RESPONSIVENESS COMPARISON OF TWO UPPER-EXTREMITY OUTCOME MEASURES

By

LEIGH A. LEHMAN

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2008

© 2008 Leigh A. Lehman

To Daddy, Moma, and Lynn
I can never thank you enough.

ACKNOWLEDGMENTS

First, I would like to thank God, who was with me and went beside me each step of the way along this journey, who in his infinite wisdom brought me to this place, this place in my journey along the path of life. Thanks to Him, who sees the whole picture and knows where each piece fits in, even at times, when it seems ambiguous to us. When the struggles and frustrations were immense and no one was physically present to encourage me, when I felt at times lonely, I knew I was not alone. I knew that God never left my side. All the countless times that I wanted to quit and give up, He renewed my strength. Over and over again, He renewed His promise, "But they that wait upon the Lord shall renew their strength; they shall mount up with wings as eagles; they shall run, and not be weary; and they shall walk, and not faint.[1]"

In addition to His deeply felt presence in my life at every moment, every second, every breath, He graced me with people in my life who never ceased to provide stability and a constant wellspring of support. I would like to thank my family, the three dearest people in the world to me (Jerry, Faye, and Lynn Lehman). I can never thank them enough! Enormous thanks go to Daddy, who twice, at two of the roughest points along this journey came to stay with me. He took a sabbatical and came to encourage me for a semester when I was doing my internships. Then again in my final month of working on my dissertation, he came to stay with me that I might reach the finish line. I will never forget the morning, after I came home ready to quit, when he ran in to my room very early, awakened me, and said, "Get up, Leigh! GO GATORS! You are going back to Gainesville and I am going with you!" Tremendous thanks go to Moma, who without any regret quit her job as an elementary school librarian to come and stay with me,

[1] Isaiah 40:31

at another low point along the way. I will never forget the day that I was so depressed, I thought I would never make it through the long hours ahead and she was there by my side and helped me through it. Undoubtedly, her commitment to my educational pursuits has sustained me during the tough times. A greater love than my parents have shown to me, I can never imagine. I would have given up on the quest for this degree a million times over without their continual love and support. My deepest thanks, to Lynn, my beloved sister, who paid my rent several semesters and never ceased to give to me, literally thousands of dollars of which I may never be able to repay, and for all the shoes, all the clothes. I am truly blessed to have been given such a sister. I am so unbelievably lucky to have been blessed with such an extraordinary family. I thank God for placing these three most special people in my life.

I would like to thank my mentor, Craig Velozo. For all the many hours he spent with me, teaching me to write and always emphasizing the correct approach to conducting research. He gave an incredible amount of time to mentoring and instructing me. I have never met anyone who works so hard and with such dedication. Without his guidance and support, I could never have made it to this point. I would like to thank my committee members, Orit Shechtman, Lorie Richards, and James Algina for their support throughout my dissertation work. I would like to thank Dennis Hart and Focus on Therapeutic Outcomes (FOTO), Inc. for providing me with the data sets for my dissertation project.

Finally, I would like to thank all of the people in the Rehabilitation Science Doctorate (RSD) program that truly became my RSD family. They opened my eyes to discover that I could do and achieve more than I ever dreamed possible. To all of them, I am forever indebted for this. Thanks to Dennis, Mike, Arlene, Sande, Jamie, Rick, Bhagwant, Inga, Patricia, Pey-Shan, Jia-Hwa (big thanks for Hawaii and all the trips!), Jessica, Megan, Michelle, Milap, Eric, Margaret,

Todd, Dr. Mann and so many more, too many to name, in the department of Occupational Therapy at UF, who showed me such kindness and expanded my life vision. Each one changed and touched my life in ways they do not even realize. May each of their journeys in life take them to the best of places!

The Florida dream, the Florida experience, now culminates in the completion of this Ph.D. degree. As the sun sets on this time in my life, I pray that I might use this degree and all that I have learned from my experiences along the way, to fulfill God's will for my life, that I might help someone, as those that have helped me along this journey. "To Him that is able to keep you from falling, and to present you faultless before the presence of his glory with exceeding joy[2]", I offer abundant thanks for the "Florida experience": for giving me this time in my life, for helping me to reach this end in the completion of my Ph.D., for placing these people in my life who are so deserving of thanks. As a new day dawns and I find myself now, humbly as Dr. Lehman, I pray for the opportunity to more fully serve Him with all the knowledge I have gained on this journey.

[2] Jude 24

TABLE OF CONTENTS

page

ACKNOWLEDGMENTS 4

LIST OF TABLES 10

LIST OF FIGURES 11

LIST OF ABBREVIATIONS 12

ABSTRACT 14

CHAPTER

1 ABILITY TO DETECT CHANGE IN PATIENT FUNCTION: RESPONSIVENESS DESIGNS AND METHODS OF CALCULATION 17

Introduction 17
Types of Responsiveness Designs 21
Single-Group Designs 21
Multiple-Group Designs 22
Discussion/Conclusions 24

2 RESPONSIVENESS OF DISABILITIES OF THE ARM, SHOULDER, AND HAND (DASH) VERSUS THE UPPER-EXTREMITY FUNCTIONAL INDEX (UEFI) 33

Introduction 33
Methods 37
Sample 37
Materials 38
Procedures 39
Statistical Analysis 39
Rasch analysis to obtain comparable measures 39
Analysis of variance (ANOVA) 41
Sensitivity and specificity 41
Correlation 42
Results 43
Discussion 44
Conclusions 47

3 ITEM CHARACTERISTICS OF THE DISABILITIES OF THE ARM, SHOULDER, AND HAND (DASH) OUTCOME QUESTIONNAIRE 53

Introduction 53
Methods 55
Sample 55

Assessment 56
Procedures 57
Data analysis 57
Unidimensionality 57
Item difficulty 59
Person-item match 59
Differential item functioning 60
Item discrimination 61
Test information 62
Results 62
Unidimensionality 62
Admission data EFA 62
Admission data CFA 63
Discharge data EFA 63
Discharge data CFA 64
Rasch-derived fit statistics and point measure correlations 64
Item Difficulty 65
Person-Item Match 65
Differential Item Functioning 66
Item Discrimination 68
Test Information 68
Discussion 69
Conclusion 73

4 CREATING A CLINICALLY USEFUL DATA COLLECTION FORM FOR THE DISABILITIES OF THE ARM, SHOULDER, AND HAND OUTCOME QUESTIONNAIRE USING THE RASCH MEASUREMENT MODEL 92

Introduction 92
Methods 95
Sample Characteristics 95
Assessment 96
Disabilities of the Arm, Shoulder, and Hand (DASH) Outcome Questionnaire 96
Procedures 97
Analyses 98
Item infit analysis and previous factor analysis with larger sample to test Rasch measurement model fit 98
Test of logic of difficulty hierarchy order 99
Obtaining a general keyform in winsteps for dash admission data 100
Creation of data collection form 100
Results 102
Item Infit Analysis and Previous Factor Analysis with Larger Sample to Test Rasch Measurement Model Fit 102
Test of Logic of Difficulty Hierarchy Order 102
Creation of Data Collection Form 103
Discussion 106

5 CONCLUSIONS 125

APPENDIX

A ITEMS ON THE DISABILITIES OF THE ARM, SHOULDER, AND HAND (DASH) OUTCOME QUESTIONNAIRE 137

B ITEMS ON THE UPPER-EXTREMITY FUNCTIONAL INDEX (UEFI) 139

C GLOBAL RATING OF CHANGE 140

LIST OF REFERENCES 141

BIOGRAPHICAL SKETCH 153

LIST OF TABLES

Table		page
1-1	Single group designs and coefficients	30
1-2	Multiple-group designs and coefficients	31
2-1	Sample of 214 demographics	49
2-2	Sensitivity, specificity, and overall error at various cutoff points for the DASH and UEFI	51
3-1	Sample of 991 demographics	75
3-2	Item factor loadings on first factor	76
3-3	Admission item factor loadings on three factors after varimax rotation	77
3-4	Discharge item factor loadings on three factors after varimax rotation	78
3-5	Infit statistics at admission and discharge	79
3-6	Point measure correlations at admission and discharge	79
3-6	Point measure correlations at admission and discharge	80
3-7	Item difficulty estimates and differential item functioning (N = 960, **p<0.002, *p<0.05).	83
3-8	Item discrimination estimates	88
4-1	Sample of 37 demographics	116
4-2	Admission item fit and difficulty estimates	117
4-3	Discharge item fit and difficulty estimates	118

LIST OF FIGURES

Figure page

2-1 Calculating sensitivity and specificity where sensitivity = a/(a +c) and specificity = d/(b+d) 50

2-2 Receiver operating characteristic (ROC) curves for the DASH (circles) and UEFI (triangles), showing the accuracy of different cutoff points (diagonal line represents assessment with 50/50 chance of detecting improvement)....52

3-1 DASH admission scree plot....81

3-2 DASH discharge scree plot....82

3-3 Admission (left) and discharge (right) person-item maps (Arranged so zeros line up to allow comparison)....84

3-4 DASH admission item calibrations vs. dash discharge item calibrations....86

3-5 With all items and with DIF items deleted mean person ability measures (y-axis, logits) at admission and discharge....87

3-6 One parameter model versus two parameter model mean person ability measures (y-axis, logits) at admission and discharge....89

3-7 Admission scale information function using one parameter model....90

3-8 Discharge scale information function using one parameter model....91

4-1 Data collection form for the DASH (Based on analysis of admission data) 119

4-2 Data collection forms with individuals' responses....120

4-3 Goal setting data collection form....123

LIST OF ABBREVIATIONS

ARAT	Action Research Arm Test
ANOVA	Analysis of Variance
AU	Area Under (Receiver Operating Characteristic Curve)
ECOS-16	Assessment of health related quality of life in osteoporosis
BQ	Boston Questionnaire
CFI	Comparative Fit Index
CFA	Confirmatory Factor Analysis
df	Degrees of freedom
DASH	Disabilities of the Arm Shoulder and Hand outcome questionnaire
ES	Effect Size
EFA	Exploratory Factor Analysis
FOTO	Focus on Therapeutic Outcomes, Inc.
FMA	Fugl-Meyer Assessment
GMAE	Gross Motor Ability Estimator
GRI	Guyatt's Responsiveness Index
ICF	International Classification of Functioning, Disability, and Health
ICIDH	International Classification of Impairments, Disabilities, and Handicaps
ICC	Intraclass Correlation Coefficient
IRT	Item Response Theory
JT-HF	Jebsen Taylor Hand Function Test
JCAHO	Joint Commission
MAP	Maximum a Posterior estimation
MLE	Maximum Likelihood Estimation
MnSq	Mean Square Standardized residuals

MHQ	Michigan Hand Outcome Questionnaire
MCID	Minimally Clinically Important Difference
MRMT	Minnesota Rate Manipulation Test
MAS	Motor Assessment Scale
PWRE	Patient-Rated Wrist Evaluation
PRO	Patient-Reported Outcome instrument
ROC	Receiver Operating Characteristic
RC	Reliable Change Index
RMSEA	Root Mean Square Error of Approximation
SD	Standard Deviation
SE	Standard Error
SEM	Standard Error of Measurement
SPADI	Shoulder Pain and Disability Index
SRM	Standardized Response Mean
SRMR	Standardized Root Mean Square Residual
TLI	Tucker-Lewis Index
UEFI	Upper Extremity Functional Index
UEFS	Upper Extremity Functional Scale
WMFT	Wolf Motor Function Test
WHO	World Health Organization

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial
Fulfillment of the Requirements for the Degree of Doctor of Philosophy

RESPONSIVENESS COMPARISON OF TWO UPPER-EXTREMITY OUTCOME MEASURES

By

Leigh A. Lehman

May 2008

Chair: Craig A. Velozo
Major: Rehabilitation Science

The primary focus in the rehabilitation clinics where patients with upper-extremity impairments are treated is enabling improvement. Not just progression at the impairment level, but advances in functional ability (i.e., changes that affect patients' daily lives). In order to ensure the achievement of this goal, assessments that are reliable, valid, and able to detect change are essential. While reliability and validity of many hand/upper-extremity assessments have been well established, this is not the case for ability to detect change, a property of instruments called responsiveness.

The responsiveness of the Disabilities of the Arm, Shoulder, and Hand (DASH) outcome questionnaire was compared to the Upper-Extremity Functional Index (UEFI). Areas under the ROC curves were almost identical for the two assessments (.67 and .65), as were correlations between global ratings and the change scores (r = 0.332 and 0.350). Given that the responsiveness calculations for both assessments are very similar, there appears to be no real advantage of one instrument over the other in detecting change. One concern highlighted in the comparison of the responsiveness of the DASH to the UEFI centered on use of a patient reported global rating of change as a "gold standard". The correlation between the global rating and the person measure change scores was in the range of what is considered low. Thus, there is a

problem with not having a "gold standard" with which to compare the change scores. In essence in this study then, both the assessment and the "gold standard" were two patient reported outcomes. This study tends to imply that patients are not accurate in reporting the amount of change that has occurred in their functional ability.

The item-level psychometrics and factor structure of the Disabilities of the Arm, Shoulder, and Hand (DASH) outcome questionnaire was investigated. Results of exploratory factor analyses (EFAs) and confirmatory factor analyses (CFAs) were inconclusive as to whether or not a one factor solution was plausible. A number of items misfit (ten at admission and eleven at discharge). DASH item point measure correlations were all above .30. However, the observation that many (19 at admission and 20 at discharge) were above .70 indicates that some of the items may be redundant (i.e., add no further information to the test than other items already obtain). Item difficulty hierarchies support the theory of motor control. Several items displayed significant differential item functioning (DIF) from admission to discharge using p value guidelines. However, using the 0.5 logit difference guideline, no items displayed significant DIF. Furthermore, mean person ability measures were not affected by the inclusion of DIF items. The intraclass correlation coefficient (ICC) was high indicating reliability in measurement between the two time-points. Since none of the item discrimination values were zero or negative, this indicates that all items contribute in a similar fashion to the overall test. Item discrimination did not appear to dramatically affect person measures, although, there is more difference between the person ability measure calculated using the different models at discharge. Similar curves for test information for admission and discharge data provide further evidence for the stability of the test at two time points.

Finally, the potential for creating clinically useful data collection forms through the use of Rasch methodologies was demonstrated. The general keyform output from Winsteps Rasch analysis program(1) was used as the basis for creating the data collection form. Creation of such a form proved to be an easy process and potentially offers several advantages over the forms used in clinics today. Instructions were placed at the top of the form and items (activities for patients to rate) were listed to the left of the possible answer choices (both numbers to circle and descriptions of what the numbers mean; for example, no difficulty placed above 5). A person measure scale ranging from zero to 100 was placed at the bottom. As depicted by forms filled out according to patients of differing admission ability levels (high, medium, and low), at a glance it is possible to get an idea of where an individual's overall ability to perform functional tasks requiring the upper-extremity would fall. Furthermore, an advantage of this type of data collection form is that it can be used in goal setting and treatment planning. By observing the location where this individual starts to have mild to moderate difficulty with a substantial number of items, the clinician can get a real sense of what activities should be set as first goals for the patient.

CHAPTER 1
ABILITY TO DETECT CHANGE IN PATIENT FUNCTION: RESPONSIVENESS DESIGNS AND METHODS OF CALCULATION

Introduction

In clinics where patients with hand impairments are treated, the primary focus is on enabling patient improvement. Not just progression at the impairment level, but advances in functional ability (i.e., changes that affect patients' daily lives). To ensure the achievement of this goal, assessments that are reliable, valid, and able to detect change are essential. While reliability and validity of many hand/upper-extremity assessments have been well established, this is not the case for ability to detect change, a property of instruments called responsiveness(2). Recently other fields have realized the importance of studying responsiveness. For instance, many studies have been conducted to assess the capacity of health-related quality of life measures to detect change(3-7). Since hand/upper-extremity clinicians spend much time and energy trying to bring about change specifically to the hand/upper-extremity ailments of their patients, using assessments that have been validated as having the ability to detect a change in the problem of interest is of foremost importance.

Outcomes of hand therapy are generally measured by therapists using three types of measures: impairment measures, performance measures, and self-report measures. For each of these measures, various authors have reported on reliability and validity, however, responsiveness has received less attention in the literature.

Measurement at the impairment level has been the primary focus of evaluating outcomes for many years in hand therapy(8). Impairment level measures most commonly used in hand therapy include manual muscle testing (MMT), grip/pinch strength, and sensibility testing. MMT standards have been well established(9-13). Cuthbert and Goodheart(14) reviewed over 100 studies providing evidence for the reliability and validity of MMT. Likewise, techniques for

grip and pinch strength measures have been well documented(15). Reliability and validity of grip strength measures have been studied by many researchers(16-26). For various grip strength measures, studies show calculations of test-retest reliability resulting in very high intraclass correlation coefficients (ICC), ranging from .95 to 99(16, 23, 25, 27). Likewise, investigators have reported concurrent validity (intraclass correlation coefficients) comparisons of various types of dynamometers to range from .98-.99(16, 27). Pinch strength measures have been also been shown to be reliable and valid(28-32). Researchers have documented interrater reliabilities ranging from .82-.99 for pinch measures (intraclass correlation coefficients)(29, 31, 32). Additionally, there is evidence for concurrent validity between manual and computerized methods of assessing pinch strength (ANOVA, $p \geq .54$)(29). Although, detailed procedures exist for performing sensibility testing, there is less evidence for the reliability and validity of these measures(8). However, tactile discrimination and object identification have been shown to be related to functional ability(33-36). Compared to the wealth of studies establishing the reliability and validity of impairment measures, there is a dearth of evidence as to their ability to detect patient change related to everyday performance of tasks.

Although perhaps used more to assess patients with stroke, performance measures are also sometimes used in hand/upper-extremity therapy clinic(37, 38). Various authors have shown commonly used performance measures to be both reliable and valid. Sabari and colleagues(39) showed the interrater reliability of the Motor Assessment Scale (MAS) to range from .95 to .99 on the three subscales (Upper-Arm Function Scale, Hand Movements Scale, and Advanced Hand Activities). Similarly, these authors showed the MAS to correlate .88 with the Fugl-Meyer Assessment (FMA). The Fugl-Meyer Assessment also has been established by many other authors to be reliable and valid(37, 40-45). The Wolf Motor Function Test (WMFT) and the

Action Research Arm Test (ARAT) have shown similar reliability and validity properties(46-48), (45, 49, 50). Another performance measure, more frequently used with problems in the hand clinic, the Jebsen Taylor Hand Function Test (JT-HF), has been shown to provide stable estimates across testing sessions(38, 51-53). Nevertheless, like assessments of impairment, there are very few studies of performance measures which assess responsiveness or ability to detect change in daily function(45).

Recently, the Food and Drug Administration has noted the importance of self-reports, termed patient-reported outcome (PRO) instruments, in measuring patient function(54). This is because several aspects of a patient's condition are only known to the patient (e.g., level of pain, satisfaction with treatment). Furthermore, the patient holds a unique perspective on the treatment process that may be useful to his/her therapist. Increasing use of self-reports is emerging in hand clinics today due to their ability to efficiently investigate patient abilities in areas of daily activity and participation in society(8). A wealth of information exists on the reliability and validity of various self-report instruments. For example, satisfactory reliability and validity of the Disabilities of the Arm, Shoulder, and Hand (DASH) outcome questionnaire has been reported in numerous studies(55-61). Beaton, Katz, Fossel, Wright, Tarasuk, and Bombardier(57) found that the DASH showed good test-rest reliability (intraclass correlation coefficient = 0.96). Additionally, these authors found the DASH to have discriminative validity, differentiating between those currently able to work and those unable to work ($t = -7.51$, $p < 0.0001$). Convergent validity was also demonstrated with all correlations with other upper-extremity measures exceeding 0.70 (Pearson). Similarly, several authors have published studies relating to the reliability and validity of the ABILHAND(62-65).

While responsiveness of a few of these self-reports has been assessed, for many of them no studies on responsiveness have been conducted. The DASH has been perhaps the most studied in terms of responsiveness. Beaton, Katz, Fossel, Wright, Tarasuk, and Bombardier(57) demonstrated the DASH's ability to reveal change in patients who were expected to change (Standardized Response Mean = 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 (Standardized Response Mean = 0.92-1.40). Similarly, Gummesson, Atroshi, and Ekdahl(56) found that the DASH was able to detect change in function before and after surgical treatment for upper-extremity conditions (Mean Score Change = 15; Effect Size = .70; Standardized Response Mean = 1.2). Likewise, Kotsis and Chung(66) found the DASH to be moderately responsive (Standardized Response Mean = .70). Much of the evidence about responsiveness of other upper-extremity assessments has come from studies comparing these assessments to the DASH. Beaton and colleagues(57) found that the DASH was as responsive as two other joint specific measures (the Shoulder Pain and Disability Index and the Brigham carpal tunnel questionnaire). Similarly, Greenslade, Mehta, Belward, and Warwick(67) found the DASH to be as responsive as the disease specific Boston questionnaire (also used with individuals with carpal tunnel syndrome). In contrast, MacDermid and colleagues(68, 69) found the Patient-Rated Wrist Evaluation (PWRE) to be more responsive than the DASH and Gay, Amadio, and Johnson(70) found the Carpal Tunnel Questionnaire to be more responsive than the DASH. The Michigan Hand Outcome Questionnaire (MHQ) has been shown to detect change in patients after carpal tunnel release (Standardized Response Mean for pain scale = 0.9; Standardized Response Mean for function scale = 0.6; Standardized Response Mean for combined function/symptom scale = 0.7).

In spite of these few studies, the study of responsiveness in upper-extremity assessments is lacking. Additionally, studies comparing the responsiveness of multiple similar assessments have been inconclusive as to which assessments are foremost in ability to detect patient change(57, 71, 72). Since the majority of a clinician's efforts are intended to effect change in patient ability, it is clear that the capacity of instruments to measure change has been under evaluated. The purpose of this review is to present the state-of-the-art responsiveness designs and methods, to highlight problems associated with comparing responsiveness coefficients, and to suggest a step toward increasing the existing knowledge of the ability of currently used hand/upper-extremity assessments to measure patient change.

Types of Responsiveness Designs

There are two categories of responsiveness designs, as outlined by Stratford and colleagues(73). These categories are based on how many groups are involved: 1) single-group designs and 2) multiple-group designs. Single-group designs include the before-after (no baseline) design and the before-after design with a baseline. While multiple-group designs include three types: 1) those that compare patients who are randomly assigned to receive either a previously proven effective treatment or a placebo, 2) those that compare two or more groups whose health status, based on prior evidence, is expected to change by different amounts, and finally, 3) those that compare two groups expected to change by different amounts based on responses to an external standard (e.g., a global rating of change score). What follows is a description of each type of design and the responsiveness coefficients that are appropriate to use with each type of design. These designs/coefficients are summarized in Tables 1-1 and 1-2.

Single-Group Designs

In single-group designs patients who are expected to undergo a change in health status may be measured before and after an intervention is implemented (before-after, no baseline designs).

Alternatively, a baseline may be taken (before intervention), followed by measurements just before and after the intervention (before-after, with baseline designs)(73). See Table 1-1. With the before-after (no baseline) design the acceptable choices for responsiveness coefficients include effect size (ES), standardized response mean (SRM), and paired *t* value(73). These coefficients represent the average change in score divided by a standard deviation (SD) (measure of average variability of scores). The difference between the formulas used for these three coefficients (ES, SRM, and paired *t* value) is the SD used. For ES the SD used is that of the initial scores, for SRM the SD used is that of the change scores, and for the paired *t* value the SD is that of the change scores divided by the square root of the number of subjects. These coefficients only detect whether or not a change has occurred and cannot be used to assess various amounts of change. In contrast to the above coefficients that are appropriate to use when there is no baseline measure taken, with the before-after designs with a baseline, the appropriate coefficient to calculate is Guyatt's Responsiveness Index (GRI), also called the responsiveness statistic or responsiveness index(72). This statistic represents the ratio of observed change after treatment compared to the random variability observed in the baseline measure.

Multiple-Group Designs

Multiple-group designs provide a more refined estimate of change over time(73). These designs include groups of patients who are expected to undergo different amounts of change. Various coefficients can be used to calculate responsiveness when using each design. Guyatt's Responsiveness Index (GRI), t value for independent change scores, Analysis of Variance (ANOVA) of change scores, Norman's S_{repeat}, Norman's S_{ancova}, and area under the Receiver Operating Characteristic (ROC) curve are all responsiveness coefficients that can be calculated when placebo and intervention groups are being compared. These coefficients with the exception of Norman's S_{ancova}, can also be used when comparing two or more groups whose

health status is expected to change by different amounts. All of these indices, in addition to correlation, can be used when comparing two groups expected to change by different amounts based on responses to an external standard (e.g., a global rating of change score obtained from the patient or therapist or the average of both ratings).

Each of the coefficients used with multiple-group designs varies in formula. First, consider the coefficients that involve two or more groups, expected to change by different amounts, which do not require an external measure of change (see Table 1-2). GRI represents the ratio of observed change in a group of patients expected to undergo a change (i.e., proven intervention group) to the variability in the group not expected to change (i.e., in this design the placebo group). In contrast, calculating a t value for independent change scores compares the change in patients who are expected to undergo an important change with the change in patients who are not expected to undergo an important change (i.e., here proven intervention to placebo group). ANOVA is similar to the calculation of a t value, only it allows for more than two groups. Both of Norman's coefficients essentially compare group variance to group and error variance. Norman's S_{repeat} represents the variance of the group x time interaction divided by the group x time interaction plus error variances. While Norman's S_{ancova} represents the group variance divided by the group plus error variance with the dependent variable being the follow-up score and the covariate being the initial score (see Table 1-2).

A special type of responsiveness coefficients used with multiple-group designs are those that require an external measure of change. With the other coefficients, there is no comparison measure, thus, when change is detected, one is still unsure whether the instrument used correctly identified change. If a strong external measure is used, this provides an advantage. However, since the external measure is considered "truth" in these designs, if the external measure is

faulty, there is much error associated with these coefficients. Two coefficients requiring an external criterion are Receiver Operating Characteristic (ROC) curves and correlation coefficients. A brief description of these two follows. First, a ROC curve graphs sensitivity on the Y-axis against 1 – specificity on the X-axis (74). Sensitivity being defined as the number of patients correctly identified by an assessment as undergoing an important change divided by all patients who undergo an important change as determined by the external criterion. (Note: Correct versus incorrect identification by an assessment is based on agreement with the external criterion. Thus, the external criterion here is being considered a "gold standard".) In contrast, specificity is defined as the number of patients who are correctly identified by an assessment as not undergoing an important change divided by the number of patients who did not undergo an important change as determined by the external criterion. The greater the area under the ROC curve, the greater the assessment's ability to distinguish patients who did and did not undergo an important change (as defined by the external criterion). Second, in calculating a correlation for this type of responsiveness design, the external standard evaluating change (e.g., global rating of change scores) is correlated with change scores from the assessment. Higher correlations indicate a stronger association between the results of the assessment and the external standard. Assuming the external standard is correct, higher correlations indicate a more responsive assessment.

Discussion/Conclusions

There are limitations associated with each of the designs presented above and problems associated with using various coefficients. Perhaps, the foremost concerns are associated with single-group designs. A disadvantage associated with the before-after, no baseline design is that if there is not a change detected, it is unclear whether the measure was unable to detect the change or whether the patients did not undergo the expected change. This design is considered

the weakest of the designs. With the before-after designs with a baseline a major disadvantage is that the period during which stability is measured (i.e., period from baseline to measurement taken just before intervention) is often shorter than the period during which change is assessed (i.e., measure taken just before intervention to measure taken after intervention). Thus, this design may underestimate the magnitude of random variability that occurs over longer periods in patients whose health status is truly stable(73).

The addition of a comparison group gives multiple group designs an advantage over single-group designs. However, each of the multiple-group designs also has limitations. One disadvantage when comparing patients who receive either the intervention or a placebo involves the difficulty finding an intervention that is known to be effective (i.e., an intervention that has a high probability of resulting in changes to the treatment group). In a similar way, the major disadvantage when comparing groups expected to change by different amounts is that often it is difficult to find groups who meet this criterion (i.e., often there are not two groups of patients available whose condition is expected to change by different amounts). When comparing two groups which have been defined by some external standard (e.g., global rating of change scores) the validity of the study may be compromised for two reasons. First, patients complete the measure under study and the criterion rating and thus, the criterion measure is not independent of the measure under study (i.e., a patient who responds to the questions on the measure indicating they have changed is also likely to respond to the criterion measure in the same manner). Secondly, when a global rating of change measure is used as the external criterion, evidence suggests that patients have difficulty recalling their initial state and consequently, may be inaccurate when assessing the amount of change that has taken place(75). Although an outside criterion provides the advantage of being a benchmark by which to assess whether change has

really occurred, accuracy of results depends on the validity of the criterion as an effective measure of change.

In addition to the limitations related to specific designs, a problem with measuring the property of responsiveness involves the discrepant results obtained when different coefficients are used. For example, when several similar instruments are compared to determine which assessment is the most responsive, the ranking of these instruments often differs depending on which design and which coefficient is used to calculate responsiveness(72, 76, 77). For instance, Wright and Young(72) compared five different assessments using five different responsiveness coefficients (GRI, SRM, relative efficiency statistic, ES, and correlation). These investigators showed that these different responsiveness coefficients provided different rank ordering of the responsiveness of the assessments. Thus, an important question to answer before beginning a responsiveness study is: What is the most appropriate responsiveness design and coefficient to use when calculating responsiveness? Since no gold standard exists for measuring responsiveness of hand/upper-extremity assessments and multiple possible responsiveness coefficients can be calculated (with conflicting results), there is a need for clarification as to what designs require the use of which coefficients. Erroneously, many researchers calculate an array of different coefficients(78). However, this approach is inconsistent with mathematical logic. Since the formulas for each of the coefficients differ slightly, the different indices can lead to conflicting results.

A more theoretically sound approach is to determine which responsiveness coefficient is most appropriate for your study design and sample. If your design involves a single group, one of the coefficients outlined above as being appropriate for single-group studies must be used. Alternatively if your design includes more than one group, you should choose from those

coefficients outlined above for use with multiple-group designs. Also, with multiple-group designs you should consider the amount of change expected of your groups. That is, what differences in change do you expect between placebo versus treatment groups, acute versus chronic groups, or between groups divided based on an outside criterion. Considering these two elements will lead away from calculations of multiple coefficients and thus, a more planned structured attack. However, it should be noted that even when the appropriate coefficients for the given number of groups and sample characteristics are used, there may be discrepancies in the calculations of different coefficients due to minor variations in the formulas for each. Highly sophisticated statistical studies are needed to help determine which results are most credible in these cases. Because hand/upper-extremity therapists' foremost concern is inciting change in the functional ability of their patients, more studies leading to higher standards of responsiveness (equal to the standards of reliability and validity) should be mandated.

One final area of concern involves confusion over terminology that exists relating to the measurement of patient change(79, 80). Sensitivity as described above as it relates to ROC curve analyses refers to an assessment's ability to obtain positive results when the condition is present(74). However, some authors use the term, sensitivity to refer to the ability to detect a change using statistical analyses, whether or not that change is clinically significant(81). Furthermore, various slightly different methods for evaluating the significance of clinical change have been proposed. Two of these methods are the Minimally Clinically Important Difference (MCID) and the Reliable Change (RC) index. The Minimally Clinically Important Difference (MCID) has been defined as "the smallest difference between the scores in a questionnaire that the patient perceives to be beneficial"(7, 82). Similarly, the Reliable Change (RC) index has been described as a method to "determine statistically reliable change that is also clinically

significant"(83). While each of these methods has advantages and disadvantages, no conclusive evidence shows one to be superior to the others (84). A brief discussion of the Minimally Clinically Important Difference (MCID) and Reliable Change index (RC) follows.

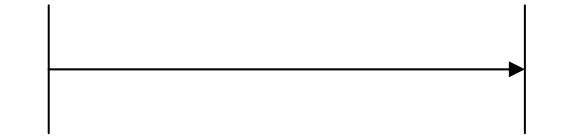
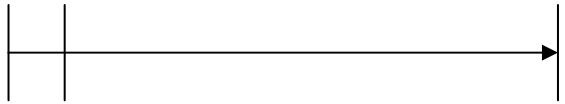
As an example of the use of the Minimum Clinically Important Difference (MCID), Badia, Diez-Perez, Lahoz, Lizan, Nogues, and Iborra(7) calculated the MCID for the Assessment of health related quality of life in osteoporosis, abbreviated ECOS-16. They looked at the change in total scores (difference between scores at baseline and those at 6 months follow-up) of those individuals that indicated on a general health status item that their condition was "slightly better". They found that a change of 0.5 points (approximately a half standard deviation) in ECOS-16 score represented the smallest amount of change that was meaningful to these patients. This is consistent with findings from other studies(85-87). Norman, Sloan, and Wyrwich(87) conducted a systematic review to identify studies that computed an MCID. In all but six out of thirty-eight studies, the MCID was close to one half standard deviation. These researchers, thus, conclude that in most circumstances, the threshold for detecting change in health status is approximately one half standard deviation. Of note, others have found that one standard error of measurement (SEM) is equivalent to an MCID of one half standard deviation(88). However, several authors(85, 89, 90) argue that the MCID is context-specific, dependent on several factors including severity of patients, individual perspectives, disease processes, and method used for determination. Beaton(85) asserts that using the half standard deviation as a rule for the MCID in all cases is "too simple", "loses the individual in the average" and necessitates the acceptance of many exceptions where the rule does not work.

