problem as better (&gt;6/10)</td>
<td>1.13</td>
<td>0.95</td>
<td>0.83</td>
</tr>
<tr>
<td>Those rating their function as better (&gt;6/10)</td>
<td>1.44</td>
<td>1.13</td>
<td>1.24</td>
</tr>
<tr>
<td><strong>Wrist/Hand patients</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observed change</td>
<td>0.74</td>
<td>0.43</td>
<td>0.76</td>
</tr>
<tr>
<td>Those rating their problem as better (&gt;6/10)</td>
<td>0.92</td>
<td>0.64</td>
<td>0.93</td>
</tr>
<tr>
<td>Those rating their function as better (&gt;6/10)</td>
<td>0.91</td>
<td>0.54</td>
<td>0.87</td>
</tr>
</tbody>
</table>

Correlations between differences in the status measures (self ratings at follow up minus self-ratings at baseline) and the change scores on the DASH were moderately high (Pearson r >.65) (see Table D.4). Those with the transitional indices (how are you now compared to
before) failed to meet our modest standard of 0.4 (correlations were in the range of 0.32 - 0.39) except in the overall rating of change in their problem (Pearson r = 0.40, Spearman r = 0.43). The correlations between the DASH scores and the differences in state measures were of the hypothesized magnitude, however those between the transitional indices and the change in DASH scores were at or just below the expected level of 0.40.

**Table D.4.** Correlation between change in DASH score and change in self-ratings of pain, function and severity of upper-limb problem measured in two ways: first as a transition question (11 point scale), second as the difference in state ratings at baseline and 12 weeks (each rated on both testing occasions using a 7 point rating scale). Pearson (parametric) and Spearman (ranked) correlations coefficients are presented.

<table>
<thead>
<tr>
<th>Correlation with change in DASH scores</th>
<th>Change in problem</th>
<th>Change in pain</th>
<th>Change in function</th>
<th>Change in problem status</th>
<th>Change in pain status</th>
<th>Change in function status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson</td>
<td>0.40</td>
<td>0.38</td>
<td>0.32</td>
<td>0.66</td>
<td>0.65</td>
<td>0.69</td>
</tr>
<tr>
<td>Spearman</td>
<td>0.43</td>
<td>0.37</td>
<td>0.39</td>
<td>0.60</td>
<td>0.62</td>
<td>0.63</td>
</tr>
</tbody>
</table>

The receiver operator characteristic curves, as shown in Table D.5 and Figure D.4, also demonstrated that the change scores in the DASH were more sensitive to ability to cope than...
just chance alone. In ROC curves 'chance' is equivalent to
the diagonal on the graph, where
sensitivity equals one minus the
specificity. Along this line the
change score is giving no more
information about who is better
and who is not than chance
along. The point furthest away
from the diagonal is often
considered the most discriminating. In our graphs, a change of at least either -15 or -20 appears
to be the most discriminative change scores for the criterion of becoming able to cope, and
correctly rated 68 and 73% of the sample respectively. Using the self-rating of the problem
being better (>6/10), as used in the statistical summary of responsiveness, much lower change
scores were found to be most discriminating (change of negative one or more had an accuracy
of 75%, sensitivity 0.87, specificity 0.44).

Figure D.4 Receiver Operator Characteristic (ROC) Curve
reflecting the ability of different change score on the DASH to
predict a persons transition from being unable to cope, to being
able to cope with their disorder. See footnote in text for
explanation of selection of these change scores.
Table D.5. Sensitivity and specificity of different levels of change when compared to the 'criterion' of saying "yes" to the question "are you able to cope with your problem and do what you would like to do?" at follow-up having said "no" to the same question before treatment. A change score of 15 or 20 was best able to discriminate between those who were better or not according to our selected criterion.

<table>
<thead>
<tr>
<th>DASH Change of ... 'x'</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Probability of having change in DASH of 'x' or lower given you are better.</td>
<td>Probability of not having change in DASH of 'x' or lower given you are not better.</td>
<td>% where change score of 'x' or lower is correctly classifying person as better or not</td>
</tr>
<tr>
<td>&lt; -1</td>
<td>0.92</td>
<td>0.29</td>
<td>48</td>
</tr>
<tr>
<td>&lt; -5</td>
<td>0.86</td>
<td>0.38</td>
<td>52</td>
</tr>
<tr>
<td>&lt; -7</td>
<td>0.82</td>
<td>0.42</td>
<td>54</td>
</tr>
<tr>
<td>&lt; -10</td>
<td>0.82</td>
<td>0.53</td>
<td>62</td>
</tr>
<tr>
<td>&lt; -15</td>
<td>0.67</td>
<td>0.68</td>
<td>68</td>
</tr>
<tr>
<td>&lt; -20</td>
<td>0.59</td>
<td>0.78</td>
<td>72</td>
</tr>
</tbody>
</table>

1. Reminder: A negative change score on the DASH means less disability, and therefore an improvement. Have a change of < -1 could mean just a little bit of improvement. A change of < -20 would mean a much larger amount of improvement according to the DASH.

Discussion

This study has provided evidence of the construct validity, test-retest reliability and responsiveness of the DASH Outcome Measure in patients undergoing treatment for either proximal or distal disorders in the upper limb.

Reliability, validity and responsiveness of the DASH

The DASH outcome measure exceeded recommended standards for test-retest reliability (McHorney et al. 1995; Nunnally et al. 1994; Lohr et al. 1996; Staquet et al. 1998; Cole et al. 1994) for both individual and group level interpretation of the scores (McHorney et al. 1995). Generally test-retest coefficients need to exceed 0.90 or 0.95 before one should consider interpreting the scores on an individual level. For group level interpretation, lower coefficients
are acceptable (approximately 0.75) (Nunnally et al. 1994; Shrout et al. 1979), our results found a coefficient of 0.96 [0.93-0.98], slightly higher than Turchin (Turchin et al. 1998) had previously shown (0.92 in stable elbow patients).

Evidence of both convergent and discriminative validity of the DASH was also found. Convergent validity was shown by demonstrating moderate to high correlations with other markers of disability and symptoms, and discriminative by demonstrating differences in DASH scores between those working or functioning and those not. The DASH validity was comparable to previous results using this questionnaire in other populations. Hudak (McConnell et al. 1999), Turchin (Turchin et al. 1998) and the results of the field testing (McConnell et al. 1999) produced similar findings, though against a smaller number of constructs.

