MEASURING POST-STROKE ARM MOTOR ABILITY: MEASUREMENT PROPERTIES OF THE FUGL-MEYER UPPER EXTREMITY ASSESSMENT EXAMINED WITH AN ITEM RESPONSE THEORY MEASUREMENT FRAMEWORK

By

MICHELLE LYNNE WOODBURY

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

2006

To Bernard Earl Watrous and Leah Michelle Smith

ACKNOWLEDGMENTS

I must first thank my dissertation co-chairs for their help and guidance throughout the process. Lorie Richards, a fellow Elizabethtown College alumnus, has been a steady source of encouragement. When I have needed her, she is there with enthusiasm, empathy and ideas. Craig Velozo is a model mentor. His integrity, clarity of thought, humility and witty demeanor make it possible for him to expect, and get, excellence from his research team. I also must thank the other members of my committee, Dena Howland and Kathye Light. Dena, an occupational therapist turned neuroscientist, somehow is able to balance her family life with a highly demanding research career. I am grateful that in the midst of chaos she always had time for me, gave me her full attention, and was able to clearly provide guidance and encouragement. Kathye is a highly intelligent wonderful teacher who has shown me how to mix academic content with a healthy dose of "real world" application and clinical intuition.

I am deeply grateful to Pam Duncan. This dissertation idea began with her; she graciously allowed access to her databases, and she has been dangerously enthusiastic about the project from the beginning!

My family has been a constant source of support. Many thanks go to my sister and brother-in-law, Melissa and Christopher Smith; my very cute niece, Leah Michelle Smith; my mother and step-father, Bonnie and Red Beuglass; my dad and step-mother, Gary and Robin Woodbury; my crazy cousin Rosemary Watrous; my wonderful "neena" Ruth Watrous; and my clever and funny grandfather Bernard Watrous whom I miss.

I must also thank my "RSD family," the fellow students that are the source of 24/7 support and wisdom. Many thanks go to the Velozo research team (Leigh, Inga, Jia-Hwa, Pey-Shan) for their assistance with Rasch and 2PL analyses. Thanks go also to Arlene, Dennis (Steve), and Mike for "forcing" me to choose UF. I am grateful for the OT RSD students who make me laugh: Megan, Jessica, Roxanna, Rick, Bhagwant, Eric, Kezia, Sandy W., Sandy P., Patricia, Christine, and Mindy.

It is easy to loose perspective in the academic rehabilitation research context. Three groups have helped me remember the truly important big picture. Prior to coming to PhD-school I was privileged to "sit at the feet" of amazing teachers: Drs. Mullen, Hulbert, Priest, Larkin, and Davidson (Columbia International University). They modeled academic and personal excellence, becoming "real life" illustrations of wisdom and knowledge. Waleed-Al Oboudi is an OTR with stunning talent whom I have been privileged to call mentor and friend. His passionate desire to advance clinically relevant stroke rehabilitation by training therapists to pursue excellence has been the goal to which I have aimed both academically and clinically. Finally, I have very clear memories of two former patients, Luke Harrell (14 years old) and Caitlin Hill (9 years old). I see them smile and laugh despite devastating brain damage. Their memories remind me, with crystal clarity, that improving outcomes for patients like Luke and Caitlin (and their families) is unquestionably the only important reason for rehabilitation research.

This dissertation work was partially supported an NIH T-32 Neuromuscular Plasticity Training Grant Fellowship (T32HD043730). The Kansas City Stroke Study was funded by the National Institute on Aging, Claude D. Pepper Center Older Americans Center (5P60AG14635).

TABLE OF CONTENTS

page

LIST OF TABLES

LIST OF FIGURES

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

MEASURING POST-STROKE ARM MOTOR ABILITY: MEASUREMENT PROPERTIES OF THE FUGL-MEYER UPPER EXTREMITY ASSESSMENT EXAMINED WITH AN ITEM RESPONSE THEORY MEASUREMENT FRAMEWORK

By

Michelle Lynne Woodbury

August 2006

Chair: Lorie Richards
Cochair: Craig Velozo.
Major Department: Occupational Therapy

There is a pressing need to ascertain whether assessment tools commonly used in post-stroke upper extremity motor rehabilitation research and clinical practice are accurately quantifying impairment, characterizing recovery and producing a score that has clear functional interpretability. The Fugl-Meyer Upper Extremity Assessment (FMA-UE) is the most widely used clinical assessment of post-stroke UE motor impairment. Because of the importance of this assessment in post-stroke rehabilitation research, it is critical to examine, and perhaps improve, the quality of its measurement properties. The overall purpose of this dissertation was to examine the measurement properties of the FMA-UE. An item-oriented measurement perspective was used to investigate (1) the validity of the FMA-UE item content, (2) the reproducibility of its item structure, and (3) interpretation of its score. Rasch analysis, a type of item response theory statistical method, was applied to FMA-UE data collected from 512 individuals in

the Kansas City Stroke Registry and Kansas City Pepper Center exercise study. In the four studies comprising this dissertation, it was determined that the items of the FMA-UE contribute to the measurement of multiple constructs and demonstrate a difficulty order that is not consistent with the item order proposed by Fugl-Meyer. The reflex items were removed, resulting in a modified 30-item FMA-UE, a unidimensional measure of post-stroke UE motor ability. It was found that the Rasch-derived item difficulty order of the 30-item FMA-UE remained consistent, independent of person ability, longitudinally across two testing occasions. The modified 30-item FMA-UE was linked to the Wolf Motor Function Test. Linking is a novel method to improve the interpretability of the FMA-UE score and clearly illustrated how changes in motor impairment can translate to improvements in activity performance. The item and person parameters estimated by Rasch analysis were validated with the two-parameter item response theory graded response model. This project is significant in that it not only establishes the quality of the 30-item FMA-UE as an instrument to measure post-stroke UE motor impairment, but also establishes the 30-item FMA-UE as an instrument with which to gather information that will advance rehabilitation science motor control theory.

CHAPTER 1
INTRODUCTION AND LITERATURE REVIEW

Each year over 700,000 individuals in the United States have a stroke [1]. Approximately three-fourths of these individuals will experience immediate flaccid paralysis of one arm and hand [2]. This flaccid period is followed by rapid recovery of arm motor function over the first 3-months post-stroke [3]. However, full recovery of arm and hand function is rare. At four years post-stroke 85% of stroke survivors report residual arm or hand motor impairment [4]. For many individuals with severe stroke the arm never becomes useful [5] and for many individuals with less severe stroke, residual arm motor impairment interferes with performance of daily tasks [4]. Stroke is the leading cause of long-term disability in the United States [1].

In recent years, progress in neuroscience and rehabilitation science has advanced the understanding of the potential for neuronal recovery after stroke-related brain damage [6]. In turn, this has led to the development of new rehabilitation interventions [7-9] and revision of traditional rehabilitation interventions [10] to restore post-stroke upper extremity (UE) motor function. The rehabilitation disciplines have adopted an evidence-based practice paradigm [11, 12], necessitating accurate measurement of motor recovery. There is a pressing need to ascertain whether assessment tools commonly used in post-stroke UE motor rehabilitation research and clinical practice are accurately quantifying impairment and characterizing recovery [13, 14].

The Fugl-Meyer Upper Extremity Assessment (FMA-UE) [15] is the most widely used clinical assessment of post-stroke UE motor impairment [14, 16]. It has excellent

intra-rater reliability (r = 0.995 [17]), inter-rater reliability (r = 0.992 [17]), test-retest reliability (Intraclass correlation coefficients 0.94 – 0.99 [18]), and internal consistency (r = 0.97 [19]. It has been used as the standard from which to establish the construct validity of other commonly used tests of post-stroke UE motor function such as the Wolf Motor Function Test and Ashworth Scale [20-26]. Researchers consistently use the FMA-UE as a descriptor of functional status following stroke [4, 20, 27]. To this end, FMA-UE scores have been used to stratify research study participants into categories of stroke severity [28], predict long-term functional participation [29], and to describe arm motor impairment [30-32], arm movement quality [21], and residual arm motor function [33-35]. Furthermore, the FMA-UE is the primary criterion for evaluating the success of novel upper extremity interventions such as rhythmic bilateral movement training [36, 37], neuromuscular stimulation [38], electromyographic triggered electrical stimulation [39, 40], Botox [41], home based exercise [30], community based exercise [42], robot aided therapy [43, 44], virtual reality [45], imagery [46], forced use [47], and modified constraint induced movement therapy [48-51]. Because of the importance of this assessment in post-stroke rehabilitation research, it is critical to examine, and perhaps improve, the quality of its measurement properties.

Table 1 presents the FMA-UE. The evaluator observes the client's performance of 33 items, scoring each item by applying a 3-point ordinal rating scale (0=the movement cannot be performed at all, 1=the movement can be partially performed, and 2=the movement can be faultlessly performed). Ratings are summed and the FMA-UE score is reported out of the maximum 66 points. Higher scores indicate more UE motor ability.

Recently the Canadian Stroke Network identified 88 stroke-related UE motor function assessments [52]. The obvious question is, which is the "best" assessment? Salter and colleagues (2005) [14] argue that evidence based practice is advanced only when researchers and clinicians choose assessments with "superior measurement properties." The question then becomes, what are the essential elements of a high quality assessment? The quality of an assessment is usually judged by its "essential" measurement properties: reliability, validity, and responsiveness [14]. These criteria assure that an assessment measures the skill that it is purported to measure (validity), produces a reproducible score (reliability), and is sensitive to change over time (responsiveness). Psychometric evaluation of an assessment's measurement properties is usually accomplished with classical test theory statistical methods such as correlating or comparing assessment scores.

The item response theory (IRT) measurement framework, e.g., Rasch measurement, offers another approach for judging the quality of an outcome measure. As the name implies, IRT methods allow one to evaluate the measurement properties of items rather than whole tests [53]. One can examine an assessment's item content (i.e., does a specific item contribute to the measurement of an intended construct?), item reproducibility (i.e., do items measure the purported construct in the same way each time the assessment is administered?), and item-level interpretation of the assessment's score (i.e., with respect to item content, what does an assessment's change score mean?). Although the measurement properties of the FMA-UE have been extensively studied using classical test theory [15, 17, 18, 20, 24, 54-57], presently there are no published studies applying IRT approaches to the FMA-UE.

The overall purpose of this dissertation is to examine the measurement properties of the FMA-UE. I will assume an item-oriented viewpoint from which to investigate (1) the validity of the FMA-UE item content, (2) the reproducibility of its item structure, and (3) interpretation of its score.

In the subsequent literature review I will develop the background to address the study's purpose. First, I will review the World Health Organization's International Classification of Function and Disability (ICF). The ICF provides an operational definition of the UE skills of interest to post-stroke UE rehabilitation researchers and models a way to organize measurement of these skills. Next, I will review traditional and contemporary concepts of UE motor control. This is important because exploring FMA-UE item-level construct validity will serve to focus our attention on motor control theory. Third, I will review the Rasch measurement model because it permits testing the theoretical claims about a construct. Finally, I will briefly review the critical differences between the Rasch model and other IRT models as they pertain to this project.

In the chapters that follow I addressed the purpose of this dissertation in four studies. The aim of study 1 was to test the validity of the traditional post-stroke UE motor control theory that underlies the FMA-UE by examining its item-level dimensionality and item difficulty hierarchy in a cross-sectional analysis. The aim of study 2 was to test the longitudinal reproducibility of the FMA-UE item structure. The aim of study 3 was to test a method for enhancing the interpretability of a FMA-UE change score by linking to another assessment with more functionally oriented items. Because Rasch analysis was the primary method used throughout these three studies, the final project was to compare the results obtained with the Rasch model to those obtained with another IRT model to

assure that my choice of statistical method was valid. The overarching purpose of this dissertation and the aims of each project are consistent with that of an overall emphasis within the field of rehabilitation research to assure that clinical assessment tools are optimal for measuring targeted outcomes [58-60] and translate to functionally interpretable measurements of change [61].

The International Classification of Function and Disability

The World Health Organization's International Classification of Function and Disability (ICF) [62] offers a systematic approach to classify and describe post-stroke UE motor function according to a theory of health and disability. The ICF provides a useful taxonomy and standard language to define UE motor function and describe its measurement. As a practical research application, the ICF describes sets of behaviors that should be included in an instrument intended to measure given domain [63].

The purpose of the ICF is to describe *components of health* in which a person's functioning is conceived as a dynamic multi-dimensional interaction between individual health attributes and contextual factors. The ICF identifies three primary health dimensions of the individual: the functioning of the body or body parts, the functioning of the individual in his/her daily activities, and the functioning of the individual in relationship to his/her participation in the socio-cultural community. These health dimensions can also be expressed in the negative to describe disability: body function/structure *impairment*, activity *limitation*, and participation *restriction*.

The ICF models the concept that stroke-related UE motor deficits affect the individual along a continuum of health dimensions; from impaired body functions to limited performance of activities and restricted participation in the community. Accordingly, post-stoke arm/hand motor function can be measured in each dimension.

Furthermore, the model suggests that altered arm motor function as measured in one dimension (e.g., body function) conceivably translates to altered arm motor function when measured in an adjacent dimension (e.g., activity performance). The purpose of this dissertation centers on UE motor body function impairment and activity restriction; thus the subsequent discussion is limited to these domains.