Jacobson, Follette, and Revenstorf(91) emphasize that clinical significant change can be conceptualized as patients entering therapy as part of a population with functional limitations and

leaving as part of one without these limitations. This implies that discharge scores on the variable of interest should fall within (or at least closer than admission scores to) what is considered to be normal(83). The Reliable Change (RC) index was developed to assess statistically significant change that is also clinically meaningful. The RC index assesses whether a patient's score at follow-up is two standard deviations better than the initial score. This degree of improvement is considered to be "true change"(92).

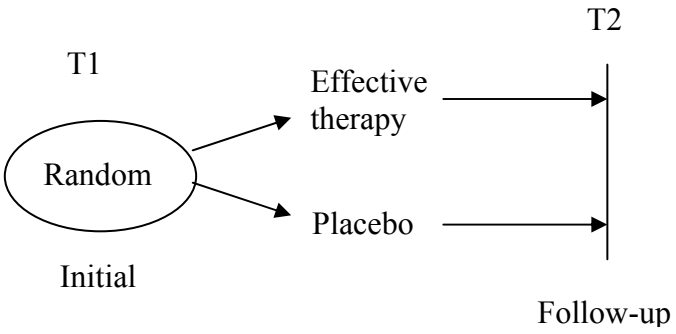
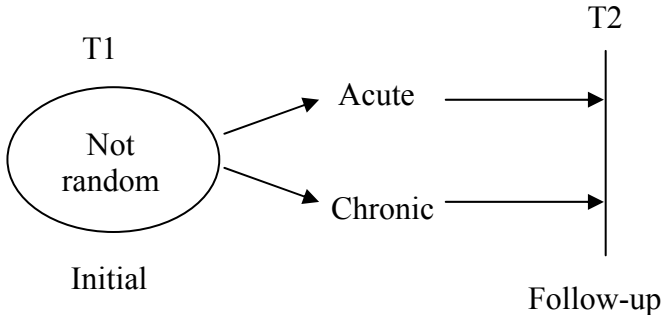
The above presentation of responsiveness opens a new and important area for studying upper-extremity measures. We need to go beyond typical reliability and validity indices in determining the most appropriate measures. The above review is intended to introduce the reader to designs and indices of responsiveness. While requiring further research, they provide us with a means to identify the measures most sensitive in assessing the effectiveness of our treatments.

Table 1-1. Single group designs and coefficients

Design	Appropriate coefficients
T1 T2 Before-after (no baseline) design  Initial Follow-up	ES = Average change / SD of initial scores SRM = Average change / SD of change scores Paired t value = Average change / SD of change scores/√n
Before-after design with a baseline T1 T2 T3  Initial Follow-up	GRI = Average change between T2 and T3 / SD of difference between T1 and T2 scores

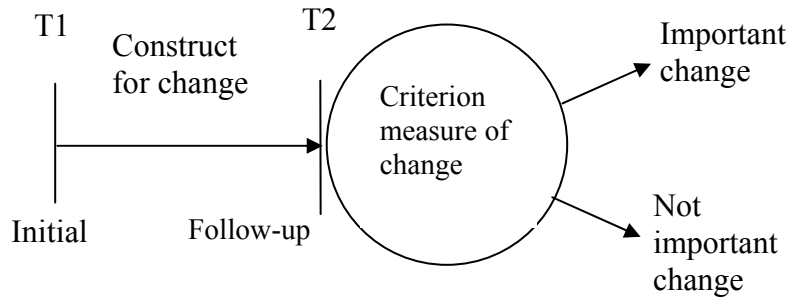
Note: ES = Effect Size, SRM = Standardized Response Mean, GRI = Guyatt's Responsiveness Index, SD = Standard Deviation, n = number of subjects, T = time period

Table 1-2. Multiple-group designs and coefficients

Design	Appropriate coefficients
Treatment and placebo groups  	GRI = Ave Δ T1 to T2/SD of Δ in Stable group t value = Ave Δ in "Important change" group - Ave Δ in "Not important change" group/√Pooled Var x ($1/n_{ic} + 1/n_{nic}$) $F_{(ANOVA)}$ = Between SS/(g-1) / Within SS/(n - g) Norman's Srepeat = Group x Time Var/Group x Time Var + Error Norman's Sancova = Group Var/Group + Error Var ROC curve = plot of sensitivity against 1-specificity
Groups expected to change by different amounts  	GRI = Ave Δ T1 to T2/SD of Δ in stable group t Value = Ave Δ in "Important change" group - Ave Δ in "Not important change" group/√Pooled Var x ($1/n_{ic} + 1/n_{nic}$) $F_{(ANOVA)}$ = Between SS/(g-1) / Within SS/(n - g) Norman's Srepeat = Group x Time Var/Group x Time Var + Error ROC curve = plot of sensitivity against 1-specificity

Note: GRI = Guyatt's Responsiveness Index, Ave = Average, Δ = Change, Var = Variance, n_{ic} = number in "Important Change Group", n_{nic} = number in "Not Important Change" Group, SD = Standard Deviation, g = number of groups, n = number of subjects, SS = sum of squares, T = time period

Table 1-2. Continued

Design	Appropriate coefficients
Groups determined by external criterion  	GRI = Ave Δ T1 to T2/SD of Δ in stable group t Value = Ave Δ in "Important change" group - Ave Δ in "Not important change" group/√Pooled Var x $(1/n_{ic} + 1/n_{nic})$ $F_{(ANOVA)}$ = Between SS/(g-1) / Within SS/(n - g) Norman's Srepeat = Group x Time Var/Group x Time Var + Error Var ROC curve = plot of sensitivity against 1-specificity Correlation coefficient = Spearman's or Pearson's can be used.

CHAPTER 2
RESPONSIVENESS OF DISABILITIES OF THE ARM, SHOULDER, AND HAND (DASH) VERSUS THE UPPER-EXTREMITY FUNCTIONAL INDEX (UEFI)

Introduction

Responsiveness, the ability of assessments to measure change, is critical to outcome measurement in the area of upper-extremity rehabilitation for three reasons. Outcomes measures need to be assessed for reliability, validity and responsiveness before they should be used in 1) research, 2) clinical applications and 3) policy decisions. While substantial evidence exists pertaining to the reliability and validity of upper-extremity assessments, more work is needed in the arena of responsiveness. First, upper-extremity rehabilitation research depends on the existence of responsive assessments. Responsive assessments allow clinical trials to be completed with fewer study participants(72, 76). This is because responsive assessments lead to increased effect size making the same power achievable with fewer participants. The necessity for fewer study participants decreases the overall cost of clinical trials and helps to expedite the availability of treatments that have evidence of their effectiveness. Furthermore, clinical trials often test interventions that may have similar outcomes (e.g., comparing various splinting techniques). Responsive assessments can determine potentially significant yet small differences in change resulting from a treatment(72). Responsive assessments are also essential for clinical applications. Responsive assessments must be available in order for patients and therapists to gauge improvement/decline in function resulting from various treatments. Finally, in order to influence healthcare policies, evidence of treatment effectiveness must be available. This requires assessments with the ability to detect clinically meaningful change(93, 94). Thus, selecting the most responsive measure is essential to upper-extremity rehabilitation.

The Disabilities of the Arm, Shoulder, and Hand (DASH) outcome questionnaire is one of the most widely used and researched assessments in upper-extremity rehabilitation(55, 57, 95-

97). The DASH is a measure of general upper-extremity function, meaning it was not designed specifically for one upper-extremity condition. There are over 47 studies on the psychometrics of the DASH(55, 57, 61, 70, 95, 97, 98). Studies have shown that the DASH has good discriminative validity (differentiates between dissimilar constructs), construct validity (accurately represents a construct), and test-retest reliability (scores can be replicated). These studies have tested the DASH on patients with a wide range of upper-extremity conditions(55, 57-61). For instance, Beaton and colleagues(57) confirmed the test-retest reliability of the DASH (intraclass correlation coefficient = 0.96). These researchers showed that the DASH had good discriminative validity, differentiating between those currently able to work and those unable to work. Convergent validity was also demonstrated with correlations to the Shoulder Pain and Disability Index (SPADI) subscales and the Brigham (carpal tunnel) questionnaire subscales exceeding 0.70.

Relevant to the present study, the DASH has demonstrated the ability to detect clinical change (responsiveness). For instance, Gummesson and colleagues(56) found that the DASH was effective in detecting change in function before and after surgical treatment for upper-extremity conditions (effect size = .70; standardized response mean = 1.2). Effect sizes of 0.2 and below are considered small, 0.5 medium, and 0.8 or higher are considered large(99). The same values hold for standardized response mean. That is, a standardized response mean of 0.2 is considered small, 0.5 is moderate, and 0.8 is large (99). Beaton and Bombardier(57) showed that in a group of patients with diverse musculo-skeletal conditions, the DASH was comparable or better at detecting clinical change than joint-specific assessments. Similarly, Kotsis and Chung(66) also found the DASH to be moderately responsive (standardized response mean = .70) when used to evaluate outcomes after carpal tunnel surgery.

Studies that have compared the DASH to disease specific assessments (as opposed to general upper-extremity function assessments) have resulted in conflicting evidence. Some of these studies have found the DASH to be as responsive as disease specific assessments(57, 67) while others have found disease specific scales to be more responsive(68-70). In support of the DASH as a responsive measure, Beaton and colleagues(57) in a sample of two hundred patients with wrist/hand or shoulder problems, found standardized response means to be equivalent or greater for the DASH compared to two other joint specific assessments (the Shoulder Pain and Disability Index and the Brigham carpal tunnel questionnaire). Likewise, with individuals with carpal tunnel syndrome, Greenslade and colleagues(67) found the DASH to be as responsive as the disease specific Boston questionnaire (BQ), reporting standardized response means of 0.66, 1.07, and 0.62 for the DASH, BQ-symptoms scale and BQ-function scale, respectively. In contrast, in support of disease specific scales being more responsive, MacDermid and colleagues(68, 69) found the Patient-Rated Wrist Evaluation (PRWE) had slightly higher responsiveness than the DASH (standardized response mean=1.51 vs. 1.37) in a cohort of 60 patients (36 hand problems, 24 wrist problems). Similarly, Gay and colleagues(70) found the Carpal Tunnel Questionnaire to be more responsive than the DASH in a sample of 34 patients after having carpal tunnel release (Carpal Tunnel Questionnaire: effect size = 1.71, standardized response means=1.66; DASH: effect size=1.01, standardized response mean=1.13).

However, there is a lack of studies comparing the responsiveness of the DASH to other self-report assessments of general upper-extremity function (those intended for any upper-extremity ailment). Despite its widespread use, the DASH may be similar or better than other self-reports at measuring change. A similar, though less researched, measure of general upper-extremity function is the Upper-Extremity Functional Index (UEFI)(71). Like the DASH, the

UEFI is a self-report assessment where individuals rate how much difficulty they have performing tasks relating to general upper-extremity function. Items on the UEFI relate to daily activities such as grooming your hair, opening doors, and cleaning. However, the UEFI is shorter than the DASH (20 items versus the 30 items on the DASH) and does not include pain items, making it more appropriate for individuals who are not experiencing the symptom of pain with their injury. Most patients complete the UEFI in three to five minutes and scoring takes only approximately 20 seconds. The UEFI has been shown to have high test-retest reliability (r = .95)(71). Additionally, the UEFI has been shown to be similar to the other upper-extremity assessments (the Upper-Extremity Functional Scale, UEFS) in its ability to detect change in function (correlation between UEFI and the UEFS change scores greater than .60)(71). If the UEFI is equivalent or better than the DASH in its ability to detect change, the quick completion and scoring times of the UEFI would make it ideal for use in the clinic.

Because these two assessments are attempting to measure the same construct (i.e., upper-extremity function) and are using similar items and rating scales (response categories) to do so, it is likely that they are similar in their ability to measure patient change. The purpose of the current study is to compare the responsiveness of the DASH and UEFI. The specific aims are: 1) compare person measure change for the DASH and UEFI, 2) identify total score change on each assessment that indicates a change in functional ability, 3) compare overall error in predicting clinically important change for the two assessments (using area under Receiver Operating Characteristic, ROC, curves), and 4) determine the correlation between person measures on these assessments and a global patient reported measure of change. The specific hypotheses are: 1) The difference between the person measure change on the DASH and UEFI will not be significant. 2) Total score change that indicates a change in upper-extremity status will be similar

for the DASH and the UEFI. 3) Overall error in predicting clinically important change (determined using area under the ROC curves) will be similar for these two assessments. 4) Correlations between person measure change on these assessments and global patient reported change will be high and similar for both assessments.

Methods

Sample

Data was collected at various outpatient clinics throughout the United States by Focus on Therapeutic Outcomes (FOTO), Inc. FOTO collects outcomes data on various assessments of physical functioning and provides outcomes reports for rehabilitation facilities. Through the use of this assessment package a large amount of data is generated for research purposes(100, 101). Patients who had completed the assessments at two time points (admission and discharge) were selected from the larger data sets. This reduced the original DASH data set from 2487 study participants to 960. For the UEFI, the original data set was reduced from 3949 to 1953. These data sets were further reduced to one data set, which included only participants that had completed both the DASH and the UEFI. This resulted in the final data set of 214. However, for analyses that required the global rating, the data set was further reduced to 178 because of missing global rating data. Two hundred and fourteen patients completed both the DASH and UEFI. Fifty one percent of the 214 patients sample were female. The mean age was 49.0 $\pm$ 14.3. Study participants most commonly experienced shoulder impairments (59%) followed by those with neck impairments (27%) and the least number of individuals with forearm impairments (1%). About 70% of the study participants had chronic conditions. Mean number of days seen in rehabilitation was 52.5 $\pm$ 33.5 and total number of treatment sessions was 14.5 $\pm$ 10.2 (See Table 2-1).

Materials

The DASH was designed to measure impairment based on patient report, as well as, to capture limitations in activities and participation imposed by single or multiple disorders of the upper-extremity(102). It is based on the World Health Organization (WHO) Model of Health, at the time of development called the International Classification of Impairments, Disabilities, and Handicaps (ICIDH) but today revised to be the International Classification of Functioning, Disability, and Health (ICF)(103). The DASH is a standardized instrument and evaluates impairments and activity limitations, as well as, participation restrictions for both leisure and work(103). It consists of two components: 30 disability/symptom questions and four optional high performance sport/music or work questions(104). The DASH includes both disability and symptom items. Examples of disability items on the DASH include: Open a tight or new jar, Write, Make a bed, and Use a knife to cut food. While examples of symptom items include: Arm, shoulder or hand pain, Weakness in your arm, shoulder or hand, and Stiffness in your arm, shoulder or hand. Response choices for disability items range from 1 to 5 (1: no difficulty to 5: unable)(104). Response choices for symptom items also use a five-point scale, but varying from none to extreme. At least 27 of the 30 disability/symptom questions must be completed for a score to be calculated(104). Typically, in order to produce a score out of one hundred, the response choices are summed and averaged, producing a score out of five(104). This value is then transformed to a score out of 100 by subtracting one and multiplying by 25(104). For example, if the patient responded to all the questions with a 5, they would have an average score of 5. You would then subtract one from 5, giving you 4 and multiply by 25. This individual, with the highest responses possible, would have a score of 100. Normally, the DASH produces scores between 0 and 100 with a higher score indicating more disability(104). For this study, however, the rating scale was recoded such that a higher score indicated more ability.

Additionally, when change scores were calculated total scores for admission and discharge were used as they were and were not converted to scores out of 100. Thus, making the range of possible scores on the DASH zero to 150. Appendix A includes a list of items on the DASH.

The Upper-Extremity Functional Index (UEFI) is also based on the International Classification of Functioning, Disability, and Health (ICF)(71). The UEFI is a self-report assessment containing twenty items assessing general upper-extremity function. Examples of items on the UEFI include: Grooming your hair, Dressing, Opening doors, Cleaning, and Throwing a ball. Participants' responses are on a 5-point scale (0: extreme difficulty or unable to perform activity to 4: no difficulty). Total UEFI scores are obtained by adding the responses to all questions, thus, total scores vary from 0 to 80 with higher scores indicating greater functional status level. Most people can complete the UEFI in three to five minutes and scoring takes about 20 seconds. Appendix B includes a list of items on the UEFI.

Procedures

When the data was obtained, in the FOTO data set response choices on the DASH had been recoded so that a higher number response indicated more ability, thus, 5 = no difficulty and 1 = unable. Additionally, in the FOTO data set the response scale on the UEFI was recoded to be more consistent with the DASH response scale. Zero was recoded to 1, 1 to 2, 2 to 3, 3 to 4, and 4 to 5. Because of the recoding of the rating scale, for this study total scores on the UEFI varied from 20 to 100. Data extraction was performed using the Statistical Package for the Social Sciences (SPSS)(105).

Statistical Analysis

Rasch analysis to obtain comparable measures

Since the DASH and UEFI have different rating scales and ranges, Rasch analysis, a one-parameter item response theory model, was used to place measures from the two instruments on

the same scale. A co-calibration analysis in which DASH and UEFI admission data were run simultaneously was conducted using Winsteps(1). Then separate analyses were run for DASH admission data, DASH discharge data, UEFI admission data, and UEFI discharge data anchoring the items and rating scale at the estimates obtained in the co-calibration analysis. Person measures (in logits) obtained during these analyses were used to obtain person measure change scores (person measure at discharge minus person measure at admission) for both the DASH and the UEFI. These person measure change scores were used in all of the analyses except for the sensitivity/specificity analyses. It was felt that since the sensitivity/specificity analyses would help to identify a cutoff score for evaluating whether or not change had occurred, it would be more useful for clinicians to use the actual assessment change scores than calculated person measure change scores.

Statistical analyses included calculating ANOVA, sensitivity/specificity, and correlations. ANOVA was calculated for the DASH and UEFI using a single group design (e.g., all DASH data was treated as a single group). For both the sensitivity/specificity and the correlational analyses, two patient groups were created based on global rating scores. For this global rating, patients rated the perceived amount they had changed since admission on a scale of -7 to +7 with a negative number indicating that the condition had worsened since admission, zero indicating no change, and positive numbers indicating varying degrees of improvement (See Appendix C). Based on previous literature(73, 106), for the current study, people with scores below +3 were considered to be different from those whose scores were +3 and above. Therefore, the groups were divided at +3, with individuals reporting a global rating of change of +3 and above placed in the "improved" group and those with less than +3 placed in the "not improved" group.

However, of note, the "not improved" group also contained those who rated their condition as being worse at discharge.

Analysis of variance (ANOVA)

A within subjects ANOVA was conducted using SPSS(105). The goal of this analysis was to determine if there was a significant difference between change in person measures on the DASH and change in person measures on the UEFI. In addition to allow further comparison, mean person measure change, range, and standard deviations were reported for the two assessments and a Pearson correlation between person measure change scores on the DASH and UEFI was conducted in SPSS(105).

Sensitivity and specificity

The sensitivity and specificity analysis was performed to identify an assessment change score that indicates a difference in functional ability. Sensitivity is defined as the ability of an assessment tool to detect a condition when it is present or to provide true positive results(107, 108). Specifically, in this study, sensitivity was calculated as the number of patients correctly identified (based on the groups created from their global ratings) by the DASH or UEFI as undergoing an important change divided by the number of patients whose global rating indicated they had actually changed (positive three or above). In contrast, specificity is defined as the ability of an assessment tool to not detect a condition when the condition does not exist, or to provide true negative results(107, 108). In the current study, specificity was calculated as the number of patients who were correctly identified by the DASH or UEFI as not undergoing an important change divided by the number of patients who had actually not changed based on their global ratings. The best cutoff point (i.e., the number of points on the DASH or UEFI that an individual's total assessment score must change to indicate a difference in functional ability) was

identified on the basis of lowest overall error rate, specificity, and sensitivity. This overall error rate was calculated using the formula [(1 – sensitivity) + (1 – specificity)]. If cutoff values had a similar error rate, the cutoff value with higher specificity was chosen, as it was considered better to err on the side of continuation of services versus discharging a patient too soon. Sensitivity, specificity, and overall error were calculated at thirty point intervals of change on the DASH from –102 points of change (minimum observed change) to 109 points of change (maximum observed change) and on the UEFI at 15 point intervals of change from –75 to 77. The optimal combination of sensitivity and specificity was confirmed by generating a Receiver Operating Characteristic (ROC) curve. This type of curve plots sensitivity on the Y-axis against 1 – specificity on the X-axis. One minus specificity represents the false positive rate. Figure 2-1 depicts how specificity and sensitivity were calculated. Areas under the ROC curves (or indices of discrimination, that is, sensitivities) were calculated by counting the number of squares beneath the curve and dividing by the total number of squares (the number of squares above the curve plus the number below). The greater the area under the ROC curve (AU), the less the overall error in predicting clinically important change. The AU for the DASH was compared to the area calculation for the UEFI.

Correlation

Correlation is another commonly used statistic for measuring responsiveness. Person measure change from admission to discharge on the DASH and UEFI were correlated with patient global rating scores(73). The magnitude of the correlations obtained for the DASH and the UEFI were compared. Values for correlations commonly considered strong are .60 and larger, moderate are between .40 and .60, and weak are below .40(109). Once again, since all of the individuals in the sample did not complete the global rating of change, the sample sizes were 178 for the DASH and 178 for the UEFI.

Results

Contradicting the first hypothesis, ANOVA between person measure change scores on the DASH and the UEFI was significant (F = 12.47, p=.001). This indicates that the person measure change scores on the two assessments significantly differ. However, the mean person measure change was only slightly larger for the UEFI (DASH mean person measure change = .80, UEFI mean person measure change = 1.1), the range was wider for the UEFI but comparable (for DASH –6.45 to 8.85, for UEFI –7.08 to 9.20), and the standard deviation was slightly larger for the UEFI (DASH standard deviation of person measure change scores = 2.30, UEFI standard deviation of person measure change scores = 2.67). Finally, the Pearson correlation between the change scores was .90, p = .000. Thus, the second hypothesis was upheld in that person measure change, indicating an alteration in ability and symptoms, was similar for the DASH and UEFI.

For the DASH, the cutoff point with the best combination of sensitivity and specificity was twenty (sensitivity = .52, specificity = .79, overall error rate = .69). For the UEFI, the cutoff point with the best combination of sensitivity and specificity was fifteen (sensitivity = .45, specificity = .77, overall error rate = .78). Table 2-2 shows sensitivity, specificity, and overall error rate for the cutoff values for the DASH and UEFI. This table illustrates the compromise between sensitivity and specificity (as sensitivity increases specificity decreases and vice versa).

Correspondingly, in support of hypothesis three, overall error in predicting clinically important change (determined using area under the ROC curves) was similar for the two assessments. Areas under the ROC curves were similar for the DASH and UEFI. For the DASH the area was approximately .67, while for the UEFI the area was approximately .65 (Figure 2-2). The area under the ROC curve is an index of the degree of separation between the distributions of true positives vs. false alarms(110). The almost identical areas under the curves for the DASH and UEFI indicate that the sensitivity of these two assessments is comparable.

Finally, correlational analyses confirmed hypothesis four. Correlations between patients' global ratings and person measure change scores from intake to discharge were similar for the two assessments. For the DASH, the correlation between the global rating scores and person measure change scores was $r = .33$, while for the UEFI the correlation was $r = .35$. While both correlations reached standards that could only be considered small (i.e., less than $r = .40$)(57), they were significant at the $p = .01$ level.

Discussion

The results of the current study indicate that for outpatients with upper-extremity problems, the DASH questionnaire and the UEFI measure patient change in upper-extremity condition very similarly. A within subjects ANOVA revealed no significant difference between person measure change between the two assessments. Additionally, areas under the ROC curves were almost identical for the two assessments (.67 and .65), as were correlations between global ratings and the change scores ($r = 0.33$ and 0.35). While both assessments identify change in patient function, since the responsiveness calculations for both assessments are very similar, there appears to be no real advantage of one instrument over the other in detecting change.

One reason that these assessments are similar in measuring change may be due to the similarity of the items making up these assessments (See Appendices A and B). For example, both have items that involve lifting objects above one's head and both have the item opening a jar lid. Furthermore, the DASH and the UEFI both include comparable items involving grooming, food preparation, and dressing. For example, the DASH includes the item "Wash or blow dry your hair", while the UEFI includes the item "Grooming your hair". Likewise, the DASH has the item "Use a knife to cut food" and the UEFI has the similar item "Preparing food (e.g., peeling, cutting)". The DASH has the item "Put on a pullover sweater" while the UEFI contains a broader item "Dressing". Because the items on these assessments are comparable, it

is likely that these two assessments discriminate between people of different abilities in a similar way.

No studies have investigated the responsiveness of the DASH and UEFI using methods comparable to the ones used in the current study. That is, no other studies have used item response theory methodologies to convert scores on these assessments to person measures and equated the two rating scales and score ranges before comparing them. However, in the literature that exists on the DASH, the findings have been comparable to those of the current study(56, 66, 111). For example, Beaton and colleagues(57) found correlations between DASH change scores and global measures of change to be r = .40 (Pearson) and R = .43 (Spearman). Also, these researchers found the area under the curve to be approximately 70%. They calculated standardized response means to be .74-.80.

The ROC curves generated in this study demonstrate the trade-off between specificity and sensitivity. A larger total change score on the DASH or UEFI (more stringent criteria) yields a smaller sensitivity and greater specificity, and vice versa; a smaller total change score (less stringent criteria) yields greater sensitivity and smaller specificity. Clinicians who use assessment change scores need to make a compromise of sorts when deciding on what type of error they are willing to make. By using a lower total score change, clinicians increase their chance of assuming a patient has changed when they have not. Conversely, using a greater total score change increases the chance of assuming a patient has not changed and continuing to treat the condition after their functional ability has returned to the desired state. Portney and Watkins(112) state that "those who use a screening tool must decide what levels of sensitivity and specificity are acceptable . . . Sensitivity is more important when the risk associated with missing a diagnosis is high, as in the case of life-threatening disease or a debilitating deformity;

specificity is more important when the cost or risks associated with further intervention are substantial." In the case of assessment change scores, it could be argued that it was better to continue to treat a patient longer than actually needed, than to discharge them too early.

Of particular concern with the findings of this study is that the correlation between the global rating and the person measure change scores was in the range of what is considered weak. Additionally, the ROC curve was really more of a line, indicating low probabilities of specificity and sensitivity, i.e., 50% chance. Thus, there is a problem with not having a "gold standard" to compare the change scores to. In essence this study is comparing two patient reported outcomes. It tends to imply that patients are not accurate in reporting the amount of change that has occurred in their functional ability. The authors felt that using total assessment change scores to calculate the best cutoff score would be more meaningful for therapists (i.e., give them an idea what actual amount of change in total score on the assessment would indicate that a patient had changed). However, since the DASH and UEFI are not on the same scale and have a different range of possible scores, this analysis is limited in the conclusions that can be drawn about the similarity of the two assessments.

There are several limitations pertaining to the statistical methods used in the calculation of responsiveness. A major disadvantage when there is no external criterion is that if a change is not detected, it is not clear whether the measure was unable to detect a change or whether the patients did not undergo the expected change(73). Correspondingly, if a statistically significant change is detected, it is uncertain whether this is due to some property of the measure (i.e., the measure is so sensitive that it has picked up a change that is so small that it is not functionally significant) or if a significant change in function was detected. This problem is somewhat solved by the use of an outside criterion (such as the global rating in the current study), as is done with

the ROC curve analysis and the correlation(73). However, if the outside criterion is not valid the results obtained will be misleading. It appears this was the case in the current study. The global rating was a very poor "gold standard" and thus, results of this study may be of questionable validity.

An additional limitation involves the sample used in this study. While large, it was a sample of convenience. Moreover, the size of the sample was much reduced once individuals who did not complete the assessments at discharge and those that did not complete the patient global rating were eliminated. However, based on previous analyses conducted by the investigators, it appeared that the reduced sample was comparable to the larger sample on all demographic characteristics. Finally, the data was collected at various clinics across the country and from patients with diverse upper-extremity diagnoses. Therefore, situations in the clinics and treatments used by therapists are likely to have differed. While this increases the generalizability of the results, it is unclear what treatments were used with which patients and contributed to the most advantageous outcomes.

Conclusions

The results of this study seem to indicate that neither the DASH nor UEFI has a clear advantage over the other when measuring responsiveness. Thus, if time is an issue, perhaps, the shorter UEFI is a better choice, even though it is a lesser known instrument. However, if comparison outcomes data is needed or communication of outcomes with other therapists is desired, the DASH may be preferred since it is the more widely used and studied. Additionally, when change scores on assessments are used careful consideration should be given to the values used (i.e., the degree of sensitivity vs. specificity desired). Further research is needed using additional samples and investigating responsiveness methods. Moreover, there is a need for the development of a "gold standard" for sensitivity and specificity analyses. Research focusing on

how to create more responsive assessments is needed so that guidelines might be established to aid future assessment development.

Table 2-1. Sample of 214 demographics

Mean age ± SD	49.0 ± 14.3 (N = 202)
Gender	
- Female	108 (50.5%)
- Male	106 (49.5%)
	(N = 214)
Primary body part injured	
- Shoulder	120 (59.4%)
- Neck	55 (27.2%)
- Elbow	9 (4.5%)
- Upper Arm	6 (3.0%)
- Wrist	6 (3.0%)
- Hand	4 (2.0%)
- Forearm	2 (1.0%)
	(N = 202)
Symptom acuity	
- Chronic	142 (70.3%)
- Subacute	41 (20.3%)
- Acute	19 (9.4%)
	(N = 202)
Mean number of days treated ± SD	52.5 ± 33.5 (N = 157)
Mean number of treatment sessions ± SD	14.5 ± 10.2 (N = 157)

Note: Many of the totals presented in table are not the same as the sample total due to missing data.

		Test (Person measure change score on DASH or UEFI)	
Global Rating (GR)	+ (GR ≥ 3)	A (True positive) (Score above cutoff)	c (False negative) (Score below cutoff)
	- (GR < 3)	B (False positive) (Score above cutoff)	d (True negative) (Score below cutoff)

Figure 2-1. Calculating sensitivity and specificity where sensitivity = a/(a +c) and specificity = d/(b+d)

Table 2-2. Sensitivity, specificity, and overall error at various cutoff points for the DASH and UEFI

DASH											
Cutoff value	-102	-70	-40	-10	20	50	80	109			
Sensitivity	0.99281	0.98561	0.93525	0.76978	0.51799	0.15827	0.05755	0			
Specificity	0	0.05128	0.12821	0.48718	0.79487	0.84615	1	1			
Overall error rate	1.00719	0.96311	0.93654	0.74304	0.68714	0.99557	0.94245	1			
UEFI											
Cutoff value	-75	-60	-45	-30	-15	0	15	30	45	60	77
Sensitivity	1	0.98561	0.96403	0.94964	0.88489	0.71223	0.45324	0.27338	0.13669	0.05036	0
Specificity	0	0.02564	0.07692	0.17949	0.35897	0.46154	0.76923	0.84615	0.87179	0.97436	1
Overall error rate	1.000	0.989	0.959	0.871	0.756	0.826	0.778	0.880	0.992	0.975	1.000

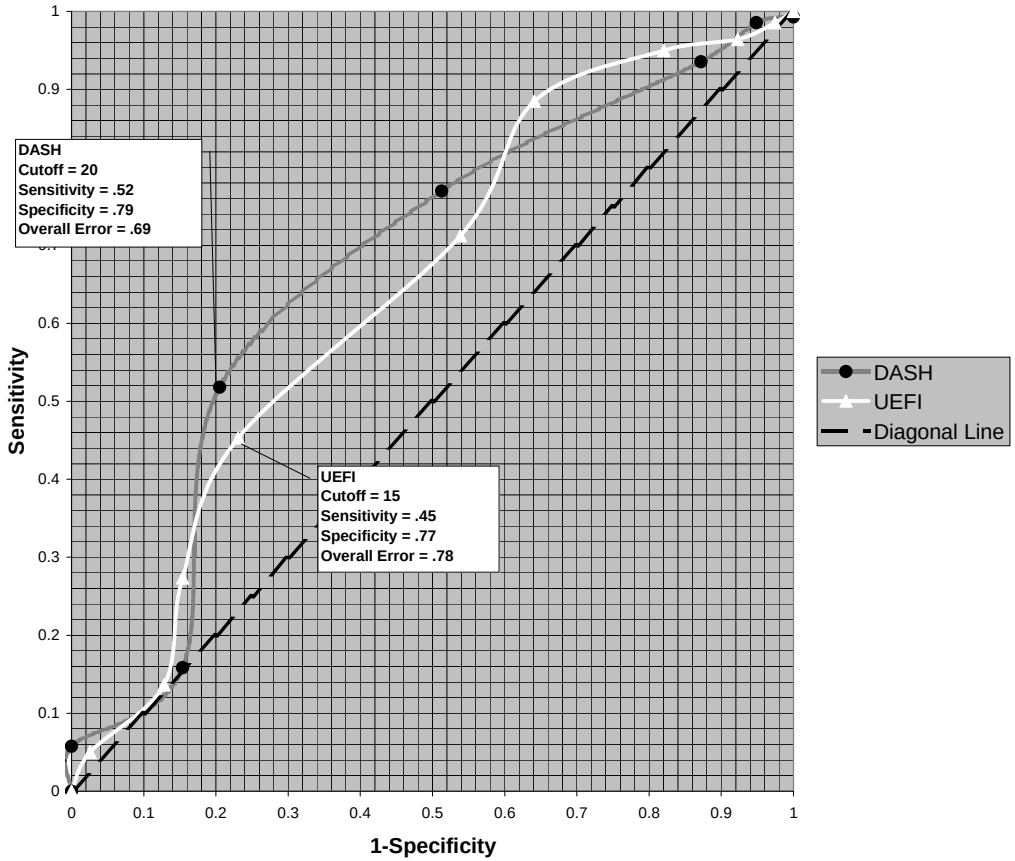


Figure 2-2. Receiver operating characteristic (ROC) curves for the DASH (circles) and UEFI (triangles), showing the accuracy of different cutoff points (diagonal line represents assessment with 50/50 chance of detecting improvement*)*

CHAPTER 3
ITEM CHARACTERISTICS OF THE DISABILITIES OF THE ARM, SHOULDER, AND HAND (DASH) OUTCOME QUESTIONNAIRE

Introduction

One of the most widely used and researched assessments in upper-extremity rehabilitation is the Disabilities of the Arm, Shoulder, and Hand (DASH) outcome questionnaire(55, 57, 95-97). There are over 47 studies on the psychometrics of the DASH(55, 57, 61, 70, 95, 97, 98). Studies have shown that the DASH has good discriminative validity (differentiates between dissimilar constructs), construct validity (accurately represents a construct), and test-retest reliability (scores can be replicated). These studies have tested the DASH on patients with a wide range of upper-extremity conditions(55, 57-61). For instance, Beaton and colleagues(57) confirmed the test-retest reliability of the DASH (intraclass correlation coefficient = 0.96). These researchers found that the DASH showed discriminative validity, differentiating between those currently able to work and those unable to work. Convergent validity was also demonstrated with correlations to the Shoulder Pain and Disability Index (SPADI) subscales and the Brigham (carpal tunnel) questionnaire subscales exceeding 0.70.