The DASH was also responsive to the different types of change designed into this study: specifically change observed before and after treatment of the target conditions, and change in those patients who said that they were better. In addition, our results demonstrated two important issues around responsiveness. First, the size of the responsiveness statistic varied with the type of change that was being quantified. SRM's for observed change were lower in magnitude than for the change in those who said that they were better. This finding supports the taxonomy for responsiveness we have presented elsewhere (Beaton et al. 1999b) that suggests instruments are "responsive to" different types of change to different degrees. They do not inherently possess a trait of being "responsive". Comparisons of the responsiveness of different instruments should only be conducted when similar types of change are being tested, or ideally when the instruments are placed in a head to head comparison. Second, we also demonstrated, as did Wright (Wright et al. 1998), that the statistic chosen will affect the description of
responsiveness. We demonstrated up to 1.33 fold variation in the responsiveness described by the effect size statistic versus the standardized response means for the same change in the same patients (0.68 versus 0.92). This difference would also span the often used (or misused) guideline for what Cohen (Cohen, 1988) called a moderate versus a large effects. However, the difference in the numeric estimates we are comparing was attributable to the statistic chosen alone.

Responsiveness was also described by correlating changes (Kirkley et al. 1998; Deyo et al. 1986) in the DASH with change in three attributes (pain, function and problem) each measured in two ways: transitional scales, and differences in status measures done at baseline and 12 weeks. Our results suggested a distinct difference between the transitional approach (correlations of 0.32-0.43) and the difference in status approach (correlations of 0.60-0.69). Given that the concepts (pain, function) being measured using the two approaches were the same, the differences might be attributed to the way the questions were asked (transition versus difference in status) leading to several possible explanations. Differences in a transitional versus differences in status over time could be due to recall bias (Herrmann, 1995; Mancuso et al. 1995), or due to a change in how people cognitively formulate a response when asked to describe a current state versus recalling a change in that state over time (Sprangers et al. 1989; Gibbons, 1999; Ross, 1989; Beaton et al. 1999a; Allison et al. 1997). Changes in how people calibrate and/or define pain, health and quality of life over time have been described in the literature (‘response shift phenomenon’ for example (Sprangers et al. 1999; Schwartz et al. 1999a; Armenakis, 1988)) and could have influenced our results (Sprangers et al. 1999; Schwartz et al. 1999a; Armenakis, 1988; Kind et al. 1999; Shaul, 1995; Reid et al. 1991). All these things could be possible reasons for the difference in the correlations between the change
in DASH score and the two approaches to determining if change had occurred (difference in serial state measures versus the transition style of constructs). We are not suggesting which is better. Arguments can be made in both directions (Armenakis et al. 1979; Armenakis, 1988; Herrmann, 1995; Howard et al. 1979a; Redelmeier et al. 1996d). Like Fischer (Fischer et al. 1999) we have demonstrated the difference in the two approaches, and like them we do not have evidence to suggest which is better.

\textit{DASH versus joint-specific measures.}

The DASH had high correlations with the two joint-specific measures, a pattern which persisted when the joint-specific measure was applied in the other region (Brigham vs. DASH in shoulder patients = 0.90; SPADI vs. DASH in wrist patients = 0.92). Therefore our hypothesis that there would be a difference in the correlations when the joint-specific measures were applied in the wrong joint was not supported. On a cross sectional basis, the joint-specific measures performed well in the ‘wrong’ joint. To our knowledge, this is the first time joint-specific measures in the upper limb have been purposively applied to the ‘wrong’ joint. Our results support our earlier findings that items from joint-specific questionnaires could not be identified as relevant to only one joint in the extremity by a panel of experienced clinicians. In that study, experts classified over 70% of the items they were given (i.e., using a phone) as being potentially affected by impairments across the upper extremity and not just to problems in one joint (Davis et al. 1999). Using a phone might be a relevant item for persons with shoulder, elbow or grasp problems. The results of our current study show similar results, but at the whole questionnaire level. The items in the Brigham appear to have been sensitive to impairments in the shoulder, and the reverse true with the SPADI in wrist patients.
The DASH, with two exceptions, also had comparable or better responsiveness to the joint-specific measures. The exceptions was in the wrist/hand group where the SRM’s for the Brigham were slightly higher (0.02 and 0.01 greater than for the DASH) for two of the three comparisons (observed and change in those who rated their problem as better respectively). Many would consider these comparable values. They would not be statistically significantly different. Nevertheless, in 16/18 comparisons between the DASH and the joint-specific measures, the DASH had larger statistics. The Brigham questionnaire was also responsive to changes in shoulder patients (though lower in magnitude than the DASH). The SPADI showed only moderate SRM’s in the wrist/hand patients (0.43 - 0.64), while both the DASH and the Brigham were larger (SRM ≥0.74). Both the Brigham and the SPADI appeared responsive outside of their region of specialty.

These results suggest that the responsiveness of the DASH was equivalent to both these joint-specific measures, and better than either in the whole cohort, affirming its utility for patients with any or multiple disorders in the upper limb. These findings are slightly different than those of Kirkley (Kirkley et al. 1998) or MacDermid (MacDermid et al. 1999a) who found the DASH slightly less responsive than a disease specific measure (though a different disease-specific measure was used in each case).

The implications of our findings are twofold. First, the DASH has potential in the role of monitoring physical function and symptoms in shoulder and wrist/hand disorders as demonstrated by its’ validity, high test-retest reliability, and responsiveness to even small changes early in the recovery process. This provides a practical solution to the problems of having to select multiple measures in patients with multiple impairments, or just of having to have multiple upper extremity measures available in a clinic or research setting. 
Second, the high levels of reliability (Cronbach's alpha was 0.97 at baseline in this study) indicate that a much shorter measure could have acceptable reliability (Nunnally et al. 1994) with less respondent burden, making it useful for epidemiological surveys or detailed outcomes studies. The longer version will likely always retain a statistical psychometric advantage, however the shorter might provide a more useful version for studies that entail a significant respondent burden (epidemiological surveys, detailed outcome studies). Recent work however has reinforced the need to maintain very high levels of reliability in order to improve precision in analysis of change, and in order to be able to use the instrument for monitoring individual clients (McHorney, 1997; McHorney, 1999). Reliability will be lost with fewer items, and the attraction of the shorter instrument may guide people to use it in individual level interpretation rather than only in large studies.

This study has demonstrated that the DASH Outcome Measure shows evidence of good construct validity, test-retest reliability and responsiveness to change (specifically change before and after four groups of treatments as well as change that was estimated to have occurred by the patients). It has done this for both proximal and distal disorders suggesting its role as a measure of physical function and symptoms in any or multiple disorders of the upper limb. Further work will focus on other patient groups, and on the interpretation of specific scores at the individual level.