In series of papers [14, 64, 65] researchers from the Canadian Stroke Network organized the most commonly used post-stroke assessments according to ICF domains for psychometric review. The FMA-UE is classified as an assessment of *body function* impairment because its item content is consistent with the ICF operational definition of this domain [14]. For example, "neuro-musculoskeletal and movement related functions" illustrate one aspect of body function. More specifically, the functions are described as "control and coordination of single and multi-joint motions" (p. 99-100). FMA-UE items fit this description.

A valuable assessment for measuring upper extremity outcome after stroke is one that provides information about the recovery spectrum, i.e., movement impairment and the functional consequences of movement impairment. The ICF models a relationship between body function impairment and activity performance. According to the ICF, it is conceivable that body function impairment impacts behaviors in the adjacent activity performance domain such as the ability to "carry, move and handle objects." More specifically, "reaching," "grasping," and " lifting," (p. 141–143). It is conceivable that an UE assessment intended to capture the functional aspects of reduced body function UE impairment should be comprised of items with this variety of content.

Although the ICF offers a broad framework to define FMA-UE item content and organize its structure, the ICF does not specify what UE motor behaviors are affected by stroke, nor does it detail expectations regarding post-stroke recovery. For that, a discussion of motor control theory is warranted.

Motor Control Theory

Motor control theory provides a conceptual framework of how the CNS organizes and controls motor function by addressing potential "rules" governing movement organization. Motor control theory models normal motor behaviors, defines injury-related motor behaviors, and predicts a post-injury course of motor recovery [66]. One way in which an abstract motor control theory becomes tangible to the rehabilitation professional is in the design, structure and scoring of a clinical assessment [67]. The theory defines "normal" and "abnormal" movement behaviors therefore, governing the content of assessment items. Moreover, the theory describes behaviors that evidence "less" and "more" motor control and, in doing so, suggests an appropriate array of items to measure a range of motor skills. Barecca et al. (2006) [68] argues that stroke rehabilitation knowledge is advanced only if the tools chosen to measure motor control are consistent with contemporary motor control theoretical expectations of deficit and recovery. Recent advances in movement science challenge the motor control theory upon which the FMA-UE was based. Prior to a review of these challenges, I will review the theoretical foundation of the FMA-UE.

Traditional Theories of Motor Control and Recovery

The characterization of post stroke recovery has been the focus of theoreticians, clinicians and researchers and represents one of the historic underpinnings of modern rehabilitation. Many of the traditional stroke rehabilitation frames of reference emanate

from Jackson's late 19th century work with persons with epilepsy and other "nervous disorders" [69]. Jackson, a neuroscientist influenced strongly by Darwin, applied the "doctrine of evolution" to his observations and ultimately posited a theoretical order of the CNS. According to his framework, the complex motor abilities seen in humans are evidence that higher motor centers have evolved out of lower centers. In the healthy nervous system these more advanced centers orchestrate and modify the reflex chains of lower centers. Jackson's observations of individuals with epilepsy led him led him to hypothesize that injury to the cortex resulted in "dissolution" of the CNS and reversal of its evolutionary structure. That is, without input from the advanced motor centers, movement control defaults to primitive lower centers evidenced in gross undifferentiated reflexive motions.

In the early 20th century Charles Sherrington and his colleagues performed a series of elegant experiments to show the existence of the reflex and describe the interaction of afferent feedback with voluntary movement. In a classic primate experiment Sherrington performed complete upper limb sensory denervation. This procedure appeared to abolish all voluntary upper limb movement. The limb was held close to the body and not used for daily tasks. Sherrington concluded that afferent information was a necessary building block of voluntary movement [70]. The work of Sherrington and his colleagues formed the experimental foundation of the classic reflex chain hypothesis [71]. The results of Sherrington and Jackson's experiments and observations were drawn together and are often referred to in the clinical literature as the reflex-hierarchical theory of motor control [66, 72].

With the advances in medicine and improved post-stroke survival in the mid 20th century, clinician-researchers applied the reflex-hierarchical theory to their clinical observations [73-75]. In a classic paper Dr. Thomas Twitchell (1951) [75] described an "orderly progression of phenomena" characterizing the course of post-stroke UE motor recovery. Initial UE flaccidity was followed by emergence of gross flexor and extensor voluntary movement synergies, gradual separation of synergy movement patterns and finally distal fine motor coordination. Twitchell attributed abnormal motor behaviors to cortical dissolution and emergence of default reflexive behaviors as per Jackson's model of the CNS. Twitchell suggested that recovery was a "re-evolution" of this hierarchy as the cortex reassumed its natural inhibition of primitive proprioceptive reflexes.

Renowned physical therapist Signe Brunnstrom furthered Twitchell's work, and made an important contribution to stroke rehabilitation by creating an assessment tool that synthesized clinical observation with motor control theory [76, 77]. She mapped post-stroke recovery into a six-stage sequence. Beginning with flaccidity (stage 1), motor recovery was first evident as return of reflexive responses (stage 2). With further recovery individuals gained voluntary motion within whole-limb flexion and/or extension synergies (stage 3), movement requiring a combination of the synergy patterns (stage 4), followed by movement independent of synergy patterns (stage 5), and finally individual joint motion and normal coordination (stage 6). The "Hemiplegia Classification and Progress Record" [76] detailed specific motor behaviors typifying each stage. The hierarchical ordering of movements documented in the Progress Record was useful to the clinician for assessment and treatment planning. Using it, therapists could identify a patient's stage of recovery by observing the movements the patient could or could not

perform. Once the client's location along the recovery continuum was known, the Progress Record enabled the clinician to map treatment to advance the client to the next stage.

Noting the dearth of quantitative assessments available to stroke rehabilitation clinicians and researchers, Fugl-Meyer et al. (1975) [15] created a rating scale assessment based, at least in part, on Brunnstrom's Progress Record and Twitchell's observations. Fugl-Meyer's theoretical framework broadly reflects the assumptions of the reflex-hierarchical motor control theory. Fugl-Meyer stated:

> The form has been constructed following the hypothesis that the restoration of motor function in hemiplegic patients follows a definable stepwise course. Thus for a patient with hemiparalysis, recurrence of reflexes always precedes volitional motor action. Thereafter through initial dependence on synergies, the active motion will become successively less dependent upon the primitive reflexes and reactions and finally complete voluntary motor function with normal muscle reflexes may be regained. [15] p. 14

Fugl-Meyer chose items to exemplify the construct of motor recovery including reflex items and voluntary movement items. Moreover, because the assessment was intended to measure recovery, Fugl-Meyer arranged the items in a hierarchical order to illustrate the recovery process. He expected motor recovery to proceed in a reflexive-to-voluntary, and synergy-to-isolated progression as the cortex resumed its ability to integrate reflexive movement components into coordinated limb motions.

Challenges to the Traditional Theories of Motor Control and Recovery

In recent years there have been many advances towards understanding how movement is controlled, organized and recovered after injury. There has been a shift away from a purely neurophysiologic explanation of motor control/recovery to a task-oriented view of motor control/recovery [66]. Contemporary motor control science posits that movement emerges from a dynamic, complex interaction of the individual's neural

and musculoskeletal apparatuses with specific tasks and environmental demands [78]. Recovery is viewed as re-acquisition of an individual's ability to successfully conquer task-specific contextual demands rather than neurophysiologic re-evolution of the cortical motor control hierarchy [79].

Recent advances in movement science challenge the "definable stepwise course" of motor recovery upon which the FMA-UE is based in two ways: (1) the place of reflex behaviors in an assessment of motor recovery, (2) the pattern of behaviors that evidence the course of recovery. These challenges will now be expanded upon.

Role of reflexes

Contemporary views of the central nervous system indicate that reflexes and voluntary movements are uniquely different types of motor behavior. A relatively simple spinal neural circuit governs the tendon tap muscle contraction, while a complex supraspinal and spinal neural network governs voluntary goal-directed movement [80]. Afferent input is not necessary for voluntary movements to occur [81], instead reflexes (e.g., the stretch reflex) appear to be important adjuncts of voluntary motor control [80, 82, 83]. The reflex motor control system is complex as evidenced by research showing that the sensitivity of the of the stretch reflex threshold is modulated according to previous muscle activity [84, 85]. The research of Wolpaw (1985) and Wolf and Segal (1986) has demonstrated that the magnitude of the stretch reflex is amendable to operant training [86, 87]. Burne and colleagues (2005) [88] found that spasticity (measured as the amplitude of the biceps stretch reflex) was different in a resting limb versus a moving limb. The researchers suggest that reduced ability to dynamically modulate reflex activity contributes more to an understanding of voluntary movement impairment than assessing reflex activity in a static condition. The research evidence suggests that the FMA-UE

reflex items may be assessing a motor behavior that is uniquely different from other items that assess voluntary movement.

The presence of the three reflex items in the FMA-UE has been long questioned. Lindmark and Hamlin (1988) [89] removed the items, altered the rating scale, and added items to form the Lindmark Functional Capacity Evaluation. Gladstone et al. (2002) [55] stated that reflexes were "over represented" in the FMA-UE total score. Whitall et al. (2004) [13] suggested that assessment of tendon-tap reflex behaviors does not contribute to an understanding of post-stroke UE motor ability. It is reasonable to question whether the presence of the three tendon tap FMA-UE reflex items confound the interpretation of the FMA-UE score and threaten the assessment's construct validity.

The pattern of upper extremity behaviors evidencing recovery

The extent to which post-stroke individuals exhibit abnormal UE flexor and/or extensor synergy as per Fugl-Meyer's descriptions [15] (adapted from Brunnstrom [76]) is debated in the literature. Table 2 reminds the reader of limb movements characterizing traditional synergy patterns. Note that these movements are items 3 – 11 on the FMA-UE (see table 1).

Recently, Welmer and colleagues (2006) [90] prospectively observed 64 consecutive patients with stroke at one week and three months post-stroke. The researchers tested the hypothesis that stroke recovery is consistent with Brunnstrom's sequence, specifically that UE flexor and extensor synergies emerge early in recovery while non-synergistic isolated movements emerge later in recovery. Subjects were evaluated by trained, study-specific, occupational and physical therapists. One-week post stroke 15% of the sample was unable to follow testing directions and 22% of the sample exhibited flaccid hemiplegia. Among the subjects who participated in voluntary UE

movement assessment, no participant exhibited only (i.e. "pure") flexor/extensor synergy movements. Welmer found that 63% of the sample demonstrated at least one UE movement inconsistent with synergy (e.g., was able to partially extend the elbow while abducting the shoulder). Three months post stroke, 13% of the sample remained unable to move. Among the others, 16% exhibited at least one UE movement typically considered part of the flexor/extensor synergy pattern (e.g., tended to flex the elbow while flexing the shoulder to or above 90 degrees) but 77% exhibited no evidence of synergy-related movements. The results suggest that some flexion/extension synergy movements occur early in recovery, but some isolated movements also occur early in recovery. The researchers concluded that individuals with stroke rarely exhibit *pure* manifestations of the flexor or extensor synergy patterns.

Welmer's work does not discount longstanding clinical observations [75, 76] that individuals with stroke exhibit impaired intra-limb control. Instead, Welmer's study suggests that the abnormal pattern is not manifest as an "all or nothing" obligatory link between all movements comprising the traditional definition of the synergy (table 2). Indeed, contemporary experimental paradigms have elucidated the specific parts of the abnormal pattern that are most likely to be present. In a series of studies Beer et al. (1999) [91] and Dewald et al. (2001) [92] have shown that individuals with stroke were unable to maintain maximal joint torque of a one joint in a single direction without producing secondary torques of an adjacent joint in an unintended direction during an isometric task. In this study, eight individuals with chronic stroke and moderate spasticity (Ashworth scores 3-5) were directed to attain and then maintain maximal shoulder flexion or shoulder abduction force. Forces were measured with load cells on the

humerus and forearm. In both the shoulder flexion and abduction conditions, all subjects concurrently produced unintended elbow flexion torque. In a similar experimental paradigm, Lum et al. (2003) [93] showed that post-stroke subjects also produced unintended humeral internal rotation torque. The results suggest some individuals with stroke exhibited abnormally constrained intra-limb motor control evident by abnormally linked shoulder flexion/abduction, elbow flexion, and internal rotation torques.

Abnormal linkages of elbow flexion with the shoulder flexor or abductor musculature are also evident during supported and unsupported reach. Beer et al. (2004) [94] studied eight persons with chronic stroke and moderate spasticity. In this study subjects performed rapid movements in the horizontal plane from a central starting point to five targets located at the outer limits of the reachable workspace. The targets were placed to require different combinations of flexion and extension ranges of motion at the elbow and shoulder. In one condition the hemiparetic limb was supported in a friction-less arm trough, in another it was not. In both conditions subjects had difficulty generate concurrent shoulder flexion and elbow extension torques necessary to reach the targets, however the deficit was greater in the unsupported versus supported condition. Kamper et al. (2002) [95] showed similar interjoint coordination deficits in an unsupported reaching condition. In this study, sixteen participants with chronic stroke were directed to point to toward a screen containing 75 targets displayed side to side across the arm reachable workspace from waist to head. Kinematic analysis showed that participants exhibited abnormally constrained elbow flexion and shoulder flexion evident in curved hand trajectories when attempting to reach to targets the farthest from the body. The deficit was similar irrespective of target locations in the ipsilesional or contralesional workspace.