The DASH has demonstrated ability to detect clinical change (responsiveness) at the level of the assessment as a whole. For instance, Gummesson and colleagues(56) found that the DASH was effective in detecting change in function before and after surgical treatment for upper-extremity conditions (effect size = .70; standardized response mean = 1.2). Effect sizes of 0.2 and below are considered small, 0.5 medium, and 0.8 is considered large(99). Beaton and Bombardier(57) showed that in a group of patients with diverse musculo-skeletal conditions, the DASH was comparable or better at detecting clinical change than joint-specific assessments. Similarly, Kotsis and Chung(66) also found the DASH to be moderately responsive (standardized response mean = .70) when used to evaluate outcomes after carpal tunnel surgery.

Despite the wealth of studies examining the properties of the DASH as a whole, no published analyses have thoroughly investigated the item characteristics of the DASH or how these item characteristics affect the ability of this assessment to measure change in patient condition. Item response theory (IRT) methodologies were used to eliminate DASH items in the creation of a Quick-DASH version of the instrument, however, in this study the only item-level statistic reported was misfit(97). Knowledge of item-level psychometrics are important for a number of reasons(113-115). First, using item response theory methodologies, the extent to which all items on a test represent the same underlying trait can be examined. Additionally, IRT analyses provide item and person ability estimates that are independent of the sample of individuals and the set of items used(114). Item difficulty estimates allow for a comparison of the challenge of items making up an assessment relative to each other. IRT allows the investigation of the equivalence of items for different samples or for the same sample at different time points(113). Estimates of item discrimination derived from item response theory analyses indicate an item's ability to distinguish between individuals who possess different amounts of the trait of interest(114). Furthermore, information curves show at which person ability levels the test is most useful. Thus, the purpose of the present study was to examine item-level characteristics of the DASH including: 1) unidimensionality of the items (i.e., do they all represent the same construct), 2) item difficulty hierarchies at admission and discharge, 3) person-item match, 4) differential item functioning (DIF) from admission to discharge, 5) item discriminations at admission and discharge, and 6) comparison of test information curves produced from admission and discharge data.

Methods

Sample

Data was collected at various outpatient clinics throughout the United States by Focus on Therapeutic Outcomes (FOTO), Inc. FOTO provides a patient evaluation tool and aggregate data management service. The Disabilities of the Arm, Shoulder, and Hand (DASH) outcome questionnaire is one of the many assessments on which FOTO collects data. FOTO has accumulated the largest external database for rehabilitation outcomes. It is included on the Joint Commission's (JCAHO) list of acceptable performance measurement systems. Large amounts of data collected by FOTO is used for research purposes(100, 101).

Data extraction was performed using the Statistical Package for the Social Sciences (SPSS)(105). Patients who had completed the assessments at two time points (admission and discharge) were selected from the larger data sets. This reduced the original DASH data set from 2487 study participants to 991. For comparison, demographic information on the larger and the reduced data sets is presented in Table 3-1.

Nine hundred and ninety-one individuals with upper-extremity conditions from various outpatient clinics throughout the United States completed the DASH. These individuals completed the assessments at two different time points (i.e., admission and discharge) as part of standard treatment procedures. Fifty-seven percent of the sample was female. Study participants most commonly experienced shoulder impairments (51%) followed by those with neck impairments (20%) and the least number of individuals with forearm impairments (1%). About half of the study participants had chronic conditions. Mean number of days seen in rehabilitation were 48.6 ± 32.2 and total number of treatment sessions were 13.3 ± 10.0. See Table 3-1.

Assessment

The Disabilities of the Arm, Shoulder and Hand (DASH) outcome questionnaire was designed to measure impairment subjectively, as well as, to capture limitations in activities and participation imposed by single or multiple disorders of the upper-extremity(102). It is based on the World Health Organization (WHO) Model of Health, at the time of development called the International Classification of Impairments, Disabilities, and Handicaps (ICIDH) but today revised to be the International Classification of Functioning, Disability, and Health (ICF)(103). The DASH is a standardized instrument and evaluates impairments and activity limitations, as well as, participation restrictions for both leisure and work(103). It consists of two components: 30 disability/symptom questions and four optional high performance sport/music or work questions(104). The DASH includes both disability and symptom items. Examples of disability items on the DASH include: Open a tight or new jar, Write, Make a bed, and Use a knife to cut food. Examples of symptom items include: Arm, shoulder or hand pain, Weakness in your arm, shoulder or hand, and Stiffness in your arm, shoulder or hand. Response choices for disability items range from 1 to 5 (1: no difficulty to 5: unable)(104). Response choices for symptom items also use a five-point scale, but varying from none to extreme. At least 27 of the 30 disability/symptom questions must be completed for a score to be calculated(104). In order to produce a score out of one hundred, the response choices are summed and averaged, producing a score out of five(104). This value is then transformed to a score out of 100 by subtracting one and multiplying by 25(104). Thus, an individual, with the highest responses possible, would have a score of 100. Normally, the DASH produces scores between 0 and 100 with a higher score indicating more disability(104). For this study, the rating scale was recoded such that a higher score indicated more ability. Appendix A includes a list of items on the DASH.

Procedures

Data analysis

Unidimensionality

In order to create a legitimate measure, all the items must contribute to the same construct(116). To test for unidimensionality, exploratory and confirmatory factor analyses (weighted least square parameter estimates using a diagonal weight matrix with standard errors and mean- and variance-adjusted chi-square test statistic that use a full weight matrix) were conducted using Mplus software(117). Polychoric correlations were factored in both exploratory and confirmatory analyses. For the exploratory factor analysis (EFA), eigenvalues for each factor, loadings on each factor, scree plots of eigenvalues, total variance associated with the factors, and correlations between the factors were investigated. Eigenvalues represent the variances extracted by the factors. Using the Kaiser criterion, only factors with eigenvalues greater than one should be retained (i.e., those factors that extract at least as much of the variance as one of the original variables)(105). The scree plot shows the eigenvalues for each factor plotted in descending order. The scree test proposes that where the line in the plot levels off is the cutoff for the number of factors to retain(105). The greater the amount of total variance accounted for by the factor, the more evidence there is for its presence(118). High correlations between factors suggest that they may be measuring the same latent trait(118).

Based on results of the EFA (number of factors present), a confirmatory factor analysis (CFA) was conducted. CFA examined model fit statistics including chi-square, Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), Standardized Root Mean Square Residual (SRMR), and Root Mean Square Error of Approximation (RMSEA). Total amount of variance was also reported. Goodness of fit statistics such as the chi-square, CFI, TLI, SRMR, and RMSEA test the model fit for a preset number of factors. With the chi-square statistic the null hypothesis

tested is that there is no difference between the data and the model being tested, while the alternative hypothesis is that the data do not support the model. Therefore, a significant chi-square indicates that the model does not fit. However, the chi-square statistic is affected by sample size. With large samples, it is difficult to show model fit with this statistic(119). For CFI, TLI, SRMR, and RMSEA there is some disagreement as to what values indicate that the model fit the data. For the current study, as suggested by Hu and Bentler(120) the values of CFI and TLI greater than 0.95 and RMSEA and SRMR below 0.06 were used.

Additionally, using Rasch analysis, the extent to which items contribute to a unidimensional construct was evaluated employing mean square standardized residuals (MnSq) produced for each item on the instrument. MnSq represents observed variance divided by expected variance(121). Consequently, the desired value of MnSq for an item is 1.0. The acceptable criteria for unidimensionality depends on the intended purpose of the measure and the degree of rigor desired. Linacre(122) and Green(123) suggest that criteria for fit statistics depends on sample size. For sample sizes close to a thousand, reasonable ranges of MnSq fit values between 0.6 and 1.1 associated with standardized Z values < 2.0 are recommended. A low MnSq value (i.e., <0.6) suggests that an item is failing to discriminate individuals with different levels of ability (i.e., people with different amount of difficulty performing upper-extremity tasks) or that an item is redundant (i.e., other items on the instrument are of similar difficulty). High values (i.e., >1.1) indicate that scores are variant or erratic, suggesting that an item does not belong with the other items on the same continuum or that the item is being misinterpreted. Items with high MnSq values represent a threat to validity and thus, are given greater consideration. Fit statistics alone should not be used to assess the dimensionality of an instrument(124, 125). Thus, results from fit and factor analyses were considered together.

As an even further test, correlations between each item and the entire instrument (point measure correlations) are often used in determining dimensionality(126). However, according to Wright (126), there are no set rules for what is an acceptable range for point measures. Negative values indicate that observed responses to the item are in opposition to the general meaning of the test but this is all that is certain. Other authors, Allen and Yen(127), recommend these correlations range from .30 to .70 with too high values seeming to suggest that the item is redundant and adds no more knowledge than is related by other items.

Item difficulty

Based on theories of upper-extremity recovery, such motor control theory(128), hypotheses were developed about the hierarchy of difficulty of the items on the DASH. For example based on the concept from motor control theory that complex items are more difficult, doing heavy household chores (e.g., wash walls, wash floors) was hypothesized to be harder than putting on a pullover sweater. Likewise, this theory contends that more challenging activities require the use of multiple joints. Thus, putting on a pullover sweater was hypothesized to be more difficult than turning a key in a lock. Item response theory methodologies provide a means to investigate the validity of a proposed hierarchy by computing item calibrations. Data was fit to a one parameter Rasch model using the Winsteps software program(1) to calculate item difficulty parameters. Separate analyses were done utilizing admission and discharge data to determine the stability of item calibrations across the two time points.

Person-item match

To further investigate the relationship between admission item difficulties and discharge item difficulties, person-item maps were produced using Winsteps(1). These maps plot person ability and item difficulty on the same scale. This illustrates how well the items are suited for the particular sample under investigation.

Differential item functioning

Differential Item Functioning (DIF) analysis was used to compare admission and discharge item difficulties. If DIF is found, this implies that the item is acting differently at admission than at discharge(129). If this is the case, change scores calculated from admission to discharge may be misleading. That is, if the order of difficulty of the items changes from admission to discharge, it would be analogous to the marks on a ruler moving from one ruler to the next ruler. Thus, it would be impossible to subtract a value associated with a mark on the first ruler from the value of another mark on the second ruler and obtain a meaningful value.

Many approaches are available to determine if items are acting differently from one time point to another(130). In this study, uniform DIF was calculated using Winsteps(1). Because the Rasch model was used, investigation of non-uniform DIF was not possible(130). The Winsteps program utilizes a relatively direct DIF procedure in which each item's difficulty is compared between time points using iterative t-tests(131). Different alpha levels have been suggested for this analysis(132, 133). Some authors have suggested using a critical value of 1.96 to detect differences in item estimates ($\alpha = .05$)(132). However, using this critical value does not correct for doing multiple comparisons, which increases the likelihood of falsely detecting significant differences(133). Acknowledging that small differences may occur by chance, a Bonferroni adjustment was made based on the number of items on the DASH (.05/30, $\alpha = .0017$). As suggested by Crane and Hart(133), both small DIF ($\alpha = .05$) and large DIF ($\alpha = .0017$) were reported. For the purposes of this study, an alpha level of .05 was considered significant while an alpha level of .0017 was considered highly significant. Nevertheless, some investigators(130) contend that when a large sample is used ($N \geq 200$) a statistically significant result may have no practical meaning. Thus, a 0.5 logit difference should be used to identify DIF items. DIF based

on these guidelines is also reported. Because the presence of DIF does not always indicate that an individual's ability is being incorrectly assessed(134), average person ability estimates derived from the analysis of versions of the DASH with and without DIF items were compared.

DIF was also portrayed using a scatter plot of item difficulty estimates. A fit of the item difficulty calibrations within the 95% confidence interval affirms item constancy(135). Items whose calibrations fell outside the 95% confidence interval were considered statistically distinct. Additionally, admission and discharge item difficulties were compared using an intraclass correlation coefficient (ICC; one-way random, 95% confidence interval, test value: 0). This type of correlation provides an indication of the reproducibility of item difficulty calibrations(136).

Item discrimination

Item discrimination gives an indication of how responses to a particular item correspond to responses to the overall assessment(137, 138). In general, high values for item discrimination are considered good, indicating that responses to that particular item reflect a similar construct as the other items on the test(138). Zero or negative discrimination values indicate that the item does not fit well with the rest of the items(137). Winsteps(1) was used to estimate item discriminations for admission and discharge data. Values obtained were examined to ensure that none were zero or negative, indicating poor items. In order to investigate the effect of item discrimination on person ability estimates, mean person ability estimates calculated utilizing a one parameter model (using Winsteps(1) software) and a two parameter model (MLE or MAP computation, graded model using Multilog(139) software) were compared. Maximum likelihood estimation (MLE) is a special case of Maximum a Posteriori (MAP) estimation(140). In MLE the most probable estimate is calculated based on the data(140). Because Winsteps estimates are in logits and Multilog estimates are in probits, estimates from the Multilog analysis were converted to logits using the conversion ratio, 1 logit = 1.7 probits(1), before comparison.

Test information

Information curves for the entire test at both admission and discharge were obtained using Winsteps Rasch analysis program(1). Information using the Rasch one parameter model (Winstesps), is simply the probability of passing an item times the probability of failing an item(141). The standard error is inversely proportional to information. Thus, less error implies more information(141). Information is somewhat more complex with the two parameter item response theory model (Multilog analysis) in that it is multiplied by item discrimination (i.e., probability of passing an item times probability of failing an item times item discrimination)(142). Since item information is an additive property, information for the entire test can be calculated(143). Test information function is related to the underlying trait of interest. With the Rasch model, the zero point is set at the mean item difficulty. Thus, using Winsteps the test information function is always centered around zero on the latent variable. Information affects the measurement of change in the following manner(144). At person abilities where scale information is high, smaller changes appear significant (result in larger changes in test scores). However at person abilities where information is low, an individual must change more on the trait measured in order to see the same change (in test score)(143).

Results

Unidimensionality

Admission data EFA

The eigenvalue for the first factor was 18.40. Retaining one factor can explain 61.32% of the total variance. All items load relatively high on one factor, ranging from .46 to .87 (See Table 3-2). The first two factors are highly correlated with each other, $r = .74$. The scree plot shown in Figure 3-1 shows a steep drop off after the mark representing the eigenvalue for the first factor. There were three factors with eigenvalues greater than one, 18.40, 1.56, and 1.54,

respectively. Loadings of each item on these three factors are shown in Table 3-3. There is a general tendency for items loading on factor one to be more gross motor, complex items, requiring more movement. Items loading on factor two tend to be items requiring only hand use, while, those loading on factor three tend to be more general items and symptom items.

Admission data CFA

Confirmatory factor analysis was run to test a one factor solution. This analysis had mixed results. The chi-square was large, 3628.84 (df = 104) and highly significant, p = .000. Consequently, this statistic did not support the one factor solution. The CFI was .77, TLI was .97, SRMR was .07, and the RMSEA was .19. Thus, only the TLI value supports the one factor model. However, the first factor accounted for a large proportion of the variance, 62.86%.

Discharge data EFA

Similar to with the admission data, the eigenvalue for the first factor was 20.23. Retaining one factor can explain 67.44% of the total variance. All items load relatively high on one factor, ranging from .68 to .89 (See Table 3-2). Again, the first two factors were highly correlated with each other, $r = .74$. The scree plot shown in Figure 3-2 illustrates a steep drop off after the mark representing the eigenvalue for the first factor. There were three factors with eigenvalues greater than one, 20.23, 1.39, and 1.16, respectively. Loadings of each item on these three factors are shown in Table 3-4. Once again there is a general tendency for items loading on factor one to be more gross motor, complex items, requiring more movement, items loading on factor two tend to require only hand use, and items loading on factor three to be more general items and symptom items.

Discharge data CFA

Also similar to with the admission data, for a one factor solution, the chi-square was large, 2677.74 (df = 103) and highly significant, p = .000. The CFI was .84, TLI was .98, SRMR was .06, and the RMSEA was .16. Thus, the TLI and SRMR supported the one-factor model.

Rasch-derived fit statistics and point measure correlations

Table 3-5 presents Rasch derived item infit statistics for the 30-item DASH at admission and discharge. At admission, ten items had high fit values (MnSq > 1.1 with ZSTD > 2.0): I feel less capable, less confident or less useful because of my arm, shoulder or hand problem; Stiffness in your arm, shoulder or hand; Open a tight or new jar; Difficulty sleeping because of pain in your arm, shoulder or hand; Interfered with your normal social activities with family, friends, neighbours or group; Tingling (pins and needles) in your arm, shoulder, or hand; Sexual activities; Write; Manage transportation needs; and Turn a key. At discharge, the following items continued to misfit: I feel less capable, less confident or less useful because of my arm, shoulder, or hand problem; Open a tight or new jar; Difficulty sleeping because of pain in your arm, shoulder or hand; Tingling (pins and needles) in your arm, shoulder, or hand; Sexual activities; Write; Manage transportation needs; and Turn a key. Additionally, the following items misfit at discharge: Wash your back; Recreation activities that require little effort; and Use a knife to cut food. Point measure correlations for admission data ranged from .46 to .82, while point measure correlations for discharge ranged from .57 to .83 (See Table 3-6). All correlations were above .30. However, the majority were above .70 (19 above .70 at admission, 20 at discharge).

Taken together, there is mixed evidence for a one factor solution. One might argue that the large number of misfit items (ten with admission data and eleven with discharge data) were due to the stringent criteria used in this study to judge fit (MnSq > 1.1). Additionally, some authors

have argued that factor analysis is not the optimal way to investigate items influenced by psychological factors (i.e., self-reports)(143). Thus, it was felt that there was enough support for a one factor solution to apply item response theory methodologies. All further analyses were continued based on the assumption that the items reflected a single construct. The inconsistent findings from the factor analysis and fit statistics will be addressed in the discussion section.

Item Difficulty

Table 3-7 presents item difficulty calibrations at admission and discharge in the order of relative challenge at admission. The most challenging items at admission (higher calibrations) are located at the top of each table and least challenging items at admission (lower calibrations) are located at the bottom of the table. Overall, item difficulties at admission and discharge were similar. DIF discussed later will compare the discharge difficulties to the admission difficulties. The four items that were most difficult at admission were: Less capable, less confident or useful, Recreational activities in which you take some force or impact, Recreational activities in which you move your arm freely, and Do heavy household chores. Turning a key, Manage transportation needs, and Write were easy items at admission. Middle difficulty items included: Open a tight or new jar lid, Wash your back, and Limited in work or other regular daily activities.

Person-Item Match

Figure 3-3 presents maps of both admission and discharge item difficulties plotted against person ability measures. These maps were generated using item difficulties obtained when admission and discharge items were run in separate analysis. The X's to the left of the vertical line symbolize the person measures, with those at the top representing individuals with the most upper-extremity ability and persons on the bottom illustrating those with the least. The items are represented to the right of the vertical line, with the most challenging items located at the top and

the least challenging at the bottom. These items are located on the map at their average measure of challenge, (i.e., the measure when individuals respond moderate difficulty). As can be seen from these maps, the sample performed better at discharge than at admission. The sample mean fell above all the items at discharge, whereas, at admission several items (I feel less capable, less confident or less useful, Recreational activities in which you take some force or impact, and Recreational activities in which you move your arm freely) were above the sample mean. At discharge 41 individuals had maximum estimated ability levels, while at admission only 2 had maximum estimated ability levels.

Differential Item Functioning

Table 3-7 presents the DIF analysis results. DASH items are arranged in decreasing order of admission item difficulty, from least difficult items at admission at the bottom to most difficult items at admission at the top. The DIF analysis compared the admission and discharge item difficulty calibrations of each item. These t-values are associated with a probability level (seen beside the t-values in the table). Five items had large DIF values ($p<.0017$): I feel less capable, less confident or less useful because of my arm, shoulder or hand problem; Difficulty sleeping because of the pain in your arm, shoulder or hand; Push open a heavy door; Tingling (pins and needles) in your arm, shoulder or hand; and Sexual activities. While four items had small DIF ($p<.05$) (Weakness in your arm, shoulder or hand; Limited in work or other regular daily activities as a result of your arm, shoulder, or hand problem; Put on a pullover sweater; and Make a bed). Of these nine items, four of these items have higher admission item difficulty than discharge (Difficulty sleeping because of the pain in your arm, shoulder or hand; Put on a pullover sweater; Make a bed; and Limited in work or other regular daily activities as a result of your arm, shoulder, or hand problem). The remaining five display lower admission item difficulty values than discharge item difficulty values (I feel less capable, less confident or less

useful because of my arm, shoulder or hand problem; Push open a heavy door; Tingling (pins and needles) in your arm, shoulder or hand; Sexual activities; and Weakness in your arm, shoulder or hand). However, using the 0.5 logit difference guidelines, none of the items displayed significant DIF.

Figure 3-4 shows a visual illustration of DIF. The item difficulty calibrations at admission are plotted on the x-axis against item difficulty calibrations at discharge on the y-axis. Lines represent where 95% confidence interval bands fall. As can be seen from the figure, among the items showing DIF, they do not fall much outside the 95% confidence interval. Labeled items in the figure are those that displayed large DIF. Comparison of admission item difficulties to discharge item difficulties using an intraclass correlation coefficient (ICC) indicated a high reliability between the two time-points (ICC = .99).

To compare the possible impact of DIF on measurement of upper-extremity condition, Rasch derived mean person ability measurements at admission and discharge were compared using two versions of the DASH: (1) the full 30-item version and (2) a 21-item version with the items displaying significant DIF removed. Figure 3-5 presents the admission to discharge (x-axis) mean person ability measurements (y-axis, logits) for the two versions. The solid line shows the change admission to discharge of person ability estimates when all the items on the DASH were used (change from .90 $\pm$ 1.37 at admission to 2.38 $\pm$ 1.68 at discharge). The dashed line illustrates the change admission to discharge of the person ability estimates when only the 21-items that did not show DIF were used (change from .93 $\pm$ 1.52 at admission to 2.49 $\pm$ 1.78 at discharge). Thus, mean person ability measures were almost identical whether or not the items showing DIF were included.

Item Discrimination

Table 3-8 presents item discrimination calibrations at admission and discharge. Values for item discrimination at admission ranged from .09-1.38. The values at discharge ranged from .28-1.36. Thus, item discrimination ranges were similar but somewhat wider for admission data. To compare the possible impact of item discrimination on measurement of upper-extremity condition, Rasch derived mean person ability measurements at admission and discharge were compared to those derived using a two parameter model. Figure 3-6 presents the admission to discharge (x-axis) mean person ability measurements (y-axis, logits) for the two models. The solid line shows the change admission to discharge of person ability estimates when the one parameter model was used (change from .91 $\pm$ 1.39 at admission to 2.59 $\pm$ 1.92 at discharge). The dashed line illustrates the change admission to discharge of the person ability estimates when the two parameter model was used (change from 1.1 $\pm$ 1.75 at admission to 2.01 $\pm$ 1.47 at discharge).

Test Information

Figure 3-7 presents the DASH information function at admission using the one parameter model, while Figure 3-8 presents the DASH information function obtained using discharge scores using the one parameter model. As shown by the figures, test information is identical at admission and discharge, thus, indicating that person ability corresponds with test information in a similar way at both admission and discharge. This implies that item difficulties (which affect the calculation of information) do not change enough from admission to discharge to affect person ability measures. Additionally, these curves illustrate how information peaks when the mean of the latent variable is zero. This is also where the standard error is the lowest.

Discussion

The purpose of this paper was to investigate the item-level characteristics of the DASH, including unidimensionality, item difficulty, person-item match, differential item functioning across admission and discharge, item discrimination, and test information. Several findings pertaining to the item-level psychometrics of the DASH are presented. Results of exploratory and confirmatory factor analyses were inconclusive as to whether or not a one factor solution was plausible. Eigenvalues for the first factor in both the admission and discharge EFAs were much greater than for the second factor, however, the second and third factor eignevalues were greater than 1.0 (criterion necessary for acceptance of the presence of a factor based on the Kaiser rule). One factor accounted for most of the variance (> 60%) for both admission and discharge data and all items loaded high on the first factor. The first two factors were highly correlated. With the admission CFA, the only goodness of fit statistic that supported a one factor model was the TLI. With the discharge CFA, the TLI and SRMR supported the one factor model. However, goodness of fit statistics should never should be used alone to determine dimensionality(143).

A number of items misfit (ten at admission and eleven at discharge). With less stringent criteria, such as that used with smaller sample sizes, it is expected that many fewer items would misfit. Preliminary analysis by the authors with data on the DASH with smaller samples sizes which allows a less stringent criteria be followed has shown only three items (at admission) and five items (at discharge) to misfit. All point measure correlations were above .30. However, the observation that many (19 at admission and 20 at discharge) were above .70 indicates that some of the items may be redundant (i.e., add no further information to the test than other items already obtain). This suggests that a short form of the DASH might be just as informative as the longer version.

Item difficulty hierarchies support the theory of motor control. Four tenants of motor control theory are: 1) complex items are more difficult, 2) environmental factors increase the challenge of items, 3) more difficult items require the use of multiple joints, and 4) isolated, abstract items are more difficult than functional tasks(128, 145, 146). Complex items were found to be more difficult. For example, doing heavy household chores (e.g., wash walls, wash floors) was harder than putting on a pullover sweater. Likewise, gardening or doing yard work was more difficult than carrying a shopping bag or briefcase. Environmental factors were shown to influence difficulty ranking. Recreational activities requiring little effort, such as card playing or knitting, were rated easier than recreational activities in which you move your arm freely, such as frisbee or badminton. Furthermore, recreational activities in which you move your arm freely were rated easier than recreational activities in which you take some force or impact through your arm, shoulder, or hand. Items requiring the use of multiple joints were more difficult items. For instance, recreational activities, such as frisbee or badminton, were rated more difficult than using a knife to cut food. Also, preparing a meal was more difficult than turning a key in a lock. Finally, isolated tasks, presented without a concrete context or purpose were rated as harder. To illustrate, carrying a heavy object was rated as harder than carrying a shopping bag.

Several items displayed significant DIF from admission to discharge. Five showed large DIF (I feel less capable, less confident or less useful because of my arm, shoulder or hand problem; Difficulty sleeping because of the pain in your arm, shoulder or hand; Push open a heavy door; Tingling (pins and needles) in your arm, shoulder or hand; and Sexual activities) and four small DIF (Weakness in your arm, shoulder or hand; Limited in work or other regular daily activities as a result of your arm, shoulder, or hand problem; Put on a pullover sweater; and

Make a bed). The four items that had a higher admission difficulty than discharge difficulty were: Difficulty sleeping because of pain in your arm, shoulder or hand; Putting on a pullover sweater; Making a bed; and Limited in work or other regular daily activities as a result of your arm, shoulder, or hand problem. Perhaps this is because at admission these activities were problematic, but at the conclusion of therapy, they had more fully resolved than other items. That is, for example, once symptoms of pain and limitations to daily activities were gone by the end of treatment they were not difficult items at all, rating even easier than would have been expected based on the admission difficulty. Five items had lower admission item difficulties than discharge. These items were: I feel less capable, less confident or less useful because of my arm, shoulder or hand problem, Push open a heavy door, Tingling (pins and needles) in your arm, shoulder or hand, Sexual activities, and Weakness in your arm, shoulder or hand. This could be because if an individual is still having problems with these at discharge, they see this as more of a problem than if they are having these issues at admission. One might consider removing DIF items to create a sounder instrument. However, using the guidelines of 0.5 logit difference to identify DIF, none of the items displayed significant DIF. Additionally, mean person ability measures were not affected by the inclusion of DIF items. Furthermore, the intraclass correlation coefficient (ICC) was high indicating reliability in measurement between the two time-points. This stability of the measure from admission to discharge is critical to the measurement of patient change.

Since none of the item discrimination values were zero or negative, this indicates that all items contribute in a similar fashion to the overall test(137, 138). Nevertheless, because these values are not equal, some might agree that analysis using a two parameter model is necessary. Yet, analysis using the one parameter model appears to be the most parsimonious solution since

item discrimination does not appear to dramatically affect person measures, although, there is more difference between the person ability measure calculated using the different models at discharge. Similar curves for test information for admission and discharge data provide further evidence for the stability of the test at two time points.

Comparison to past literature is complicated by the lack of studies investigating item-level psychometrics of the DASH. The one study investigating item-level properties of the DASH(97) in order to create a shortened version, found only four items to misfit (weakness, tingling, sexual activities and write). However, they were using a less stringent criteria (MnSq less than 1.3). Likewise, one study has investigated the factor structure of the DASH. Liang and colleagues(147) conducted a principal axis factor analysis. These researchers' conclude that there is support for a one factor structure. They found that the first factor had an eigenvalue of 13.51 and explained 45% of the total variance. Loadings on the first factor in this study ranged from .43 to .89. None of their findings were indicative of a two factor solution.

There are several limitations of this study. One involves inconclusive evidence as to whether the DASH should be considered unidimensional (i.e., measuring one construct). Considering that the fit statistic criteria used was very stringent (MnSq > 1.1) and past studies have determined that goodness of fit statistics may be misleading(119), a unidimensional solution was considered to be plausible. Yet, other guidelines, such as the Kaiser rule (eigenvalues greater than one) supported a three factor solution. Examination of which items loaded highest on each of the factors in the three factor solution, separates the items into three possible constructs. One (those that load highest on the first factor) representing more gross motor, complex activities, a second (those that load highest on the second factor) being solely hand activities, and third (those that load highest on the third factor) involving more general

items or symptom items. Thus, the decision to treat the DASH as unidimensional is debatable. Perhaps, dividing the DASH into three separate scales with three total scores would be more meaningful. Using multiple scales, instead of one with debatable unidimensionality might provide a clearer picture of patient progress from admission to discharge.

A second limitation involves the sample used. While the sample used in this study was large, it was a sample of convenience. Moreover, the size of the sample was much reduced once individuals who did not complete the assessments at discharge were eliminated. Despite that the demographics presented were comparable, there is the potential that participants who completed the discharge surveys differed on some significant variables from the original larger sample. Third, the data was collected at various clinics across the country and from patients with diverse upper-extremity diagnoses. Therefore, situations in the clinics and treatments used by therapists differed. Some treatments could have been more effective. While this increases the generalizability of the results, because there is no information on the treatment used in each case, it is unclear which treatments are most optimal.

As a final limitation, item discrimination did not affect person ability estimates dramatically, but especially at discharge there was a notable difference. Therefore, a thorough analysis using the two parameter model is indicated. Additional tests of model fit are also indicated. Because item psychometrics is a new area of investigation for the DASH, studies should be conducted with other populations and at other time-points throughout the treatment process as well to provide conclusive evidence as to the properties of this assessment.

Conclusion

The evidence presented in this paper suggests that the psychometrics of the DASH could be improved upon. On the positive side, the item hierarchy is supported by motor control theory, DIF, although present does not appear to affect the measurement of person ability, and there is

some support for unidimensionality. However, there is also evidence that the DASH items may represent three different constructs: a more gross/complex construct, a hand construct, and a general/symptom construct. Also, removing DIF items might produce a sounder instrument. Due to the limited number of studies investigating the item-level psychometrics of this assessment, more studies are needed using different IRT models and with different populations and settings to confirm these findings. Perhaps, since unidimensionality is an assumption of IRT models, dividing the DASH into three different constructs would be superior methodologically based on the highest loadings of the items on three different, conceptually logical factors in the EFA.