Acknowledgments:
The authors would like to acknowledge Ms. Elaine Hamiman in Toronto and Ms. Joanne Nicklas in Boston who implemented the study protocol. Thanks go to Dr Robin Richards, Dr
James Mahoney and Dr Mike McKee for allowing us to access their patients, often through their clinics; to Dr Louis Bessette, who contributed to the conceptualization and design of this project; and to the patients who participated in the project. This work was supported by a research grant from the American Society of Surgery of the Hand, and the Institute for Work & Health. Ms Beaton is supported by a PhD Fellowship (Health Research) from the Medical Research Council of Canada. Dr Wright is the R.B.Salter Chair in Surgical Research and supported as a Scientist of the Medical Research Council of Canada. Dr Katz is supported through a grant from the NIH #AR36308 and the US National Arthritis Foundation.
Appendix E:

INDICES USED AS EXTERNAL MARKERS FOR CHANGE IN THE COHORT STUDY.

Testa's model for responsiveness (Testa et al. 1996) clearly indicates the need for some sort of standard that indicates that change has occurred, the "ΔC" in the model. This is consistent with our definition of responsiveness: "the ability to accurately detect change when it has occurred" (De Bruin et al. 1997). This appendix will summarize the markers for change used in the cohort study (results described in Chapters five and Appendix D), and the distribution of responses on each of them found at the four week and 12-week follow-up assessment. As described in Chapter five, the four week follow-up was conducted in the Toronto portion of the cohort only and thus is a subset of those continuing on to 12 weeks (n=93 of the 172 at 12 follow-up).

Far from being "gold standards" or criterion measures of change, these indices reflect constructs -- that is, they set up theoretical situations where change has likely occurred. We have thus called them external markers of change rather than criterion markers.

Two different conceptual approaches were used to estimate the occurrence of change in the study participants. First, serial measures of an attribute (i.e., pain) were obtained by asking for a rating of current status at each testing time. Four or twelve week and baseline ratings were then subtracted and the difference considered to be the indicator of change. The second approach used a retrospective evaluation of change in the attribute -- asking them to give a rating of the change they experienced between current (four or twelve weeks) and baseline states. We anchored all of the recall back to the baseline assessment (pre-treatment) primarily because it was decided this would be the baseline for any assessment of responsiveness, but
also because it may have provided a clear marker for their recall (just before treatment began).

Four different attributes were measured in these two manners: rating of overall problem, intensity of pain, ability to function in daily activities, and ability to work. Others were also included (see Appendix C for a sample follow-up questionnaire package which included these indices).

Others have compared these two different conceptual approaches to estimating the occurrence of change. In particular, Fischer et al. (Fischer et al. 1999) recently published a comparison of these two approaches showing the differences in information gained from each. They also found that the highest disagreement between approaches was found in those patients with small or modest change in disease markers (Fischer et al. 1999). There is no clear consensus as to which approach is better (more valid). The literature has evidence supporting the advantage of either perspective, some supporting the use of recalled health states such as would be used to construct the transitional approach (Schwartz et al. 1999b; Schwartz et al. 1999a; Sprangers et al. 1999). Others raise concerns, favoring serial measures of current health state. For example, many cite examples of poor recall of past health or symptoms (Herrmann, 1995; Linton et al. 1982; Redelmeier et al. 1996d), and Ross (Ross, 1989) describes how past health states are likely constructed cognitively using implicit theories (this is how I am now, and I am better, so this is what I must have been like) (Ross, 1989). Still arguments are found in favor of either. As others have concluded (Fischer et al. 1999), and as we demonstrated in Chapter five, the two approaches gather different kinds of information. In our work the changes in the DASH score had higher correlations (Spearman) with the serially measured health states \( r_x = 0.48\) to \( 0.63 \) than the transitional scales \( r_t = 0.37\) to \( 0.43 \).

In addition to these four attributes measured in two ways, we also include the
description of the response to two threshold type of questions (can you cope? can you do what you want to do?). These were measured at each time, much like the serial state descriptors. Shifting from a no to a yes over time on these indices was considered indicative of an improvement.

The following pages will show the exact items used in the cohort study (and subsequently in the analysis of responsiveness in Chapter five) along with the percent of the four week and twelve week follow-up sample that fell into each response category. The transitional items appeared after the serial status items in the four and twelve week packages.

A. Changes from patients’ perspective using differences in serial measures.

Rating of problem.

| How would you rate your shoulder problem over the last week? (Check one) |
|------------------------|--------|--------|--------|--------|--------|--------|--------|
| □                      | □      | □      | □      | □      | □      | □      | □      |
| 1                      | 2      | 3      | 4      | 5      | 6      | 7      |
| Very mild              | Neither severe nor mild | Very severe |

Distribution of difference scores on the rating of overall problem (4 & 12 week minus baseline). Negative scores = improvement (less severity).

<table>
<thead>
<tr>
<th></th>
<th>Improved (%)</th>
<th>No change</th>
<th>Worse</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-6</td>
<td>-5</td>
<td>-4</td>
</tr>
<tr>
<td>4 weeks</td>
<td>1.2</td>
<td>3.5</td>
<td>3.5</td>
</tr>
<tr>
<td>12 weeks</td>
<td>1.2</td>
<td>4.8</td>
<td>9.0</td>
</tr>
</tbody>
</table>
Intensity of pain.

How would you rate the intensity or severity of your shoulder pain or discomfort over the last week? (Check one box)

<p>| | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

No pain

Pain as bad as it can be

Distribution of difference scores on the pain intensity ratings (4 & 12 week minus baseline). Negative scores = improvement (less pain).

<table>
<thead>
<tr>
<th>Improved (%)</th>
<th>No change</th>
<th>Worse</th>
</tr>
</thead>
<tbody>
<tr>
<td>-10 -9 -8 -7 -6 -5 -4 -3 -2</td>
<td>-1 0 1</td>
<td>2 3 4 5</td>
</tr>
</tbody>
</table>

4 wk

<table>
<thead>
<tr>
<th>-10</th>
<th>-9</th>
<th>-8</th>
<th>-7</th>
<th>-6</th>
<th>-5</th>
<th>-4</th>
<th>-3</th>
<th>-2</th>
<th>12.8</th>
<th>17.4</th>
<th>14.3</th>
<th>9.3</th>
<th>1.2</th>
<th>2.3</th>
<th>1.2</th>
</tr>
</thead>
</table>

12wk

<table>
<thead>
<tr>
<th>-10</th>
<th>-9</th>
<th>-8</th>
<th>-7</th>
<th>-6</th>
<th>-5</th>
<th>-4</th>
<th>-3</th>
<th>-2</th>
<th>12.8</th>
<th>17.4</th>
<th>14.3</th>
<th>9.3</th>
<th>1.2</th>
<th>2.3</th>
<th>1.2</th>
</tr>
</thead>
</table>

Ability to function.