Dewald, Beer and Kamper's results are consistent with those of Levin et al. (1996) [35]. In this study, the researchers studied reaching trajectories of 10 participants with moderate post-stroke UE impairment (moderate spasticity measured on a scale designed by the investigator and baseline FMA-UE scores 19 – 66). Targets were arranged on a table so that subjects were required to utilize muscle activation patterns within and outside of the traditional synergy configurations. For example, to reach a target placed contralateral to the affected limb, the participant was required to activate the extensor synergy musculature; shoulder adductors, internal rotators, and elbow extensors. The researchers reasoned if participant's movements were strictly constrained into a flexor or extensor synergy pattern, participants would be less able to reach targets outside this pattern. That is, if extensor synergy were dominant, reach to a contralateral target would be less impaired than reach to an ipsilateral target. If the flexor synergy were dominant, reach to a near target would be less impaired than reach to a far target. The results showed that this was not the case. Reaching trajectories were less smooth and more curved for individuals with stroke when compared to healthy individuals, however post-stroke movements within a typical synergy pattern were no less impaired (i.e., were not smoother and straighter) than movements to targets outside the synergy pattern.

In an earlier study Trombly (1992) [96] realized a similar phenomena. In this study, 5 subjects with left hemiparesis were directed to reach to three targets placed so as to require movement into an extensor synergy pattern or away from this pattern. Kinematic and electromyographic (EMG) data were recorded. All clients demonstrated impaired kinematics (less smooth trajectories, more movement units, increased movement time, and earlier time to peak velocity) and altered EMG (higher percentage of maximum

voluntary output) in the impaired UE as compared to the unimpaired UE. However, the biceps-anterior deltoid coactivity indices were not significantly different between the limbs indicating that an abnormal synergistic constraint was not hampering the movement.

The work of Trombly, Levin, Kamper, Dewald, Lum and Beer provide evidence that the post-stroke reach is impaired by an abnormally constrained inter-joint coordination pattern involving a tendency to flex the elbow and internally rotate the humerus when attempting to lift the humerus against gravity. Tromby [96] suggested that this observable pattern is not directly attributable to abnormal co-activity of the "flexor synergy" musculature, but may evidence a motor control strategy employed to reach targets despite muscle weakness. When put together with the work of Welmer et al. [90], it is apparent that the observed post-stroke UE motor deficit is not fully explainable by traditional definitions of synergy. Full flexor or extensor synergy patterns are not always present, at least in the subjects studied by the above labs. It is possible that the other muscles typically considered part of the synergy patterns, specifically the scapular retractors, scapular elevators, shoulder external rotators and forearm rotators have not been studied due to limitations in collecting surface EMG from deep muscles and limitations in UE kinematic and kinetic modeling. However, at least to date, the evidence suggests pure synergies (as per the traditional definition) do not appear to be obligatory patterns for all individuals with stroke.

The importance of this research lies in its potential impact on the interpretation of the FMA-UE score. Fugl-Meyer designed the scale with an expected regularity of recovery in mind [15]. The evidence only partially supports this expected synergistic

recovery pattern. Therefore, it is possible that the measurement properties of the FMA-UE with regards to the interpretation of its score may be affected.

Researchers use the FMA-UE score to characterize post-stroke recovery. One interpretation of the FMA-UE score is consistent with traditional expectations regarding the pattern of recovery; specifically low scores represent synergy-dominated movement abilities whereas higher scores represent more isolated movement abilities. For example, Chae et al. (2003) [97] interpreted the FMA-UE score in this way *"it (the FMA-UE) measures evolving synergy patterns with subjects with voluntary movements limited to proximal joints and synergy patterns exhibiting the lowest scores and with subjects demonstrating isolated movements in distal joints, with normal reflexes exhibiting the highest scores"* (p. 7). Some researchers have used categories as a context from which to interpret the FMA-UE score. For example, Michaelson et al. (2001) [98] stated that FMA-UE scores of 0–19 points represented severe UE motor impairment, FMA-UE scores of 20–64 points represented moderate UE motor impairment, and FMA-UE scores of 65-66 represented normal UE motor ability. Pang [42] stated that FMA-UE scores of 0−27 points represented severe UE motor impairment, FMA-UE scores of 28−57 points represented moderate UE motor impairment, and FMA-UE scores of 58−66 represented mild UE motor impairment. Sometimes researchers designate a FMA-UE score to indicate the boundary between synergy-dominated and non-synergy dominated movements. In the original article, Fugl-Meyer provided evidence that a score approximately equal to 60% of the total shoulder/arm score, a score of 13 out of 22 points, was the boundary between synergy and non-synergy movements [15]. Finley et al. (2005) [44] utilized a maximum FMA-UE score of 15 as inclusion criteria for a study of

robotic training because *"the score represented severe motor impairment in the paretic arm, as demonstrated by limited movement within synergy patterns and no voluntary wrist or hand function"* (p. 684). The ranges of FMA-UE scores and their relationship to the categories are frequently used in clinical research as a reference point for describing participants' motor behavior and interpret outcomes. However, whether the scores are being interpreted correctly is not known.

Contemporary Theories of Motor Control and Recovery

To this point I have argued that the validity of the FMA-UE item content and the interpretation of the FMA-UE score may be threatened by challenges to traditional motor control theory upon which the assessment was founded. The purpose of this dissertation includes enhancing the measurement properties of the FMA-UE if necessary and possible. This necessitates further exploration of contemporary theory to provide a better understanding of the UE motor recovery construct than was provided by traditional theories.

UE motor impairment

Bernstein (1967) [99] observed that any given motor task could be achieved using many different combinations of muscle activation patterns and limb configurations. The motor system chooses from the vast number of "degrees of freedom" available to it depending on the contextual demands of the movement task (e.g., influence of gravity, inter-segmental joint torques). Bernstein argued that it is unlikely that the motor control system exerts executive control over each movement possibility (i.e., one to one control of each degree of freedom) because a vast amount of storage capacity would be required of the cortex. Instead, Bernstein proposed "functional synergies" as the more probable motor control strategy. "Functional synergies" are defined as "*highly evolved task-*

specific ensembles of neuromuscular and skeletal components constrained to act as a single unit so as to reduce the computational burden associated with controlling voluntary movements in multisegmented limbs" [99] p. 47.

In the healthy CNS, functional synergies appear to be flexibly organized according to specific task requirements and goals. Using kinematic assessment, researchers have shown that distinctive motor patterns emerge in response to task and contextual demands [100-105]. For example, Trombly's lab demonstrated that neurologically intact individuals displayed unique kinematic patterns of movement when asked to scoop actual coins from a table versus scooping imagined coins [100] and smoother reaching trajectories when reaching to a food chopper containing actual versus imagined food [101]. Trombly suggested that the perceptual and cognitive environments in the imagined versus real conditions were different 'sets' affording movement responses particularly patterned to the respective context.

Functional synergies also appear to be dynamically organized according to certain task-specific mechanical factors. For example, Murray and Johnson (2004) [106] found that healthy individuals scaled UE joint torques according to the number of joints involved in a task. For example, healthy individuals generated less joint torque during functional tasks requiring primarily a single joint movement such as feeding, when compared to tasks requiring multiple joint movements such as simulated hair brushing.

This research suggests that healthy individuals dynamically, and flexibly organize and execute motor programs in accordance with visuo-perceptual and mechanical task-demands. Do individuals with stroke retain the capacity to flexibly organize and apply "functional synergies" in task-specific ways?

Latash and Nicholas (1996) [107] suggested that an “abundant” motor control system enables an individual to adapt to specific environmental conditions or states of injury. He argued that the “abnormal” synergy is actually a “normal” attempt to overcome muscle weakness and stroke-related neural damage in order to successfully accomplish a task (i.e., lift the arm). Latash notes that the changed movement pattern reflects the admirable adaptability of the human CNS [108].

Research supports the presence of post-stroke UE weakness as a possible mechanism underlying the abnormal flexor synergy pattern. In a paradigm similar to that of the Dewald and Beer lab, Lum et al. (2003) [93] found that individuals with stroke producing unintended secondary torques in patterns consistent with traditional definitions of flexor synergy. However, unlike the Dewald and Beer studies, Lum collected EMG recordings of shoulder and elbow agonists/antagonists during the experiment. Subjects producing the largest secondary torques (i.e., less able to dissociate elbow flexion from shoulder abduction) also exhibited the greatest force production deficits. According to Lum, the results suggested that abnormal post-stroke synergy patterns are, at least in part, as a consequence of stroke-related UE weakness.

Mercier et al. (2005) [109] examined single and multi-joint isometric muscle activity requiring different combinations of shoulder and elbow flexion/extension. Subjects were included in the study if their UE movements were consistent with Chedoke–McMaster Stage 3 [23] (similar to Brunnstrom’s flexion synergy pattern; participants demonstrated active UE movement, but mostly in flexion). Joint torques were measured with force transducers placed at the forearm and humeral shaft. The results showed that subjects demonstrated overall decreases in force production at the shoulder.

However, the deficits were similar irrespective of the direction to which the force was aimed.

In subjects with stroke, Zackowski et al. (2004) [110] found that a movement requiring simultaneous shoulder flexion and elbow *flexion* was performed faster and more accurately than an arm movement requiring shoulder flexion and elbow *extension*. Zackowski suggested that activation of the biceps to flex the shoulder might be a strategy to compensate for weakness of the anterior deltoid.

McCrea and colleagues (2005) [111] demonstrated that the anterior deltoid reached maximal activation ("saturation") in persons with stroke much sooner than healthy controls while abducting the shoulder. As saturation was achieved, subjects flexed the elbow and internal rotated the humerus to reach the target. McCrea found that the impairments were worse in individuals with more severe stroke compared to less severe stroke.

Taken together, these studies suggest that shoulder weakness contributes to the observed "flexor synergy" movement pattern. Following stroke the motor control system has reduced resources with which to successfully interact with (i.e., adapt to) the environment [112]. Cortical stroke damages the corticospinal pathway and may unmask spared secondary motor tracts (e.g., vestibulospinal and rubrospinal) [113]. Stroke reduces the capacity to recruit motor units [114], thereby reducing strength [115] , altering patterns of agonist and antagonist recruitment [33] and reducing interjoint control [116]. Latash and Anson (1996) [108] suggested that flexing the elbow (i.e., the observed abnormal flexor synergy pattern) may afford a mechanical advantage by

reducing the length of the lever arm and allowing for the individual to adapt to the post-stroke state yet still accomplish a task goal (i.e., flex the shoulder).

Extending this argument, Carr and Shepherd (2002) [78] propose that post-stroke recovery can be viewed from a similar task-specific perspective. Individuals with stroke-related motor control deficit are able to accomplish tasks if their motor resources are sufficient to satisfy the task-specific demands placed on the motor system. Individuals with less UE motor ability would be more likely to accomplish an "easy" versus a "more challenging" task. As these individuals recover they gain the capacity to accomplish both easy and more challenging motor tasks.

Post-stroke UE recovery

According to Carr and Shepherd, motor recovery following stroke is "fundamentally" a process of relearning how to move [78]. Most contemporary rehabilitation frameworks are predicated on the assumption that in order for a patient to improve his/her motor skill (i.e., recover from stroke-related hemiparesis) the patient must practice motor skills [79]. Principles of neural plasticity first researched in animal models translated to humans suggest that healthy brain neuronal connections and cortical maps are dynamically remodeled by experience [6, 117]. Post-stroke UE recovery may be enhanced with intense task specific-practice that engages the system to actively solve motor problems [118, 119].

Note the conceptual dissimilarity between this contemporary understanding of recovery and its more traditional counterpart. Traditionally post-stroke recovery was viewed as a passive (from the viewpoint of the patient) process. Rehabilitation interventions were provided to patients who largely assumed the role of passive recipient of the intervention, in order to facilitate (or inhibit) an internal neurophysiological

process [120]. In contrast current motor control researchers and theorists view post-stroke recovery as an active process in which the individual, given his/her internal neural and muscular deficits, re-learns how to interact with the external environment to satisfy the demands of the activity [70, 121].

This may signal that recovery is characterized by enhanced ability to execute movement patterns with the appropriate trajectory, acceleration and relative muscle forces necessary to meet the temporal and spatial requirements of a task [66]. Roher (2002) [122] demonstrated that with one month of robotic training, individuals with chronic stroke were able to reach to a target with a smoother hand trajectory velocity profile, suggesting increased ability to move independently of the elbow-shoulder flexor "synergy". Ellis and Dewald (2005) [123] found that with practice, individuals with chronic stroke were able to modify the abnormal coupling of isometric shoulder and elbow joint torques evidenced by increased ability to isolate joint torque production at a single joint during a multijoint task. Michaelson et al. (2006, 2004) [9, 124] and Thielman et al. (2004) [125] have shown that with training, individuals with severe to moderate stroke demonstrate improved shoulder-elbow inter-joint coordination evidenced by decreased trunk recruitment and shorter (straighter) hand path trajectories during reaching tasks in various parts of the reachable workspace.