Table 3-1. Sample of 991 demographics

	Total Sample	Sample with Intake and Discharge Data
Mean age ± SD 95% Confidence Interval (CI)	47.51 ± 13.90 (N = 2441) CI = 47.10, 48.83	48.84 ± 13.88 (N = 934) CI = 48.18, 50.44
Gender - Female - Male	 1427 (57.4%) 1057 (42.5%) (N = 2484)	 543 (56.6%) 417 (43.3%) (N = 960)
Primary body part injured - Shoulder - Neck - Hand - Elbow - Wrist - Upper Arm - Forearm	 1078 (43.3%) 474 (19.1%) 334 (13.4%) 237 (9.5%) 255 (10.3%) 63 (2.5%) 46 (1.8%) (N = 2487)	 496 (51.7%) 192 (20.0%) 87 (9.1%) 79 (8.2%) 69 (7.2%) 24 (2.5%) 13 (1.4%) (N = 960)
Symptom acuity - Chronic - Subacute - Acute	 1253 (50.4%) 814 (32.7%) 414 (16.6%) (N = 2481)	 540 (56.3%) 302 (31.5%) 116 (12.1%) (N = 960)
Mean number of days treated ± SD 95% Confidence Interval (CI)	42.64 ± 31.33 (N = 990) CI = 41.08, 45.12	48.63 ± 32.22 (N = 535) CI = 46.22, 51.90
Mean number of treatment sessions ± SD 95% Confidence Interval (CI)	11.43 ± 9.06 (N = 990) CI = 10.87, 11.99	13.30 ± 9.98 (N = 535) CI = 12.12, 13.79

Table 3-2. Item factor loadings on first factor

Item	Admission factor loadings	Discharge factor loadings
I feel less capable, less confident or less useful because of my arm, shoulder or hand problem	0.660	0.704
Recreational activities in which you take some force or impact through your arm, shoulder or hand	0.805	0.848
Recreational activities in which you move your arm freely (e.g., playing frisbee, badminton, etc.)	0.838	0.839
Do heavy household chores (e.g., wash walls, wash floors)	0.861	0.876
Arm, shoulder or hand pain when you performed a specific activity	0.734	0.841
Garden or do yard work	0.856	0.861
Arm, shoulder or hand pain	0.671	0.827
Carry a heavy object (over 10 lbs)	0.825	0.838
Weakness in your arm, shoulder or hand	0.744	0.809
Change a lightbulb overhead	0.802	0.823
Stiffness in your arm, shoulder or hand	0.692	0.738
Place an object on a shelf above your head	0.808	0.813
Open a tight or new jar	0.757	0.767
Wash your back	0.800	0.802
Limited in work or other regular daily activities as a result of your arm, shoulder, or hand problem	0.833	0.865
Difficulty sleeping because of the pain in your arm, shoulder or hand	0.581	0.755
Carry a shopping bag or briefcase	0.861	0.859
Put on a pullover sweater	0.773	0.824
Wash or blow dry your hair	0.814	0.845
Push open a heavy door	0.825	0.838
Make a bed	0.865	0.875
Interferred with your normal social activities with family, friends, neighbours or group	0.750	0.853
Tingling (pins and needles) in your arm, shoulder or hand	0.464	0.682
Use a knife to cut food	0.845	0.839
Prepare a meal	0.871	0.888
Sexual activities	0.714	0.757
Recreational activities which require little effort (e.g., cardplaying, knitting, etc.)	0.694	0.715
Write	0.797	0.819
Manage transportation needs (getting from one place to another)	0.787	0.810
Turn a key	0.773	0.777

Note: Items are arranged in decreasing order of admission difficulty.

Table 3-3. Admission item factor loadings on three factors after varimax rotation

Item	Factor 1	Factor 2	Factor 3
I feel less capable, less confident or less useful because of my arm, shoulder or hand problem	0.347	0.314	0.511
Recreational activities in which you take some force or impact through your arm, shoulder or hand	0.585	0.429	0.355
Recreational activities in which you move your arm freely (e.g., playing frisbee, badminton, etc.)	0.701	0.351	0.364
Do heavy household chores (e.g., wash walls, wash floors)	0.735	0.415	0.294
Arm, shoulder or hand pain when you performed a specific activity	0.340	0.267	0.736
Garden or do yard work	0.665	0.464	0.315
Arm, shoulder or hand pain	0.283	0.203	0.763
Carry a heavy object (over 10 lbs)	0.576	0.549	0.271
Weakness in your arm, shoulder or hand	0.334	0.451	0.540
Change a lightbulb overhead	0.820	0.257	0.257
Stiffness in your arm, shoulder or hand	0.352	0.358	0.518
Place an object on a shelf above your head	0.804	0.247	0.299
Open a tight or new jar	0.329	0.692	0.296
Wash your back	0.725	0.286	0.335
Limited in work or other regular daily activities as a result of your arm, shoulder, or hand problem	0.447	0.450	0.570
Difficulty sleeping because of the pain in your arm, shoulder or hand	0.335	0.098	0.627
Carry a shopping bag or briefcase	0.589	0.568	0.302
Put on a pullover sweater	0.720	0.237	0.345
Wash or blow dry your hair	0.560	0.539	0.280
Push open a heavy door	0.716	0.366	0.303
Make a bed	0.644	0.505	0.313
Interferred with your normal social activities with family, friends, neighbours or group	0.408	0.329	0.597
Tingling (pins and needles) in your arm, shoulder or hand	0.065	0.286	0.523
Use a knife to cut food	0.392	0.757	0.317
Prepare a meal	0.448	0.718	0.338
Sexual activities	0.466	0.401	0.360
Recreational activities which require little effort (e.g., cardplaying, knitting, etc.)	0.217	0.738	0.268
Write	0.340	0.694	0.360
Manage transportation needs (getting from one place to another)	0.505	0.481	0.362
Turn a key	0.314	0.818	0.212

Note: Items are arranged in decreasing order of admission difficulty.

Table 3-4. Discharge item factor loadings on three factors after varimax rotation

Item	Factor 1	Factor 2	Factor 3
I feel less capable, less confident or less useful because of my arm, shoulder or hand problem	0.368	0.329	0.532
Recreational activities in which you take some force or impact through your arm, shoulder or hand	0.666	0.401	0.380
Recreational activities in which you move your arm freely (e.g., playing frisbee, badminton, etc.)	0.710	0.338	0.381
Do heavy household chores (e.g., wash walls, wash floors)	0.746	0.431	0.314
Arm, shoulder or hand pain when you performed a specific activity	0.409	0.352	0.720
Garden or do yard work	0.687	0.469	0.311
Arm, shoulder or hand pain	0.364	0.368	0.730
Carry a heavy object (over 10 lbs)	0.644	0.563	0.225
Weakness in your arm, shoulder or hand	0.415	0.480	0.515
Change a lightbulb overhead	0.791	0.259	0.345
Stiffness in your arm, shoulder or hand	0.388	0.322	0.580
Place an object on a shelf above your head	0.736	0.236	0.412
Open a tight or new jar	0.391	0.678	0.267
Wash your back	0.637	0.248	0.490
Limited in work or other regular daily activities as a result of your arm, shoulder, or hand problem	0.481	0.457	0.566
Difficulty sleeping because of the pain in your arm, shoulder or hand	0.362	0.245	0.727
Carry a shopping bag or briefcase	0.570	0.573	0.333
Put on a pullover sweater	0.656	0.222	0.537
Wash or blow dry your hair	0.611	0.504	0.332
Push open a heavy door	0.647	0.363	0.423
Make a bed	0.620	0.514	0.367
Interferred with your normal social activities with family, friends, neighbours or group	0.444	0.431	0.613
Tingling (pins and needles) in your arm, shoulder or hand	0.202	0.421	0.591
Use a knife to cut food	0.356	0.714	0.402
Prepare a meal	0.457	0.661	0.426
Sexual activities	0.461	0.425	0.422
Recreational activities which require little effort (e.g., cardplaying, knitting, etc.)	0.221	0.734	0.314
Write	0.404	0.682	0.343
Manage transportation needs (getting from one place to another)	0.422	0.486	0.501
Turn a key	0.242	0.828	0.312

Note: Items are arranged in decreasing order of admission difficulty.

Table 3-5. Infit statistics at admission and discharge

Item	Admision MNSQ	Admission ZSTD	Discharge MNSQ	Discharge ZSTD
I feel less capable, less confident or less useful because of my arm, shoulder or hand problem	1.16	3.40	1.57	9.90
Recreational activities in which you take some force or impact through your arm, shoulder or hand	0.94	-1.30	0.84	-3.70
Recreational activities in which you move your arm freely (e.g., playing frisbee, badminton, etc.)	0.86	-3.20	1.01	0.20
Do heavy household chores (e.g., wash walls, wash floors)	0.71	-7.20	0.70	-7.20
Arm, shoulder or hand pain when you performed a specific activity	0.73	-6.60	0.67	-7.90
Garden or do yard work	0.73	-6.80	0.83	-3.80
Arm, shoulder or hand pain	0.85	-3.60	0.71	-6.90
Carry a heavy object (over 10 lbs)	0.85	-3.60	0.92	-1.80
Weakness in your arm, shoulder or hand	0.94	-1.40	0.84	-3.50
Change a lightbulb overhead	1.00	0.10	1.04	0.90
Stiffness in your arm, shoulder or hand	1.12	2.70	1.02	0.50
Place an object on a shelf above your head	0.86	-3.30	0.90	-2.01
Open a tight or new jar	1.23	4.80	1.36	6.60
Wash your back	1.07	1.50	1.13	2.60
Limited in work or other regular daily activities as a result of your arm, shoulder, or hand problem	0.78	-5.40	0.83	-3.60
Difficulty sleeping because of the pain in your arm, shoulder or hand	1.42	8.20	1.21	3.70
Carry a shopping bag or briefcase	0.72	-6.80	0.81	-3.80
Put on a pullover sweater	0.91	-2.00	0.90	-1.90
Wash or blow dry your hair	0.88	-2.70	1.04	0.70
Push open a heavy door	0.85	-3.30	0.81	-4.00
Make a bed	0.74	-6.10	0.84	-3.00
Interferred with your normal social activities with family, friends, neighbours or group	1.17	3.40	1.06	1.00
Tingling (pins and needles) in your arm, shoulder or hand	1.91	9.90	1.46	7.40
Use a knife to cut food	1.03	0.60	1.22	3.30
Prepare a meal	0.72	-5.90	0.80	-3.40
Sexual activities	1.62	9.90	2.09	9.90
Recreational activities which require little effort (e.g., cardplaying, knitting, etc.)	1.09	1.60	1.21	2.90
Write	1.38	6.40	1.62	7.90
Manage transportation needs (getting from one place to another)	1.22	3.70	1.29	3.80
Turn a key	1.19	3.30	1.22	3.00

Note: Items are arranged in decreasing order of admission difficulty.

Table 3-6. Point measure correlations at admission and discharge

Item	Admission correlations	Discharge correlations
I feel less capable, less confident or less useful because of my arm, shoulder or hand problem	0.63	0.63
Recreational activities in which you take some force or impact through your arm, shoulder or hand	0.77	0.8
Recreational activities in which you move your arm freely (e.g., playing frisbee, badminton, etc.)	0.8	0.8
Do heavy household chores (e.g., wash walls, wash floors)	0.82	0.83
Arm, shoulder or hand pain when you performed a specific activity	0.71	0.78
Garden or do yard work	0.81	0.81
Arm, shoulder or hand pain	0.66	0.76
Carry a heavy object (over 10 lbs)	0.78	0.79
Weakness in your arm, shoulder or hand	0.72	0.75
Change a lightbulb overhead	0.77	0.77
Stiffness in your arm, shoulder or hand	0.68	0.69
Place an object on a shelf above your head	0.78	0.77
Open a tight or new jar	0.71	0.71
Wash your back	0.76	0.75
Limited in work or other regular daily activities as a result of your arm, shoulder, or hand problem	0.79	0.81
Difficulty sleeping because of the pain in your arm, shoulder or hand	0.58	0.69
Carry a shopping bag or briefcase	0.8	0.79
Put on a pullover sweater	0.73	0.76
Wash or blow dry your hair	0.76	0.78
Push open a heavy door	0.77	0.75
Make a bed	0.79	0.79
Interferred with your normal social activities with family, friends, neighbours or group	0.7	0.77
Tingling (pins and needles) in your arm, shoulder or hand	0.46	0.61
Use a knife to cut food	0.73	0.7
Prepare a meal	0.76	0.76
Sexual activities	0.61	0.61
Recreational activities which require little effort (e.g., cardplaying, knitting, etc.)	0.61	0.57
Write	0.69	0.67
Manage transportation needs (getting from one place to another)	0.66	0.64
Turn a key	0.65	0.62

Note: Items are arranged in decreasing order of admission difficulty.

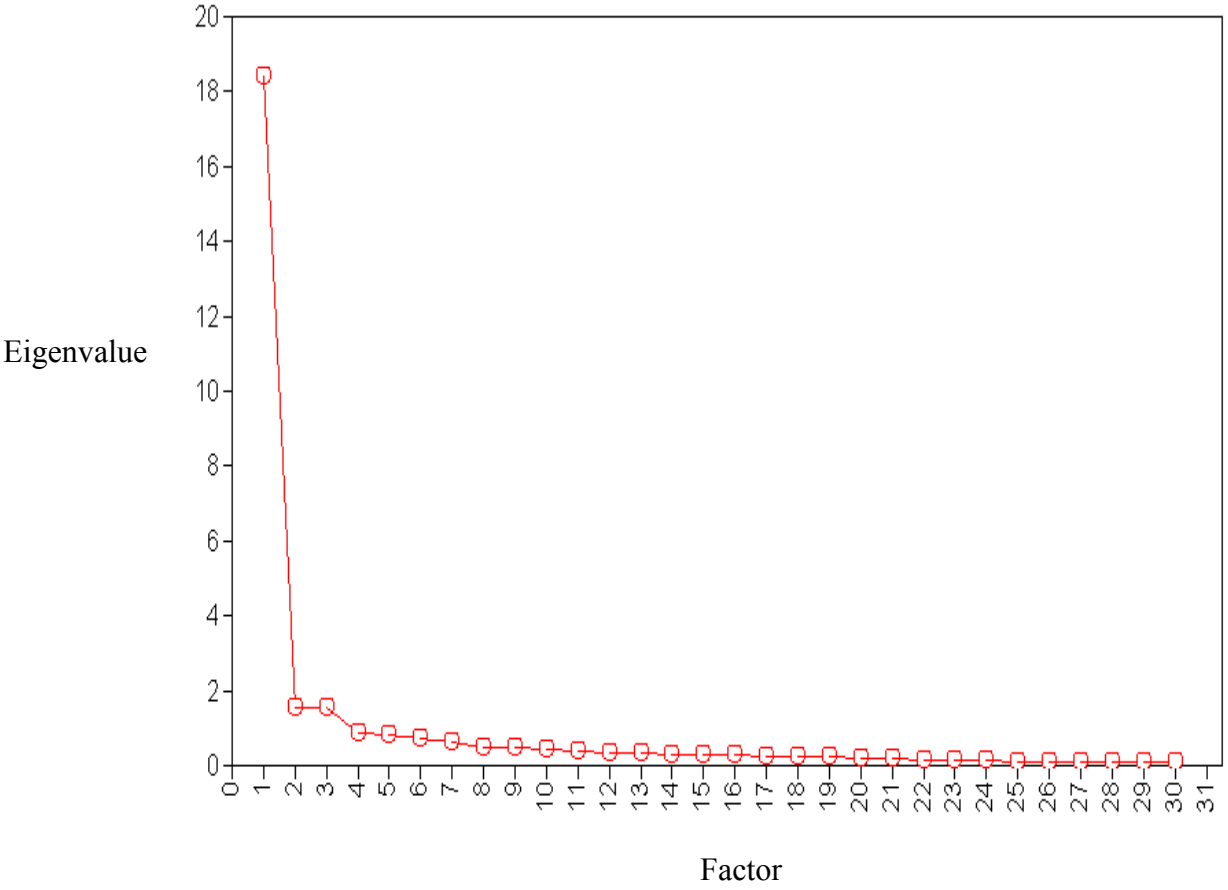


Figure 3-1. DASH admission scree plot

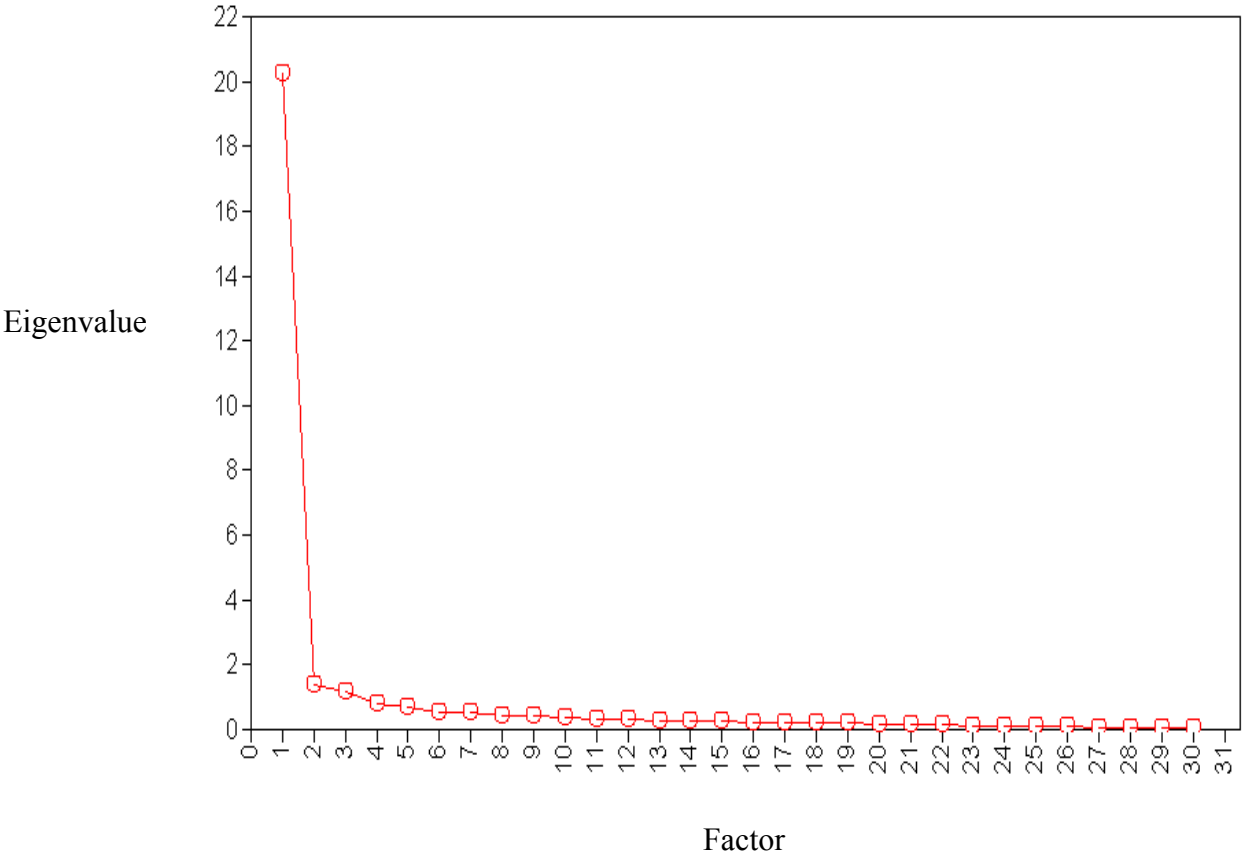


Figure 3-2. DASH discharge scree plot

Table 3-7. Item difficulty estimates and differential item functioning (N = 960, **p<0.002, *p<0.05).

Item	Admission item difficulty	Error	Discharge item difficulty	Error	DIF (calculated t-value)	p-value	Significance
I feel less capable, less confident or less useful because of my arm, shoulder or hand problem	1.95	0.04	2.13	0.04	3.64	0.0003	**
Recreational activities in which you take some force or impact through your arm, shoulder or hand	1.36	0.04	1.61	0.04	0.70	0.4831	
Recreational activities in which you move your arm freely (e.g., playing frisbee, badminton, etc.)	1.16	0.04	1.32	0.05	1.46	0.1446	
Do heavy household chores (e.g., wash walls, wash floors)	0.86	0.04	1.14	0.05	-1.18	0.2389	
Arm, shoulder or hand pain when you performed a specific activity	0.83	0.04	1.01	0.05	0.05	0.9620	
Garden or do yard work	0.82	0.04	1.07	0.05	-0.99	0.3225	
Arm, shoulder or hand pain	0.64	0.04	0.86	0.05	-1.09	0.2741	
Carry a heavy object (over 10 lbs)	0.53	0.04	0.72	0.05	-0.95	0.3428	
Weakness in your arm, shoulder or hand	0.48	0.04	0.73	0.05	-1.98	0.0480	*
Change a lightbulb overhead	0.44	0.04	0.47	0.05	0.94	0.3471	
Stiffness in your arm, shoulder or hand	0.44	0.04	0.62	0.05	-1.29	0.1977	
Place an object on a shelf above your head	0.43	0.04	0.55	0.05	-0.39	0.6975	
Open a tight or new jar	0.42	0.04	0.65	0.05	-1.90	0.0575	
Wash your back	0.39	0.04	0.43	0.05	0.71	0.4783	
Limited in work or other regular daily activities as a result of your arm, shoulder, or hand problem	0.20	0.04	0.09	0.05	2.17	0.0301	*
Difficulty sleeping because of the pain in your arm, shoulder or hand	0.07	0.04	-0.19	0.05	3.69	0.0002	**
Carry a shopping bag or briefcase	-0.14	0.04	-0.22	0.05	0.67	0.5050	
Put on a pullover sweater	-0.16	0.04	-0.37	0.06	2.43	0.0152	*
Wash or blow dry your hair	-0.28	0.04	-0.47	0.06	1.70	0.0899	
Push open a heavy door	-0.28	0.04	0.04	0.05	-5.09	0.0000	**
Make a bed	-0.42	0.04	-0.67	0.06	1.99	0.0469	*
Interferred with your normal social activities with family, friends, neighbours or group	-0.49	0.04	-0.66	0.06	0.82	0.4107	
Tingling (pins and needles) in your arm, shoulder or hand	-0.69	0.04	-0.4	0.06	-5.62	0.0000	**
Use a knife to cut food	-1.06	0.05	-1.37	0.07	0.88	0.3815	
Prepare a meal	-1.09	0.05	-1.29	0.07	-0.32	0.7491	
Sexual activities	-1.12	0.05	-1.07	0.06	-3.33	0.0009	**
Recreational activities which require little effort (e.g., cardplaying, knitting, etc.)	-1.24	0.05	-1.61	0.07	1.16	0.2446	
Write	-1.24	0.05	-1.49	0.07	-0.25	0.8050	
Manage transportation needs (getting from one place to another)	-1.37	0.05	-1.84	0.08	1.79	0.0737	
Turn a key	-1.46	0.05	-1.79	0.08	0.15	0.8785	

Note: Items are arranged in decreasing order of admission difficulty.

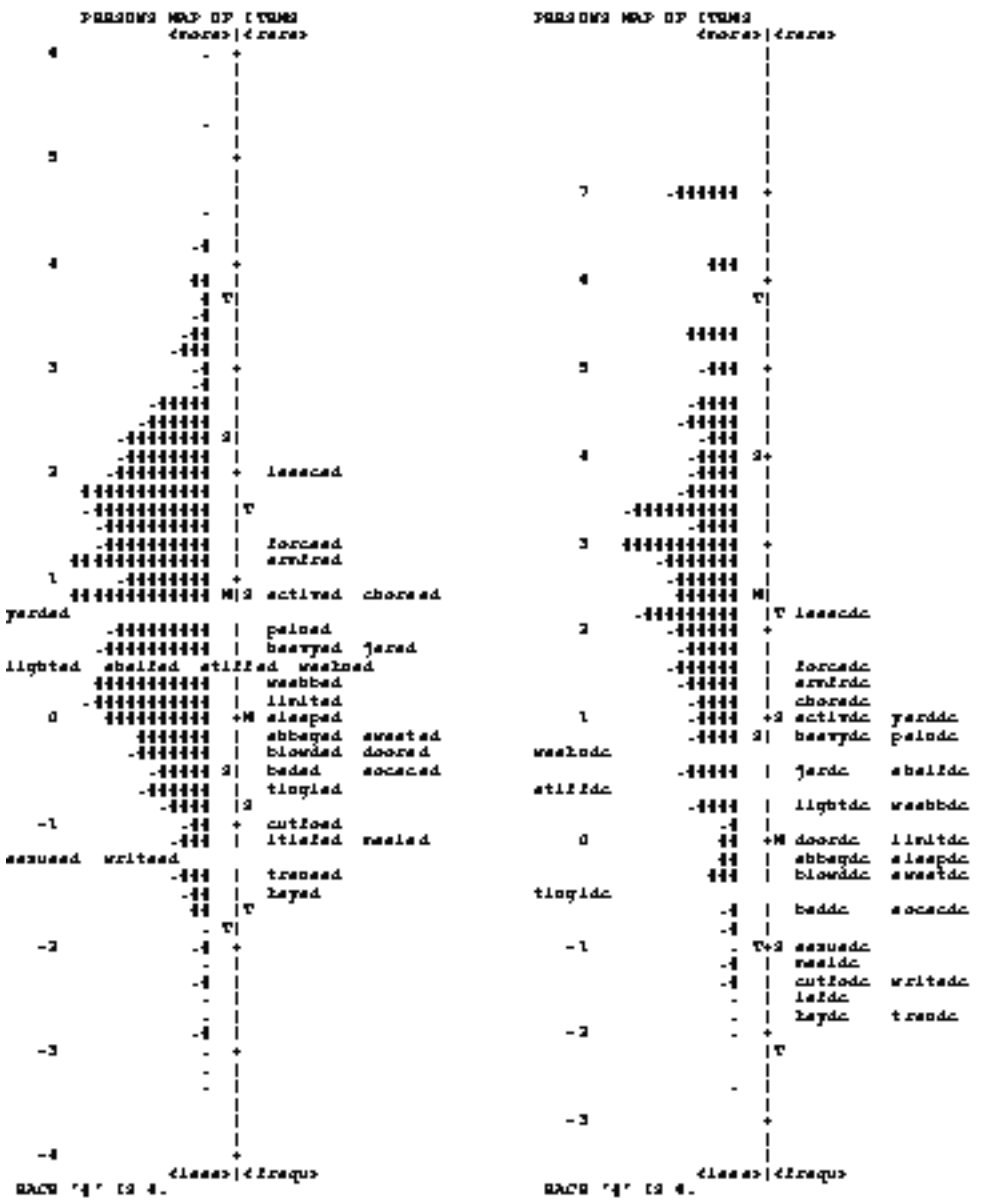


Figure 3-3. Admission (left) and discharge (right) person-item maps (Arranged so zeros line up to allow comparison)

Note: On Admission Map (left): activad = Limited in your work or other regular daily activities at admission; forcead = Recreational activities in which you take some force or impact at admission; heavyad = Carry a heavy object (over 10 lbs) at admission; armfrad = Recreational activities in which you move your arm freely at admission; lesscad = I feel less capable, less confident or less useful at admission; jarad = Open a tight or new jar at admission; yardad = Garden or do yard work at admission; limitad = Limited in your work or other regular daily activities at admission; painad = Arm, shoulder or hand pain at admission; weaknad = Weakness

in your arm, shoulder or hand at admission; stiffad = Stiffness in your arm, shoulder or hand at admission; choread = Do heavy household chores (e.g., wash walls, wash floors) at admission; washbad = Wash your back at admission; socacad = Interfered with your normal social activities at admission; bedad = Make a bed at admission; shelfad = Place an object on a shelf above your head at admission; shbagad = Carry a shopping bag or briefcase at admission; sleepad = Difficulty Sleeping at admission; lightad = Change a lightbulb overhead at admission; doorad = Push open a heavy door at admission; blowdad = Wash or blow dry your hair at admission; cutfoad = Use a knife to cut food at admission; mealad = Prepare a meal at admission; tinglad = Tingling (pins and needles) in your arm, shoulder or hand at admission; sexuaad = Sexual activities at admission; ltlefad = Recreational activities which require little effort at admission; sweatad = Put on a pullover sweater at admission; writead = Write at admission; transad = Manage transportation needs at admission; keyad = Turn a key at admission
On Discharge Map (right): activdc = Limited in your work or other regular daily activities at discharge; forcedc = Recreational activities in which you take some force or impact at discharge; heavydc = Carry a heavy object (over 10 lbs) at discharge; armfrdc = Recreational activities in which you move your arm freely at discharge; lesscdc = I feel less capable, less confident or less useful at discharge; jardc = Open a tight or new jar at discharge; yarddc = Garden or do yard work at discharge; limitdc = Limited in your work or other regular daily activities at discharge; paindc = Arm, shoulder or hand pain at discharge; weakndc = Weakness in your arm, shoulder or hand at discharge; stiffdc = Stiffness in your arm, shoulder or hand at discharge; choredc = Do heavy household chores (e.g., wash walls, wash floors) at discharge; washbdc = Wash your back at discharge; socacdc = Interfered with your normal social activities at discharge; beddc = Make a bed at discharge; shelfdc = Place an object on a shelf above your head at discharge; shbagdc = Carry a shopping bag or briefcase at discharge; sleepdc = Difficulty Sleeping at discharge; lightdc = Change a lightbulb overhead at discharge; doordc = Push open a heavy door at discharge; blowddc = Wash or blow dry your hair at discharge; cutfodc = Use a knife to cut food at discharge; mealdc = Prepare a meal at discharge; tingldc = Tingling (pins and needles) in your arm, shoulder or hand at discharge; sexuadc = Sexual activities at discharge; ltlefdc = Recreational activities which require little effort at discharge; sweatdc = Put on a pullover sweater at discharge; writedc = Write at discharge; transdc = Manage transportation needs at discharge; keydc = Turn a key at discharge

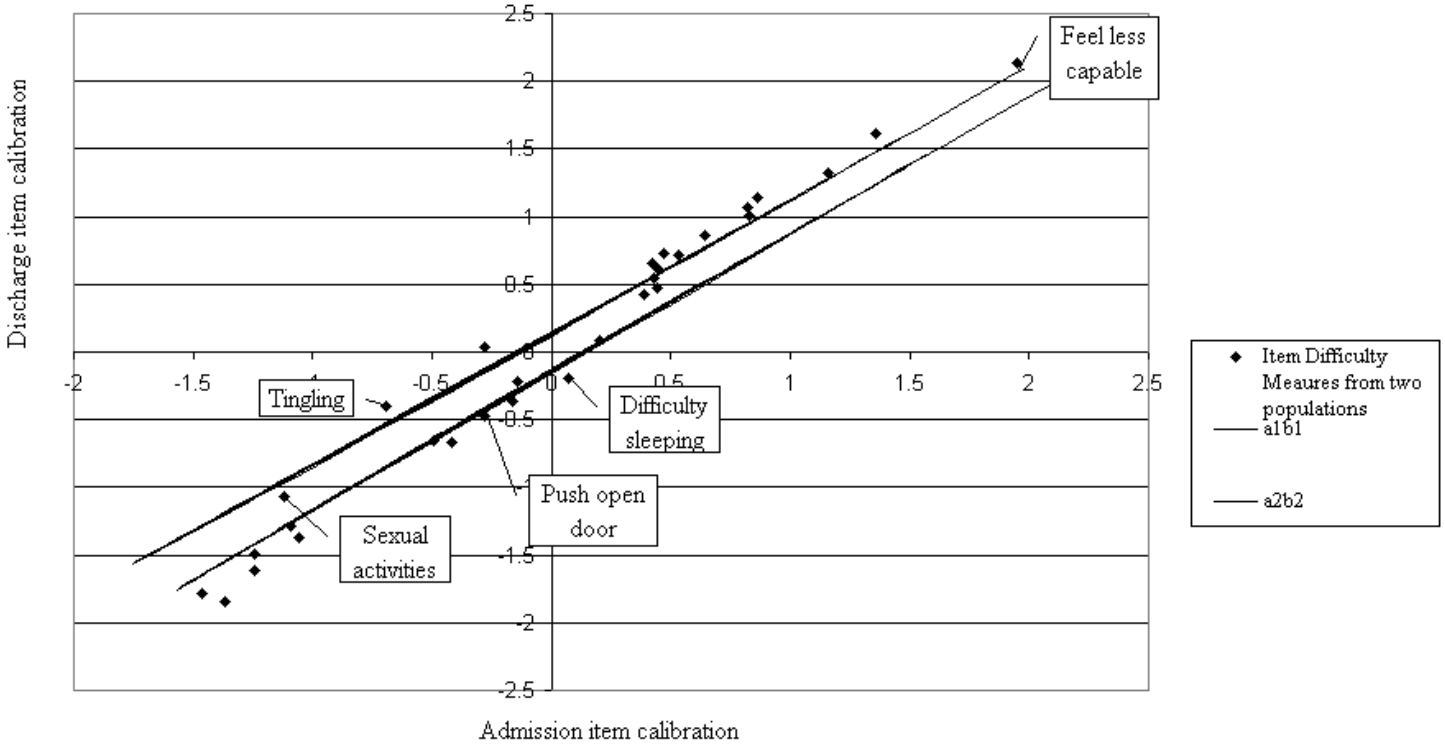


Figure 3-4. DASH admission item calibrations vs. dash discharge item calibrations

Note: Labeled items in the figure showed large DIF in Winsteps analysis.