The items given as examples in this question were those that received the highest ‘importance + difficulty’ rating by patients in a study that was conducted to select items to be included in the DASH outcome measure (Marx et al. 1999) (McConnell et al. 1999).

How would you rate your ability to do you usual daily activities (such as opening jars, writing, doing household chores) over the last week? (Check one)

<p>| | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Not Very difficult

Difficult
Distribution of difference scores on the ability to do daily activities (four or 12 week minus baseline). Negative scores = improvement (less difficulty).

<table>
<thead>
<tr>
<th>Improved (%)</th>
<th>No change</th>
<th>Worse</th>
</tr>
</thead>
<tbody>
<tr>
<td>-6 -5 -4 -3 -2</td>
<td>-1 0 1</td>
<td>2 3 4 5 6</td>
</tr>
<tr>
<td>4 weeks</td>
<td>0 1.2 3.5 7.1 5.9</td>
<td>12.9 30.6 10.6</td>
</tr>
<tr>
<td>12 weeks</td>
<td>1.8 2.4 4.9 11.6 12.8</td>
<td>22.6 27.4 8.5</td>
</tr>
</tbody>
</table>

**Ability to work**

How would you rate your ability to do your paid work over the last week? (Check one)

- ☐ ☐ ☐ ☐ ☐ ☐ ☐
- 1 2 3 4 5 6 7
- Not Difficult
- Very difficult

Distribution of difference scores on the ability to work ratings (four or 12 week minus baseline). Negative scores = improvement (less difficulty).

<table>
<thead>
<tr>
<th>Improved (%)</th>
<th>No change</th>
<th>Worse</th>
</tr>
</thead>
<tbody>
<tr>
<td>-6 -5 -4 -3 -2</td>
<td>-1 0 1</td>
<td>2 3 4 5 6</td>
</tr>
<tr>
<td>4 wk</td>
<td>0 0 0 1.8 8.8</td>
<td>3.5 36.8 15.8</td>
</tr>
<tr>
<td>12 wk</td>
<td>.9 0 8.4 5.6 6.5</td>
<td>21.5 39.3 9.3</td>
</tr>
</tbody>
</table>

**Ability to do what you want to do and forget about the injured part.**

In the qualitative study described in Chapter three, threshold concepts of change were also described. For instance being better meant that you could finally do things without thinking about the injured extremity. To capture these concepts we included dichotomized items worded
based on the concepts from qualitative work.

In these dichotomized items, we considered people to be 'improved' if they made the transition from a problematic response (no, I can not cope) to an improved one (I can cope).

Can you do all the activities you would like to do? (Check one)
- Yes, I can forget about my problem and do what I want to do
- No. If no, are you having to think about your problem for
  - only a few activities (e.g. only specific tasks)
  - about half your activities
  - all activities

Distribution of responses for this item.

<table>
<thead>
<tr>
<th>4-wk follow-up</th>
<th>Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes, I can do all</td>
<td>12</td>
</tr>
<tr>
<td>No, I can’t do all</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>12-wk follow-up</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes, I can do all</td>
</tr>
<tr>
<td>No, I can’t do all</td>
</tr>
</tbody>
</table>

* those considered improved by our definition

Ability to cope with problem.

Is your shoulder problem at a level now where you can ignore or cope with it and do whatever it is you have to do in your daily life? (Check one)
- Yes, I can cope with it at the level it is at now.
- No, it is not at a level where I can cope with it.
Distribution of responses for this item.

<table>
<thead>
<tr>
<th>4-wk follow-up</th>
<th>Baseline</th>
<th>12-wk follow-up</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes, I can cope</td>
<td>15</td>
<td>60</td>
</tr>
<tr>
<td>No, I can’t cope</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>Yes, I can cope</td>
<td>17*</td>
<td></td>
</tr>
<tr>
<td>No, I can’t cope</td>
<td>44</td>
<td></td>
</tr>
</tbody>
</table>

* those considered improved by our definition

B. Transitional scales representing a retrospective recall of change in health, measured at 4 and 12 weeks.

Change in overall rating of problem.

Think about your shoulder problem now compared to when you started your treatment (or before your surgery). How would you rate the change in your shoulder problem? (Check one)

<table>
<thead>
<tr>
<th>□</th>
<th>□</th>
<th>□</th>
<th>□</th>
<th>□</th>
<th>□</th>
<th>□</th>
<th>□</th>
<th>□</th>
<th>□</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>Much worse</td>
<td>No change</td>
<td>A lot better</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Responses to this scale were as follows (shown for 4 and 12 week follow ups).

<table>
<thead>
<tr>
<th>Worse (%)</th>
<th>No change</th>
<th>Improved</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 1 2 3</td>
<td>4 5 6</td>
<td>7 8 9 10</td>
</tr>
<tr>
<td>4 wk</td>
<td>8 0 3.4 2.3</td>
<td>10.3 18.4 11.5 23</td>
</tr>
<tr>
<td>12 wk</td>
<td>.6 2.4 1.2 0</td>
<td>6.1 17.1 12.8 37.8</td>
</tr>
</tbody>
</table>
## Change in severity of pain.

Again, compared to before your treatment/surgery, how would you rate the change in the intensity or severity of your shoulder pain or symptoms? (Check one)

<p>| | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>Much worse</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No change</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A lot better</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Responses to this scale were as follows (shown for four and 12 week follow ups).

<table>
<thead>
<tr>
<th>Worst (%)</th>
<th>No change</th>
<th>Improved</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>4 wk</td>
<td>2.2</td>
<td>7.9</td>
</tr>
<tr>
<td>12 wk</td>
<td>3.7</td>
<td>7.3</td>
</tr>
</tbody>
</table>

## Change in ability to function.

Compared to before your treatment/surgery, how would you rate the change in your ability to do the things you want or need to do (such as opening jars, writing, preparing meals, doing heavy chores, carrying bags, lifting children)? (check one)

<p>| | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>Much worse</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No change</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A lot better</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Responses to this scale were as follows (shown for four and 12 week follow ups).

<table>
<thead>
<tr>
<th>Worst (%)</th>
<th>No change</th>
<th>Improved</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>4 wk</td>
<td>0</td>
<td>18.3</td>
</tr>
<tr>
<td>12 wk</td>
<td>2.5</td>
<td>21.5</td>
</tr>
</tbody>
</table>
Change in ability to work.

Compared to before your treatment/surgery, how would you rate the change in your ability to do your usual work (including homemaking if that is your primary role)? (Check one)

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Much worse</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>No change</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A lot better</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Responses to this scale in percent of sample were as follows (shown for 4 and 12 week follow ups).