Taken together, the above studies suggest that post-stroke recovery is characterized, at least in part, by the patient relearning the "rules that govern motion" [107]. Individuals with stroke become more able to control each joint during a multijoint task. Patients improve, using ICF terminology, the ability to "coordinate single and multiple joint motions" [62].

Post-stroke recovery of functional UE motor abilities

The research presented to this point supports the view that post-stroke recovery is a process in which individuals gain more UE motor ability with regards to overcoming task demands. More functionally oriented activities requiring interaction with objects and at increased speeds of movements. Task-specific increases in neural, mechanical and/or contextual requirements concomitantly increase the computational demands on the motor system [112]. Interacting with an object not only increases the mechanical load on the limb but also increases the perceptual processing requirements of the task, e.g., attention to the characteristics of the object [126], feedforward adjustments of posture [127] and increased reliance on feedback to correct movement errors [70]. It is likely that these tasks are more difficult and recover later than movements without functional endpoint. This appears to be the expectation of Fugl-Meyer who found a strong correlation between the FMA-UE score and ADL performance [15]. However, evidence suggests that the progression from reduced body function impairment (i.e., improved intra-limb coordination) to recovery of activity performance is neither linear nor clearly understood [128].

The association between motor function impairment assessed by the FMA-UE and functional assessments of activities of daily living is both supported and debated in the literature. For example, both Filiatrault et al. (1991) [21] and Platz et al. (2005) [56] have shown a low correlation (Spearman r = 0.04 and 0.08) between the FMA-UE score and the score on the Barthel Index [129], an assessment of ADL performance. Francisco et al. (1998) [39] found that four individuals with stroke, all of whom exhibited palpable wrist extension at baseline, demonstrated concurrent gains in FMA-UE and Functional Independence Measure (now known as the FIM™) motor scores following inpatient

rehabilitation with usual care augmented by wrist EMG electrical stimulation. In contrast, Winstein et al. (2004) [130] showed no relationship between gains on any impairment measure (e.g., the FMA-UE) and changes in the FIM™ after a two week intervention. Shelton (2001) [131] showed that admission motor impairment (measured by the FMA-UE) predicted discharge disability (measured by the FIM™), however the researchers note that much of the gain in FIM™ score came as a result of increased use of the less affected limb to perform the tasks.

Shelton's observation is consistent with Winstein's argument [59, 130] that the FIM™ like other global assessments of ADL such as the Barthel Index, are not sensitive to impairment level changes in UE motor ability. This is because neither the FIM™ nor Barthel directly assess the function of the hemiparetic arm. A client is able to use compensatory one-handed strategies to accomplish tasks with the unaffected limb and still receive a high score on these assessments. The association between the FMA-UE score and scores of functionally oriented assessments where the assessment requires use of the hemiparetic arm is stronger. The FMA-UE was highly correlated (Spearman r = 0.92) with scores on the Action Research Arm Test [56], the Wolf Motor Function Test (Spearman r = 0.57 – 0.68) [26], the Arm Motor Ability Test (Spearman r = 0.92) [97] and the Lindmark (1988) [89] scale. Each of these assessments requires the patient to use the hemiparetic arm for functional reaching or grasping tasks.

Mercier et al. (2004) [132] notes that it is possible that the relationship between body function impairment and functional may not be linear across different levels of ability. For example, she suggests that a task requiring grip strength (e.g., lifting a pitcher filled with water) will be performed at a lower stage of the recovery process than a task

requiring a precision grip (e.g., picking up a small object from a table). Winstein (2004) [130] found that participants in the less severe group (stratified by Orpington Prognostic Scale score [133]) showed greater improvements in the Functional Test for the Hemiparetic UE (an assessment comprised of 17 graded tasks assessing "integrated function" of the UE) [134] than clients in the more severe group. Feyes et al. (1998) [135] found that clients with more severe motor impairment demonstrated significant gains on the FMA-UE but the results did not "generalize" to concurrent changes in the Action Research Arm Test or Barthel Index.

It is possible that individuals with less UE motor ability do not have enough motor ability to perform even the simplest functional tasks. Chae et al. (2003) [97] found that individuals with higher scores on the FMA-UE hand subscale demonstrated a strong correlation with the overall AMAT score. Chae suggests that these individuals most likely had enough UE motor ability that enabled them to successfully complete AMAT items (e.g., cutting meat, dialing a telephone). In contrast individuals with less UE motor ability (a lower FMA-UE score) likely did not have enough motor capacity to accomplish the more functional AMAT items. Pang et al. (2006) [42] found a similar relationship between participant ability and functional performance. In this study, more severely impaired clients showed significant pre to post intervention changes in FMA-UE scores (2.5 ± 2.3 points), but no change in WMFT score (-0.3 ± 0.5 seconds). Moderately impaired clients showed concurrent changes in both assessments (FMA-UE 7.4 ± 3.8 points and WMFT -1.5 ± 6.2 seconds). The least severely impaired clients demonstrated little increase in FMA-UE score (7.4 ± 3.8 points) but the greatest gains on the WMFT

test (-6.6 ±1 6.4 seconds). In general, the evidence suggests that individuals with greater ability will show changes in both domains, less able people will not.

Clinical significance of the FMA-UE score

The importance of the studies reviewed above is that the FMA-UE score is difficult to interpret with regards to its functional relevance. That is, although individuals with stroke may recover the ability to "coordinate single and multiple joint movements" as evidence by improved FMA-UE score, the functional meaning of this improvement is not always clear.

Issues of clinically significant or clinically meaningful changes in assessment scores have become paramount in rehabilitation research [61, 136, 137]. The relationship between "statistically significant" research results, and "clinically significant" client-oriented functional improvement is not always apparent [136]. Bonnifer and colleagues (2005) articulated this concern;

> Although this study showed significant increases in FMA-UE (scores) after the constraint induced movement therapy (CIMT) intervention, these findings do not suggest that a 3-week CIMT program restores motor ability to pre-stroke levels. It is not clear how increases in scores on any of these measures translate to real-world functional abilities. For instance, a peak score of 33 on the FMA-UE score does not suggest that a person can perform functional tasks using the involved upper extremity. On the contrary, the arm is still quite impaired and most likely used as an assist at most for some activities. [51]

Conceptually, there are two approaches for determining what constitutes "significant" change: distribution and anchor (or criterion) based approaches [61, 137, 138]. "Distribution-based" methods define an important change score as one that exceeds a pre-defined level of error. These techniques used to calculate significant change include minimal detectable change, the minimal important difference, effect size and/or one of numerous responsiveness statistics. The drawback of all of the se

approaches is that the value obtained in the analysis is not a fixed property of an assessment, instead varying with the setting of the study and characteristics of the sample. Different values of "significance" might be obtained in different groups of patients based on their initial level of impairment. Ceiling or floor effects can mask change and effect the calculation of a responsiveness or effect size value. Moreover, these methods require a judgment of what is important change. But, who should be the judge, client, therapist, physician, payer or test developer? Therefore it is unlikely that there exists a single score to define "significant" change across all samples and contexts.

In contrast, "anchor-based" methods relate the change score to a functionally relevant criterion pre-determined by client or clinician. This criterion "anchors" the meaning (i.e., significance) of the change. Both approaches have been applied to interpretation of FMA-UE change scores [61].

According to Van der Lee et. al. (1999) [47] the literature contains no criteria for interpreting a FMA-UE change score with regards to its "minimal clinically important difference (MCID)". Van der Lee et al. (2001) [24] determined that a 10% change in the overall FMA-UE score made measurement error unlikely thus allowing for the conclusion that real change occurred. Using Van der Lee's criteria, McCombe-Waller and Whitall (2005) [139] interpreted their 8% gain in FMA-UE score as "approaching clinical relevance". Based on an effect size calculated in a pilot study, Feyes et al. (1998) [135] determined that a 10% change in the overall FMA-UE score was "clinically relevant" meaning that participants moved from a more impaired to a less impaired stage of recovery. Pang et al. (2006) [42] explored the clinical significance of a FMA-UE change score using external criterion. In this study the researchers administered a

satisfaction survey to participants in a community based exercise program. Clients in this intervention showed a 0.71 – 0.81 effect size for pre to post test FMA-UE score change, and concurrently reported increased ability to perform simple functional tasks such as "now able to switch on/off lights". Pang notes that the intervention related changes appear to be clinically significant from the patient's perspective.

Summary

In the section above I have identified three concerns with the FMA-UE. First, it is possible that some items, specifically the reflex items, are not consistent with the construct represented by the other items (i.e., voluntary UE motor ability). Second, many assumptions have been proposed regarding the meaning of the FMA-UE score. These assumptions have gone untested. Finally, the relationship between the FMA-UE score and performance of functional "reaching, grasping, lifting" activities with the more affected UE is not obvious.

Item Response Theory offers a method to test specific hypotheses addressing each of these concerns. The following section will review Item Response Theory in general, the Rasch measurement model specifically, to establish the background for using this measurement framework to explore these areas.

The Rasch Measurement Model

Measurement

Measurement is defined as "determining the amount or quantity of something". Mass, length, and time are considered fundamental quantities of measurement in the physical sciences [140] and are measured with devices accepted as "objective" measurement tools. An objective measurement remains constant and unchanging across

the persons measured and is described in a unit that maintains its size as the unit is repeated [141].

Ideally the process of measuring a client's UE motor function should be similar to the process of measuring length, temperature or distance. That is, UE motor function should be quantifiable using an assessment tool that is structured to measure UE motor skill in the same way that a ruler is structured to measure length or that a thermometer is structured to measure temperature. Bond and Fox argue that in order for a clinical assessment to objectively measure a skill, e.g., UE motor function, the assessment tool should display markings (e.g., assessment items) that clearly describe "less" and "more" of the skill. The order of the markings should represent the acquisition of the skill in a pattern consistent with its recovery or development [142]. Also, an "objective" measurement of UE motor function should connect the numbers produced by an assessment with its content [143].

Item Response Theory

The increasing need for meaningful and psychometrically sound rehabilitation assessments has led many to apply modern measurement theoretical statistical methods (IRT) to the construction, re-design, and scoring of clinical assessments. IRT measurement models are comprised of a family of mathematical models [53]. Each model estimates an item difficulty parameter. The two-parameter logistic models (2PL) also estimate an item discrimination parameter, and the three-parameter logistic models (3PL) estimate item discrimination and guessing parameters [144]. The Rasch model is usually considered a one-parameter IRT model, although the technicalities of this are the subject of ongoing debates in the literature.

IRT, as its name implies allows for an item-oriented rather than test-oriented analysis of an assessment's measurement properties. The appeal of IRT is that person-ability and item-difficulty are calculated on the same measurement scale which means that a score can be directly related to the probability of item responses [53].

Potential Advantages for Using the Rasch Model to Explore the FMA-UE Measurement Properties

Rasch analysis has been widely utilized to construct and/or refine clinical health care assessment tools [for example [145-152]]. The model is a powerful framework with which to address the measurement properties of the FMA-UE: specifically to test (1) the validity of the FMA-UE item content, (2) the reproducibility of its item structure, and (3) functional interpretation of its score.

Validity of the item content

Using Rasch analysis an item-by-item analysis is conducted to test whether each item contributes to the measurement of UE motor ability. This is accomplished by examining the dimensionality of item set and the item difficulty hierarchy.

In order to create a legitimate measure, all items must contribute to the same construct, that is, the assessment must be unidimensional [143]. Unidimensionality means that a single trait or skill accounts for performance of assessment items [153, 154]. Unidimensionality should remain invariant each time the assessment is used because without unidimensionality it is not clear what construct the items define or what the score means [153]. All IRT models assume that the data undergoing analysis are unidimensional [53].

There is no single satisfactory method to assess the unidimensionality of an item set and so researchers often use a variety of techniques to assess it [154-156]. Principal

Components Analysis (PCA), a classical test theory statistical method, is often partnered with Rasch analysis for this purpose. The intent of PCA is to reorganize multivariate data into a limited number of components (factors or dimensions) so that each component captures a substantial amount of the overall variance within the dataset [157]. PCA can be applied to either the original raw score data or the Rasch-derived residuals [154].

PCA produces an index of variance statistic for each factor, an eigenvalue, to describe how well items measure a single construct. Techniques for interpreting the eigenvalue(s) vary. The Kaiser rule [157] is typically employed as the criteria for determining the number of factors to retain. According to the Kaiser rule, factors with eigenvalues > 1.0 are retained, and each is thought to represent a distinct construct. Cattell's scree test is also used to interpret eigenvalues [157, 158]. The scree plot is a graph of all eigenvalues in their decreasing order. The plot looks like the side of a mountain, and "scree" refers to the debris fallen from a mountain and lying at its base. According to the scree test, factors are retained if their data points occur prior to the point at which the "mountain" ends and the "debris" begins. Eigenvalues of "scree" are considered trivial because they are thought to detect "noise" (e.g., measurement error) rather than actual multidimensionality. Only points that markedly deviate from a linear trend are valuable to the analysis [157]. McHorney and Cohen (2000) [159] suggest that if greater than 20% of the variance is accounted for by the first factor in a PCA, unidimensionality is established. However, Cook et al. (2003) [160] argued that this often cited criterion is a "perpetuation of an incorrect interpretation" of an earlier work concerning the stability of item parameter estimates in multidimensional tests rather than suggesting a criterion for establishing unidimensionality. Finlayson and Mallinson (2005)

[161] suggest that a factor must contribute a minimum of 10% of explanatory variance to be considered a unique factor in modeling the data, while Haley and colleagues [162] suggest that meaningful factors are those that each account for 5% of the variance. Parallel analysis provides an alternative, empirically derived criterion [163]. In parallel analysis dimensionality is determined by retaining only the number of eigenvalues from the real data that exceed the eigenvalues estimated from a parallel data set of the same size that is randomly simulated from a Rasch model [162].