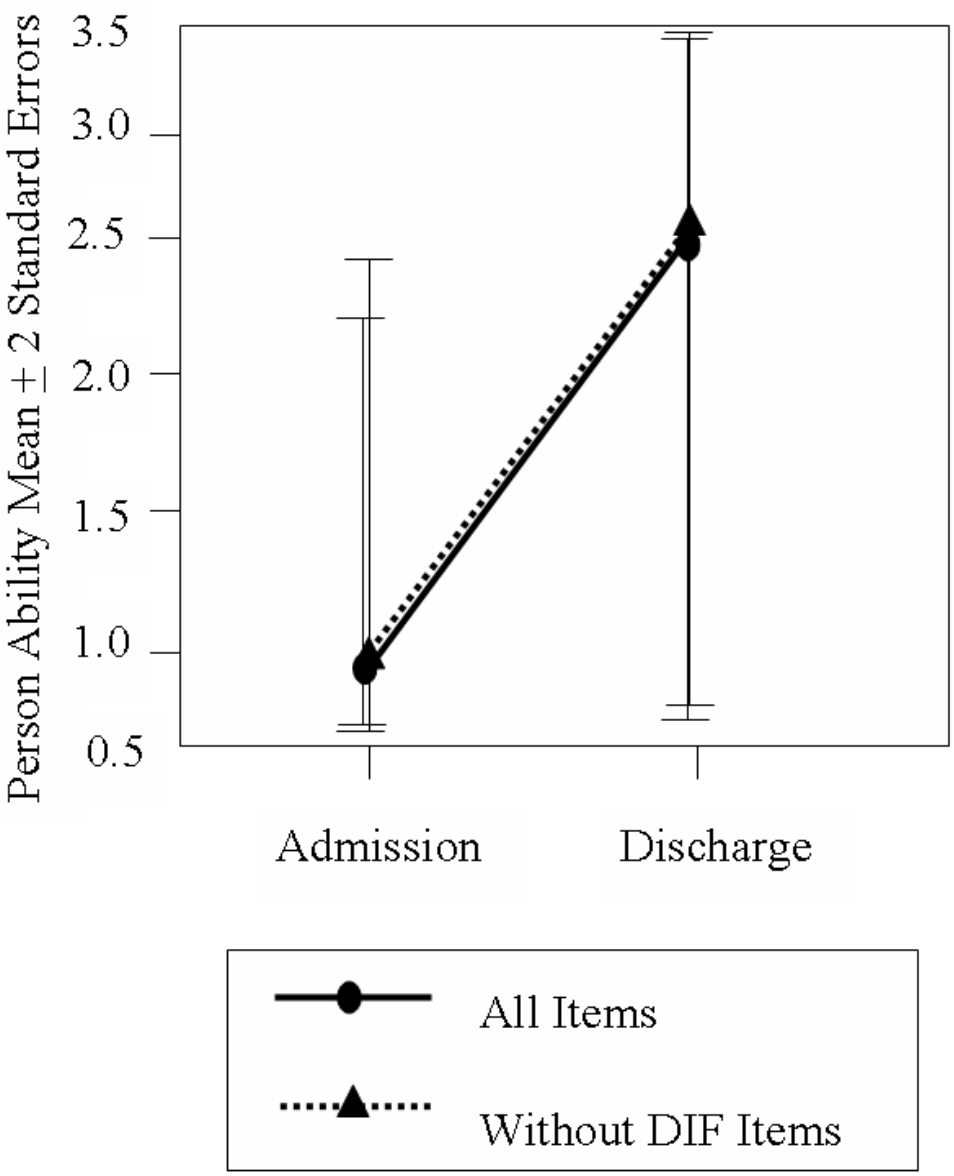


Figure 3-5. With all items and with DIF items deleted mean person ability measures (y-axis, logits) at admission and discharge

Table 3-8. Item discrimination estimates

Item	Admission	Discharge
I feel less capable, less confident or less useful because of my arm, shoulder or hand problem	0.73	0.28
Recreational activities in which you take some force or impact through your arm, shoulder or hand	1.11	1.22
Recreational activities in which you move your arm freely (e.g., playing frisbee, badminton, etc.)	1.24	1.09
Do heavy household chores (e.g., wash walls, wash floors)	1.38	1.36
Arm, shoulder or hand pain when you performed a specific activity	1.04	1.23
Garden or do yard work	1.35	1.24
Arm, shoulder or hand pain	0.84	1.12
Carry a heavy object (over 10 lbs)	1.20	1.11
Weakness in your arm, shoulder or hand	1.00	1.06
Change a lightbulb overhead	1.06	1.05
Stiffness in your arm, shoulder or hand	0.77	0.82
Place an object on a shelf above your head	1.17	1.10
Open a tight or new jar	0.78	0.71
Wash your back	1.01	0.92
Limited in work or other regular daily activities as a result of your arm, shoulder, or hand problem	1.27	1.21
Difficulty sleeping because of the pain in your arm, shoulder or hand	0.39	0.82
Carry a shopping bag or briefcase	1.34	1.18
Put on a pullover sweater	1.05	1.08
Wash or blow dry your hair	1.15	1.17
Push open a heavy door	1.18	1.07
Make a bed	1.31	1.21
Interferred with your normal social activities with family, friends, neighbours or group	0.90	1.09
Tingling (pins and needles) in your arm, shoulder or hand	0.09	0.59
Use a knife to cut food	1.15	1.03
Prepare a meal	1.29	1.21
Sexual activities	0.70	0.57
Recreational activities which require little effort (e.g., cardplaying, knitting, etc.)	0.79	0.77
Write	1.04	1.01
Manage transportation needs (getting from one place to another)	1.00	0.99
Turn a key	1.01	0.96

Note: Items are arranged in decreasing order of admission difficulty.

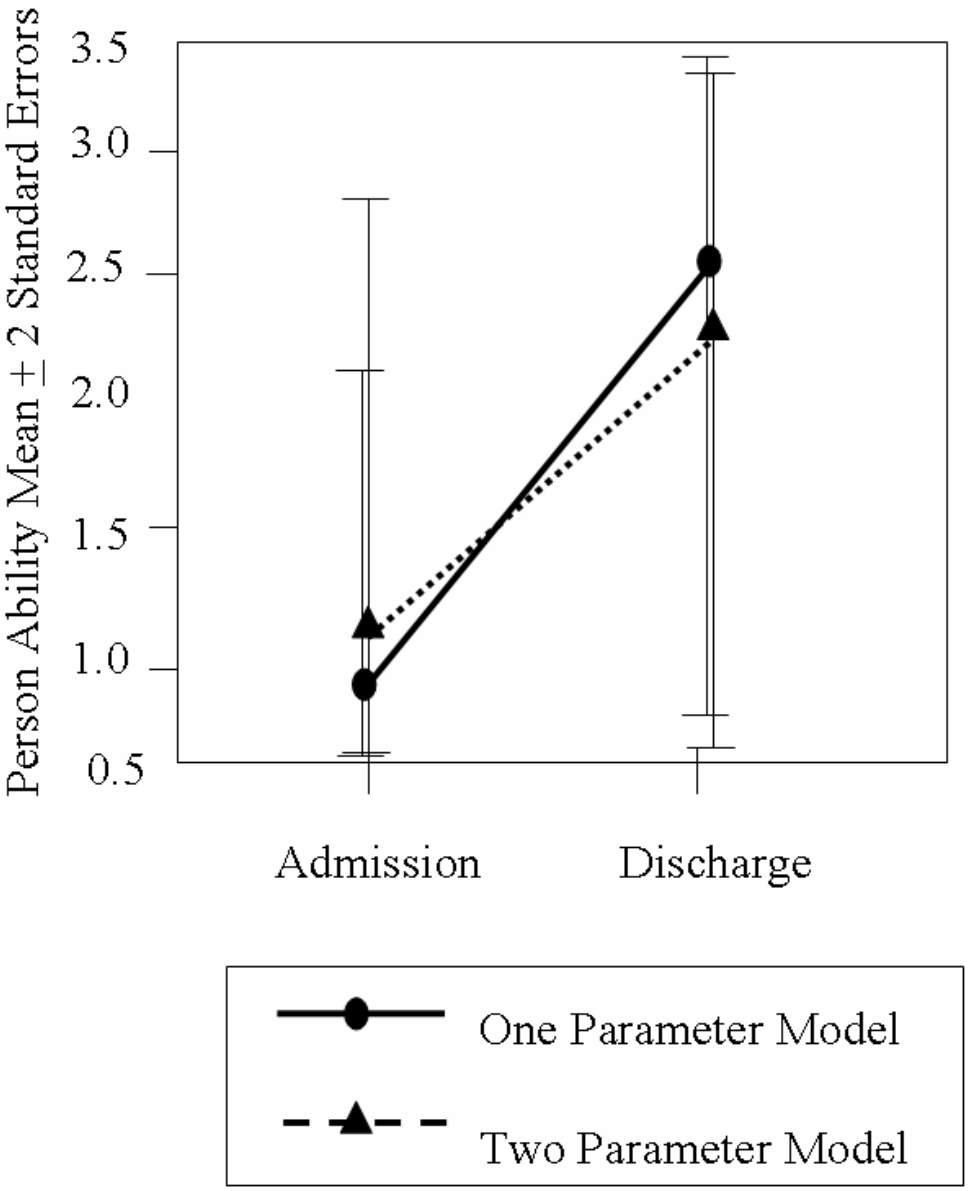


Figure 3-6. One parameter model versus two parameter model mean person ability measures (y-axis, logits) at admission and discharge

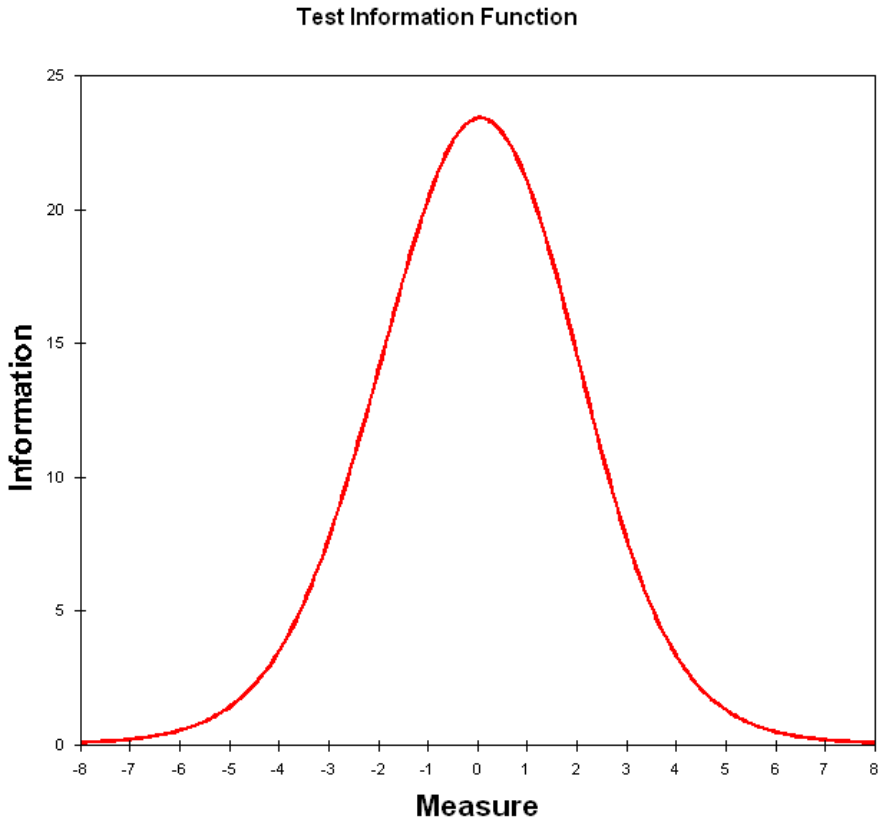


Figure 3-7. Admission scale information function using one parameter model

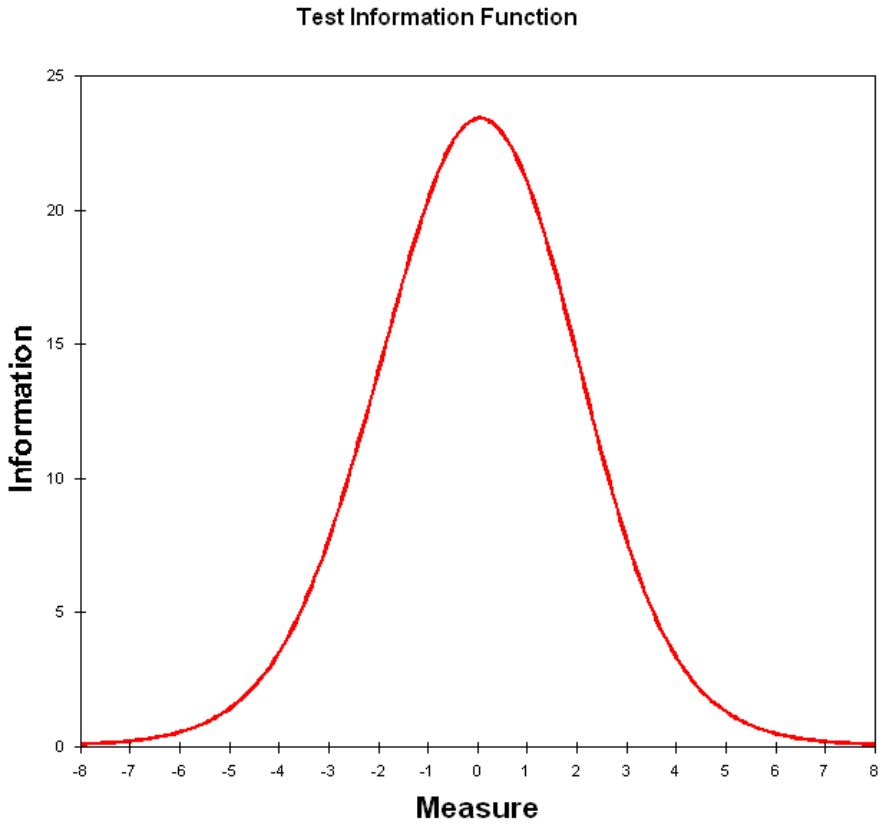


Figure 3-8. Discharge scale information function using one parameter model

CHAPTER 4
CREATING A CLINICALLY USEFUL DATA COLLECTION FORM FOR THE DISABILITIES OF THE ARM, SHOULDER, AND HAND OUTCOME QUESTIONNAIRE USING THE RASCH MEASUREMENT MODEL

Introduction

Numerous assessments exist which evaluate upper-extremity function. Therapists evaluating upper-extremity ability can chose from multiple performance measures, such as the Motor Assessment Scale (MAS)(148), Fugl-Meyer Assessment (FMA)(41), and Wolf Motor Function Test (WMFT)(47). Alternatively, self-reports of daily functional ability of the upper-extremity, such as the Disabilities of the Arm, Shoulder and Hand (DASH) outcome questionnaire(57), Upper-Extremity Functional Index (UEFI)(71), or Shoulder Pain and Disability Index (SPDI)(149), may be utilized. Yet another option is to assess capacity at the impairment level using measures such as the Minnesota Rate Manipulation Test (MRMT) or Purdue pegboard test(150).

Although various assessments and even types of assessments of the upper-extremity exist and have been shown to be reliable and valid(45, 47, 150-153), often they are not used in the clinic. This may be because results from these assessments do little to inform day-to-day clinical practice(154). When they are used, frequently this is because they are required by accrediting agencies or are necessary for reimbursement(154). In contrast, to being seen as useful tools to guide therapy, clinicians often view the completion of such assessments as taking valuable time away from treatment.

One reason that assessments fail to advise clinical practice is that they do not enhance clinicians' knowledge of what activities individuals can actually perform and what tasks should be set as the next best goal for a patient. Frequently assessment total scores are summed from patients' responses to the items on the assessment. This sum is not directly interpretable, i.e.,

when standing alone it only represents a mere number. The actual tasks that the individual is able or unable to complete are obscured when only the total score is known. By and large, assessments do not clearly quantify how well an individual will perform on specific items. Thus, they provide a vague picture of what goals should be set for a patient. In order to be useful in informing clinical decision-making (i.e., goal setting, treatment planning), an assessment must provide a directly interpretable description of a patient's abilities.

To illustrate how assessment total scores are unclear about actual ability to perform tasks, imagine a patient with a score of 70 on the Disabilities of the Arm, Shoulder, and Hand (DASH) outcome questionnaire(154). This score of 70 tells the clinician little about the problems the patient is experiencing in daily life and what goals should be set to help conquer these challenges. Can this individual carry a shopping bag without difficulty? Is opening a jar lid troublesome? Or is it only when this individual engages in recreational activities requiring force or impact to their arm that problems arise? A total score provides little information about the patient's present abilities and provides virtually no insight into his/her potential for improvement.

Moreover, when total scores are used, missing data becomes problematic. If several items do not apply and thus, are not answered by a patient, naturally the total score will be lower. However, one cannot assume this individual is less able than someone who answered all the questions and got a higher total score. In order to make results comparable even when items are not answered (i.e., do not apply to a given individual) requires time for additional mathematical calculations.

The Rasch measurement model, the one parameter item response theory model, provides a means through which assessments can be modified to more effectively inform clinical practice

(i.e., goal setting and treatment planning)(155). This method involves redesigning assessments into meaningful data collection forms. Rasch analysis makes the creation of such data collection forms possible by placing an individual's ability measures and item difficulty estimates on the same continuum (i.e., calibrating the item difficulty estimates relative to the sample's ability or put another way, arriving at difficulty estimates based on person responses). On the data collection forms, items are ordered according to their difficulties relative to each other (most difficult items at the top, followed by less difficult ones). Person ability measures are transformed to range from zero to 100. Patients circle rating scale choices for each item (ranging from unable to no difficulty), which are arranged according to where they fall in relation to the person ability scale (0-100). Someone looking at such a data collection form is able to quickly evaluate what items an individual is having challenges with and which they are not(156). This paper intends to generate and show the application of Rasch based data collection forms using the Disabilities of the Arm, Shoulder, and Hand (DASH) outcome questionnaire. The output from Rasch analysis that assists in the creation of these data collection forms has been termed a "keyform" by Linacre(156) who first proposed the idea. Thus, the term "keyform" will be used in this paper to refer to the raw output from the statistical program. The term "data collection form" will be used to refer to a more clinically useful modification of this output.

To further clarify how the Rasch measurement model makes possible the creation of such a data collection form, requires an explanation of several concepts that are foundational to this model. First, it must be conceivable that all the items on an assessment contribute to one idea or construct (for instance, all items relate to upper-extremity function). If this assumption (i.e., that all items contribute to one construct) is upheld, Rasch analysis can be used to turn ordinal data into equal interval data, which has a unit of measurement called a "logit" (log-odds unit). All the

items on the assessment can be seen as making up a "ruler" with each item's difficulty estimate (logit measure) falling at a specific marking. Person ability estimates (also calculated in logits based on the individual's responses to the items) can be placed on the same continuum allowing for the direct comparison of person ability to item difficulty(155). The formation of this person ability/item difficulty continuum is a second concept of importance in the creation of efficacious data collection forms. Person ability estimates indicate that someone possesses more or less of the construct being measured. For instance, on the DASH, a patient with a lower person measure would have less upper-extremity ability than one with a higher person measure. This holds regardless of whether or not all the items applied to these individuals (i.e., all the items were completed by both individuals).

The purpose of this study is three-fold: (1) To demonstrate, using data collected on the DASH, how Rasch methodologies can be used to generate a clinically useful data collection form, (2) To show how different ability study participants' responses (at admission and discharge) would look on the data collection form, and finally (3) To demonstrate how these forms can be useful in setting patient goals. Prior to addressing these goals, analyses will be presented to determine if two assumptions of the Rasch model are met. First, to determine as accurately as possible with the given sample size whether the required assumption of unidimensionality is met (i.e., whether all the items can be conceptualized as representing one construct or trait) and second, to determine whether items on the DASH follow a logical progression of difficulty.

Methods

Sample Characteristics

Thirty-seven participants with upper-extremity impairments were recruited from outpatient clinics in Gainesville, Florida including the Malcom Randall Veterans Affairs Medical Center

(VAMC), Orthopedic Center at the University of Florida, Student Health Care Center at the University of Florida, as well as, other outpatient clinics in the University of Florida/Shands Healthcare system. Table 4-1 presents demographic information for these participants. Admission and discharge data was obtained from all study participants. The sample included slightly more females (54.1%) than males. The mean age was 39.0 years ranging from 19 to 67 years. The mean number of medications the participants were currently taking was 3.3 and the mean number of surgeries was 1.5. Most participants reported that they exercise one or two times a week (43.2%), followed by exercising at least three times a week (29.7%), and seldom or never exercising (27.0%). Patient diagnoses included a wide variety of ailments such as wrist tendonitis, lateral and medial epicondylitis, basal joint arthritis, finger fractures, olecranon bursitis, humeral fracture, shoulder dislocation, Dupytren's, and carpal tunnel syndrome.

Assessment

Disabilities of the Arm, Shoulder, and Hand (DASH) Outcome Questionnaire

The Disabilities of the Arm, Shoulder, and Hand (DASH) outcome questionnaire was designed to measure impairment subjectively, as well as, to capture limitations in activities and participation imposed by single or multiple disorders of the upper-extremity(102). It is based on the World Health Organization (WHO) Model of Health, at the time of development called the International Classification of Impairments, Disabilities, and Handicaps (ICIDH) (revised to be the International Classification of Functioning, Disability and Health (ICF))(103). The DASH is a standardized instrument and evaluates impairments and activity limitations, as well as, participation restrictions in both leisure and work (103). It consists of two components: 30 disability/symptom questions and four optional high performance sport/music or work questions(104). Examples of items on the DASH include: Open a tight or new jar, Write, Make a bed, and Use a knife to cut food. Response choices for disability items range from 1 to 5 (1: no

difficulty to 5: unable)(104). Response choices for symptom items also use a five-point scale, but vary from none to extreme. Guidelines for the original DASH, suggest that at least 27 of the 30 disability/symptom questions must be completed for a score to be calculated(104). General procedures for scoring the DASH involve summing the response choices and averaging them to produce a score out of five(104). This value is then transformed to a score out of 100 by subtracting one and multiplying by 25(104). For example, if the patient responded to all the questions with a 5, they would have an average score of 5. You would then subtract one from 5, giving you 4 and multiply by 25. This individual, with the highest responses possible, would have a score of 100. Normally, this is the method used to produce DASH total scores between 0 and 100 with a higher score indicating more disability(104). For this study, however, the rating scale was recoded such that a higher score indicated more ability. Thus, 1 was unable (for disability items) or extreme (for symptom items) and 5 no difficulty (for disability items) or none (for symptom items). Rasch methods, outlined in the analysis section, were used to produce person measure scores ranging from 0 to 100. Only the 30 disability/symptom questions were used for this study (i.e., items from the four optional high performance sport/music sections were not completed by individuals in the current sample). Appendix A includes a list of all the items on the DASH. Note the DASH presented in the appendix is in the form most commonly used in clinics.

Procedures

Participant inclusion criteria were upper-extremity impairment and receiving treatment for the condition. Study participants were given $10 for completing the DASH at two time points (admission and discharge). This study was approved by the Institutional Review Board (IRB) at the University of Florida.

Analyses

Microsoft Access(157) and SPSS(105) were used for data management and descriptive statistics. Winsteps(1) was used for the Rasch analysis. To test the Rasch model assumption of unidimensionality, goodness of fit statistics were obtained. Given the small number of study participants, factor analysis was not possible with this sample, however, results of a previous factor analysis of the DASH with a much larger sample is reported to provide further information regarding the unidimensionality of the items. The difficulty of each item was calibrated and the order was evaluated based on tenants of motor control theory to assess whether the hierarchy was logical (i.e., intuitively believable). Once model assumptions were tested, a general keyform was produced, using Winsteps(1), based on the current sample data. From this keyform, a data collection form was created. Responses at admission and discharge from three patients of differing ability levels (high, medium, and low) were circled on the data collection form to illustrate how the forms can be used for informing goal setting and treatment planning.

Item infit analysis and previous factor analysis with larger sample to test Rasch measurement model fit

All of the items on a measure must contribute to a single construct or idea(116). Using Rasch analysis, the extent to which items represent a unidimensional construct is evaluated employing mean square standardized residuals (MnSq) produced for each item on the instrument. MnSq represents observed variance divided by expected variance(121). Consequently, the desired value of MnSq for an item is 1.0. The acceptable criterion depends on the intended purpose of the measure and the degree of rigor desired. For surveys using rating scales (such as the DASH), Wright and Linacre(158) suggest reasonable ranges of MnSq fit values between 0.6 and 1.4 associated with standardized Z values < 2.0. A low MnSq value (i.e., <0.6) suggests that an item is failing to discriminate individuals with different levels of ability

(i.e., people with different amounts of difficulty performing upper-extremity tasks) or that an item is redundant (i.e., other items on the instrument are of similar difficulty). High values (i.e., >1.4) indicate that scores are variant or erratic, suggesting that an item does not belong with the other items on the same continuum or that the item is being misinterpreted. Items with high MnSq values represent a threat to validity and thus, are given greater consideration. For the current study, misfit items will be considered those with infit MnSq values greater than 1.4 and standardized Z values (ZSTD) > 2.0. Separate analyses were conducted for admission and discharge data.

To further test for unidimensionality, principal components or factor analyses should also be conducted(124, 125). With the sample size in this study (N = 37), obtaining accurate results from such analyses is not possible. However, previously, with a much larger sample of DASH data (N = 991) with individuals with similar demographic characteristics, the authors conducted exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) on both admission and discharge data. Although the results of this analysis were somewhat mixed, the authors concluded that there was enough support for the unidimensionality of the DASH to proceed with the Rasch analysis (See Chapter 3).

Test of logic of difficulty hierarchy order

Based on the theory of motor control(128) hypotheses were developed about the hierarchy of difficulty of the items on the DASH. For example, based on the concept from motor control theory that complex items are more difficult, doing heavy household chores (e.g., wash walls, wash floors) was hypothesized to be harder than putting on a pullover sweater. Likewise, this theory contends that more challenging activities require the use of multiple joints. Thus, putting on a pullover sweater was hypothesized to be more difficult than turning a key in a lock. Rasch methodologies provide a means to investigate the validity of a proposed hierarchy by computing

item calibrations. Data was fit to a one parameter Rasch model using the Winsteps software program(1) to calculate item difficulty parameters. Separate analyses were done utilizing admission and discharge data. The resulting item difficulty order was examined to determine if it supported the basic tenants of motor control theory.

Obtaining a general keyform in winsteps for dash admission data

Winsteps Rasch analysis program(1) was used to obtain a general keyform for the DASH admission data. To make this scale more clinically useful, Winsteps commands (UMEAN and USCALE) were used to convert the person measures obtained from the raw data into person measures on a scale from 0 to 100. The UMEAN and USCALE values applied to achieve a 100 point scale were 48.6 and 13.0, respectively. Additionally, to obtain person measures at discharge, item average difficulties (i.e., difficulty of responding moderate difficulty to an item) and average rating scale step estimates (i.e., the estimated value for the transition from one rating scale choice to the next, for example, from mild to moderate difficulty) were anchored at admission values. This was to make person measures obtained at discharge comparable to those obtained at admission, thus, allowing for any changes in ability to be observed.

Creation of data collection form

The general keyform output from Winsteps was used as the basis for the creation of a data collection form. Patient instructions were added to the top, items were moved to the left of the response choices, and descriptions were added above the response choice numbers. A percentage person ability scale ranging from 0-100 was placed at the bottom of the form. Shading was added to improve readability of the data collection form.

Using the admission data, three patients were chosen: a high ability patient (i.e., one with a high person measure score), a medium ability patient (i.e., one with a medium person measure score), and a low ability patient (i.e., one with a low person measure score). These individuals'

responses were circled on data collection forms to illustrate how completed forms would look. Additionally, using Winsteps(1), person fit indices were obtained for these patients. Person fit indices, similar to those for item fit, are reported in terms of mean square standardized residuals (MnSq), observed variance divided by expected variance(121, 158). As suggested by Wright and Linacre(158), high infit MnSq values were considered those greater than 1.4 with standardized Z values greater than 2.0. For misfitting individuals, triangles were placed around unexpected responses (i.e., patient responses that were either two standard deviations higher or lower than would be expected based on his/her pattern of responses).

Patient responses at both admission and discharge for each of these three patients were circled on a form to show how patient progress might be observed using the form. On the forms, the patients' person ability measures (calculated in Winsteps using the conversion to a scale out of 100) were found at the bottom of the scale. A solid line was drawn up at this point to represent where the patient's ability measure fell in relation to his/her response to each item. Dotted lines were plotted on each side of the solid line to represent two standard errors associated with the person ability measure. These dotted lines were omitted only when they went outside the scale (0-100).

Finally, two sets of data collection forms (admission and discharge for selected patients) were created with patient goals indicated. One set for a high ability patient at admission and one for a low ability patient at admission. A box was placed around potential shorter term goal activities and potential longer term goal activities on the admission form. The discharge form illustrates whether these goals had been achieved.

Results

Item Infit Analysis and Previous Factor Analysis with Larger Sample to Test Rasch Measurement Model Fit

To provide some evidence that the requirement of unidimensionality was met with this sample (despite the small sample size), infit statistics from Rasch analysis were calculated. Overall the items worked well together with 90% of the items having a mean infit within the acceptable criterion at admission and 83% at discharge. However, at admission 10% (3/30) of the items were not within the acceptable criterion (MnSq ≤ 1.4; ZSTD < 2.0) and at discharge 17% (5/30) of the items were outside the acceptable values. At admission items with high infit included: Difficulty sleeping because of the pain in your arm, shoulder or hand, Recreational activities which require little effort (e.g., cardplaying, knitting, etc.), and Write, while at discharge the items with high infit were: Arm, shoulder or hand pain when you performed a specific activity, I feel less capable, less confident or less useful because of my arm, shoulder or hand problem, Stiffness in your arm, shoulder or hand, Recreational activities in which you move your arm freely (e.g., playing frisbee, badminton, etc.), and Write. The admission items were 53%-173% more erratic than expected and the discharge items were 58%-135% more erratic than expected. See Tables 4-2 and 4-3.

Test of Logic of Difficulty Hierarchy Order

Item difficulty estimates were investigated to determine if the items followed a logical progression of upper-extremity function difficulty, which could be used as a basis for goal setting and treatment intervention. Table 4-2 presents DASH item difficulty calibrations at admission and Table 4-3 presents DASH item difficulty calibrations at discharge in the order of relative challenge at admission. The most challenging items based on admission estimates are located at the top and the least challenging at the bottom. The most difficult items at admission

were: Recreational activities in which you take some force or impact through your arm, shoulder or hand, Arm, shoulder or hand pain when you performed a specific activity, and I feel less capable, less confident or less useful because of my arm, shoulder or hand problem, while the most difficult items at discharge were: Recreational activities in which you take some force or impact through your arm, shoulder or hand, I feel less capable, less confident or less useful because of my arm, shoulder or hand problem, and Open a tight or new jar. The least challenging items at admission included: Turn a key, Manage transportation needs, and Write. The least challenging items at discharge included: Turn a key, Manage transportation needs, and Sexual Activities.

Creation of Data Collection Form

From the general keyform produced by Winsteps Rasch analysis program(1), a data collection form was created. This form is presented in Figure 4-1. Patient instructions: Please rate your ability to do the following activities in the last week by circling the number below the appropriate response, are presented at the top of the form. Items are located at the left in order of decreasing difficulty at admission. Rating scale choices (1-5) are presented to the right of each item with descriptions of what ratings mean above the number responses. These rating scale choices are placed at a location relative to the person ability scale at the bottom. That is, a person with a person ability measure of 50, would be most likely to rate the item Make a bed- mild or moderate difficulty and to rate Carry a heavy object- moderate difficulty. In contrast, someone with a person ability measure of 30, would be most likely to rate the item Make a bed- severe difficulty and would most likely rating the item Carry a heavy object- unable. In contrast an individual with a person ability measure of 90, would most likely rate both items– no difficulty. Shading was included on the data collection form to improve readability.

Figure 4-2 presents six data collection forms that have been filled out according to responses of three different individuals at admission and discharge. Figure 4-2a shows responses of a patient who began treatment at a high ability level, Figure 4-2b illustrates responses of a patient who began treatment at a medium ability level, and Figure 4-2c contains responses of a patient who began treatment at a low ability level. The forms on the left have admission responses of these patients circled, while those on the right have discharge responses circled. Triangles are placed around unexpected responses (i.e., those where patients indicated a response that was either higher or lower than expected based on his/her other responses).

As can be seen in the first two forms in Figure 4-2a, the high ability individual entered treatment with a person measure of 71.8 ± 7.0. At this time, the individual had no difficulty with the majority of the easier items and moderate-mild difficulty on the majority of the more difficult items. At discharge, his/her person ability measure had risen to 97.9 $\pm$ 15.4. Now, this individual had no difficulty with all but three items (mild difficulty) and these three items had a tendency to be the more difficult items. This person did not have any responses that were unexpected based on his/her pattern of responses.

The patient who began treatment with medium ability (Figure 4-2b; person measure of 51.2 $\pm$ 4.6) had a pattern of responses that was somewhat unexpected, answering that they were unable to perform many of the easier items. This individual had a high misfit value associated with his/her pattern of responses (MnSq = 1.93, ZSTD = 3.3). The triangles around responses to the items I feel less capable, less confident or less useful because of my arm, shoulder or hand problem, Wash back, Carry a shopping bag or briefcase, and Write indicate that these responses were unexpected. On several of the easy items, this individual had mild to no difficulty. With the middle level items, this individual generally reported having mild difficulty and with hard

items generally reported being unable to perform the activity or able to perform the activity with mild difficulty. At discharge, this patient reported having no difficulty with all of the items except one, Arm, shoulder or hand pain when you performed a specific activity. His/her person measure at this time reached the maximum of 100 $\pm$ 26.2.