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Worse</td>
<td>11.0</td>
<td>6.1</td>
<td>9.8</td>
<td>3.7</td>
<td>4.9</td>
<td>18.3</td>
<td>9.8</td>
<td>7.3</td>
<td>8.5</td>
<td>7.3</td>
<td>13.4</td>
</tr>
<tr>
<td>No change</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Improved</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Importance of that change.

Study subjects were then asked to describe the importance of the change experienced to them.

They were asked in the following manner:

How important was this experience of change for you in your life?

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not very important</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Fairly important</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very important</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extremely important</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This type of question was therefore asked for each of the transitional questions above. The desire was to obtain information about the patients rating of the importance of a change. A composite variable was created profiling the importance ratings for each individual across the
different transitional questions. We did this by multiplying the first importance rating by 10,000 then adding to it the second importance rating multiplied by 1,000, the third by 100 plus the fourth times 10 and finally the fifth importance rating was added. This provides a variable that visually displays the pattern of response (11114 = no important change in the first four, and extremely important rating in the last one). What we were looking for was whether the individuals discriminated across the attributes being measured in terms of their importance ratings. The following table summarizes the results for the five transitional scores (one, frequency of pain was not presented above). The composite represents the importance score given for: overall problem, severity of pain, frequency of pain, ability to function, and ability to work — in that order.

As can be seen on the table (Table E.1, last page of this appendix), the majority of people rated their change as important - with 43% of the sample responding "extremely important" to all five questions. There was a tendency to respond in the same manner across questions at other response levels: 1 = 2%, 2 = 7.4% and 3 = 21.6%. The highest "mixed" rating of importance was 33332, where everything but work was rated as important but only 3 people selected that pattern.

The fact that a total of 74% of our sample gave identical "importance" ratings across five different attributes could reflect one of two scenarios. First, we could have misled people in terms of

![importance of change across self-ratings of change in functional ability](image)

**Figure E.1.** Distribution of importance ratings (divisions within bars) across responses to transitional question on pain (asked at 12 weeks). Note that change is rated as extremely important across the range of transitional change ratings.
understanding and thinking about the concept of importance in the questionnaire wording and format. It is unlikely that such of large percent would have this difficulty. The second possibility is that this is a very accurate reflection of importance, and that the value of importance of change goes across the magnitude of change in specific attributes. In some way it might be an appraisal of their whole experience and not the parts of it. This is also supported by the fact that the correlation between the magnitude of a change, and the importance was low ($r = 0.3 - 0.45$) (see figure E.1 as well). Further research into this concept should be undertaken.

Summary

This appendix has shown the specific response patterns to the external markers of change used in our cohort. There were differences in the occurrence of change described by the different conceptual approaches, and by the different attributes. This reinforces that these are markers, not criterion, for change.

Differences were also found between the two conceptual approaches used to find the occurrence of change: difference in serial scores, and the transitional indices. Fischer et al. provided evidence of this in a recent publication (Fischer et al. 1999), and our evidence would support the differences. Unlike Fischer, we were able to correlate the two approaches with change in the DASH scores, and we found higher correlations with the markers that corresponded to the style of measurement for the DASH, the serial state measures. Given concerns that have been raised regarding recall bias (Herrmann, 1995; Linton et al. 1982; Ross, 1989), it could be that this is a better indicator of change if the within-person change in serial measures of health is the targeted comparison for the index. In that situation (which is typical for assessments of responsiveness), the index of change and the health status measures
are being gathered in the same manner, perhaps reducing a source of measurement error.

**Table E.1** Response patterns to five questions of importance (at 12 weeks only). Shaded boxes indicate where subjects gave the same rating to each of the five attributes. (Missing: n = 24)

<table>
<thead>
<tr>
<th>Composite variable</th>
<th># subjects with this response</th>
<th>% of subjects with this response</th>
<th>Composite variable</th>
<th># subjects with this response</th>
<th>% of subjects with this response</th>
</tr>
</thead>
<tbody>
<tr>
<td>11111</td>
<td>3</td>
<td>2</td>
<td>33333</td>
<td>32</td>
<td>21.6</td>
</tr>
<tr>
<td>11211</td>
<td>1</td>
<td>.7</td>
<td>33334</td>
<td>1</td>
<td>.7</td>
</tr>
<tr>
<td>12222</td>
<td>1</td>
<td>.7</td>
<td>33433</td>
<td>3</td>
<td>2.0</td>
</tr>
<tr>
<td>21112</td>
<td>1</td>
<td>.7</td>
<td>33444</td>
<td>2</td>
<td>1.4</td>
</tr>
<tr>
<td>22211</td>
<td>2</td>
<td>1.4</td>
<td>34413</td>
<td>1</td>
<td>.7</td>
</tr>
<tr>
<td>22222</td>
<td>11</td>
<td>7.4</td>
<td>34433</td>
<td>1</td>
<td>.7</td>
</tr>
<tr>
<td>23333</td>
<td>1</td>
<td>.7</td>
<td>34434</td>
<td>1</td>
<td>.7</td>
</tr>
<tr>
<td>23222</td>
<td>1</td>
<td>.7</td>
<td>34443</td>
<td>1</td>
<td>.7</td>
</tr>
<tr>
<td>23322</td>
<td>1</td>
<td>.7</td>
<td>34444</td>
<td>2</td>
<td>1.4</td>
</tr>
<tr>
<td>23333</td>
<td>2</td>
<td>1.4</td>
<td>43333</td>
<td>1</td>
<td>.7</td>
</tr>
<tr>
<td>31111</td>
<td>1</td>
<td>.7</td>
<td>43444</td>
<td>1</td>
<td>.7</td>
</tr>
<tr>
<td>33222</td>
<td>2</td>
<td>1.4</td>
<td>44343</td>
<td>1</td>
<td>.7</td>
</tr>
<tr>
<td>33223</td>
<td>1</td>
<td>.7</td>
<td>44411</td>
<td>1</td>
<td>.7</td>
</tr>
<tr>
<td>33323</td>
<td>1</td>
<td>.7</td>
<td>44413</td>
<td>1</td>
<td>.7</td>
</tr>
<tr>
<td>33331</td>
<td>1</td>
<td>.7</td>
<td>44434</td>
<td>2</td>
<td>1.4</td>
</tr>
<tr>
<td>33332</td>
<td>3</td>
<td>2.0</td>
<td>44444</td>
<td>64</td>
<td>43.2</td>
</tr>
</tbody>
</table>
Appendix F:

APPLICATION OF THE TAXONOMY TO OTHER MEASURES FIELDED IN THIS STUDY.