Once the number of factors has been identified the meaning of the factors must be determined. PCA typically involves a second step, factor rotation. Factor rotation is a mathematical technique intended to re-express factors in a more simple structure to facilitate their interpretation. A factor is interpreted (i.e., subjectively labeled as per the construct it represents) by determining what trait is shared among the items that load strongly onto the factor [164].

An application of PCA to the Rasch-derived standardized residuals is also intended to elucidate potential multidimensionality in the data. Residuals are the data not explained by the Rasch model. If the data are unidimensional there should be no dimensionality remaining in the residuals once the Rasch model has extracted the dominant factor [154]. The presence of distinct factors within the data is determined by visually inspecting the factor structure plot produced by the analysis [165]. This diagram cross plots the standardized residual loading on each factor against the item difficulty calibration. If the data is unidimensional the graph should show a random distribution of the items. Multidimensionality is evident when items cluster together in “obvious” groupings [166].

Smith (2002) [153] acknowledges that defining unidimensionality as the existence of one (and only one) skill accounting for performance on assessment items is likely too strict of a definition. Smith recognizes the likelihood of many factors (e.g., cognitive, psychological or physical) influencing test-takers' responses during assessment. He states that "unidimensionality will hold" as long as the same factors affect the performance on each item consistently throughout an assessment. Muraki and colleagues (2000) [167] agree, and state that it is difficult to satisfy unidimensionality when IRT models are applied to performance assessments, because performance assessments, by their very nature, require examinees to complete tasks that demonstrate ability to apply skills to a "real life" situations. These situations are complex, thus it is likely that the data is influenced by "construct-irrelevant" variance such as environmental effects, practice effects and examinee fatigue. Murakai argues that unidimensionality is violated to some extent in every dataset, and even more in data from performance assessments. The authors suggest that as long as the violations are not "extreme" IRT analysis appears to be robust.

One way to interpret the apparent presence of more than one factor in data was exemplified in a recent study by Hart and colleagues (2006) [168]. In this study the researchers used factor analysis to examine the dimensionality of the Moss Attention Rating Scale (MARS). The analysis suggested the presence of 3 factors in the data. The authors interpret the results as meaning the instrument measures a single overarching construct, attention, but is comprised of 3 inter-correlated sub-dimensions; initiation; restlessness/distractibility; and sustained/consistent attention. The authors argue that the

sub-dimensions are not unique constructs, but reflect different ways of expressing the overarching construct.

Rasch analysis takes the analysis of unidimensionality one-step farther in the sense that it tests the extent to which persons have responded to assessment items in an expected manner. Using Rasch analysis, the unidimensionality of an item is evaluated with infit and outfit statistics. Under the Rasch model, each person with a given level of UE motor ability is expected to do well (score higher) on relatively easy tasks, and have more difficulty (progressively lower scores) on tasks of increasing difficulty [142]. The model expectations are assessed through item goodness-of-fit statistics. Fit statistics are reported as the mean square (MnSq) of the item residuals, where a residual is the difference between a person's actual response to an assessment item and the response expected by the Rasch model. Consequently the MnSq statistic has an ideal value of 1.0 with departures from this value indicating potential departures from unidimensionality. Fit statistics are also reported in a standardized form (ZSTD), which is the transformation of the mean square of the residuals to a statistic with a z-distribution [153]. Two types of fit statistics are reported. The infit statistic is sensitive to items that depart from model expectations when the item difficulty is of a similar level as person ability. The outfit statistic is sensitive to items that depart from model expectations when the item is much harder or easier than a person's ability level (i.e., outliers). The acceptable criteria for unidimensionality depends on the intended purpose of the measure and the degree of rigor desired. For clinical observations, Wright and Linacre [169] suggest reasonable ranges of MnSq fit values between 0.5 and 1.7 associated with standardized z-values (ZSTD) less than 2.0 for clinical observations. High values indicate that scores are

variant or erratic, suggesting that an item belongs to a construct that is different from that represented by the other items of the instrument. High values could also indicate that the item is being inaccurately scored. A low MnSq value suggests that an item is failing to discriminate individuals with different levels of ability (i.e., with different amounts of motor ability) or that an item is redundant (i.e., other items represent similar motor challenge). Items with high MnSq values represent a greater threat to construct validity.

Item difficulty hierarchy

An assessment intends to measure a skill. Rasch analysis paints a picture of the skill by arranging the assessment items along a hierarchical continuum (an "item map") from "least difficult" to "most difficult". The item difficulty hierarchy is therefore a visual representation of the abstract skill the assessment is intended to measure. Ryall et al. (2003) [170] argues that examining the agreement between the a-priori predicted hierarchical ordering of items and the empirical item order is one way to test the validity of the theoretical construct underlying the assessment. The empirical disordering of one or more items may indicate that the item(s) contain(s) flaws, or that the theoretical construct governing item order is "deficient" [171].

Rasch analysis places person ability and item difficulty on the same linear continuum measured in "logits". A logit is the natural logarithm of the odds of a person being able to perform a particular task to the probability of being able to perform the task [172]. Logits of greater magnitude represent increasing item difficulty (and person ability). Because items and people are measured on the same scale, person ability scores reflect what a person can or cannot do on the basis of item content.

The item hierarchy affords a method to clearly interpret the assessment score which is beneficial for measuring person ability and informing theory [173]. Because the items

are ordered according to their difficulty level, the expected "next step" of the client's progress is clearly visible. The Gross Motor Function Test (a test of motor development) item map presented in Russell et al. (2000) [174] provides an example of how the item map connects a score to a description of behavior. In this example, a GMFM score of "80" is located towards the higher end of the developmental continuum. Because developmental behaviors (i.e., assessment items) are also positioned on the continuum, the score of "80" marks the location at which the child is able to accomplish the behaviors below that level (i.e., easier items) but is not yet capable of accomplishing the behaviors above that level (i.e., harder items). Similar item maps (also termed "keyforms") are used for scoring the Pediatric Evaluation and Disability Inventory [175] and the ABLEHAND [176] questionnaire.

The item hierarchy is useful to ascertain whether an assessment score is interpreted in the way expected. White and Velozo (2002) [145] applied Rasch analysis to the Owestry Low Back Pain Disability Questionnaire to determine if the often used Owestry classification scheme was empirically supported. By matching assessment scores to the item hierarchy, the researchers learned that the Owestry classification categories did not describe actual client behaviors in the way intended by the classification taxonomy. The researchers suggest re-categorization of the response scale to improve the meaningfulness of the score.

Sabari and colleagues (2005) [177] applied Rasch analysis to the Motor Assessment Scale [178]. The researchers found that the Rasch-derived item hierarchy both supported and challenged the underlying conceptual foundations of the assessment. For example, the item hierarchy revealed that overhead reach was less difficult in supine

than while standing, a finding consistent with contemporary motor control principles that movement difficulty is governed by external mechanical factors, e.g., gravitational torque. The item hierarchy also showed discrete, proximal hand movements were of similar difficulty as some functional activities requiring simultaneous control over multiple movements at a variety of joints, a finding that challenged theoretical expectations. Chiu and colleagues (2006) [179] applied Rasch analysis to the Dynamic Gait Index. The hierarchical ordering of locomotor behaviors was consistent with clinical expectations. Penta et al. (1998) [176] utilized Rasch analysis to develop the ABILHAND questionnaire. The item hierarchy "defined" manual ability, a trait previously not well explored in motor control theory. For example, the most difficult items elucidated characteristic behaviors (strength, mental effort, dexterity) that added to an understanding of the construct.

For the construct underlying the FMA-UE, we might expect flexor synergy items to represent easy items along the continuum of UE motor ability, while "hand" items which involve more intricate prehension patterns and isolated joint movements to represent more challenging items across the continuum. Comparing the Rasch-derived item hierarchy to Fugl-Meyer's expected course of recovery will provide an opportunity to test traditional theoretical expectations of recovery.

Reproducibility of the item structure

Rehabilitation researchers are often interested in whether an intervention is effective. To this end, researchers choose an appropriate assessment to score an outcome of interest and compare scores between and/or within groups. Regardless of the statistic used to assess effectiveness (e.g., comparison of group means, correlation test, etc.) it is implicitly assumed that the measurement properties of the assessment are identical each

time the assessment is used. Gluck (2001) [180] points out that this often untested assumption may negatively impact the reliability of an assessment.

Bingenheimer et al. (2005) [181] argues that *measurement equivalence* is necessary for valid interpretation of assessment scores. Specifically, an instrument should display *item* and *scalar* equivalence meaning that an assessment's items should assess the same trait each time it is given (item equivalence), and identical scores should always have the same interpretation (scalar equivalence). The items of an assessment are the operational definition of the trait the assessment is intended to measure, and the score quantifies how much of the trait the test-taker possesses. If an item varies with regards to its meaning or the score varies with regards to its interpretation, the assessment is not an accurate measure of the trait. Bingenheimer argues that without convincing evidence of item and scalar equivalence, there is no basis for comparing scores on an assessment between or within groups.

Differential Item Functioning (DIF) is a statistical procedure in which the items of a test are examined, one at a time, to ascertain whether test-item characteristics are consistent irrespective of test-taker characteristics. The presence of DIF means that a test item measures a unique feature (trait) of one group of people when compared to another group of people, or that the test item was not administered correctly [182].

There are various approaches for examining DIF (see [183] for review), the most commonly used is the Mantel-Haenszel (MH) procedure. MH begins with the assumption that the ratio of correct to incorrect responses is the same between the groups for each item. To test this hypothesis, test-takers are split into two groups (e.g., time 1 and time 2) stratified by ability level (test score). Item difficulties between identical strata of each

group are compared, item-by-item, using the odds ratio statistic. An odds ratio value of 1 means that the item displays no DIF. Based on the magnitude of the value's deviation from 1, items are rated as having negligible, intermediate or large DIF. Proponents of the MH procedure support excluding items showing large DIF, while leaving to clinical judgment decisions regarding the inclusion or exclusion of items with intermediate DIF.

Rasch analysis employs a relatively straightforward DIF procedure in which the item difficulty parameters between two groups (or two time points) are directly compared. This procedure is similar to the above MH procedure, but is independent of sample-dependent person ability (i.e., individual raw scores are not used in the mathematical calculation [184]). The Rasch procedure has reliably detected DIF with small sample sizes ($N < 200$) [183].

In the healthcare literature DIF analyses have been useful to study item equivalence (i.e., to detect potential test item bias) among persons of different cultures and diagnostic groups. For example, using DIF analyses, Azocar and colleagues (2003) [185] learned that Latino-Americans were more likely to receive a higher score on some items of the Beck Depression Inventory than non Latino-Americans, potentially flagging a distinct Latino cultural interpretation of the items. Azocar recommended a reconsideration of the assessment's Spanish translation in order to facilitate cross-cultural comparisons of assessment scores. Tennant and colleagues (2004) [186] found that certain Functional Independence Measure self-care items displayed DIF. The researchers learned that culturally unique ways of bathing and grooming influenced the way the items were administered among European countries. To facilitate cross-country comparisons of scores, the assessment scoring procedure was adjusted for those items. Roorda and

colleagues (2004) [187] found that the hierarchical item ordering of a stair climbing assessment was different for amputees versus non-amputees. The researchers recommend using a shortened assessment, with DIF items removed, if the intention is to compare scores between the two diagnostic groups. Dallmeijer and colleagues (2005) detected DIF in the Functional Independence Measure [188]. When comparing Functional Independence Measure scores among individuals with various neurological diagnoses (stroke, multiple Sclerosis, and TBI) DIF was detected in 7of 11 motor items and 4 of 5 cognitive items. For example, persons with MS found the item "dressing lower body" more difficult than did persons with stroke, and persons with stroke found the item "problem solving" much more difficult than did persons with MS. The authors caution that FIM scores cannot be compared across these patient groups.

Instead of testing the measurement equivalence of an assessment between cultures or diagnostic groups, it is also useful to employ DIF analyses to compare the measurement equivalence of an assessment between testing occasions within a single group. The use of DIF analyses for this purpose has precedence in the literature. Tesio et al. (2003) [182] unexpectedly detected DIF among several FIM mobility items (transfer to a chair, locomotion, and stairs) when admission scores were compared to discharge scores. The items were "easy" for a group of orthopedic patients with mobility restrictions at admission (e.g., no lower extremity weightbearing due to orthopedic precautions). However, the items became "more difficult" for this group at discharge *after* the mobility restrictions were removed. This counterintuitive result led the staff at this facility to examine the manner in which the assessment was scored. They learned that the admission score was based on a prediction of the patient's expected outcome, while

the discharge score was based on actual observation of patient performance. In essence, the assessment was measuring a different construct at patient admission (e.g., ability of the staff to predict patient status) versus patient discharge.