Finally, the patient who had low ability at admission (Figure 4-2c; person measure score of 30.5 $\pm$ 5.8), was having severe difficulty with five of the activities, moderate difficulty with five of the activities, and was unable to complete the rest of the activities. Three of this patient's responses were unexpected (as shown by the triangles around these responses). Write was answered by this individual as being easier than expected, as were Arm, shoulder or hand pain, and Limited in your work or other regular activities as a result of your arm, shoulder, or hand problem. At discharge, this individual's person measure had increased to 80.9 $\pm$ 8.8 and reported having mild difficulty with eight of the items, moderate difficulty with one item, and no difficulty with the rest of the items.

Figure 4-3 illustrates how goals might be set for patients of differing admission ability levels (a high ability patient at admission and a low ability patient at admission) using this type of data collection form. For the patient with high ability at admission (Figure 4-3a), shorter term goals might include increasing range of motion for back washing, increasing strength for carrying heavy objects, and interventions to decrease pain and tingling sensations. Longer term goals for this patient might include increasing hand strength for activities such as opening a jar lid and further decreasing pain.

Using a functional approach to treatment (such as one used in Constraint Induced Movement Therapy, CIMT), this patient might be asked to perform activities such as combing his/her hair, reaching overhead to get objects out of a cabinet, carrying objects from one side of

the room to another, and opening different sized containers. At discharge, as seen in Figure 4-3a, this patient had met all his/her shorter term goals, but was still having mild difficulty with pain (longer term goal).

Figure 4-3b illustrates goal setting for a patient with low ability at admission. This individual's responses are circled at admission on the left form and at discharge on the right form. With this individual, shorter term goals might include increasing shoulder range of motion for completing tasks such as putting on a pullover sweater, washing and blow drying his/her hair, changing a light bulb overhead, and placing an object on a shelf above his/her head. Fine motor ability might be addressed to help ease tasks such as using a knife to cut food. Perhaps, longer term goals for this patient would include further increasing range of shoulder motion, so that washing his/her back, recreational activities in which you move your arm freely, and increasing strength for heavy household chores, such as washing walls and floors, were easier.

Functional activities to work on with this individual might include placing various objects on shelves above his/her head to increase shoulder range of motion. Improving fine motor ability through practice picking up various sized coins. Finally, lifting and moving heavy books or grocery bags might improve strength. At discharge, this individual had met all of these shorter and longer term goals.

Discussion

In summary, the goals of this study were to demonstrate, using data collected on the DASH, how Rasch methodologies can be used to create a clinically useful data collection form, to illustrate how different ability study participants' responses would look on the data collection form at admission and discharge, and to demonstrate how these forms can be used in setting patient goals. Prior to addressing these goals, psychometric evidence was presented to support the use of Rasch analysis for the creation of this data collection form (i.e., that necessary

assumptions were met). The two assumptions that were tested were model fit (i.e., that all items contribute to a single underlying trait) and the logic (i.e., believability) of the hierarchy of item difficulty.

In regards to the infit analysis, with the admission data, only three items misfit: Difficulty sleeping because of the pain in your arm, shoulder or hand, Recreational activities which require little effort (e.g., cardplaying, knitting, etc.), and Write. Perhaps, these items misfit because many of the individuals did not have pain to the extent that it interfered with their sleeping or did not have fine motor issues that would have affected activities such as writing. Therefore, they were uncertain how to respond to these items. At discharge five items misfit. Two of these were symptom items: Arm, shoulder or hand pain when you performed a specific activity and Stiffness in your arm, shoulder or hand. Once again, many individuals may not have been experiencing these symptoms. Another item that misfit at discharge was a more general item that may have been affected by life issues other than the patients' upper-extremity problems (I feel less capable, less confident or less useful). The small number of misfit items seems to indicate that the items could be conceived as representing one construct. Nevertheless, some authors suggest that it is necessary for less than 5% of the items to misfit to assume unidimensionality of a measure(159). This was not the case here, since 10% (3/30) of the items misfit with the admission data and 17% (5/30) misfit with the discharge data.

With the larger data set from the previous study, results of exploratory and confirmatory factor analyses showed some support for a one factor solution. Eigenvalues for the first factor in both the admission and discharge exploratory factor analyses were substantially greater than for the second factor. One factor accounted for most of the variance (> 60%) for both admission and

discharge data and all items loaded high on the first factor. The first two factors were highly correlated.

In general, the item difficulty hierarchies supported the theory of motor control. Four tenants of motor control theory are: 1) complex items are more difficult, 2) environmental factors increase the challenge of items, 3) more difficult items require the use of multiple joints, and 4) isolated, abstract items are more difficult than functional tasks(128, 145, 146). Complex items were found to be more difficult. For example, doing heavy household chores (e.g., wash walls, wash floors) was harder than putting on a pullover sweater. Likewise, gardening or doing yard work was more difficult than carrying a shopping bag or briefcase. Environmental factors were shown to influence difficulty ranking. Recreational activities requiring little effort, such as card playing or knitting, were rated easier than recreational activities in which you move your arm freely, such as frisbee or badminton. Furthermore, recreational activities in which you move your arm freely were rated easier than recreational activities in which you take some force or impact through your arm, shoulder, or hand. Items requiring the use of multiple joints were more difficult than items requiring the use of fewer joints. For instance, recreational activities, such as frisbee or badminton, were rated more difficult than using a knife to cut food. Finally, isolated tasks, presented without a concrete context or purpose were rated as harder. To illustrate, carrying an unspecified heavy object was rated as harder than carrying a heavy shopping bag or briefcase.

The general keyform output from Winsteps Rasch analysis program(1) was used as the basis for creating the data collection form. Creation of such a form proved to be an easy process and potentially offers several advantages over the forms used in clinics today. Instructions were placed at the top of the form and items (activities for patients to rate) were listed to the left of the

possible answer choices (both numbers to circle and descriptions of what the numbers mean; for example, no difficulty placed above 5). A person measure scale ranging from zero to 100 was placed at the bottom of the form.

As depicted by the forms filled out according to patients of differing admission ability levels (high, medium, and low), at a glance it is possible to get a "picture" of an individual's overall ability to perform functional tasks requiring the upper-extremity. The person ability measure scale at the bottom of the form aids in this estimation. On the forms presented with study participant data, a line was drawn up from the actual person measure estimate (i.e., the estimate of person ability derived for that person in Winsteps). It has been suggested that when computer software is not available to estimate this measure (as is most often the case in the clinic), a line can be drawn by visually determining where most of the responses fall(156). Such a line could be estimated even with missing responses to some questions. The down side of this is that the accuracy of such a line drawn by sight is questionable. As seen in the admission form illustrating a person of medium ability (left form in Figure 4-2b), his/her responses are somewhat inconsistent and scattered. Determining an estimate of this person's ability may be challenging due to the variability of his/her responses.

In addition, some patients' response patterns differ from what would be expected based on the obtained hierarchy(155, 156). Consider once again the medium ability individual (left form in Figure 4-2b). His/her pattern of responses at admission does not follow the hierarchy of item difficulty very well (i.e., easiest items being rated less challenging than more difficult items). However, many of the easy items, with which this patient is having a lot of difficulty, require shoulder range of motion (changing a light bulb overhead, putting on a pullover shoulder). In

this way, unusual patterns in the data can be diagnostic. Possibly, this individual has some type of shoulder impairment.

Another advantage of this type of data collection form is that they can be used in conjunction with impairment and performance assessments to aid goal setting and treatment planning. The forms filled out with actual patient data illustrate this benefit. Examining such a form gives a sense of what activities and issues an individual is struggling with completing. For example, on the forms created and used to demonstrate goal setting (see Figure 4-3), with the patient with high ability at admission, it is possible to see that this individual potentially has some range of motion issues (because of difficulty with the item, Wash your back), some problems with strength (because of indicated difficulty with the item, Carry a heavy object (over 10 lbs)), and some active symptoms, because of indicated difficulty with the items Tingling (pins and needles) in your arm, shoulder or hand and Difficulty sleeping because of pain in your arm, shoulder or hand. Using this information, the therapist is able to set goals that will help achieve improved range of motion and strength and decreased negative symptoms (i.e., pain and tingling). On re-evaluation, the clinician once again can see at a glance whether or not these items have improved and areas where new goals might be set (where lower numbers are still being circled). New goals for this patient might include increasing hand strength for activities such as opening a jar lid and further decreasing pain. Finally, at discharge treatment effectiveness can be determined by whether or not the patient is circling response choices to the right of the form (4's and 5's). With the high ability patient in Figure 4-3a, all shorter term goals were met with only one longer term goal not met (Arm, shoulder, or hand pain).

The patient with low ability at admission also provides a dramatic example of successful treatment. Looking at the admission form filled out by this patient (Figure 4-3b) one could

easily see that this individual seems to be having some difficulty with shoulder range of motion and could set shorter term goals to improve this resulting in less difficulty with daily tasks such as pullover sweater, washing and blow drying his/her hair, changing a light bulb overhead, and placing an object on a shelf above his/her head. Additionally, a goal to improve fine motor skills might be set in order to make daily tasks such as using a knife to cut food easier. Once these goals were obtained, longer term goals for this patient might include further increasing range of shoulder motion, so that washing his/her back, recreational activities in which you move your arm freely, and heavy household tasks were easier. At discharge, this patient had met all of these goals.

Several studies exist which have discussed the creation of keyforms utilizing Rasch methodologies. These studies have taken a slightly different focus than the current study. Earlier work by Linacre(156) and Kielhofner(155), focused on the keyform as a way of obtaining "instantaneous measurement". These authors center their discussion on how using the keyform, without the use of computer software, a line can be drawn to represent an overall person measure for a patient and how patterns of responses can be used diagnostically. Linacre(156) created a keyform based on data on the Functional Independence Measure (FIM), while Kielhofner used data from the Occupational Performance History Interview –2nd Version (OPHI-II). Another approach in using keyforms was demonstrated by Woodbury and colleagues(160). Keyforms created from data on the Fugl-Meyer Assessment of the upper-extremity were used to validate the consistency of the pattern of responses across study participants. Using the keyform, it was possible to verify that participants of differing abilities were scoring higher on easier items and lower on harder items. Evidence of this scoring pattern indicates that the Fugl-Meyer Assessment retains its structure when measuring subjects across the ability range.

Work by Avery and colleagues(159) using data collected on the Gross Motor Ability Estimator (GMAE) demonstrated how a computer program could be used to generate output in a form similar to that of a keyform produced by Winsteps. The item hierarchy obtained by these authors shows a clear pattern of gross motor development. For instance, the hierarchy found by these authors showed that a baby can lift its head before pulling to a sit, crawling follows attaining a sitting position, and walking and hopping come later in development. In order to obtain the keyform output described in this study, clinicians would be required to enter data collected on an assessment into a computer. However, these authors report that over 50 therapists in Ontario said they would be interested in using such a program(159). The advantage of a data collection form designed using the same methodologies, however, is that no computer entry would be required of the therapists. This study represents the first time a data collection form has been created for the DASH based on keyforms generated through Rasch methodologies. Moreover, there are no published studies on keyforms that present longitudinal data (i.e., data at both admission and discharge).

There are several limitations to this study. Although, the previous study with a larger sample, showed some support for the unidimensionality of the DASH, other evidence from that study was inconclusive. Perhaps, as indicated by the items loading highest on three different factors in previous work, the DASH should be split into three separate subscales. This would provide for a clearer item difficulty hierarchy and thus, a more obvious progression a patient function. Factor analyses suggested the three constructs present could represent: (1) gross motor, complex items, requiring more movement, (2) items requiring hand use, and (3) general/ symptom items (see Chapter 3). Conceptually, it is reasonable to assume that an individual with pain as a symptom of his/her condition would rate pain items harder than other individuals.

Likewise, one with fine motor problems would rate items requiring hand use harder than more gross motor items. Thus, having items on one instrument that inquire about pain and other symptoms, fine motor tasks, and more gross motor tasks may in some ways confuses the item difficulty order. Possibly, then three separate data collection forms should be created for the DASH. Nonetheless, dividing up the DASH, would result in subscales with a small number of items and unknown psychometric properties.

This leads to another issue regarding the creation of a data collection form for the DASH or any other well established instrument. The reliability and validity of the original DASH has been well-established(55, 57, 95, 102, 161). If the items are reordered according to relative challenge and the rating scale modified so that higher numbers indicate more ability instead of more disability, how would this affect previously established psychometric properties of the assessment? Of concern is that the ordering of items with the most difficult at the top (read and answered first by the patient) and least difficult items at the bottom (read and answered last by the patient), might influence how the questions are perceived and answered. Such an influence of item ordering has been a well known and studied phenomenon in psychological literature for many years(162, 163). If an individual unconsciously recognizes that the activities in question are getting easier, they may start to respond accordingly without critically thinking about how challenging the task really is for them.

Of concern is the issue that the general keyform produced using Rasch measurement methodologies and the difficulty hierarchies ascertained differ slightly when produced using admission data from when produced using discharge data. If the difficulty ordering of items on an assessment were to vary widely from admission to discharge this could compromise the usefulness (and validity) of a data collection form produced using the methods in this study.

However, previous analyses by the authors (see Chapter 3) showed that with a larger sample size DASH item difficulties from admission to discharge were stable enough to not affect person ability estimates. Additionally, in the current study, the concern of differing admission and discharge difficulty estimates, was handled by obtaining person measures at discharge after anchoring on the basis of item and step difficulties at admission. Then, the admission general keyform was used to produce the data collection form. This was based on the thinking that therapists would desire to see how much a patient's ability changed from the point of admission. One further problem with this anchoring method was that since difficulties were anchored based on admission data, many of the participants ended up with maximum person ability estimates at discharge.

A final limitation in this study, was the small sample size. This has already been discussed in terms of establishing unidimensionality of the DASH. Beyond this, with such a small sample, the hierarchy arrived at and used for the creation of the data collection form, may not be stable (i.e., might vary from sample to sample). Preliminary analysis by the authors using a much larger sample of DASH data resulted in the production of a very similar keyform. Nevertheless, collection of more data, re-creation of the data collection form, and comparison to the one designed in this study is advised.

In conclusion, the positive implications of a data collection form like the one presented in this paper are numerous. First, as demonstrated by the completed forms in this study, this type of assessment form could be clinically useful to therapists, aiding in goal setting and treatment planning. Second, once such a form was adopted and widely used in clinics, it could be further used to justify reimbursement by insurance companies and by accreditation agencies. The clear illustration of improvement shown by this type of form would make an impressive statement

about the possibilities of making functional gains from treatment. Finally, if outcomes of standardized assessments were designed, like the ones presented in this study, and became useful tools for clinicians, this would lead to many more possibilities for comparison of treatment effectiveness between therapists at different sites.

Future work comparing the psychometric properties of the created DASH data collection form to already established reliability, validity, and responsiveness studies of the original DASH is necessary. Additionally, therapist perceptions on the usefulness of such data collection forms would be beneficial. Perhaps, clinicians might suggest further modifications that would aid the treatment planning process. Actual trials in the clinic with these forms are needed to determine if these data collection forms are useful in actual practice. Moreover, the creation and testing of data collection forms using Rasch methodologies with other assessments would enhance the evidence supporting the feasibility of such forms.

Table 4-1. Sample of 37 demographics

Mean age ± SD	39.0 ± 15.3 (N = 37)
95% Confidence Interval (CI)	CI = 33.4, 43.9
Gender	
- Female	20 (54.1%)
- Male	17 (45.9%)
	(N = 37)
Mean number of medications taking ± SD	3.3 ± 4.6 (N = 36*)
95% Confidence Interval (CI)	CI = 1.6, 4.0
Mean number of surgeries + SD	1.5 ± 2.1 (N = 35*)
95% Confidence Interval (CI)	CI = 0.8, 2.2
Exercise history	
- At least three times a week	11 (29.7%)
- One or two times a week	16 (43.2%)
- Seldom or never	10 (27.0%)
	N = 37

Note: Several of the totals presented in table do not total 37 due to missing data.

Table 4-2. Admission item fit and difficulty estimates

Item	Item fit (MNSQ)	ZSTD	Admission difficulty	Error
Recreational activities in which you take some force or impact through your arm, shoulder or hand	0.71	-1.40	0.93	0.18
Arm, shoulder or hand pain when you performed a specific activity	0.71	-1.40	0.79	0.18
I feel less capable, less confident or less useful because of my arm, shoulder or hand problem	1.49	1.60	0.70	0.22
Carry a heavy object (over 10 lbs)	0.87	-0.50	0.53	0.18
Garden or do yard work	0.68	-1.60	0.50	0.18
Open a tight or new jar	1.15	0.70	0.41	0.18
Limited in your work or other regular daily activities as a result of your arm, shoulder, or hand problem	0.69	-1.50	0.38	0.18
Arm, shoulder or hand pain	0.86	-0.60	0.38	0.18
Weakness in your arm, shoulder or hand	0.99	0.00	0.36	0.18
Do heavy household chores (e.g., wash walls, wash floors)	0.66	-1.70	0.35	0.18
Stiffness in your arm, shoulder or hand	1.30	1.30	0.27	0.18
Recreational activities in which you move your arm freely (e.g., playing frisbee, badminton, etc.)	1.19	0.90	0.25	0.18
Wash your back	0.94	-0.20	0.06	0.18
Tingling (pins and needles) in your arm, shoulder or hand	1.01	0.10	0.02	0.19
Difficulty sleeping because of the pain in your arm, shoulder or hand	1.59	2.10	-0.08	0.20
Carry a shopping bag or briefcase	1.15	0.70	-0.11	0.18
Interferred with your normal social activities with family, friends, neighbours or groups	0.73	-1.20	-0.11	0.19
Place an object on a shelf above your head	1.05	0.30	-0.14	0.18
Make a bed	0.76	-1.00	-0.21	0.19
Push open a heavy door	1.07	0.40	-0.24	0.19
Change a lightbulb overhead	0.80	-0.80	-0.24	0.19
Wash or blow dry your hair	0.69	-1.40	-0.35	0.19
Put on a pullover sweater	0.92	-0.30	-0.35	0.19
Use a knife to cut food	0.68	-1.40	-0.38	0.19
Sexual activities	0.98	0.00	0.42	0.19
Recreational activities which require little effort (e.g., cardplaying, knitting, etc.)	1.53	2.00	-0.5	0.19
Prepare a meal	0.74	-1.10	-0.53	0.19
Write	2.73	4.90	-0.69	0.20
Manage transportation needs (getting from one place to another)	1.52	1.90	-0.77	0.20
Turn a key	1.19	0.80	-0.81	0.21

Note: Items are arranged in decreasing order of admission difficulty.

Table 4-3. Discharge item fit and difficulty estimates

Item	Item fit (MNSQ)	ZSTD	Discharge difficulty	Error
Recreational activities in which you take some force or impact through your arm, shoulder or hand	0.99	0.00	1.40	0.22
Arm, shoulder or hand pain when you performed a specific activity	1.69	2.30	0.77	0.22
I feel less capable, less confident or less useful because of my arm, shoulder or hand problem	2.35	3.40	0.88	0.25
Carry a heavy object (over 10 lbs)	0.82	-0.70	0.62	0.23
Garden or do yard work	0.58	-1.70	0.35	0.23
Open a tight or new jar	0.59	-1.80	0.87	0.22
Limited in your work or other regular daily activities as a result of your arm, shoulder, or hand problem	0.61	-1.60	0.51	0.23
Arm, shoulder or hand pain	0.89	-0.30	0.24	0.24
Weakness in your arm, shoulder or hand	0.86	-0.40	0.27	0.24
Do heavy household chores (e.g., wash walls, wash floors)	0.38	-2.90	0.24	0.24
Stiffness in your arm, shoulder or hand	1.88	2.50	-0.03	0.25
Recreational activities in which you move your arm freely (e.g., playing frisbee, badminton, etc.)	1.58	2.00	0.67	0.23
Wash your back	0.57	-1.70	-0.05	0.25
Tingling (pins and needles) in your arm, shoulder or hand	1.24	0.90	-0.05	0.25
Difficulty sleeping because of the pain in your arm, shoulder or hand	1.22	0.80	0.10	0.25
Carry a shopping bag or briefcase	1.06	0.30	-0.24	0.25
Interferred with your normal social activities with family, friends, neighbours or groups	0.70	-1.10	0.24	0.24
Place an object on a shelf above your head	1.07	0.40	0.30	0.24
Make a bed	0.95	-0.10	-0.44	0.26
Push open a heavy door	0.55	-1.90	0.13	0.24
Change a lightbulb overhead	1.56	1.70	-0.37	0.26
Wash or blow dry your hair	0.76	-0.80	-0.18	0.25
Put on a pullover sweater	0.65	-1.20	-0.58	0.27
Use a knife to cut food	0.61	-1.30	-0.73	0.28
Sexual activities	1.54	1.60	-0.88	0.28
Recreational activities which require little effort (e.g., cardplaying, knitting, etc.)	0.70	-0.90	-0.80	0.28
Prepare a meal	1.08	0.40	-0.58	0.27
Write	1.78	2.20	-0.65	0.27
Manage transportation needs (getting from one place to another)	0.84	-0.40	-0.96	0.29
Turn a key	1.02	0.20	-1.05	0.29

Note: Items are arranged in decreasing order of admission difficulty.

INSTRUCTIONS: Please rate your ability to do the following activities in the last week by circling the number below the appropriate response.					
Activity					
	Unable	Severe Difficulty	Moderate Difficulty	Mild Difficulty	No Difficulty
Recreational activities in which you take some force or impact	1	2	3	4	5
	Extreme	Severe	Moderate	Mild	None
Arm, shoulder or hand pain when you performed a specific activity	1	2	3	4	5
	Agree	Neither Agree or Disagree			Disagree
I feel less capable, less confident or less useful because of my arm, shoulder or hand problem	1	2	3	4	5
	Unable	Severe Difficulty	Moderate Difficulty	Mild Difficulty	No Difficulty
Carry a heavy object (over 10 lbs)	1	2	3	4	5
Garden or do yard work	1	2	3	4	5
Open a tight or new jar	1	2	3	4	5
	Unable	Very Limited	Moderately Limited	Slightly Limited	Not At All Limited
Limited in your work or other regular daily activities as a result of your arm, shoulder, or hand problem	1	2	3	4	5
	Extreme	Severe	Moderate	Mild	None
Arm, shoulder or hand pain	1	2	3	4	5
Weakness in your arm, shoulder or hand	1	2	3	4	5
	Unable	Severe Difficulty	Moderate Difficulty	Mild Difficulty	No Difficulty
Do heavy household chores (e.g., wash walls, wash floors)	1	2	3	4	5
	Extreme	Severe	Moderate	Mild	None
Stiffness in your arm, shoulder or hand	1	2	3	4	5
	Unable	Severe Difficulty	Moderate Difficulty	Mild Difficulty	No Difficulty
Recreational activities in which you move your arm freely	1	2	3	4	5
Wash your back	1	2	3	4	5
	Extreme	Severe	Moderate	Mild	None
Tingling (pins and needles) in your arm, shoulder or hand	1	2	3	4	5
	Unable	Severe Difficulty	Moderate Difficulty	Mild Difficulty	No Difficulty
Difficulty sleeping because of the pain in your arm, shoulder or hand	1	2	3	4	5
Carry a shopping bag or briefcase	1	2	3	4	5
	Extremely	Quite a Bit	Moderately	Slightly	Not At All
Interfered with your normal social activities with family, friends, neighbours or groups	1	2	3	4	5
	Unable	Severe Difficulty	Moderate Difficulty	Mild Difficulty	No Difficulty
Place an object on a shelf above your head	1	2	3	4	5
Make a bed	1	2	3	4	5
Push open a heavy door	1	2	3	4	5
Change a lightbulb overhead	1	2	3	4	5
Wash or blow dry your hair	1	2	3	4	5
Put on a pullover sweater	1	2	3	4	5
Use a knife to cut food	1	2	3	4	5
Sexual activities	1	2	3	4	5
Recreational activities which require little effort (e.g., cardplaying, knitting)	1	2	3	4	5
Prepare a meal	1	2	3	4	5
Write	1	2	3	4	5
Manage transportation needs (getting from one place to another)	1	2	3	4	5
Turn a key	1	2	3	4	5
Percentage Ability Measure out of 100	0	10	20	30	40	50	60	70	80	90	100

Figure 4-1. Data collection form for the DASH (Based on analysis of admission data)

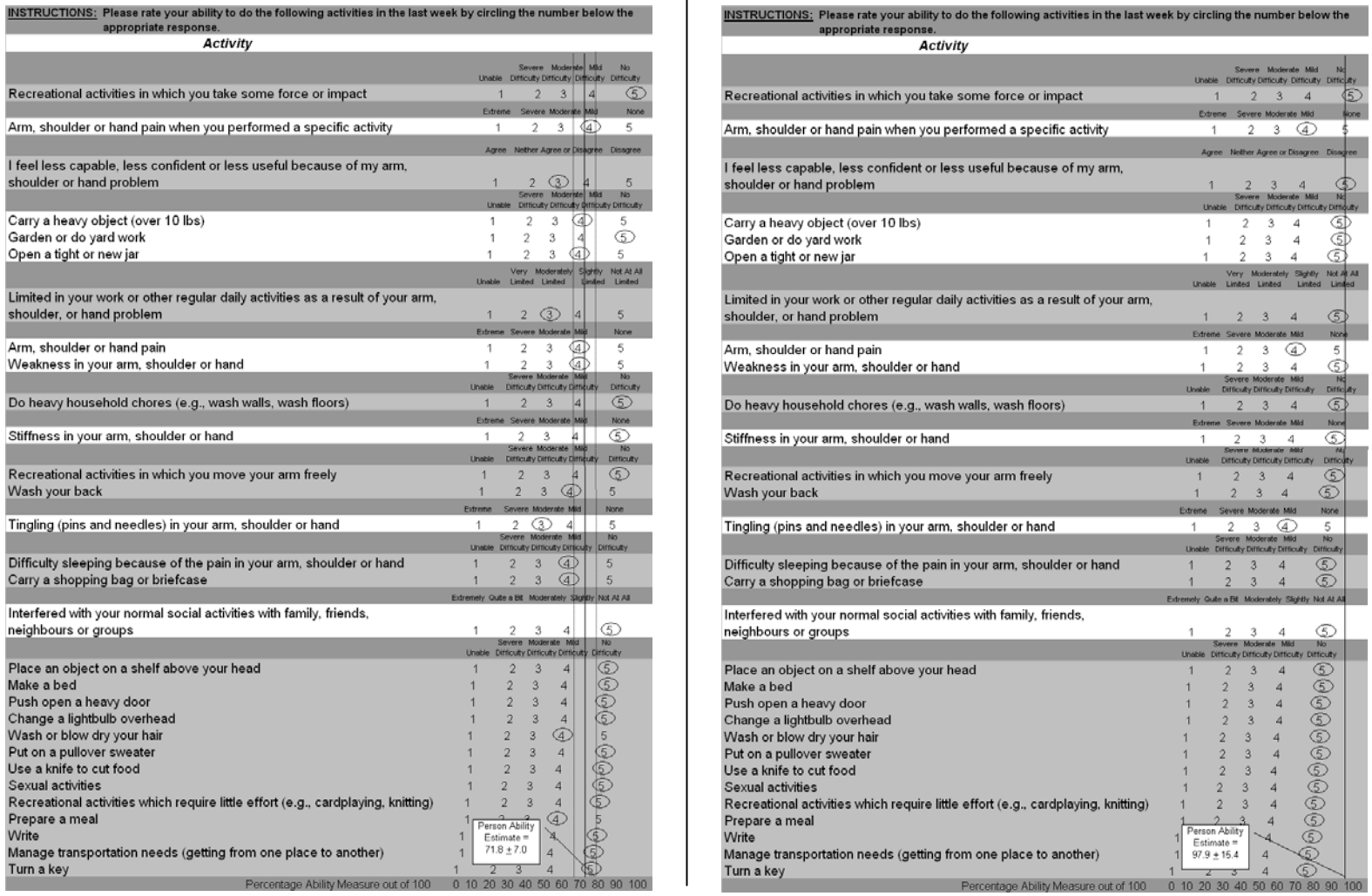
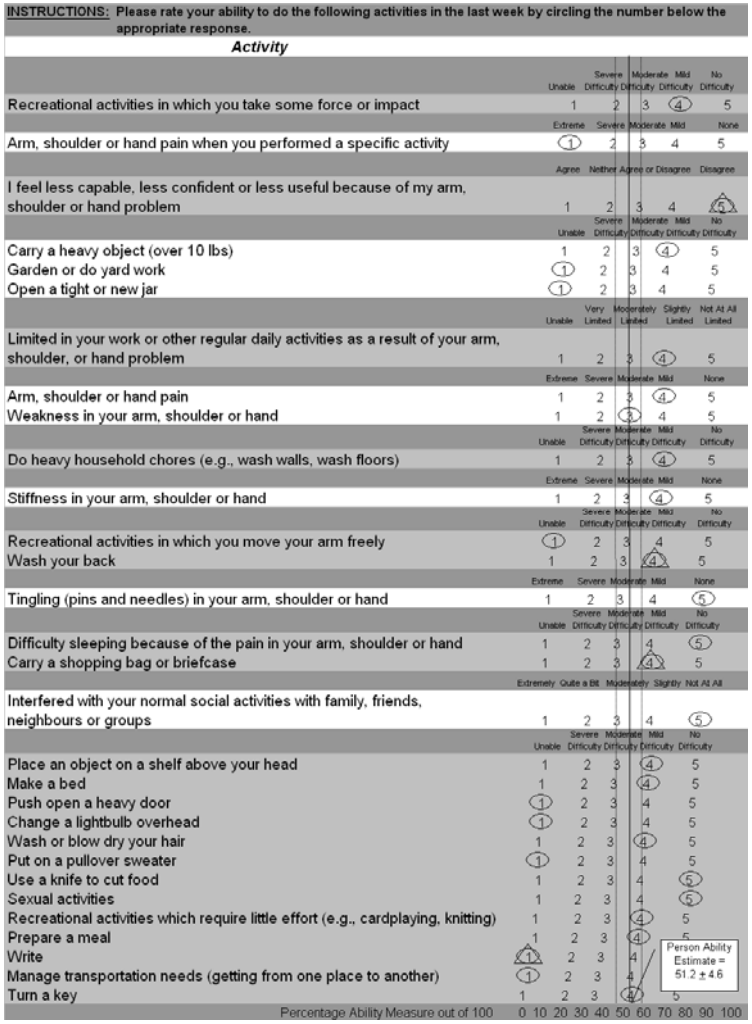
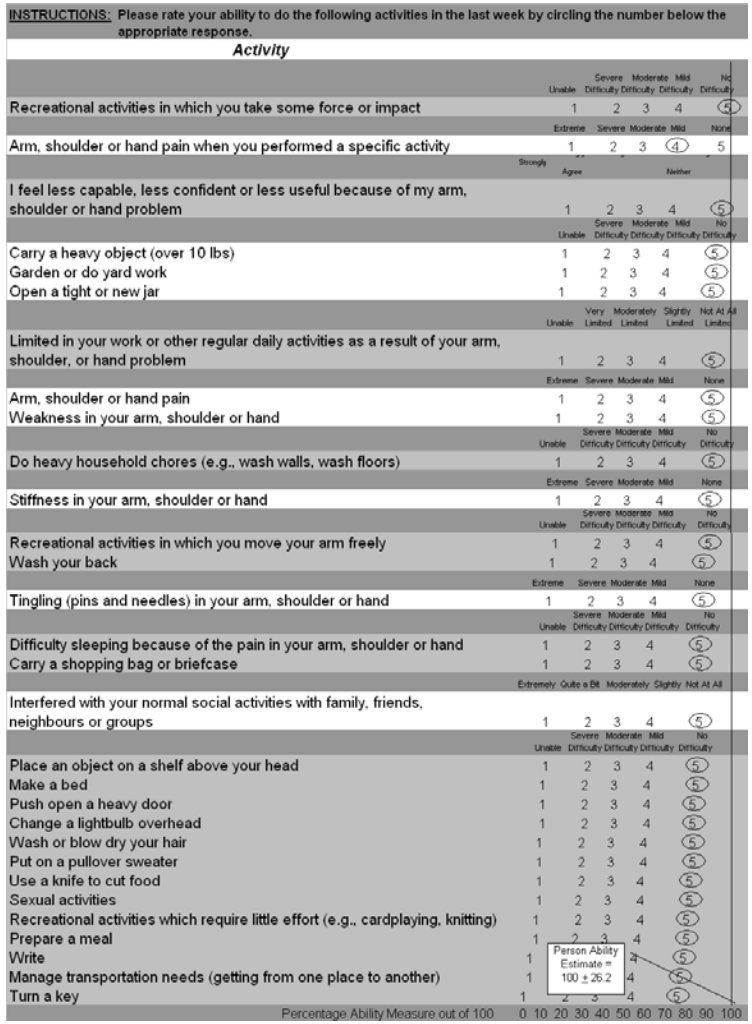


Figure 4-2. Data collection forms with individuals' responses (a. High ability individual at admission, b. Medium ability individual at admission, c. Low ability individual at admission)

Note: Triangles indicate unexpected responses. Data collection forms are shown at admission (left) and at discharge (right).