The participants in the cohort study completed a number of different questionnaires at each data collection point. The focus of Chapter five was to present the results for one measure, the DASH outcome measure, across cells in the taxonomy of responsiveness. This focused the attention on the differences related to altering the categories in the taxonomy of responsiveness (different kinds of responsiveness), rather than the differences between instruments. The DASH was chosen for that task because it was shown to be applicable across the extremity (Appendix D), and as a specific measure (region-specific) it would be more sensitive to change in health than the broader generic measures also included (Patrick et al. 1989; Beaton et al. 1996a). This appendix will summarize the results found for the other instruments. The same categories of responsiveness will be tested as described in Chapter five and the results will be summarized across instruments using the standardized response mean statistic (mean change divided by the standard deviation of the change).

Other measures used in the cohort study.

The following paragraphs summarize the other measures fielded in the cohort.

SPADI: Shoulder pain and disability index. Roach et al introduced this measure in 1991 (Roach et al. 1991), using visual analog scales (11 cm) to rate various attributes of pain and disability. The SPADI has since been shown to be a valid, reliable, and responsive measure of
shoulder disability (Beaton et al. 1996b; Beaton et al. 1998; Heald et al. 1997; Roach et al. 1991; William et al. 1995). The SPADI is scored as two separate scales: pain (5 items) and disability (8 items), and then an overall score is also obtained by averaging the two dimensions scores (Roach et al. 1991). In this thesis we have focused on the function/disability subscale of the SPADI, not the overall score. With permission of the developers (personal communication) we shortened the lines to 10 cm visual analog scales for this study. Heald (Heald, 1997) also demonstrated minimal variation in the SPADI scores when comparing a visual analog scale version and a numeric rating scale.

**Brigham Carpal Tunnel Questionnaire:** In 1993 Levine et al. introduced the Brigham Carpal Tunnel Questionnaire (Levine et al. 1993). It is a widely used scale designed to measure symptoms and function on two scales (Amadio et al. 1996a; Amadio et al. 1996b; Katz et al. 1995; Levine et al. 1993; Katz et al. 1994). In this study it was considered a joint-specific measure for the subjects with wrist/hand problems. Of these subjects (n=62), many were in fact carpal tunnel (n=27) the remaining being some form of wrist or hand tendinitis. Others have also used the Brigham outside of the specific diagnosis of carpal tunnel syndrome (Amadio et al. 1996b).

**Short Form - 36 (SF-36) Acute:** The SF-36 Acute version was used as a measure of overall general health. Previous work has suggested that the SF-36, when examined as a profile of the subscale scores, shows some evidence of reliability and validity (general health) within patients with musculoskeletal disorders (Beaton et al. 1996a; Beaton et al. 1997; Rozencwaig et al. 1998).
1998; Gartsman et al. 1998a). However, in the past it has also been shown that the SF-36 is not sensitive to changes specifically within upper limb disorders (Beaton et al. 1996b; Beaton et al. 1997; Beaton et al. 1998; MacDermid et al. 1999a; Gartsman et al. 1998a; Gartsman et al. 1998b). Low responsiveness is appropriate in some dimensions where change is not expected with treatment of an isolated musculoskeletal disorder (such as mental health, or social functioning), however there were very small changes in some dimensions expected to change (physical function). The content of the physical functioning dimension of the SF-36 leaves little room for a person with an isolated upper-limb problem to change. Only four items are likely responsive to changes in upper limb function, the remaining are geared towards mobility. The implication of this is that it may under represent upper limb function if the SF-36 is used to make comparisons of treatment effectiveness across disease groups (for instance total shoulder vs. total hip replacement). In this study we will continue comparing the SF-36 Acute with measures of upper limb function to gain further insight into their relative performance.

In this study the results will also be presented for the physical component scores for the SF-36 (Ware, Jr. et al. 1994). These scores were developed based on factor analysis and standardization against American general population data. Two summary scores, a mental and a physical component score accounted for more than 80% of the variance in the sample (Ware, Jr. et al. 1994). The physical component score is derived by the factor loadings of physical function, role-physical, bodily pain, general health perceptions, vitality, and social functioning on the factor (Ware, Jr. et al. 1994). The last three dimensions also contribute to the mental component score. The standardization results in the health of the general population equaling 50, with a standard deviation of 10. It should be noted that there are some concerns regarding
the components scores, both in terms of the way they were generated (Davis, 1997), and the applicability of scores standardized to United States general population data to other settings. The physical component score (PCS) will be tested in terms of its responsiveness along with the physical functioning dimension of the SF-36. It is hypothesized that although there are concerns about its construct validity, it may be more responsive than the physical functioning dimension alone because the pain dimension is integrated into the physical component score. They are not expected to be as responsive as the joint-specific measures. The other dimensions will not be presented as these two were felt to be most comparable with the other measures and most likely to be responsive (along with bodily pain) to changes in these patients (musculoskeletal disorders).

**EQ-5D:** The Euroqol group (EuroQol Group, 1990) has introduced the EQ-5D which has the goal of being an indirect measure of patients values for different health states. It is a measure of value, rather than utility of the health states. Patients complete a brief questionnaire containing five items (three response options each). A profile is obtained for the patient’s health state across these five items. This is then transformed into a score (0-1) using values derived for each of the 216 possible health states through research by the developers (Johnson et al. 1998; EuroQol Group, 1990). In a recent American study valuations for 45 of the health states were found to have comparable weights to the British data (Johnson et al. 1998). The study had a poor response rate, and therefore the authors caution against generalizability of their findings (Johnson et al. 1998). The weights that were used in this thesis were the original British weights. Others have tested reliability, validity and responsiveness of the EQ-5D (Hurst et al.
The table (Table F.1) below summarizes the performance of the different instruments when the type of change (one axis on the taxonomy) is altered. The same types of change described in Chapter five are repeated here: the observed change in the whole cohort and various subgroups, the estimated change using serially measured states or transitional indices, and important change (which added the question of importance to the four transitional indices). Both types of estimated change and the important change asked questions of overall rating of the problem, pain, function and ability to work. In addition two threshold type questions: are you able to cope with your upper limb problem?, and are you able to do what you want/need to do and forget about your arm? were also asked. Patients moving from saying no to either of these at baseline, to saying yes at follow up were considered improved. The data gathered at the 12 week follow-up is presented. The responsiveness is summarized using the standardized response mean (mean change score divided by the standard deviation of the change).
Table F.1. Standardized response means (SRM) across instruments reflecting each instrument’s responsiveness to the different types of change at 12 weeks of follow-up. The data for the DASH is the same as found in table 5.9. All SRM’s adjusted so that a positive reflects a positive change in health on that instrument.