Finlayson and Mallinson (2005) [161] applied DIF analysis to the primary functional outcome measure used in a large longitudinal study of aging and ADL performance in various healthcare settings. Rasch analysis of a 24-item ADL assessment was conducted at three time points over the 13-year study. Although DIF was detected among 4 items, the researchers were unable to discover a specific reason for it. Because statistically significantly differences in item difficulty do not always translate into practical differences in person ability scores [189], the researchers measured ADL ability using an assessment in which the 4 unstable items were removed and compared the results to measurement using the full assessment. The analysis suggested that DIF had only minimal impact on the Rasch-ability measures over time or across settings, leading the researchers to continue using the unadjusted entire assessment.

The presence of DIF can have significant impact on the validity of an assessment for measuring person ability. Linacre's (1994) [190] study provided a classic example of the impact of item instability on the measurement of patient change. Using Rasch analyses, Linacre found that when the 13 motor and 5 cognitive items of the Functional Independence Measure were combined into a single 18-item assessment, the admission and discharge item difficulty calibrations of the items were not equivalent. He suggests that the presence of variant item difficulty measures indicated that a single pattern of disability could not be identified when comparing patient admission scores to discharge scores. That is, the meaning of the score was dependent on the time point (admission

versus discharge) at which the assessment was administered. Linacre demonstrated that the statistical validity of the scale was improved if the motor and cognitive items were separated into sub-scales.

Functional interpretation of the FMA-UE score

As previously noted, one of the primary problems facing uses of the FMA-UE is the functional interpretation of its score. IRT provides the theory and method for linking the FMA-UE to a more functionally oriented assessment [143, 191, 192]. Linking is a method that re-calibrates separate assessments onto a common measurement scale. The theory behind this process has been termed "scale free measurement" [143]. Scale free measurement means that what one measures is independent of the instrument used to make the measurement. Applied to post-stroke UE rehabilitation, scale-free measurement theorizes that an abstract trait (e.g., UE motor ability) can be measured by an "infinite" number of items representing the trait. Any assessment used to measure this trait is simply a subsample of this item pool.

Item pools (termed an "item bank"[193]) are the foundation of Computerized Adaptive Testing. This testing procedure is considered "adaptive" because different test-takers answer different sets of questions depending on their level of skill with regards to the content (construct) being tested. The computer selects question sets from the item pool, which are items that have been previously validated and calibrated. A person's test score is derived from responses to the items administered meaning that test takers are not required to complete the same items to receive comparable scores. Item banking and computerized adaptive testing is widely used in educational research (the Graduate Record Exam is a computerized adaptive test) and has been increasingly applied to health care outcomes research [159, 194].

Items within an item pool are calibrated to a common metric using one of two processes. The first involves combining the items together for a single analysis that concurrently calibrates difficulty parameters. A variation of this method is to estimate difficulty parameters for one set of items first, then use these difficulty estimates to fix (anchor) this set of items in a combined analysis with the remaining items [193]. Hanson et al. (2002) [195] suggests that, at least for 2PL analyses, concurrent calibration realizes less error in item parameter calculation than separate calibrations. Kim and Cohen et al. (1998) [196] suggest that the two methods yield nearly identical results in situations where a large percentage of items are common to the participant sample.

Linking two existing assessments is another way to create a common item pool. If two assessments measure the same underlying trait, then the items of each assessment are derived from the common item pool and therefore the assessments can be mathematically linked. Linking establishes the relationship between different assessments, therefore allowing direct comparison of items and scores [167]. Fisher (1995) [192] linked two assessments of physical function to a common scale with Rasch analysis. Thirteen motor FIM™ items and Patient Evaluation and Conference System (PECS) items were co-calibrated to single scale that allowed a patient's score on one assessment to be directly translated to a score on the other assessment. Fisher suggested that the assessment scores could be reported in "rehabits", a unit of measurement representing the common measurement scale to which the assessments were linked. Costner and colleagues (2004) [148] linked five existing widely used ADL and IADL assessments (FIM™, Minimum Data Set (MDS), MDS-post acute care, OASIS and Physical Functioning –10) for the purpose of tracking recovery across the continuum of care. The conceptual basis

underlying the link was that the assessments each contain daily living items sharing a common requirement for upper limb and hand skills. The researchers concurrently calibrated the items from the existing assessments, retaining 62 items in a unidimensional item pool.

Fisher and Eubanks (1997) [197] concurrently calibrated the physical functioning subscales of the Medical Outcomes Study Short Form 36 physical functioning subscale (PF-10) and the Louisiana State University Health Status Instruments Physical Functioning Scale (PFS) to a common metric using Rasch analysis, demonstrating that the method was feasible. Segal and colleagues (1997) [198] used Rasch methodology to link the PF-10 and Functional Independence Measure motor items. Six PF-10 and 4 FIM™ items were co-calibrated to form a new assessment for post-stroke outpatients. As separate assessments floor effects in this population hampered the PF-10 and ceiling effects hampered the FIM™. However, the linked assessment demonstrated excellent measurement properties for measuring long term physical functioning in this population.

The ICF provides a conceptual framework that identifies UE motor behaviors relevant to a single domain (i.e., body function and/or activity performance) and specifies a relationship between domains. This operationalizes health care dimensions and provides a framework for rehabilitation outcome measurement within and across dimensions [199]. Linking a post-stroke UE motor assessment from the body function domain to an assessment from the activity domain sharing similar UE motor behaviors is a way to test whether the domains are distinct or related constructs in post-stroke recovery. Moreover, linking is a way to directly investigate the functional relevance of

reduced body function impairment because linking connects the assessments to a common metric.

The Wolf Motor Function Test (WMFT) [26, 200] is a logical choice for linking to the FMA-UE. Both the FMA-UE and the WMFT measure UE motor ability: the FMA-UE is a measure of UE body function impairment, while the WMFT is a measure of UE activity performance restriction. The WMFT is an assessment of post-stroke UE motor function [201]. It is intended to measure a "wide range of functional UE tasks" in individuals with mild to moderate stroke-related UE motor impairment [201]. Its items are arranged in a sequential order according to joints involved (shoulder to hand) and level of difficulty (gross movements to fine motor movements) [26]. The first seven items involve simple limb movements without functional endpoint (e.g., *extend elbow to the side on a table top, lift hand to table*), and the remaining eight items involve performance of common tasks (e.g., *lift soda can, flip cards*). The WMFT has been shown to have high interrater reliability (intraclass correlation coefficients, ≥ 0.88), internal consistency (Cronbach $\alpha \geq 0.86$), and test-retest reliability ($r \geq 0.90$) [26, 201].

Placing the FMA-UE and WMFT items on the same "ruler" as client ability allows the individual's Rasch ability score to be directly related to his/her probability of performing a particular set of movements or tasks. This permits one to describe a person's motor ability in terms of specific movements or tasks that he can/cannot perform. Furthermore, changes in score can be directly linked to acquisition of more difficult movements or tasks. By linking the FMA-UE and WMFT in this manner one should be able to convert improvement scores into performance of specific UE movements and tasks (e.g., moving from elbow extension to grasping a soda can) thereby

identifying a critical threshold at which impairment reduction translates to functional ability. Translating improvements in UE ability from impairment to function is the foundation of documenting the clinical effectiveness of restorative rehabilitation.

Criticisms of the Rash Measurement Model

Sample size

The purpose of a sample size justification is to predict how many subjects are needed to satisfactorily answer the research objective. In traditional statistics, a power analysis guides the researcher in determining the appropriate sample size for the objectives of the study. With IRT analyses sample size guidelines are not as straightforward. Linacre (1994) [202] suggests that as few as 50 subjects may be an appropriate sample size for a valid analysis using the one parameter Rasch model, however Hambleton (1989) [203] suggests that at least 200 subjects are necessary for Rasch analyses. Embretson and Reise (2000) [204] indicate that the two-parameter model may require 250 to 500 participants while Hambleton indicates that a minimum of 500 subjects are necessary. Hambleton also suggests a minimum sample size of 1000 subjects if the researcher intends to use a three-parameter IRT model.

The purpose of the study affects the sample size recommendation. According to Reeve and Fayers (2005) [205], if the aim of a study is to examine the properties of an existing assessment, one does not need a large sample size to describe it clearly. However if the study purpose is to calibrate items for test development then larger (over 500 subjects) sample sizes are needed. Lai et al. (2003) [206] recommends that test equating requires a sample size of 400 for the Rasch model and 1,500 for a three-parameter model.

There are very few empirical studies influencing IRT sample size recommendations. Holman's (2003) [207] simulation study results suggests 450, 90, and 40 participants are needed in each arm of a randomized trial in order to detect an effect size of 0.2, 0.5, or 0.8 respectively with a significance of 0.05 and a power of 80% using a two parameter IRT analysis. In comparison, a t-test powered to detect the same effects would require 394, 64, or 26 patients per arm of the study. Holman also indicates that the number of "informative" items in the assessment affects the sample size. That is, more items with more precise information require lower sample size.

Obviously these sample size guidelines may make using IRT methods prohibitive for many health care researchers. In Rasch measurement, the researcher seeks to obtain stable person, item and rating scale estimates. A stable parameter is one that has low standard error. Linacre indicated that ± 0.3 logits is the best that can be expected for most variables and that the sample size needed to have 99% confidence that no item calibration is more than 1 logit away from its stable value is 50 [202]. In a Monte Carlo simulation study, Wang and Chen (2005) [208] found that with sample sizes of 5 – 20 individuals, item parameters were unstable at either the higher or lower end of the item difficulty hierarchy in short tests (10 to 20 items). However, the parameters were stable with longer tests and the same sample size. The researchers conclude that longer assessments contain more information and thus stabilize item parameter generation in a small sample size.

Should data fit the model or visa versa?

The Rasch model is criticized because it assumes equal item discrimination. Item discrimination refers to an item's ability to differentiate between test takers of different levels of ability. As stated, the two-parameter IRT model contains an item discrimination parameter while the one parameter Rasch model does not. The Rasch model constrains

the item discrimination value at “1”. According to McHorney [209] it is critically important to examine the item discrimination capabilities of assessment items. It is possible that equally valid items will discriminate in various ways.

To this end, the Rasch model software calculates an item discrimination index [148]. However, in contrast to the 2-parameter IRT model, this index does not enter into the original person scoring. The index describes whether items are close to or depart from the ideal discrimination value of “1”. In other words, the index reveals how consistent items are with the unidimensional expectations of the Rasch model. Low values indicate that an item fails to discriminate as well as other items; high values may indicate that the item is idiosyncratic in some way. Misfitting items are reexamined for redesign if necessary. The Rasch model uncovers possible anomalies in item design thereby enabling further investigation of and modifications to the test [210].

The underlying conflict between the Rasch model and other IRT models has to do with views of how data provide information [211]. In short: should a measurement model explain data, or should a measurement model disclose information perhaps hidden in data? Andrich (2004) [211] frames this controversy using the language of mathematician-philosopher Thomas Kuhn as a clash of “incompatible paradigms”.

Kuhn stated that the purpose of measurement is to disclose anomalies in data that cannot be explained by the theory underlying the measurement model with which the data is analyzed. With this approach one is able to ask questions as to why data does not fit a given theory thereby potentially learning new information that contributes to advancement of the theory. Kuhn’s view is in stark contrast to traditional measurement

theory that asserts data should fit a model so as to support the theory underlying the model. This issue is notably evident in controversies surrounding the Rasch model [211].

According to CTT and non-Rasch IRT models, data is best interpreted according to a mathematical model found to explain the largest amount of variance in the data. Models are tried and abandoned depending on how well the model explains the data. For example, a recent article eloquently advocates IRT methods to measure clinical cognitive change [212]. The researcher describes how various models were tried and discarded until he settled on a 2-parameter IRT model because it provided a "better fit" than the other models. Kirsci et al. (2006) [213] employed significance testing of chi-square statistics as a method for comparing one and two parameter IRT models as best fitting the data. The point is that the a-priori assumption in the above paradigms was that a mathematical model should fit the data.

The Rasch model has the inverse assumption. That is, data should be viewed with the assumption that it either fits or does not fit the model. If it does not fit, then, rather than abandoning the model in favor of another, the data is examined for "anomalies" [211]. According to Andrich [211], the process of examining misfitting data (either items or persons) "uncovers" information that would have been hidden otherwise. New information may lead to advances in theory. As Wright states: "Misfitting items can be redesigned. Misfitting populations can be reassessed" [214].

Conclusion

There exists an interdependent link between theory and measurement. The theory guides the design of the measurement tool. In turn, the information gathered from the measurement tool advances the theory. This concept was recently articulated by Dennis Tate in the John Stanley Memorial Lecture to the American Congress of Rehabilitation

Medicine [215]. He stated “the process by which we measure rehabilitation domains and test our theories about reality shapes what constitutes rehabilitation knowledge.”