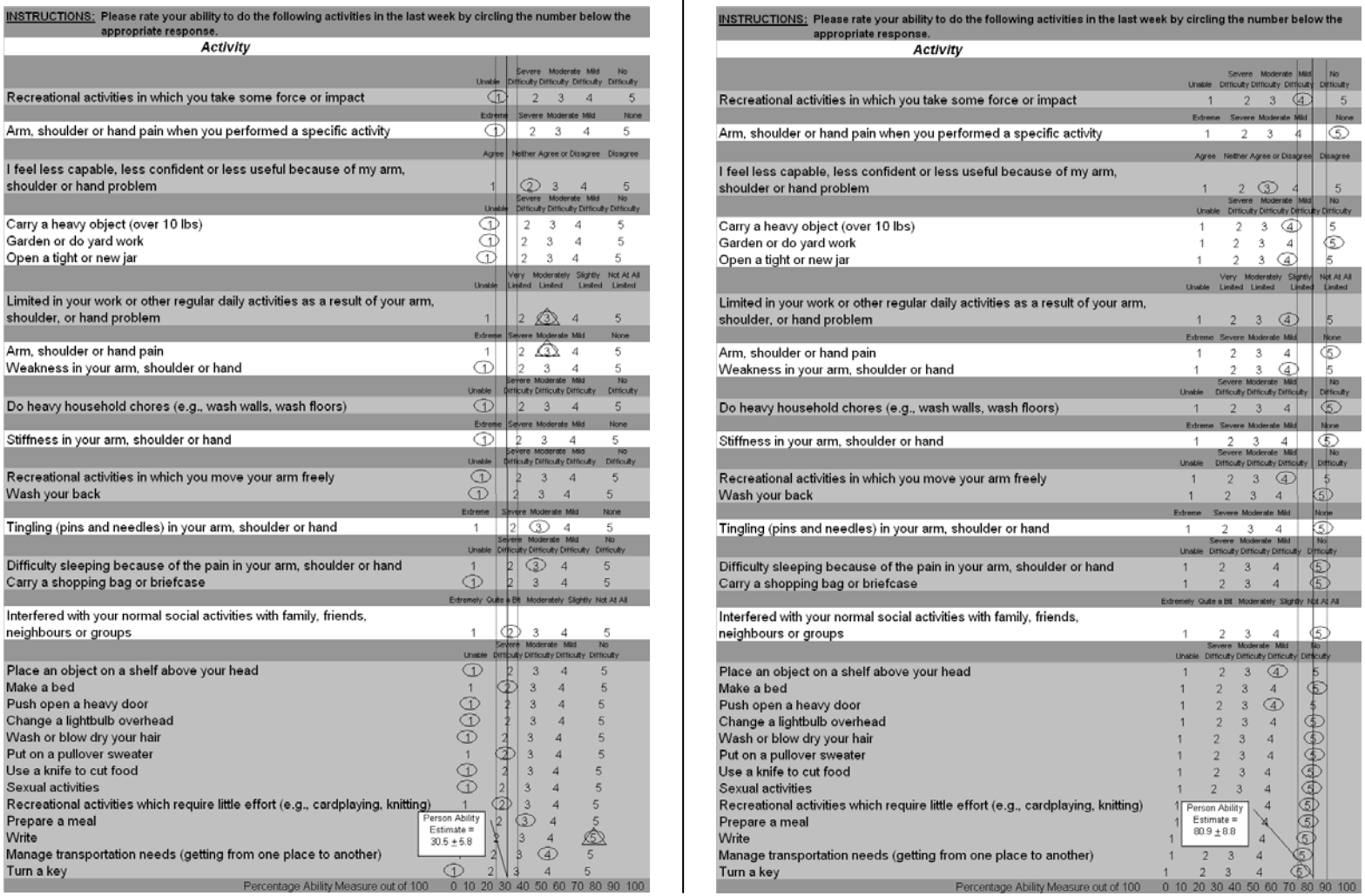




(b)

Figure 4-2. Continued

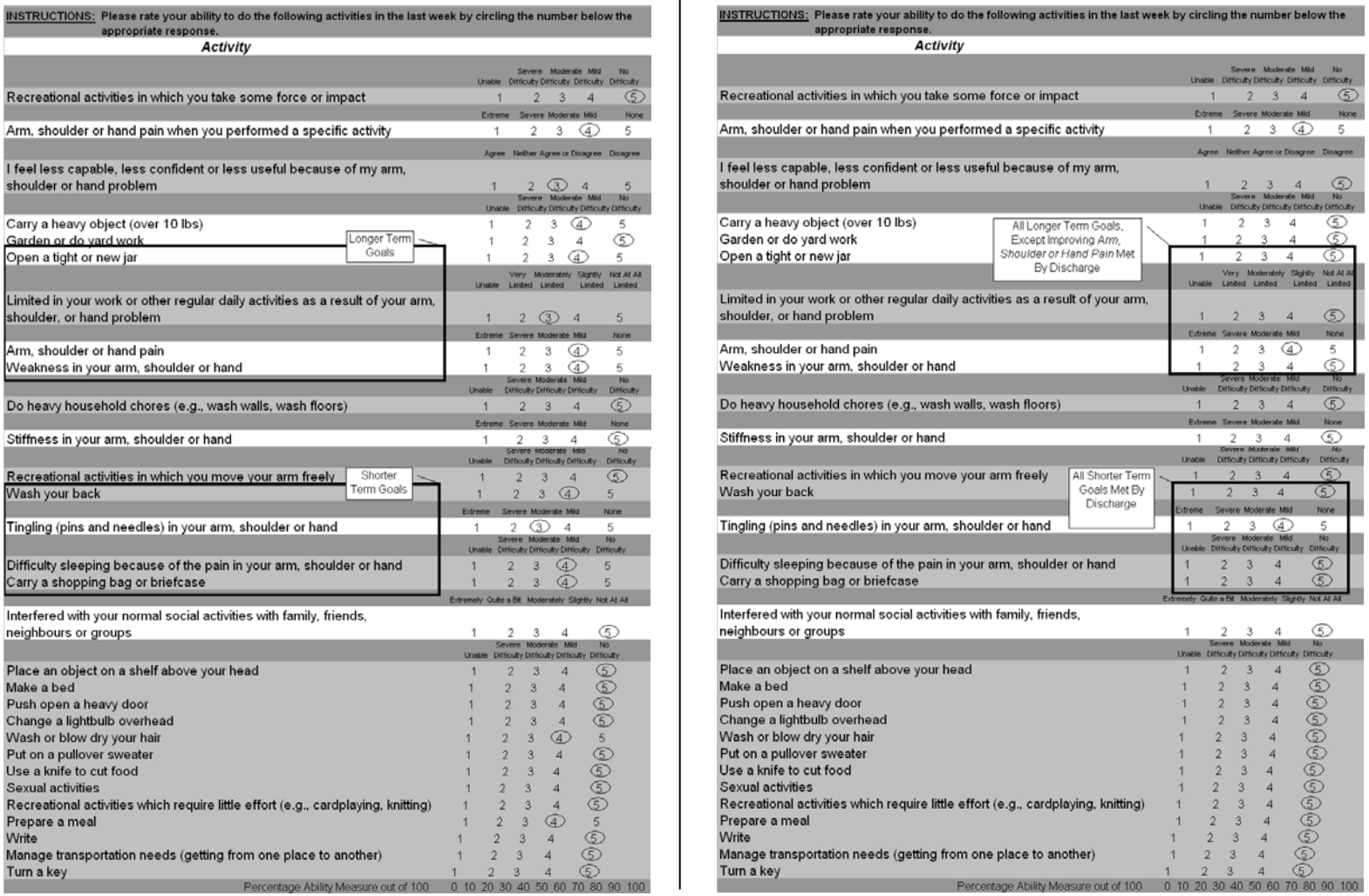
Note: Triangles indicate unexpected responses. Data collection forms are shown at admission (left) and at discharge (right).



(c)

Figure 4-2. Continued

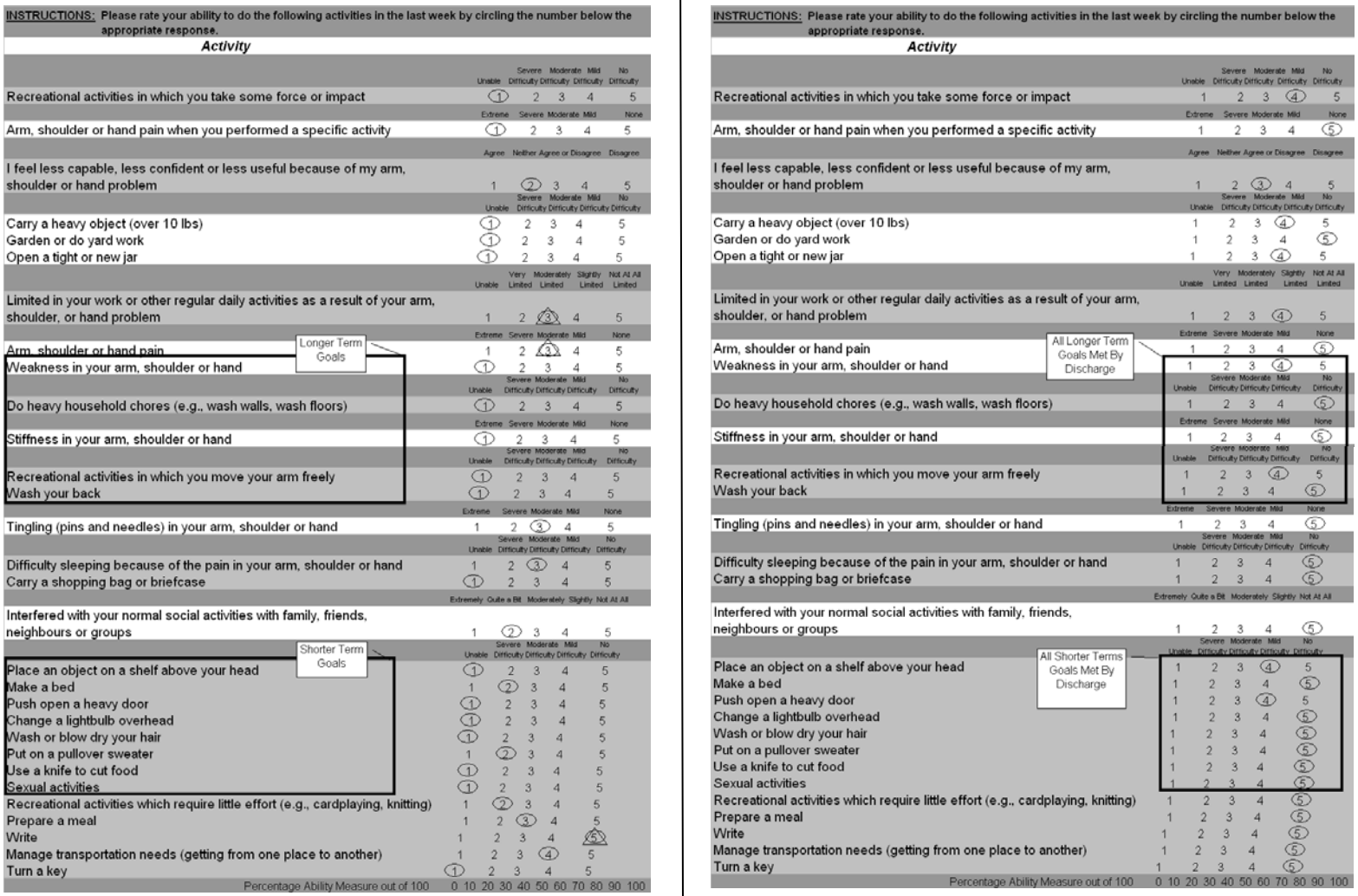
Note: Triangles indicate unexpected responses. Data collection forms are shown at admission (left) and at discharge (right).



(a)

Figure 4-3. Goal setting data collection form (a. High ability individual at admission, b. Low ability individual at admission)

Note: Triangles indicate unexpected responses. Data collection forms are shown at admission (left) and at discharge (right).



(b)

Figure 4-3. Continued

Note: Triangles indicate unexpected responses. Data collection forms are shown at admission (left) and at discharge (right).

CHAPTER 5
CONCLUSIONS

In the rehabilitation clinics where patients with upper-extremity impairments are treated, the primary focus is on enabling improvement. Not just progression at the impairment level, but advances in functional ability (i.e., changes that affect patients' daily lives). In order to ensure the achievement of this goal, assessments that are reliable, valid, and able to detect change are essential. The reliability and validity of many hand/upper-extremity assessments have been well established. However, this is not the case for ability to detect change. The ability to detect change is a property of instruments called responsiveness(2). Recently in various health-related areas investigators have acknowledged the importance of studying responsiveness. For instance, many studies have been conducted to assess the capacity of health-related quality of life measures to detect change(3-7). Hand therapists spend considerable time and energy attempting to bring about change, specifically to the hand/upper-extremity ailments of their patients. Thus, using assessments that are able to detect change is of foremost importance.

Outcomes of hand/upper-extremity therapy are generally evaluated by therapists using three types of measures: (1) impairment measures, (2) performance measures, and (3) self-report measures. For each of these measures, various authors have reported on reliability and validity, however, responsiveness has received less attention in the literature. Consequently, although there are a few studies examining responsiveness, the investigation is far from complete. Additionally, studies comparing the responsiveness of similar assessments have been inconclusive about which ones are superior in ability to detect patient change(57, 71, 72). Since much of a clinician's efforts are intended to effect change in patient function, it is clear that the capacity of instruments to measure change has been under studied.

The purpose of this project was four-fold: (1) To present the state-of-the-art responsiveness designs and methods, highlight problems with coefficients, and suggest a step toward increasing the existing knowledge of the measurement of change. (2) To compare the responsiveness of a well known upper-extremity assessment, the Disabilities of the Arm, Shoulder and Hand (DASH) outcome questionnaire to a lesser known and researched assessment, the Upper-Extremity Functional Index (UEFI). (3) To use item response theory methodologies to assess the item content of the DASH and (4) To demonstrate how Rasch methodologies can be used to create a clinically useful data collection form, and illustrate the benefits of such a data collection form in setting patient goals.

Chapter 1 provided background detail on responsiveness designs and methods. Two categories of responsiveness designs, as outlined by Stratford and colleagues (73) were presented. These categories are based on how many groups are involved: 1) single-group designs and 2) multiple-group designs. Single-group designs include the before-after (no baseline) design and the before-after design with a baseline, while multiple-group designs include three types: 1) those that compare patients who are randomly assigned to receive either a previously proven effective treatment or a placebo, 2) those that compare two or more groups whose health status, based on prior evidence, is expected to change by different amounts, and lastly, 3) those that compare two groups expected to change by different amounts based on responses to an outside criterion (e.g., a global rating of change score).

There are limitations associated with each of the designs and problems associated with using various coefficients. Perhaps, the foremost concerns are associated with single-group designs. A disadvantage associated with the before-after, no baseline design is that if there is not a change detected, it is unclear whether the measure was unable to detect the change or whether

the patients did not undergo the expected change. This design is considered the weakest of the designs. With the before-after designs with a baseline, a major disadvantage is that the period during which stability is measured (i.e., period from baseline to measurement taken just before intervention) is often shorter than the period during which change is assessed (i.e., measure taken just before intervention to measure taken after intervention). Thus, this design may underestimate the magnitude of random variability that occurs over longer periods in patients whose health status is truly stable(73).

The addition of a comparison group gives multiple group designs an advantage over single-group designs. However, each of the multiple-group designs also has limitations. One disadvantage when comparing patients who receive either the intervention or a placebo involves the difficulty finding an intervention that is known to be effective (i.e., an intervention that is highly likely to result in changes to the treatment group). In a similar way, the major disadvantage when comparing groups expected to change by different amounts is that often it is difficult to find groups who meet this criterion (i.e., often there are not two groups of patients available whose condition is expected to change by different amounts). When comparing two groups which have been defined by some external standard (e.g., global rating of change scores) the soundness of the study may be compromised for two reasons. First, patients complete the measure under study (e.g., the Disabilities of the Arm, Shoulder, and Hand outcome questionnaire) and the criterion rating (e.g., a global rating of change). Thus, the criterion measure is not independent of the measure under study (i.e., a patient who responds to the questions on the measure indicating they have changed is also likely to respond to the criterion measure in the same manner). Secondly, when a global rating of change measure is used as the external criterion, evidence suggests that patients have difficulty recalling their initial state and

consequently, may be inaccurate when assessing the amount of change that has taken place(75, 164, 165). Although an outside criterion provides the advantage of being a benchmark by which to assess whether change has occurred, if the outside criterion is not a genuine measure of change the results will not provide a good representation of how much change has actually occurred.

In addition to the limitations related to specific designs, a problem with measuring the property of responsiveness involves the discrepant results obtained when different coefficients are used. That is, when several similar instruments are compared to determine which assessment is the most responsive, the ranking of these instruments often differs depending on which design and which coefficient is used to calculate responsiveness(72, 76, 77). For instance, Wright and Young(72) found different rankings of five different assessments when using five different responsiveness coefficients (Guyatt's Responsiveness Index, standardized response mean, relative efficiency statistic, effect size, and correlation). Thus, an important question to answer before beginning a responsiveness study is: What is the most appropriate responsiveness design and coefficient to use when calculating responsiveness? Since no standardized methods exists for calculating responsiveness and multiple possible responsiveness coefficients can be computed (with conflicting results), there is a need for clarification as to what designs require the use of which coefficients. Erroneously, many researchers calculate an array of different coefficients(78). Since the formulas for each of the coefficients differ, the indices can lead to conflicting results.

A more theoretically sound approach is to determine which responsiveness coefficient is most appropriate for the study design and sample. If the design involves a single group, one of the coefficients appropriate for single-group studies should be utilized (effect size (ES),

standardized response mean (SRM), paired *t* value, or Guyatt's Responsiveness Index (GRI)). Alternatively if the design includes more than one group, you should choose from those coefficients appropriate for use with multiple-group designs (Guyatt's Responsiveness Index (GRI), *t* value for independent change scores, Analysis of Variance (ANOVA) of change scores, Norman's S_{repeat}, Norman's S_{ancova}, area under the Receiver Operating Characteristic (ROC) curve, or correlation). Also, with multiple-group designs one should consider the amount of change expected of your groups. That is, what differences in change is expected between placebo versus treatment groups, acute versus chronic groups, or between groups divided based on an outside criterion. Considering these two elements can eliminate the possibility of using several of the coefficients and thus, a more planned structured approach. However, it should be noted that even when the appropriate coefficients for the given number of groups and sample characteristics is used, there may be discrepancies in the calculations of different coefficients due to variations in the formulas. Because hand/upper-extremity therapists' foremost concern is inciting change in the functional ability of their patients, more studies leading to higher standards of responsiveness (equal to the standards of reliability and validity) should be mandated.

Chapter 2 presented the comparison of the responsiveness of the Disabilities of the Arm, Shoulder, and Hand (DASH) outcome questionnaire, a well known upper-extremity assessment, to the Upper-Extremity Functional Index (UEFI), a lesser known and researched assessment. The results of this comparison indicated that neither instrument has a clear advantage over the other when measuring responsiveness. The DASH questionnaire and the UEFI measure patient change in upper-extremity condition very similarly. A within subjects ANOVA revealed a significant difference between person measure change between the two assessments. However, areas under the ROC curves were almost identical for the two assessments (.67 and .65), as were

correlations between global ratings and the change scores (r = 0.33 and 0.35). Given that the responsiveness calculations for both assessments are very similar, there appears to be no real advantage of one instrument over the other in detecting change. Thus, if time is an issue, perhaps, the shorter UEFI is a better choice, even though it is a lesser known instrument. However, if comparison outcome data is needed or communication of outcomes with other therapists is desired the DASH may be preferred since it is the more widely used and studied.

One concern highlighted in the comparison of the responsiveness of the DASH to the UEFI centered on use of a patient reported global rating of change as a "gold standard". The correlations between the global rating and the person measure change scores were only r = 0.33 for the DASH and r = 0.35 for the UEFI. Additionally, the ROC curve was really more of a line, indicating low probabilities of specificity and sensitivity, i.e., 50% chance. Thus, there is a problem with not having a "gold standard" with which to compare DASH and UEFI change scores. In essence in this study, the assessment and "gold standard" are patient reported outcomes. This study tends to imply that patients are not accurate in reporting the amount of change that has occurred in their functional ability. Moreover, the wording of the global rating of change question used in this study (Please rate on a scale from –7 to +7 how much you think your condition has changed since your first therapy session. –7 indicates that your condition is much worse, while +7 indicates that your condition is much better. Please fill in the circle above your answer choice.), may have influenced how individuals responded.

Even when an appropriate external criterion has been established, there are other important decisions that need to be addressed. One involves considering sensitivity and specificity in determining a cutoff score for patients who have changed versus those who have not changed. This was demonstrated through the ROC curves generated in this study. The issue involves a

trade-off between specificity and sensitivity. A larger total change score on the DASH or UEFI (more stringent criteria) yields a smaller sensitivity and greater specificity, and vice versa; a smaller total change score (less stringent criteria) yields greater sensitivity and smaller specificity. Clinicians who use assessment change scores need to make a compromise of sorts when deciding on what type of error they are willing to make. By accepting a lower total score change as a cut off for dividing improved and not improved groups, clinicians increase their chance of assuming a patient has changed when they have not. Conversely, using a greater total score change increases the chance of assuming a patient has not changed and continuing to treat the condition after their functional ability has returned to the desired state. Portney and Watkins(112) state that "those who use a screening tool must decide what levels of sensitivity and specificity are acceptable. Sensitivity is more important when the risk associated with missing a diagnosis is high, as in the case of life-threatening disease or a debilitating deformity; specificity is more important when the cost or risks associated with further intervention are substantial." In the case of assessing change resulting from rehabilitation, sensitivity may be chosen over specificity since it could be argued that it is better to continue to treat a patient longer than actually needed, than to discharge them too early.

Chapter 3 investigated the item-level psychometrics and factor structure of the Disabilities of the Arm, Shoulder, and Hand (DASH) outcome questionnaire. Results of exploratory factor analyses (EFAs) and confirmatory factor analyses (CFAs) were inconclusive as to whether or not a one factor solution was plausible. One factor accounted for most of the variance (> 60%) for both admission and discharge data with the next two factors each only adding approximately five percent more variance. All items loaded high on the first factor and the first two factors were highly correlated. Eigenvalues for the first factor in both the admission and discharge EFAs

were much greater than for the second factor, however, the second and third factor eignevalues were greater than 1.0 (criterion necessary for acceptance of the presence of a factor based on the Kaiser rule). Examination of which items loaded highest on each of the factors in a three factor solution, separates the items into three possible constructs. One (those that load highest on the first factor) representing more gross motor, complex activities, a second (those that load highest on the second factor) being solely hand activities, and third (those that load highest on the third factor) involving more general functioning or symptom items. Thus, the decision to treat the DASH as unidimensional is debatable. Perhaps, dividing the DASH into three separate scales with three total scores would be more meaningful. Using multiple scales, instead of one multidimensional scale might be more informative for treatment (e.g., when specifically treating a symptom) and provide a clearer picture of patient progress from admission to discharge.

CFAs with both admission and discharge data were conducted to test a one factor solution. The results revealed that with admission data, the only goodness of fit statistic that supported the model, was the Tucker-Lewis Index (TLI). With the discharge data, the TLI and Standardized Root Mean Square Residual (SRMR) supported the one factor solution. A number of items misfit (ten at admission and eleven at discharge). However, the criterion used with such a large sample (close to one thousand) is very stringent (MnSq fit values between 0.6 and 1.1(122, 123)). Moreover, DASH item point measure correlations (i.e., correlations between each item and the entire instrument) were all above .30. Thus, the evidence is mixed regarding the dimensionality of the DASH.

Maps of the person-item match, revealed that the sample performed better at discharge than at admission. The sample mean fell above all the items at discharge, whereas, at admission several items (I feel less capable, less confident or less useful, Recreational activities in which

you take some force or impact, and Recreational activities in which you move your arm freely) were above the sample mean. At discharge 41 individuals had maximum estimated ability levels, in contrast to only two at admission.

Item difficulty hierarchies support the theory of motor control. Four tenants of motor control theory are: 1) complex items are more difficult, 2) environmental factors increase the challenge of items, 3) more difficult items require the use of multiple joints, and 4) isolated, abstract items are more difficult than functional tasks(128, 145, 146). Complex items were found to be more difficult. For example, doing heavy household chores (e.g., wash walls, wash floors) was harder than putting on a pullover sweater. Likewise, gardening or doing yard work was more difficult than carrying a shopping bag or briefcase. Environmental factors were shown to influence difficulty ranking. Recreational activities requiring little effort, such as card playing or knitting, were rated easier than recreational activities in which you move your arm freely, such as frisbee or badminton. Furthermore, recreational activities in which you move your arm freely were rated easier than recreational activities in which you take some force or impact through your arm, shoulder, or hand. Items requiring the use of multiple joints were more difficult items. For instance, recreational activities, such as frisbee or badminton, were rated more difficult than using a knife to cut food. Also, preparing a meal was more difficult than turning a key in a lock. Finally, tasks presented out of context or without clear purpose were rated as harder. To illustrate, carrying a heavy object was rated as harder than carrying a shopping bag.

There was evidence that item difficulties were sufficiently stable from admission to discharge. Several items displayed significant differential item functioning (DIF) from admission to discharge. Five items had large DIF values ($p<.0017$): I feel less capable, less

confident or less useful because of my arm, shoulder or hand problem; Difficulty sleeping because of the pain in your arm, shoulder or hand; Push open a heavy door; Tingling (pins and needles) in your arm, shoulder or hand; and Sexual activities. While four items had small DIF ($p<.05$) (Weakness in your arm, shoulder or hand; Limited in work or other regular daily activities as a result of your arm, shoulder, or hand problem; Put on a pullover sweater; and Make a bed). However, none of the items calibrations from admission to discharge differed by 0.5 logits and mean person ability measures were not affected by the inclusion of DIF items. Furthermore, the intraclass correlation coefficient (ICC) was high indicating reliability in measurement between the two time-points. Similar curves for test information for admission and discharge data provide further evidence for the stability of the test at two time points. This stability of the measure from admission to discharge is important to the measurement of patient change.

A further item characteristic of importance is item discrimination. This gives an indication of how responses to a particular item correspond to responses to the overall assessment(137, 138). In general, high values for item discrimination are considered good, indicating that responses to that particular item reflect a similar construct as the other items on the test(138). Zero or negative discrimination values indicate that the item does not fit well with the rest of the items(137). Winsteps(1) was used to estimate item discrimination for admission and discharge data. Since none of the item discrimination values were zero or negative; this indicates that all items contribute in a similar fashion to the overall test(137, 138).

Item discrimination estimates are used in the calculation of item difficulty and person ability with the two parameter item response theory model. However, item discrimination is not used in these calculations with the one parameter item response model. With the one parameter

item response model, it is assumed that item discrimination is negligible and thus, is set at one for all items. Since item discrimination values were not equal in this study, some might agree that analysis using a two parameter model is necessary. Yet, analysis using the one parameter model appears to be the most parsimonious solution since comparison of person measures calculated using both models revealed little difference in these measures.

Finally, Chapter 4 demonstrated the potential for creating clinically useful data collection forms through the use of Rasch methodologies. The general keyform output from Winsteps Rasch analysis program(1) was used as the basis for creating the data collection form. Creation of such a form proved to be relatively easy process and offers several advantages over data-collection forms used in clinics today. On the data collection form, items are listed in hierarchical difficulty order. This helps to reveal patterns of scoring (i.e., doing better on easier items than more difficulty items). A person measure scale ranging from zero to 100 is placed at the bottom. A line can be drawn through the point on the form where most of the responses fall(156). The corresponding person measure at the bottom can help to determine the overall ability of the patient. Such a line can be estimated even with missing responses to some items.

As depicted by forms filled out according to patients of differing admission ability levels (high, medium, and low), at a glance it is possible to estimate an individual's overall ability to perform functional tasks requiring the upper-extremity. Furthermore, these forms can be used in goal setting and treatment planning. By observing the location where this individual starts to have mild to moderate difficulty with a substantial number of items, the clinician can determine what activities should be set as short and long-term goals for the patient.

The positive implications of this type of data collection form are numerous. First, this type of assessment form could be useful to therapists, aiding in goal setting and treatment planning.

The location of the circled responses on the form indicating that the patient is starting to have mild to moderate difficulty with a substantial number of items, provides information on what activities should be set as goals for the patient. Second, such a form could be used to justify reimbursement by insurance companies and by accreditation agencies. The illustration of patterns of functional improvement shown by this type of form could make an impressive statement about the possibilities of making functional gains from treatment. Finally, with the advantages provided by these forms, clinicians may be more likely to adopt standardized assessment in practice. Adoption of standardized assessments in clinical practice could lead to more possibilities for comparison of treatment effectiveness between different types of interventions and across different clinical sites.

In conclusion, the goals of this project were to present responsiveness designs and methods, compare the responsiveness of the DASH and UEFI, using item response theory methodologies to assess the DASH, and to demonstrate how a clinically useful data collection form can be designed. Clearly the study of responsiveness in clinical assessments is lacking. This may be due in part to the need for clarity about correct methods, designs, and details involved in the study of change in patient ability. Furthermore, the item characteristics of many upper-extremity assessments have not been investigated. Information about item characteristics may lead to further understanding of responsiveness. Finally, the use of standardized assessments in the clinics is lacking due to their inability to inform the treatment process. Chapter 4 attempted to illustrate a method to design a more clinically useful assessment. The hope of this project was to move upper-extremity assessment a small step towards becoming more clinically meaningful.

APPENDIX A
ITEMS ON THE DISABILITIES OF THE ARM, SHOULDER, AND HAND (DASH) OUTCOME QUESTIONNAIRE

The Disabilities of the Arm, Shoulder, and Hand (DASH) Outcome Questionnaire

Please rate your ability to do the following activities in the last week by circling the number below the appropriate response.

	No Difficulty	Mild Difficulty	Moderate Difficulty	Severe Difficulty	Unable
1. Open a tight or new jar.	1	2	3	4	5
2. Write.	1	2	3	4	5
3. Turn a key.	1	2	3	4	5
4. Prepare a meal.	1	2	3	4	5
5. Push open a heavy door.	1	2	3	4	5
6. Place an object on a shelf above your head	1	2	3	4	5
7. Do heavy household chores (e.g., wash walls, wash floors).	1	2	3	4	5
8. Garden or do yard work.	1	2	3	4	5
9. Make a bed.	1	2	3	4	5
10. Carry a shopping bag or briefcase.	1	2	3	4	5
11. Carry a heavy object (over 10 lbs)	1	2	3	4	5
12. Change a lightbulb overhead.	1	2	3	4	5
13. Wash or blow dry your hair.	1	2	3	4	5
14. Wash your back.	1	2	3	4	5
15. Put on a pullover sweater.	1	2	3	4	5
16. Use a knife to cut food.	1	2	3	4	5
17. Recreational activities which require little effort (e.g., cardplaying, knitting, etc.).	1	2	3	4	5
18. Recreational activities in which you take some force or impact through your arm, shoulder or hand (e.g., golf hammering, tennis, etc.).	1	2	3	4	5
19. Recreational activities in which you move your arm freely (e.g., playing frisbee, badminton, etc.).	1	2	3	4	5
20. Manage transportation needs (getting from one place to another).	1	2	3	4	5
21. Sexual activities.	1	2	3	4	5
	Not At All	Slightly	Moderately	Quite a Bit	Extremely
22. During the past week to what extent has your arm, shoulder or hand problem intered with your normal social activities with family, friends, neighbours or groups?	1	2	3	4	5
	Not Limited At All	Slightly Limited	Moderately Limited	Very Limited	Unable
23. During the past week, were you limited in your work or other regular daily activities as a result of your arm, shoulder, or hand problem?	1	2	3	4	5
	None	Mild	Moderate	Severe	Extreme
24. Arm, shoulder or hand pain.	1	2	3	4	5
25. Arm, shoulder or hand pain when you performed a specific activity.	1	2	3	4	5
26. Tingling (pins and needles) in your arm, shoulder or hand.	1	2	3	4	5
27. Weakness in your arm, shoulder or hand.	1	2	3	4	5
28. Stiffness in your arm, shoulder or hand.	1	2	3	4	5
	No Difficulty	Mild Difficulty	Moderate Difficulty	Severe Difficulty	So Much Difficulty That I Can't Sleep
29. During the past week, how much difficulty have you had sleeping because of the pain in your arm, shoulder or hand?	1	2	3	4	5
	Strongly Disagree	Disagree	Neither Agree Nor Disagree	Agree	Strongly Agree
30. I feel less capable, less confident or less useful because of my arm, shoulder or hand problem.	1	2	3	4	5

WORK MODULE (OPTIONAL)					
	No Difficulty	Mild Difficulty	Moderate Difficulty	Severe Difficulty	So Much Difficulty That I Can't Sleep
1. Using your usual technique for your work?	1	2	3	4	5
2. Doing your usual work because of your arm, shoulder or hand pain?	1	2	3	4	5
3. Doing your work as well as you would like?	1	2	3	4	5
4. Spending your usual amount of time doing your work?	1	2	3	4	5

SPORTS/PERFORMING ARTS MODULE (OPTIONAL)					
	No Difficulty	Mild Difficulty	Moderate Difficulty	Severe Difficulty	So Much Difficulty That I Can't Sleep
1. Using your usual technique for playing your instrument or sport?	1	2	3	4	5
2. Playing your musical instrument or sport because of arm, shoulder or hand pain?	1	2	3	4	5
3. Playing your musical intrument or sport as well as you would like?	1	2	3	4	5
4. Spending your usual amount of time practicing or playing your instrument or sport?	1	2	3	4	5

APPENDIX B
ITEMS ON THE UPPER-EXTREMITY FUNCTIONAL INDEX (UEFI)

The Upper-Extremity Functional Index (UEFI)
We are interested in knowing whether you are having any difficulty at all with the activities listed below because of your upper limb problem for which you are currently seeking attention. Please provide an answer for each activity. Today, do you or would you have any difficulty with:

(Circle one number on each line)

Activities	Extreme Difficulty or Unable to Perform Activity	Quite a bit of Difficulty	Moderate Difficulty	A Little bit of Difficulty	No Difficulty
1. Any of your usual work, household or school activities	0	1	2	3	4
2. Your usual hobbies, recreational or sporting activities	0	1	2	3	4
3. Lifting a bag of groceries to waist level	0	1	2	3	4
4. Lifting a bag of groceries above your head	0	1	2	3	4
5. Grooming your hair	0	1	2	3	4
6. Preparing food (e.g., peeling, cutting)	0	1	2	3	4
7. Pushing up on your hands (e.g., from bathtub or chair)	0	1	2	3	4
8. Driving	0	1	2	3	4
9. Vacuuming, sweeping or raking	0	1	2	3	4
10. Dressing	0	1	2	3	4
11. Doing up buttons	0	1	2	3	4
12. Using tools or appliances	0	1	2	3	4
13. Opening doors	0	1	2	3	4
14. Cleaning	0	1	2	3	4
15. Tying or lacing shoes	0	1	2	3	4
16. Sleeping	0	1	2	3	4
17. Laundering clothes (e.g., washing, ironing, folding)	0	1	2	3	4
18. Opening a jar	0	1	2	3	4
19. Throwing a ball	0	1	2	3	4
20. Carrying a small suitcase with your affected limb	0	1	2	3	4

APPENDIX C
GLOBAL RATING OF CHANGE

Please rate on a scale from –7 to +7 how much you think your condition has changed since your first therapy session. –7 indicates that your condition is much worse, while +7 indicates that your condition is much better. Please fill in the circle above your answer choice.