<table>
<thead>
<tr>
<th></th>
<th>DASH</th>
<th>SPADI function</th>
<th>Brigham function</th>
<th>SF-36 PF*</th>
<th>SF-36 PCS**</th>
<th>EQ-5D</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Observed change</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- all patients</td>
<td>0.78</td>
<td>0.62</td>
<td>0.64</td>
<td>0.14</td>
<td>0.45</td>
<td>0.43</td>
</tr>
<tr>
<td>- all shoulder</td>
<td>0.80</td>
<td>0.71</td>
<td>0.59</td>
<td>0.16</td>
<td>0.26</td>
<td>0.42</td>
</tr>
<tr>
<td>- all wrist/hand</td>
<td>0.71</td>
<td>0.43</td>
<td>0.76</td>
<td>0.11</td>
<td>0.41</td>
<td>0.45</td>
</tr>
<tr>
<td>- all TSA</td>
<td>0.98</td>
<td>0.80</td>
<td>1.09</td>
<td>0.50</td>
<td>0.53</td>
<td>1.02</td>
</tr>
<tr>
<td>- all RCT</td>
<td>0.82</td>
<td>0.73</td>
<td>0.52</td>
<td>0.06</td>
<td>0.19</td>
<td>0.29</td>
</tr>
<tr>
<td>- all CTS</td>
<td>0.54</td>
<td>0.38</td>
<td>0.72</td>
<td>0.04</td>
<td>0.39</td>
<td>0.44</td>
</tr>
<tr>
<td>- all tendinitis (w/h)</td>
<td>1.12</td>
<td>0.54</td>
<td>0.89</td>
<td>0.24</td>
<td>0.42</td>
<td>0.63</td>
</tr>
<tr>
<td>- all operative</td>
<td>1.04</td>
<td>0.92</td>
<td>0.89</td>
<td>0.22</td>
<td>0.43</td>
<td>0.67</td>
</tr>
<tr>
<td>- all non-operative</td>
<td>0.39</td>
<td>0.11</td>
<td>0.24</td>
<td>-0.02 ‡</td>
<td>0.08</td>
<td>-0.02 ‡</td>
</tr>
<tr>
<td><strong>Estimated change</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(difference in states)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- rating of problem</td>
<td>1.36</td>
<td>0.54</td>
<td>1.04</td>
<td>0.20</td>
<td>0.48</td>
<td>0.86</td>
</tr>
<tr>
<td>- pain</td>
<td>1.29</td>
<td>1.12</td>
<td>1.08</td>
<td>0.25</td>
<td>0.49</td>
<td>0.72</td>
</tr>
<tr>
<td>- function</td>
<td>1.47</td>
<td>1.21</td>
<td>1.12</td>
<td>0.55</td>
<td>0.61</td>
<td>0.79</td>
</tr>
<tr>
<td>- ability to work</td>
<td>1.69</td>
<td>1.02</td>
<td>1.39</td>
<td>0.88</td>
<td>1.22</td>
<td>0.99</td>
</tr>
<tr>
<td>- ability to cope</td>
<td>1.28</td>
<td>1.19</td>
<td>1.23</td>
<td>0.39</td>
<td>0.64</td>
<td>0.68</td>
</tr>
<tr>
<td>- able to forget about arm</td>
<td>1.36</td>
<td>1.02</td>
<td>1.28</td>
<td>0.39</td>
<td>0.67</td>
<td>0.57</td>
</tr>
<tr>
<td><strong>Estimated change</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(transitional indices)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- rating of problem</td>
<td>1.05</td>
<td>0.84</td>
<td>0.87</td>
<td>0.22</td>
<td>0.42</td>
<td>0.56</td>
</tr>
<tr>
<td>- pain</td>
<td>0.99</td>
<td>0.81</td>
<td>0.85</td>
<td>0.22</td>
<td>0.41</td>
<td>0.58</td>
</tr>
<tr>
<td>- function</td>
<td>1.19</td>
<td>0.86</td>
<td>1.07</td>
<td>0.27</td>
<td>0.52</td>
<td>0.66</td>
</tr>
<tr>
<td>- ability to work</td>
<td>1.04</td>
<td>0.78</td>
<td>0.92</td>
<td>0.30</td>
<td>0.44</td>
<td>0.60</td>
</tr>
<tr>
<td><strong>Important change</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(in)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- problem</td>
<td>1.15</td>
<td>0.88</td>
<td>0.90</td>
<td>0.22</td>
<td>0.39</td>
<td>0.62</td>
</tr>
<tr>
<td>- pain</td>
<td>1.06</td>
<td>0.85</td>
<td>0.89</td>
<td>0.23</td>
<td>0.40</td>
<td>0.63</td>
</tr>
<tr>
<td>- function</td>
<td>1.27</td>
<td>0.94</td>
<td>1.14</td>
<td>0.28</td>
<td>0.54</td>
<td>0.75</td>
</tr>
<tr>
<td>- work ability</td>
<td>1.12</td>
<td>0.81</td>
<td>0.95</td>
<td>0.28</td>
<td>0.46</td>
<td>0.65</td>
</tr>
</tbody>
</table>

‡ These two SRM’s suggested a decline in health (negative mean change for the group).
Discussion

These results demonstrate two important points. First, there is a considerable difference in responsiveness described for generic (SF-36) versus disease-specific measures of health state even in these early stages of recovery. This supports the literature which would guide us to expect larger responsiveness statistics in the specific over the generic measures (Patrick et al. 1989). Of interest is the fact that the EQ-5D was more responsive to the changes in these patient groups than was the SF-36. It could be possible explained by the fact that most of the “ill health” on the EQ-5D appeared in the pain dimension plus the anxiety (remembering that at baseline most of the subjects were close to having surgery for this problem). Many people moved from a response of 3 to 1, which mathematically in the scoring algorithm, would lead to a dramatic change in score. This finding is different than that anticipated by Brazier et al. in 1993 who felt that a high ceiling effect in the EQ-5D scores would indicate the SF-36 would be more responsive (Brazier et al. 1993). The difference in findings could well be explained by the difference in respondent groups: Brazier used a general population postal survey (Brazier et al. 1993), whereas our data is based on patients attending tertiary level care facilities.