As recent advances in neuroscience and rehabilitation science have increased knowledge about the potential for upper limb motor recovery after stroke, recognition of the need to adequately measure impairment and the functional implications of impairment has also increased. Stroke rehabilitation professionals face an urgent need to assure that assessments used to establish the efficacy and effectiveness of translational intervention are of the highest quality for measuring client ability, change with intervention, informing the clinical decision-making process and providing valuable quantitative information about impairment and recovery.

The FMA-UE is considered the “gold standard” clinical assessment of post-stroke UE motor impairment, however its item-level measurement properties have not been well studied with contemporary measurement models, e.g., item response theory. Several deficiencies potentially exist in FMA-UE item content and structure that could impair its use as an outcome measure; (1) some items may not be consistent with current motor control theoretical expectations of post-stroke impairment and recovery, (2) the reproducibility of the construct “recovery” illustrated by its item hierarchy has not been established across subjects of various ability levels, nor within subjects over time, and (3) in its current format, the FMA-UE score has no clear functional meaning. The aims of the subsequent studies address these deficiencies and serve the purpose of exploring and potentially enhancing the measurement properties of this important and long-standing stroke rehabilitation measurement tool.

Table 1: The Fugl-Meyer Upper Extremity Assessment

FMA-UE Item Number	FMA-UE Item Description
	Shoulder/Elbow/Forearm
Item 1	Biceps reflex elicited
Item 2	Triceps reflex elicited
Item 3	Scapular elevation
Item 4	Scapular retraction
Item 5	Shoulder abduction
Item 6	Shoulder external rotation
Item 7	Elbow flexion
Item 8	Forearm supination
Item 9	Shoulder adduction with internal rotation
Item 10	Elbow extension
Item 11	Forearm pronation
Item 12	Hand to lumbar spine
Item 13	Shoulder flexion to 90° degrees with elbow extended
Item 14	Pronation–supination of forearm with elbow at 90°
Item 15	Shoulder abduction to 90° with elbow extended
Item 16	Shoulder flexion to 90°-180° with elbow extended
Item 17	Pronation–supination of forearm with elbow extended
Item 18	Normal reflex activity
	Wrist
Item 19	Wrist stable with elbow at 90°
Item 20	Wrist flexion-extension with elbow at 90°
Item 21	Wrist stable with elbow extended and shoulder at 30°
Item 22	Wrist flexion-extension with elbow extended and shoulder at 30°
Item 23	Wrist circumduction
	Hand
Item 24	Finger mass flexion
Item 25	Finger mass extension
Item 26	Hook grasp (MPs extended, PIPs and DIPs flexed)
Item 27	Lateral prehension (thumb adduction to hold paper)
Item 28	Palmar pinch (thumb to index finger to hold pencil)
Item 29	Cylindrical grasp (hold small can)
Item 30	Spherical grasp (hold tennis ball at fingertips)
	Coordination/Speed
Item 31	Movement without tremor
Item 32	Movement without dysmetria
Item 33	Movement with normal speed

Table 2: Descriptions of the traditional upper extremity limb synergies

Flexor Synergy	**Extensor Synergy**
Shoulder retraction	Shoulder adduction with internal rotation
Shoulder elevation	Elbow extension
Shoulder abduction to 90 degrees	Forearm pronation
Shoulder external rotation	
Elbow flexion	
Forearm supination	

CHAPTER 2
USING THE RASCH MEASUREMENT MODEL TO EXAMINE POST-STROKE UPPER EXTREMITY RECOVERY

The characterization of post-stroke upper extremity (UE) motor recovery has long been of the focus of rehabilitation clinicians and researchers. An assessment tool that quantifies UE motor impairment can identify the amount of UE motor recovery, i.e., locate a patient along the motor recovery progression. Accurate measurement of UE motor impairment is important to capture the effects of translational interventions, explore optimal training parameters of existing interventions, and to predict future UE motor function. There is a pressing need to ascertain whether assessment tools commonly used in post-stroke UE motor recovery research are accurately quantifying impairment and characterizing recovery.

The Fugl-Meyer Upper Extremity Assessment (FMA-UE) [15] is the most widely used clinical assessment of post-stroke upper extremity motor impairment [16]. It has excellent inter-rater reliability (Pearson Product-Moment Correlations r = 0.97-0.99 [17]) and test-retest reliability (Intraclass correlation coefficients 0.94 – 0.99 [18]). It has been used as the standard from which to establish the validity of other commonly used tests of upper extremity motor function such as the Wolf Motor Function Test and Ashworth Scale [20-26]. Researchers consistently use the FMA-UE as a descriptor of functional status following stroke [4, 20, 27]. To this end, the FMA-UE scores have been used to stratify research study participants into categories of stroke severity [28], predict long-term functional participation [29], and to describe arm motor impairment [30-32], arm

movement quality [21], and residual arm motor function [33-35]. Furthermore, the FMA-UE is the primary criterion for evaluating the success of novel upper extremity interventions such as rhythmic bilateral movement training [36, 37], neuromuscular stimulation [38], electromyographic triggered electrical stimulation [39, 40], Botox [41], home based exercise [30], community based exercise [42], robot aided therapy [43, 44], virtual reality [45], imagery [46], forced use [47], and modified constraint induced movement therapy [48-51]. Because of the importance of this assessment in post-stroke rehabilitation research, it is critical to examine, and perhaps improve, the quality of its measurement properties.

The measurement properties of an assessment tool are usually studied with traditional psychometric methods such as reliability, validity, and responsiveness. The item response theory (IRT) measurement framework offers another way to examine an assessment's measurement properties. Stroke rehabilitation researchers have increasingly used IRT methods, e.g. Rasch analysis♦, to develop new assessments [176] or redesign existing assessments [177]. Rasch analysis offers the advantage of examining an instrument at the item level rather than as a whole. With this approach one can explore item content (i.e., does a specific item contribute to the measurement of an intended construct?), and item structure (i.e., does the item difficulty hierarchy progress from "less of" to "more of" the intended trait?). Although the measurement properties of the FMA-

♦ Some authors use the terms Item Response Theory (IRT) and Rasch analysis interchangeably. However, the measurement models are different. The Rasch model requires that items have equal discrimination, whereas IRT models include a parameter for items to have different discriminations. The merits of each model are the subject of ongoing debates. For the purposes of this article the simpler of the models, the Rasch model, is used.

UE have been extensively studied using traditional methods [15, 17, 18, 20, 24, 54-56], presently there are no published studies applying IRT approaches to the FMA-UE.

McDowell and Newell (1996) suggest that health care instruments be founded on a specific theoretical framework [67]. Fugl-Meyer et al.'s (1975) [15] theoretical framework broadly reflects the observations of Twitchell (1951) [75] and Brunnstrom (1966) [76] regarding the stages of post-stroke UE sensorimotor recovery. Fugl-Meyer stated:

> The form has been constructed following the hypothesis that the restoration of motor function in hemiplegic patients follows a definable stepwise course. Thus for a patient with hemiparalysis, recurrence of reflexes always precedes volitional motor action. Thereafter through initial dependence on synergies, the active motion will become successively less dependent upon the primitive reflexes and reactions and finally complete voluntary motor function with normal muscle reflexes may be regained (p. 14).

Fugl-Meyer chose items to exemplify the construct of motor recovery including reflex items and voluntary movement items. Moreover, because the assessment is intended to measure *recovery*, Fugl-Meyer arranged the items, from easy to hard to map the recovery process. It is clear that Fugl-Meyer expected motor recovery to proceed in a reflexive-to-voluntary, and synergy-to-isolated progression. We may therefore expect the FMA-UE items to be ordered according to this hypothesis.

Recent advances in movement science challenge the "definable stepwise course of recovery" upon which the FMA-UE is based. Contemporary views of the central nervous system suggest that the reflex items measure a different behavior than other items. A tendon tap reveals the integrity of a relatively simple spinal neural control circuit, while a goal directed voluntary movement reveals the integrity of a more complex supraspinal neural network [80]. Furthermore, recent studies of UE motor control suggest that coordinated arm motion and patterns of muscle recruitment are influenced by task-

specific mechanics [98, 216-218] and environmental demands [100]. For example, Michaelson et al. (2001) [98] suggested that trunk restraint "normalized" post-stroke reaching kinematics. Wu and Trombly (2000) [100] found that characteristics of the task-specific target influenced reaching kinematics in both healthy individuals and individuals with stroke. While Fugl-Meyer's recovery model may provide a neurophysiological description of post-stroke UE motor impairment, task-specific interactions of neural, biomechanical and contextual factors may also influence the course of post-stroke UE motor recovery.

The purpose of this study is to investigate the dimensionality and construct validity of the FMA-UE using Rasch analysis. The aims of this study are to (1) determine if all items of the FMA-UE contribute to the measurement of a single construct (dimensionality), and (2) determine if the items are ordered according to Fugl-Meyer's expected "stepwise" sequence (construct validity).

Methods

Dimensionality

In order to create a legitimate measure, all items must contribute to the same construct [143]. Using Rasch analysis, the extent to which items contribute to a unidimensional construct is evaluated employing infit and outfit statistics. The infit statistic is most sensitive to ratings on items that are closely matched to subjects' ability, the outfit statistic is most sensitive to ratings on items that are much easier or much harder than subjects' ability [142]. Fit statistics are reported as mean square standardized residuals (MnSq) produced for each item of the instrument. MnSq represents observed variance divided by expected variance [172]. 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. For surveys using ordinal rating scales, Wright and Linacre [169] suggest reasonable ranges of MnSq fit values between 0.5 and 1.7 associated with standardized Z values (ZSTD) less than 2.0. High values indicate that scores are variant or erratic, suggesting that an item belongs to a construct that is different from that represented by the other items of the instrument. High values could also indicate that the item is being inaccurately scored. A low MnSq value suggests that an item is failing to discriminate individuals with different levels of ability (i.e., with different amounts of motor ability) or that an item is redundant (i.e., other items represent similar motor challenge). In this analysis we focused on items with high MnSq values, because they represent a greater threat to construct validity.

Fit statistics alone are inadequate to determine unidimensionality [154-156]. A principal components analysis (PCA) was performed using the SAS Factor procedure (SAS v 8.2) with oblique rotation. The intent of this PCA was to reorganize the multivariate data (participants' FMA-UE item ratings) into a limited number of components (factors) so that each component captures a substantial amount of the overall variance within the dataset. That is, the purpose of the PCA was to detect "obvious" factors within the FMA-UE. The PCA was carried out in two steps: an initial analysis to examine the dimensionality of the assessment, and an oblique rotation to examine if traditional subscale divisions were empirically supported.

We hypothesized that the assessment items illustrated a single factor, "UE motor ability". To test our hypothesis we examined PCA eigenvalues and factor loading statistics. We employed the Kaiser rule [157] as the criteria for retaining principal components, i.e., we examined only the components with an eigenvalue > 1.0. We

expected one eigenvalue would explain the majority of the variance in the data. Furthermore we expected that the majority of FMA-UE assessment items would correlate with a single principal component as evidenced by factor loadings on the first component > 0.40 [219].

Fugl-Meyer suggested dividing the assessment into four subscales: *arm* (items 1 – 18), *wrist* (items 19 - 23), *hand* (items 24 - 30), and *coordination/speed* (items 31-33) [15]. Each of Fugl-Meyer's subscales may represent a unique motor control construct. To examine if there was empirical evidence for this subscale division, we executed a second step in the PCA analysis. We performed an oblique rotation (obvarimax) of four principal components. We chose an oblique rotation because we anticipated inter-factor correlations, i.e., movements of the hand were expected to correlate with movements of the arm. Factor rotation is a mathematical technique in which the multi-dimensional axes of initial PCA factor loadings (the number of axes matches the number of retained factors) are rotated to model the data. The rotation re-expresses the factors as a less complex structure to facilitate interpretation [164]. This procedure produces a factor matrix with factor loading values (equivalent to regression coefficients) indicating the "weight" given to the item when interpreting the factor [157]. We considered factor loadings meaningful if the absolute value exceeded 0.40 [219]. We interpreted the factors (i.e., we labeled the factor as per the construct it represents) by subjectively determining what trait was shared among the items loading strongly onto that factor. Because the purpose of our factor rotation was to explore whether Fugl-Meyer's suggested subscale pattern was present in our data, we examined the groups of items comprising a factor to determine if they matched Fugl-Meyer's suggested item groupings.

Construct Validity

The foundation of "objective" measurement is to connect the numbers produced by an instrument with its content [141]. Through Rasch measurement, items represent "difficulty" markings (calibrated in log-equivalent units called "logits") along the continuum of a construct. For example, for the construct underlying the FMA-UE, we may expect flexor synergy items to represent easy items along the continuum of motor ability, while hand items which involve more intricate prehension patterns to represent more challenging items across the continuum. For purposes of this study, we postulated that the item hierarchy would illustrate the trait "UE motor ability". Analysis of data from a cross section of individuals with different abilities post-stroke may provide an initial description of the hierarchy.