○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7
WORSE														BETTER

LIST OF REFERENCES

1. Linacre JM. WINSTEPS Rasch Measurement. Version 3.57.2 Date: 05-01-2005 ed; 2005.

2. Beaton DE, Bombardier C, Katz JN, Wright JG. A taxonomy for responsiveness. J Clin Epidemiol. 2001 Dec;54(12):1204-17.

3. Tidermark J, Bergstrom G, Svensson O, Tornkvist H, Ponzer S. Responsiveness of the EuroQol (EQ 5-D) and the SF-36 in elderly patients with displaced femoral neck fractures. Qual Life Res. 2003 Dec;12(8):1069-79.

4. De La Loge C, Trudeau E, Marquis P, Revicki DA, Rentz AM, Stanghellini V, et al. Responsiveness and interpretation of a quality of life questionnaire specific to upper gastrointestinal disorders. Clin Gastroenterol Hepatol. 2004 Sep;2(9):778-86.

5. Reine G, Simeoni MC, Auquier P, Loundou A, Aghababian V, Lancon C. Assessing health-related quality of life in patients suffering from schizophrenia: a comparison of instruments. Eur Psychiatry. 2005 Aug 30.

6. Eurich DT, Johnson JA, Reid KJ, Spertus JA. Assessing responsiveness of generic and specific health related quality of life measures in heart failure. Health Qual Life Outcomes. 2006;4:89.

7. Badia X, Diez-Perez A, Lahoz R, Lizan L, Nogues X, Iborra J. The ECOS-16 questionnaire for the evaluation of health related quality of life in post-menopausal women with osteoporosis. Health Qual Life Outcomes. 2004 Aug 3;2:41.

8. MacDermid JC. Measurement of health outcomes following tendon and nerve repair. J Hand Ther. 2005 Apr-Jun;18(2):297-312.

9. Cutter N, Kevorkian C, editors. Handbook of Manual Muscle Testing. New York: McGraw-Hill; 1999.

10. Hislop H, Montgomery J. Daniels and Worthingham's Muscle Testing. Techniques of Manual Examination. Philadelphia: W.B. Saunders; 2002.

11. Kendall F, McCreary E, Provance P. Muscles Testing and Function. 4th ed. ed. Philadelphia: Lippincott Williams & Wilkins; 1993.

12. Reese N. Muscle and Sensory Testing. Philadelphia: W.B. Saunders; 1999.

13. Dvir Z. Grade 4 in manual muscle testing: the problem with submaximal strength assessment. Clin Rehabil. 1997 Feb;11(1):36-41.

14. Cuthbert SC, Goodheart GJ, Jr. On the reliability and validity of manual muscle testing: a literature review. Chiropr Osteopat. 2007 Mar 6;15(1):4.

15. Fess E. Grip Strength. In: Casanova J, editor. Clinical Assessment Recommendations. Chicago: American Society of Hand Therapists; 1992. p. 41-6.

16. Bellace JV, Healy D, Besser MP, Byron T, Hohman L. Validity of the Dexter Evaluation System's Jamar dynamometer attachment for assessment of hand grip strength in a normal population. J Hand Ther. 2000 Jan-Mar;13(1):46-51.

17. Hamilton A, Balnave R, Adams R. Grip strength testing reliability. J Hand Ther. 1994 Jul-Sep;7(3):163-70.

18. Lagerstrom C, Nordgren B. On the reliability and usefulness of methods for grip strength measurement. Scand J Rehabil Med. 1998 Jun;30(2):113-9.

19. Lagerstrom C, Nordgren B, Olerud C. Evaluation of grip strength measurements after Colles' fracture: a methodological study. Scand J Rehabil Med. 1999 Mar;31(1):49-54.

20. MacDermid JC, Lee A, Richards RS, Roth JH. Individual finger strength: are the ulnar digits "powerful"? J Hand Ther. 2004 Jul-Sep;17(3):364-7.

21. Nitschke JE, McMeeken JM, Burry HC, Matyas TA. When is a change a genuine change? A clinically meaningful interpretation of grip strength measurements in healthy and disabled women. J Hand Ther. 1999 Jan-Mar;12(1):25-30.

22. Pincus T, Brooks RH, Callahan LF. Reliability of grip strength, walking time and button test performed according to a standard protocol. J Rheumatol. 1991 Jul;18(7):997-1000.

23. Shechtman O, Davenport R, Malcolm M, Nabavi D. Reliability and validity of the BTE-Primus grip tool. J Hand Ther. 2003 Jan-Mar;16(1):36-42.

24. Spijkerman DC, Snijders CJ, Stijnen T, Lankhorst GJ. Standardization of grip strength measurements. Effects on repeatability and peak force. Scand J Rehabil Med. 1991;23(4):203-6.

25. Stephens JL, Pratt N, Michlovitz S. The reliability and validity of the Tekdyne hand dynamometer: Part II. J Hand Ther. 1996 Jan-Mar;9(1):18-26.

26. Stephens JL, Pratt N, Parks B. The reliability and validity of the Tekdyne hand dynamometer: Part I. J Hand Ther. 1996 Jan-Mar;9(1):10-7.

27. Shechtman O, Gestewitz L, Kimble C. Reliability and validity of the DynEx dynamometer. J Hand Ther. 2005 Jul-Sep;18(3):339-47.

28. Agre JC, Magness JL, Hull SZ, Wright KC, Baxter TL, Patterson R, et al. Strength testing with a portable dynamometer: reliability for upper and lower extremities. Arch Phys Med Rehabil. 1987 Jul;68(7):454-8.

29. Brown A, Cramer LD, Eckhaus D, Schmidt J, Ware L, MacKenzie E. Validity and reliability of the dexter hand evaluation and therapy system in hand-injured patients. J Hand Ther. 2000 Jan-Mar;13(1):37-45.

30. MacDermid JC, Evenhuis W, Louzon M. Inter-instrument reliability of pinch strength scores. J Hand Ther. 2001 Jan-Mar;14(1):36-42.

31. MacDermid JC, Kramer JF, Woodbury MG, McFarlane RM, Roth JH. Interrater reliability of pinch and grip strength measurements in patients with cumulative trauma disorders. J Hand Ther. 1994 Jan-Mar;7(1):10-4.

32. Schreuders TA, Roebroeck ME, Goumans J, van Nieuwenhuijzen JF, Stijnen TH, Stam HJ. Measurement error in grip and pinch force measurements in patients with hand injuries. Phys Ther. 2003 Sep;83(9):806-15.

33. Novak CB, Mackinnon SE, Williams JI, Kelly L. Establishment of reliability in the evaluation of hand sensibility. Plast Reconstr Surg. 1993 Aug;92(2):311-22.

34. Novak CB, Mackinnon SE, Kelly L. Correlation of two-point discrimination and hand function following median nerve injury. Ann Plast Surg. 1993 Dec;31(6):495-8.

35. King PM. Sensory function assessment. A pilot comparison study of touch pressure threshold with texture and tactile discrimination. J Hand Ther. 1997 Jan-Mar;10(1):24-8.

36. Novak CB, Kelly L, Mackinnon SE. Sensory recovery after median nerve grafting. J Hand Surg [Am]. 1992 Jan;17(1):59-68.

37. Sanford J, Moreland J, Swanson LR, Stratford PW, Gowland C. Reliability of the Fugl-Meyer assessment for testing motor performance in patients following stroke. Phys Ther. 1993 Jul;73(7):447-54.

38. Jebsen RH, Taylor N, Trieschmann RB, Trotter MJ, Howard LA. An objective and standardized test of hand function. Arch Phys Med Rehabil. 1969 Jun;50(6):311-9.

39. Sabari JS, Lim AL, Velozo CA, Lehman L, Kieran O, Lai JS. Assessing arm and hand function after stroke: a validity test of the hierarchical scoring system used in the motor assessment scale for stroke. Arch Phys Med Rehabil. 2005 Aug;86(8):1609-15.

40. Fugl-Meyer AR. Post-stroke hemiplegia assessment of physical properties. Scand J Rehabil Med Suppl. 1980;7:85-93.

41. Fugl-Meyer AR, Jaasko L, Leyman I, Olsson S, Steglind S. The post-stroke hemiplegic patient. 1. a method for evaluation of physical performance. Scand J Rehabil Med. 1975;7(1):13-31.

42. Platz T, Pinkowski C, van Wijck FM, Kim I-H, di Bella P, Johnson G. Reliability and validity of the arm function assessment with standardized guidelines for the Fugl-Meyer Test, Action Research Arm Test and Box and Block Test: A mutlicentre study. Clinical Rehabilitation. 2005;19:404-11.

43. van Wijck FM, Pandyan AD, Johnson GR, Barnes MP. Assessing motor deficits in neurological rehabilitation: patterns of instrument usage. Neurorehabil Neural Repair. 2001;15(1):23-30.

44. Poole JL, Cordova KJ, Brower LM. Reliability and validity of a self-report of hand function in persons with rheumatoid arthritis. J Hand Ther. 2006 Jan-Mar;19(1):12-6, quiz 7.

45. Poole JL, Whitney SL. Motor assessment scale for stroke patients: concurrent validity and interrater reliability. Arch Phys Med Rehabil. 1988 Mar;69(3 Pt 1):195-7.

46. Morris DM, Uswatte G, Crago JE, Cook EW, 3rd, Taub E. The reliability of the wolf motor function test for assessing upper extremity function after stroke. Arch Phys Med Rehabil. 2001 Jun;82(6):750-5.

47. Wolf SL, Catlin PA, Ellis M, Archer AL, Morgan B, Piacentino A. Assessing Wolf motor function test as outcome measure for research in patients after stroke. Stroke. 2001 Jul;32(7):1635-9.

48. Wolf SL, Lecraw D, Barton L, Jann B. Forced use of hemiplegic upper extremities to reverse the effect of learned nonuse among chronic stroke and head-injured patients. Exp Neurol. 1989;104:125-32.

49. Barreca SR, Stratford PW, Lambert CL, Masters LM, Streiner DL. Test-retest reliability, validity, and sensitivity of the Chedoke arm and hand activity inventory: a new measure of upper-limb function for survivors of stroke. Arch Phys Med Rehabil. 2005 Aug;86(8):1616-22.

50. Hsieh CL, Hsueh IP, Chiang F, Lin P. Inter-rater reliability and validity of the Action Research Arm Test in stroke patients. Age Ageing. 1998;27:107-14.

51. Stamm TA, Cieza A, Coenen M, Machold KP, Nell VP, Smolen JS, et al. Validating the International Classification of Functioning, Disability and Health Comprehensive Core Set for Rheumatoid Arthritis from the patient perspective: a qualitative study. Arthritis Rheum. 2005 Jun 15;53(3):431-9.

52. Hunter J, Mackin E, Callahan A, editors. Rehabilitation of the Hand: Surgery and Therapy. Fourth Edition ed. St. Louis: Mosby; 1995.

53. Stern EB. Stability of the Jebsen-Taylor Hand Function Test across three test sessions. Am J Occup Ther. 1992 Jul;46(7):647-9.

54. U.S. Department of Health and Human Services FaDA. Guidance for Industry, Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. Draft Guidance; February 2006.

55. Atroshi I, Gummesson C, Andersson B, Dahlgren E, Johansson A. The disabilities of the arm, shoulder and hand (DASH) outcome questionnaire: reliability and validity of the Swedish version evaluated in 176 patients. Acta Orthop Scand. 2000 Dec;71(6):613-8.

56. Gummesson C, Atroshi I, Ekdahl C. The disabilities of the arm, shoulder and hand (DASH) outcome questionnaire: longitudinal construct validity and measuring self-rated health change after surgery. BMC Musculoskelet Disord. 2003 Jun 16;4:11.

57. Beaton DE, Katz JN, Fossel AH, Wright JG, Tarasuk V, Bombardier C. Measuring the whole or the parts? Validity, reliability, and responsiveness of the Disabilities of the Arm, Shoulder and Hand outcome measure in different regions of the upper extremity. J Hand Ther. 2001 Apr-Jun;14(2):128-46.

58. Rosales RS, Delgado EB, Diez de la Lastra-Bosch I. Evaluation of the Spanish version of the DASH and carpal tunnel syndrome health-related quality-of-life instruments: cross-cultural adaptation process and reliability. J Hand Surg [Am]. 2002 Mar;27(2):334-43.

59. Dubert T, Voche P, Dumontier C, Dinh A. [The DASH questionnaire. French translation of a trans-cultural adaptation]. Chir Main. 2001 Aug;20(4):294-302.

60. Davis AM, Beaton DE, Hudak P, Amadio P, Bombardier C, Cole D, et al. Measuring disability of the upper extremity: a rationale supporting the use of a regional outcome measure. J Hand Ther. 1999 Oct-Dec;12(4):269-74.

61. Navsarikar A, Gladman DD, Husted JA, Cook RJ. Validity assessment of the disabilities of arm, shoulder, and hand questionnaire (DASH) for patients with psoriatic arthritis. J Rheumatol. 1999 Oct;26(10):2191-4.

62. Arnould C, Penta M, Renders A, Thonnard JL. ABILHAND-Kids: a measure of manual ability in children with cerebral palsy. Neurology. 2004 Sep 28;63(6):1045-52.

63. Gustafsson S, Sunnerhagen K, Dahlin-Ivanoff S. Occupational Therapists' and Patients' Perceptions of ABILHAND, a New Assessment Tool for Measuring Manual Ability. Scandinavian Journal of Occupational Therapy. 2004;11:107-17.

64. Penta M, Tesio L, Arnould C, Zancan A, Thonnard J-L. The ABILHAND Questionnaire as a Measure of Manual Ability in Chronic Stroke Patients. Stroke. 2001;32:1627-34.

65. Penta M, Thonnard JL, Tesio L. ABILHAND: a Rasch-built measure of manual ability. Arch Phys Med Rehabil. 1998 Sep;79(9):1038-42.

66. Kotsis SV, Chung KC. Responsiveness of the Michigan Hand Outcomes Questionnaire and the Disabilities of the Arm, Shoulder and Hand questionnaire in carpal tunnel surgery. J Hand Surg [Am]. 2005 Jan;30(1):81-6.

67. Greenslade JR, Mehta RL, Belward P, Warwick DJ. Dash and Boston questionnaire assessment of carpal tunnel syndrome outcome: what is the responsiveness of an outcome questionnaire? J Hand Surg [Br]. 2004 Apr;29(2):159-64.

68. MacDermid JC, Tottenham V. Responsiveness of the disability of the arm, shoulder, and hand (DASH) and patient-rated wrist/hand evaluation (PRWHE) in evaluating change after hand therapy. J Hand Ther. 2004 Jan-Mar;17(1):18-23.

69. MacDermid JC, Richards RS, Donner A, Bellamy N, Roth JH. Responsiveness of the short form-36, disability of the arm, shoulder, and hand questionnaire, patient-rated wrist evaluation, and physical impairment measurements in evaluating recovery after a distal radius fracture. J Hand Surg [Am]. 2000 Mar;25(2):330-40.

70. Gay RE, Amadio PC, Johnson JC. Comparative responsiveness of the disabilities of the arm, shoulder, and hand, the carpal tunnel questionnaire, and the SF-36 to clinical change after carpal tunnel release. J Hand Surg [Am]. 2003 Mar;28(2):250-4.

71. Stratford, Binkley JM, Stratford. Development and initial validation of the upper extremity functional index. Physiotherapy Canada. 2001;Fall 2001.

72. Wright J, Young N. A comparison of different indices of responsiveness. J Clin Epidemiol. 1997 Mar;50(3):239-46.

73. Stratford, Binkley F, Riddle D. Health status measures: strategies and analytic methods for assessing change scores. Phys Ther. 1996 Oct;76(10):1109-23.

74. Portney LG, Watkins MP. Foundations of clinical research : applications to practice. 2nd ed. Upper Saddle River, NJ: Prentice Hall; 2000.

75. Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. J Clin Epidemiol. 1997 Aug;50(8):869-79.

76. Wallace D, Duncan PW, Lai SM. Comparison of the responsiveness of the Barthel Index and the motor component of the Functional Independence Measure in stroke: the impact of using different methods for measuring responsiveness. J Clin Epidemiol. 2002 Sep;55(9):922-8.

77. Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. J Chronic Dis. 1986;39(11):897-906.

78. Stratford PW, Riddle DL. Assessing sensitivity to change: choosing the appropriate change coefficient. Health Qual Life Outcomes. 2005 Apr 5;3(1):23.

79. Liang MH. Evaluating measurement responsiveness. J Rheumatol. 1995 Jun;22(6):1191-2.

80. Liang MH, Lew RA, Stucki G, Fortin PR, Daltroy L. Measuring clinically important changes with patient-oriented questionnaires. Med Care. 2002 Apr;40(4 Suppl):II45-51.

81. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chronic Dis. 1987;40(2):171-8.

82. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials. 1989 Dec;10(4):407-15.

83. Jerrell JM. Behavior and symptom identification scale 32: sensitivity to change over time. J Behav Health Serv Res. 2005 Jul-Sep;32(3):341-6.

84. Redelmeier DA, Guyatt GH, Goldstein RS. On the debate over methods for estimating the clinically important difference. J Clin Epidemiol. 1996 Nov;49(11):1223-4.

85. Beaton DE. Simple as possible? Or too simple? Possible limits to the universality of the one half standard deviation. Med Care. 2003 May;41(5):593-6.

86. Norman GR, Sloan JA, Wyrwich KW. Is it simple or simplistic? Med Care. 2003 May;41(5):599-600.

87. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Med Care. 2003 May;41(5):582-92.

88. Wyrwich KW, Tierney WM, Wolinsky FD. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. J Clin Epidemiol. 1999 Sep;52(9):861-73.

89. Wright J. The minimal important difference: who's to say what is important? Journal of Clinical Epidemiology. 1996;49:1221-2.

90. Redelmeier DA, Guyatt GH, Goldstein RS. Assessing the minimal important difference in symptoms: a comparison of two techniques. J Clin Epidemiol. 1996 Nov;49(11):1215-9.

91. Jacobson N, Follette W, Revenstorf D. Psychotherapy outcome research: methods for reporting variability and evaluating clinical significance. Behavior Therapy. 1984;15:336-52.

92. Wieduwilt K, Jerrell JM. Measuring sensitivity to change of the Role Functioning Scale using a RC_{ID} index. International Journal of Methods in Psychiatric Research. 1998;7(4):163-70.

93. Black N. Evidence based policy: proceed with care. Bmj. 2001 Aug 4;323(7307):275-9.

94. Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol. 2003 May;56(5):395-407.

95. De Smet L. Responsiveness of the DASH score in surgically treated basal joint arthritis of the thumb: preliminary results. Clin Rheumatol. 2004 Jun;23(3):223-4.

96. Davidson J. A comparison of upper limb amputees and patients with upper limb injuries using the Disability of the Arm, Shoulder and Hand (DASH). Disabil Rehabil. 2004 Jul 22-Aug 5;26(14-15):917-23.

97. Beaton DE, Wright JG, Katz JN. Development of the QuickDASH: comparison of three item-reduction approaches. J Bone Joint Surg Am. 2005 May;87(5):1038-46.

98. Jester A, Harth A, Germann G. Measuring levels of upper-extremity disability in employed adults using the DASH Questionnaire. J Hand Surg [Am]. 2005 Sep;30(5):1074 e1-e10.

99. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, N.J.: L. Erlbaum Associates; 1988.

100. Dobrzykowski E, Nance T. The Focus On Therapeutic Outcomes (FOTO) outpatient orthopedic rehabilitation database: results of 1994-1996. Journal of Rehabilitation Outcomes Measurement. 1997;1:56-60.

101. Swinkels I, van den Ende C, de Bakker D, van der Wees J, Hart D, Deutscher D, et al. Clinical databases in physical therapy. Physiotherapy Theory and Practice. 2007;23(3):153-67.

102. Hudak PL, Amadio PC, Bombardier C. Development of an upper extremity outcome measure: the DASH (disabilities of the arm, shoulder and hand) [corrected]. The Upper Extremity Collaborative Group (UECG). Am J Ind Med. 1996 Jun;29(6):602-8.

103. Jester A, Harth A, Wind G, Germann G, Sauerbier M. Disabilities of the arm, shoulder and hand (DASH) questionnaire: Determining functional activity profiles in patients with upper extremity disorders. J Hand Surg [Br]. 2005 Feb;30(1):23-8.

104. Solway S, Beaton DE, McConnell S, Bombardier C. The DASH Outcome Measure User's manual. Toronto, Ontario: Institute for Work & Health; 2002.

105. SPSS. 13.0 ed: SPSS Inc., 1989-2004; 2004.

106. Juniper EF, Guyatt GH, Willan A, Griffith LE. Determining a minimal important change in a disease-specific Quality of Life Questionnaire. J Clin Epidemiol. 1994 Jan;47(1):81-7.

107. Shechtman O. The Use of Rapid Exchange Grip Test in Detecting Sincerity of Effort, Part II: Validity of the Test. Journal of Hand Therapy. 2000;13:203-10.

108. Sackett D, Strauss S, Richardson W, Rosenberg W, Haynes R. Evidence Based Medicine. 2nd ed. Edinburgh, UK: Churchill Livingstone; 2000.

109. Kielhofner G. Research in Occupational Therapy: Methods of Inquiry for Enhancing Practice. Philadelphia: F.A. Davis Company; 2006.

110. Shechtman O. The Coefficient of Variation as a Measure of Sincerity of Effort of Grip Strength, Part II: Sensitivity and Specificity. Journal of Hand Therapy. 2001(14):188-94.

111. Beaton DE, Hogg-Johnson S, Bombardier C. Evaluating changes in health status: reliability and responsiveness of five generic health status measures in workers with musculoskeletal disorders. J Clin Epidemiol. 1997 Jan;50(1):79-93.

112. Portney LG, Watkins MP. Foundations of Clinical Research: Applications to Practice. Norwalk, Conn: Apppleton & Lange; 1993. p. 81.

113. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Med Care. 2000 Sep;38(9 Suppl):II28-42.

114. McHorney C. Ten recommendations for Advancing Patient-Centered Outcomes Measurement for Older Persons. Ann Intern Med. 2003;139:403-9.

115. Ware J, Jr. Item Response Theory and Computerized Adaptive Testing: Implications for Outcomes Measurement in Rehabilitation. Rehabilitative Psychology. 2005;50(1):71-8.

116. Thurstone L. Measurement of Social Attitudes. Journal of Abnormal and Social Psychology. 1931;26:249-69.

117. Muthen L, Muthen B. Mplus. Fourth Edition ed. Los Angeles, CA: Muthen & Muthen; 1998-2007.

118. Kline R. Principles and practices of structural equation modeling. 2nd edition ed. New York: Guilford Publications, Inc.; 2005.

119. Ajiferuke I. Sample size and informetric model goodness-of-fit outcomes: a search engine log case study. Journal of Information Science. 2006;32(3):212-22.

120. Hu L, Bentler P. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling. 1999;6(1):1-55.

121. Wright BD, Stone MH. Best Test Design. Chicago: MESA Press; 1979.

122. Linacre J. Size vs. Significance: Standardized Chi-Square Fit Statistic. Rasch Measurement Transactions. 2003;17(1):918.

123. Green K, Franton C. Survey Development and Validation with the Rasch Model. International Conference on Questionnaire Development, Evaluation, and Testing; 2002 November 14-17, 2002; Charleston, SC; 2002.

124. Smith EV. Attention Deficit Hyperactivity Disorder: Scaling and Standard Setting Using Rasch Measurement. Journal of Applied Measurement. 2000;1(1):3-24.

125. Linacre JM. Structure in Rasch Residuals: Why Principal Components Analysis? Measurement Transactions. 1998;12:636.

126. Wright B. Point-biserial Correlations and Item Fits. Rasch Measurement Transactions. 1992;5(4):174.

127. Allen M, Yen W. Introduction to Measurement Theory. Prospect Heights, IL: Waveland Press, Inc.; 1979.

128. Shumway-Cook A, Woollacott M. Motor Control: Theory and Practical Applications. 2nd edition ed. Philadelphia: Lippincott Williams & Wilkins; 2000.

129. Tesio L. Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. J Rehabil Med. 2003 May;35(3):105-15.

130. Lai J, Teresi J, Gershon R. Procedures for the analysis of differential item functioning (DIF) for small sample sizes. Eval Health Prof. 2005;28(3):283-94.

131. Wright BD, Masters GN. Rating Scale Analysis. Chicago, IL: MESA Press; 1982.

132. Kothari DH, Haley SM, Gill-Body KM, Dumas HM. Measuring functional change in children with acquired brain injury (ABI): comparison of generic and ABI-specific scales using the Pediatric Evaluation of Disability Inventory (PEDI). Phys Ther. 2003 Sep;83(9):776-85.

133. Crane P, Hart D, Gibbons L, Cook K. A 37-item shoulder functional status item pool had negligible differential item functioning. Journal of Clinical Epidemiology. 2006;59(5):478-84.

134. Roznowski M, Reith J. Examining the measurement quality of tests containing Differentially Functioning Items: Do biased items result in poor measurement? Educational and Psychological Measurement. 1999;59(2):248-69.

135. Wright B. Time 1 to Time 2. Rasch Measurement Transactions. 1996;10(1):478-9.

136. Chang WC, Chan C. Rasch analysis for outcomes measures: some methodological considerations. Arch Phys Med Rehabil. 1995 Oct;76(10):934-9.

137. Kelley T, Ebel R, Linacre J. Item Discrimination Indices. Rasch Measurement Transactions. 2002;16(3):883-4.

138. Lopez A. The Item Discrimination Index: Does it Work? Rasch Measurement Transactions. 1998;12(1):626.

139. MULTILOG for Windows. 7.0.2327.3 ed: Scientific Software International, Inc.; January 2003.

140. http://campar.in.tum.de/twiki/pub/Chair/TeachingWs05ComputerVision/3DCV_reminder_mle_000.pdf. 9.6 Maximum Likelihood Estimation. [cited 2007 November 8, 2007]; Available from:

141. Hambleton R, Swaminathan H, Rogers H. Fundamentals of Item Response Theory. Newbury Park, CA: Sage Press; 1991.

142. Andrich D. Controversy and the Rasch model: a characteristic of incompatible paradigms? Med Care. 2004 Jan;42(1 Suppl):I7-16.

143. Reise SP, Haviland MG. Item response theory and the measurement of clinical change. J Pers Assess. 2005 Jun;84(3):228-38.

144. Linacre J. Re: RMT Request. In: Lehman L, editor.; 2007. p. Email.

145. Alexander NB, Galecki AT, Grenier ML, Nyquist LV, Hofmeyer MR, Grunawalt JC, et al. Task-specific resistance training to improve the ability of activities of daily living-impaired older adults to rise from a bed and from a chair. J Am Geriatr Soc. 2001 Nov;49(11):1418-27.

146. Collin C, Wade D. Assessing motor impairment after stroke: a pilot reliability study. J Neurol Neurosurg Psychiatry. 1990 Jul;53(7):576-9.

147. Liang HW, Wang HK, Yao G, Horng YS, Hou SM. Psychometric evaluation of the Taiwan version of the Disability of the Arm, Shoulder, and Hand (DASH) questionnaire. J Formos Med Assoc. 2004 Oct;103(10):773-9.

148. Carr JH, Shepherd RB, Nordholm L, Lynne D. Investigation of a new motor assessment scale for stroke patients. Phys Ther. 1985 Feb;65(2):175-80.

149. Bot SD, Terwee CB, van der Windt DA, Bouter LM, Dekker J, de Vet HC. Clinimetric evaluation of shoulder disability questionnaires: a systematic review of the literature. Ann Rheum Dis. 2004 Apr;63(4):335-41.

150. Tiffin J, Asher E. The Purdue Pegboard: norms and studies of reliability and validity. Journal of Applied Psychology. 1948;32:234.

151. Malouin F, Pichard L, Bonneau C, Durand A, Corriveau D. Evaluating motor recovery early after stroke: comparison of the Fugl-Meyer Assessment and the Motor Assessment Scale. Arch Phys Med Rehabil. 1994 Nov;75(11):1206-12.

152. Schulzer M, Mak E, Calne SM. The psychometric properties of the Parkinson's Impact Scale (PIMS) as a measure of quality of life in Parkinson's disease. Parkinsonism Relat Disord. 2003 Jun;9(5):291-4.

153. van der Lee JH, Beckerman H, Lankhorst GJ, Bouter LM. The responsiveness of the Action Research Arm test and the Fugl-Meyer Assessment scale in chronic stroke patients. J Rehabil Med. 2001 Mar;33(3):110-3.

154. Woodbury M, Velozo CA. Potential for outcomes to influence practice and support clinical competency [Electronic Version]. OT Practice. 2005;10(10):7-8.

155. Kielhofner G, Dobria L, Forsyth K, Basu S. The Construction of Keyforms for Obtaining Instantaneous Measures From the Occupational Performance History Interview Rating Scales. Occupational Therapy Journal of Research: Occupation, Participation, and Health. 2005;25(1):1-10.

156. Linacre J. Instantaneous Measurement and Diagnosis. Physical Medicine and Rehabilitation: State of the Art Reviews. 1997;11(2):315-24.

157. Microsoft Access. 2003 ed: Microsoft Corporation; 2003.

158. Wright BD, Linacre JM. Reasonable Item Mean-Square Fit Values. Rasch Measurement Transactions. 1994;8:370.

159. Avery LM, Russell DJ, Raina PS, Walter SD, Rosenbaum PL. Rasch analysis of the Gross Motor Function Measure: validating the assumptions of the Rasch model to create an interval-level measure. Arch Phys Med Rehabil. 2003 May;84(5):697-705.

160. Woodbury M, Velozo C, Richards L, Duncan P, Studenski S, Lai S. Dimensionality and Construct Validity of the Fugl-Meyer

Assessment of the Upper Extremity. Archives of Physical Medicine and Rehabilitation. 2007;88:715-23.

161. Schuind FA, Mouraux D, Robert C, Brassinne E, Remy P, Salvia P, et al. Functional and outcome evaluation of the hand and wrist. Hand Clin. 2003 Aug;19(3):361-9.

162. Stone L. Social desirability and order of item presentation in the MMPI. Psychological Reports. 1965;17(2):518.

163. Guertin W. The effect of instructions and item order on the arithmetic subtest of the Wechsler-Bellevue. Journal of Genetic Psychology. 1954;85:79-83.

164. Kairiss EW, Miranker WL. Cortical memory dynamics. Biol Cybern. 1998 Apr;78(4):277-92.

165. Hayes JA, Black NA, Jenkinson C, Young JD, Rowan KM, Daly K, et al. Outcome measures for adult critical care: a systematic review. Health Technol Assess. 2000;4(24):1-111.

BIOGRAPHICAL SKETCH

Leigh Lehman graduated from Boiling Springs High School (Boiling Springs, SC) in 1993. She completed her Bachelor of Science degree in psychology and biology at the University of South Carolina Upstate (Spartanburg, SC) in 1998. In 2005, she received a Master of Health Science degree in occupational therapy from the University of Florida (UF; Gainesville, FL). She graduated with a Ph.D. in rehabilitation science from UF in 2008.