•SF-36 PF = physical functioning dimension of the Short-Form 36 health status questionnaire
•SF-36 PCS = physical component score of the SF-36.
The second point is that the description of the responsiveness for the same instrument varied depending on the type of change being examined. This supports the findings in Chapter five, but also shows that the difference found in the DASH was replicated in other joint-specific measures (the Brigham and the SPADI function scales). These specific measures had responsiveness that varied considerably (DASH SRM's varied from 0.39 to 1.69; SPADI from 0.11 to 1.21; Brigham from 0.24 to 1.39). The two generic measures did not vary as much across the different categories of responsiveness tested, although still were variable (EQ-5D SRM’s ranged from -0.02 to +1.02; SF-36 PF dimension from -0.02 to +0.55 and the SF-36 PCS from 0.08 to 1.22). Two SRM’s stand out in this group the EQ-5D in the total shoulders which was more responsive than two of the three disease specific measures, and the negative values for the SF-36 PF and the EQ-5D in the non-operative group (which was also the lowest SRM for the other measures). Further work needs to be done to try to understand these, in particular the total shoulder group finding.

![Comparison of responsiveness across change, for each questionnaire](image)

Figure F.1 Graphic depiction of responsiveness of four instruments to different types of change. Vertical bar represent SRM of 0.80.

The SF-36 physical component score was more responsive than the physical function dimension, as we hypothesized, perhaps suggesting that the inclusion of pain in the scoring of the former helped with the responsiveness in these patients.

As mentioned in Chapter five, each of these standardized response means
could be considered in and of themselves, to be the findings of a reasonable study of that questionnaire's responsiveness. Therefore, the variability of these results reinforce the conclusions of Chapter five, that is that the responsiveness described for a given instrument will vary with changes in the category of responsiveness being tested.
Appendix G: Guidelines for Questions to Ask of a Study of Responsiveness.

This appendix provides a guide for those using the literature as a source of evidence about the responsiveness of a given measure for their (targeted) purpose(s).

**Step One:** Clarification of the type of responsiveness in target application. Define the patient group, the treatment situation & the type of change under consideration

**Step Two:** appraise whether this study offers you the right kind of information by asking these questions:

<table>
<thead>
<tr>
<th>Question</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Are the patients similar enough to my own?</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>2. Are they looking at a similar type of treatment?</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>3. What category of responsiveness is being studied?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A) Who is the focus of the analysis? Are results presented for individuals or groups of patients?</td>
<td>□ group level</td>
<td>□ individual level</td>
</tr>
<tr>
<td>B) When is the comparison taking place?</td>
<td>□ over time (same patients, over time)</td>
<td>□ one point in time (between persons)</td>
</tr>
<tr>
<td></td>
<td>□ both, hybrid (between group differences of within person change)</td>
<td></td>
</tr>
<tr>
<td>C) What is the nature of the change being quantified?</td>
<td>□ minimum change potentially detectable by the instrument</td>
<td></td>
</tr>
<tr>
<td></td>
<td>□ minimum change detectable given the measurement error</td>
<td></td>
</tr>
<tr>
<td></td>
<td>□ observed change in a given population</td>
<td></td>
</tr>
<tr>
<td></td>
<td>□ observed change in those deemed to have improved (estimated change) according to.....</td>
<td></td>
</tr>
<tr>
<td></td>
<td>□ patient</td>
<td></td>
</tr>
<tr>
<td></td>
<td>□ clinician/researcher</td>
<td></td>
</tr>
<tr>
<td></td>
<td>□ payer</td>
<td></td>
</tr>
<tr>
<td></td>
<td>□ society</td>
<td></td>
</tr>
<tr>
<td></td>
<td>□ observed change in those deemed to have an important change according to....</td>
<td></td>
</tr>
<tr>
<td></td>
<td>□ patient</td>
<td></td>
</tr>
<tr>
<td></td>
<td>□ clinician/researcher</td>
<td></td>
</tr>
<tr>
<td></td>
<td>□ payer</td>
<td></td>
</tr>
<tr>
<td></td>
<td>□ society</td>
<td></td>
</tr>
</tbody>
</table>

Before findings in the literature are generalized to the target application, there should be a definite match in type of change, as described by the taxonomy, and in this form.
Appendix H. Qualitative Interviews and Documentation.

Topics Covered.
The following topic headings served as guides for the qualitative interviews. As is typical of grounded theory methodology, the exact content of the interview is not set. The questions used in an interview varied over time in response to the developing theory. In general, the interview covered the following areas:

A) Tell me about your arm discomfort.
   - Where do you have pain, and what has it been like?
   - Are you better now?
   - Who do you go to for support or advice?

B) Tell me about what worked? / didn’t work for your discomfort? (treatment successes/fails)

C) How did you know it worked? How did you know that you are better or worse?

D) What would be an indication that you are better or worse from how you are today?

Face Sheets
Following each interview a face sheet was completed where the following were recorded:

1) What were the main issues or themes that stand out from this interview?
2) Summarize the information that I got, or failed to get for each target question:
   - treatment choices (successes, influencing forces)
   - defining change (important change, description of construct, meaning of being better).
3) Did anything else strike me as interesting or important with this participant?
4) What new (or remaining) questions do I have to think about for the next interview? (Including who should be interviewed next).

Focused Coding.
At the end of the study, twelve transcripts were recoded using themes that were included in the theory. New face sheets were created describing findings from each of these transcripts in the following areas:

Disorder experience: impairments, capacity, role performance.

Appraisal of that experience: Am I sick? Is my pain a problem?

Mediators of the response: social comparison (internal vs. external comparisons documented), issues of legitimacy, coping style.

Resultant state: document nature of response: resolution of the disorder, readjustment, redefinition and any contextual facts.
Reference List


Christensen L. and Mendoza, J. L. (1986b) A method of assessing change in a single subject: an alteration of the RC
index. *Beh Ther.* 17; 305-8.


Cohen J. (1990) Things I have learned so far. *American Psychologist* 45(12); 1304-12.


Cronbach L. J. and Furby, L. (1970) How we should measure "change" - or should we? *Psychol Bull* 74(1); 68-80.


Jette A. M. (1995) Outcomes research: Shifting the dominant research paradigm in physical therapy. *Phys Ther* 75(11); 965-70.


Lachs M. S. (1993) The more things change ... *J Clin Epidemiol* 46(10); 1091-2.


Lydick E. and Epstein, R. S. (1993) Interpretation of quality of life changes. *Quality of Life Research* 2; 221.


MacDermid J. C., Richards, R. S., Donner, A., Bellamy, N., and Roth, J. H. (1999a) Responsiveness of the SF-36, DASH. Patient-rated wrist evaluation and physical impairments in evaluating recovery after a distal radius fracture. Unpublished manuscript


life. Medical Care 27(3)Suppl.; S217-S232.


Wood-Dauphinee S. (1999) Assessing quality of life in clinical research: From where have we come and where are we going? *J Clin Epidemiol* 52(4); 355-63.


World Health Organization. 1999. *International Classification of Impairment, Disability and Handicap - 2 (beta version).* WHO Website