By placing item difficulties and person abilities on the same continuum, Rasch analysis can be used to match the difficulty of the assessment items to the ability of the sample tested. In general, the analysis will tell us what items are easy enough for individuals with "poor" arm motor skills (e.g. individuals who have little arm movement) and what items are challenging enough for individuals with "good" arm motor skill (e.g. individuals who have fine-motor coordination). The placement of item difficulty and person ability on the same continuum also provides information on how *specific* individuals are expected to respond to particular items. For example, if a person receives a logit measure of 6.0 and the item *shoulder flexion to 90 degrees with elbow extended* also receives a logit measure of 6.0, then it would be expected that that person would have a 50% probability of being able to perform this movement successfully. This individual would be expected to have a *greater* than 50% probability of being able to accomplish easier items such as those representing the "flexor synergy" items.

Furthermore, he/she would be expected have *less* than 50% probability of being able to accomplish more difficult types of movement, such as *palmar prehension* (grasping a pencil with the pads of the thumb and index finger).

Results

Participants

A secondary analysis was performed on data pooled from two studies. The first dataset consisted of 100 persons enrolled in a randomized clinical trial of therapeutic exercise [220]. FMA-UE data was collected prior to the exercise intervention by therapists trained in standard administration of the assessment. Persons with stroke were included in this study if they met the following inclusion criteria; "(1) Stroke within 30 to 50 days, (2) ability to ambulate 25 feet independently, (3) mild to moderate stroke deficits defined by a total Fugl-Meyer score of 27-90 (upper and lower extremities) and Orpington Prognostic Scale score of 2.0 to 5.2, (4) palpable wrist extension on the involved side, and (5) Folstein Mini-Mental Status exam score of greater than 16" [220]. The second dataset consisted of 459 individuals enrolled in the Kansas City Stroke Study [221]. Participants were included if they met the following criteria; (1) stroke onset within 0 – 14 days prior to enrollment, (2) 18 years of age or older, (3) ischemic stroke as diagnosed by physician, (4) lived in the community prior to stroke onset, and (5) able to participate in baseline testing. Study staff trained in the administration of the FMA-UE evaluated patients. Pooling of these data resulted in a 559-person dataset. However, because lesion location may effect neural reorganization and clinically observed patterns of UE recovery [222], subjects with non-cortical (e.g., brainstem or cerebellum) stroke were not selected from the pooled dataset. This resulted in a final dataset of 512 participants. Characteristics of the 512-person sample are presented in Table 3.

Fit Statistics

Item Fit Statistics from the initial analysis of the entire 33-item FMA-UE are presented in Table 4. Two items (*biceps reflex and triceps reflex)*, shaded in gray, show infit values that are beyond the acceptable ranges described by Wright and Linacre (1994) [169]. These items have infit statistics that are above 1.7 with a standard deviation greater than 2.0. These items also showed abnormally high outfit statistics. In addition, the two items showed relatively low score correlations (0.36 and 0.26). Infit and outfit statistics are within acceptable ranges for all other items.

Principal Components Analysis

The PCA retained 4 factors having eigenvalues > 1.0 (Kaiser rule [157]). See Table 5. A single component explained 68% of the variance in the data. Three other factors explained an additional 5%, 4%, and 3% of the variance. Table 6 displays the factor loadings onto the first principal component. Thirty of 33 items loaded strongly onto the first component (factor loading values ranging from 0.75 – 0.91). Three items *(biceps reflex, triceps reflex, normal reflex activity)* had a poor to moderate loadings (r = 0.14 – 0.53) with this factor.

Rotated factor loadings for the 4 components with eigenvalues > 1 are presented in Table 7. Primary loadings (factor loadings with an absolute value > 0.40) for each component are shaded in gray. The values are sorted into descending order of magnitude and grouped into columns to facilitate interpretation of the item groupings. Several items do not load strongly onto any factor. For example, *pronation-supination with elbow at 90*, and *movement without dysmetria* do not display any factor loading greater than 0.40. Five items "cross-load" onto more than one factor. For example, *wrist flexion-extension with elbow extended* and *wrist stable with elbow extended* each load onto factor 2 (0.41

and 0.46) and factor 3 (0.54 and 0.52). Items with no factor loadings and those that cross-load are grouped near the bottom of the chart and marked with a double asterisk.

Examination of the shared traits among items that load onto each factor may illustrate the motor behavior represented by each factor. For example the items, *elbow flexion, shoulder abduction, scapular elevation, scapular retraction, shoulder external rotation*, and *forearm supination* load onto the first factor. In addition, this factor contains *shoulder adduction with internal rotation* and *forearm pronation*. These items are consistent with the traditional description of the post-stroke UE flexor and extensor synergies as described by Twitchell and Brunnstrom. Therefore this factor may pick up a "synergy trait" underlying the FMA-UE. The grasp and prehension items load onto the second factor. Therefore this factor may detect a "hand trait" evident in the data. This item grouping roughly matches the hand subscale division suggested by Fugl-Meyer. Items requiring movements away from the body using elbow extension (e.g., *shoulder flexion to 180 with elbow extended, pronation-supination with elbow extended*) load strongly onto the third factor. However, it is difficult to interpret this factor because it contains five cross-loaded items. Finally, *triceps reflex* and *biceps reflex* load onto the fourth factor, perhaps detecting a "reflex" trait. Although the items located in factors 1 and 2 are roughly similar to Fugl-Meyer's arm and hand subscales, the other item groupings do not match Fugl-Meyer's FMA-UE subscale suggestions.

Although the PCA suggests that four motor control constructs (four eigenvalues > 1) may be present, the majority of the variance in the data is explained by a single principal component. However, poor factor loadings onto the first principal component, and abnormal fit statistics suggest that the reflex items do not fit with the intended

measurement construct of the assessment; therefore these items were removed from subsequent Rasch analysis. Fit statistics, the large first eigenvalue and strong principal component factor loadings support the unidimensionality of the remaining items. Because the intent of our study is to look at the relationship of items across the entire upper limb we chose not to subdivide the assessment.

Item fit statistics the revised 30-item FMA-UE (reflex items removed) are presented in Table 8. All 30 items fit the unidimensional assumptions of the Rasch model. The revised 30-item FMA-UE showed good internal consistency. The person reliability index, analogous to coefficient alpha, was 0.96. The instrument divided the sample into 7 statistically significant strata (Number of Strata =[4(Gp = 4.67) + 1]/3, where GP= person separation [223]) indicating a good person to item match.

Item Hierarchy

Table 9 presents items in order of decreasing challenge. FMA-UE items at the bottom of the center column indicate the least challenging items; those at the top represent most challenging items. *Elbow flexion (measure = -1.76 ± 0.13 logits)* and *shoulder adduction with internal rotation (measure = -1.56 ± 0.12 logits)* were the easiest items for this sample to perform, and *wrist circumduction (measure = 1.67 ± 0.10 logits), and hook grasp (measure = 1.33 ± 0.10 logits)* were the most difficult items for this sample to perform. Five of the nine "flexor synergy" and "extensor synergy" items (items 3 through 11) group towards the easy end of the hierarchy, while some flexor and extensor synergy items span the item difficulty hierarchy. For example, *elbow flexion* is the easiest item in the hierarchy (–1.76 ± 0.13 logits), *scapular elevation* is slightly more

difficult (–1.40 ± 0.12 logits), and *forearm supination* is moderately difficult (0.65 ± 0.10 logits).

Movements requiring combinations of shoulder flexion and elbow extension are much more difficult than movements demanding no shoulder motion with the elbow positioned at 90 degrees. For example, p*ronation-supination with the elbow at 90 degrees*, *wrist flexion-extension with the elbow at 90 degrees*, and *wrist stable with the elbow at 90 degrees* are among the moderately difficult items. *Shoulder flexion to 180° with the elbow extended*, *wrist flexion-extension with the elbow extended*, and *pronation-supination with the elbow extended* are among the most difficult items.

In addition, although the majority of the "hand" items (items 24 through 30) group at the "difficult" end of the hierarchy, some of these items also span the item difficulty hierarchy. For example, h*ook grasp* (1.33 ± 0.10 logits), *spherical grasp* (1.25 ± 0.10 logits), and *lateral prehension* (1.08 ± 0.10 logits) are among the most difficult items, while *finger mass flexion* (-1.44 ± 0.12 logits) is among the easiest items and *palmar prehension* (0.06 ± 0.10 logits) is calibrated near the middle of the scale. Surprisingly, *finger mass extension* (-1.25 ± 0.12 logits) calibrates as an easy item for this sample to perform (see discussion).

Keyforms

While the above results show the overall hierarchical pattern for the sample under study, a critical question is the consistency of this pattern across individual subjects. Similar to the fit statistics that are produced for items are fit statistics for persons. Ninety-eight percent of the sample showed acceptable infit statistics (MnSq <1.7 and ZStd <2.0), suggesting that individual subjects in the sample are responding similarly to the item

difficulty hierarchy. Figure 1 demonstrates the item difficulty hierarchy relative to the scoring pattern of three clients representing different UE abilities. Items are listed in terms of increasing difficulty level (i.e., *elbow flexion* being the easiest item at the bottom of each panel, *wrist circumduction* representing the hardest item at the top of each panel). The 3-point rating scales for each item are presented to the left of each panel. As the item difficulty increases, the rating scale stair steps to the right. The bottom scale represents item difficulty. For example, for the item *elbow flexion*, a rating of "1" is at a difficulty of approximately –2.0 logits. This bottom scale also reflects person ability (i.e., the solid vertical line crossing the scale represents the individual's person ability measure, the dotted vertical lines represent the 95% confidence interval around this ability measure). Circled numbers represent the actual ratings for the individual. The triangle shapes represent ratings that are statistically erratic (p=0.05).

A person of low ability (-1.25 logits) is represented on the top left panel (Person A), of moderate ability (0.20 logits) is represented on the top right panel (Person B), and of high ability (3.88 logits) (Person C) is represented on the bottom left panel. The individual of low ability (Person A) has a tendency to receive ratings of "2" (faultless performance) and "1" (partial performance) on the easy items (*elbow flexion, shoulder adduction with internal rotation, finger mass flexion*), and receive ratings of "0" (unable to perform) on the most difficult items (*shoulder flexion to 90 degrees with elbow extended, hook grasp, and wrist circumduction*). The individual of moderate ability (Person B) has a tendency to receive ratings of "2" (faultless performance) on the easy items, and receive ratings of "1" (partial performance) on the moderately difficult items (*scapular retraction, shoulder external rotation*), and receives ratings of "1" or "0"

(unable to perform) on the most difficult items. The individual of high ability (Person C) has a tendency to receive ratings of "2" (faultless performance) on the easy items, and the same the moderately difficult items. However this person receives ratings of "1" (partial performance) on the most difficult items (*shoulder flexion to 180 degrees with elbow extended, wrist circumduction*). In general, while these individuals are of different abilities their scoring pattern is the same; all three score higher on easier items and lower on harder items. The pattern retains its structure when measuring a low ability, moderate ability, and high ability person.

There are exceptions to this pattern as evidenced by the ratings denoted with a triangle. Person A with lower ability receives a higher than expected rating of "2" on a moderately challenging item (*scapular retraction*). Person C of high ability receives a lower than expected rating of "1" on a moderately difficult item (lateral prehension). Overall, 12 out of 512 people (2%) showed erratic scores based on infit statistics. Among these 12 individuals, only 2-5 items were statistically erratic.

Discussion

The purpose of this study was to investigate the dimensionality and construct validity of the FMA-UE. Using item response theory statistical techniques we determined that the items of the FMA-UE contribute to the measurement of multiple constructs and demonstrate a difficulty order that does not reflect with the item order proposed by Fugl-Meyer. Furthermore, we determined that this item difficulty order remains consistent, independent of person ability. The above results will be discussed relative to traditional and contemporary motor control theories.

Dimensionality

The initial finding from both the Rasch Analysis and the more traditional PCA revealed that the reflex items empirically appear to be disconnected from the remainder of the FMA-UE. While, according to Fugl-Meyer's suggested "stepwise" course of recovery, it would be expected that persons of little to no UE motor ability would be the only individuals without reflex behavior, the high infit and outfit statistics indicate that at least some individuals of high ability also show this behavioral response. Multiple statistical findings (i.e. low PCA weights, low correlations and high fit statistics) suggest that the reflex items are contributing little explanatory variance for the FMA-UE and support our decision to remove these items from further analysis.

The three reflex items are exceptional to the assessment in two ways. First, the items evaluate a passive response rather than active motion. Specifically, the behavior of the phasic stretch reflex is assessed through tendon taps of the biceps (items 1 and 18), triceps (items 2 and 18) or finger flexors (item 18). Secondly, the rating scales for the reflex items are different than the rest of the assessment. For example, items 1 and 2 utilize a two-point scale (*0 = No reflex activity elicited*, and *2 = Reflex activity elicited*), and item 18 utilizes a uniquely defined three-point scale (*0 = At least two of three phasic reflexes are hyperactive*, or *1 = One reflex is hyperactive or two reflexes are lively* or 2 = *No more than one reflex is lively and none are hyperactive).* These procedural and scoring differences from that of the majority of the assessment may cause the items to misfit and correlate as a separate factor.

While the above statistical findings may be a product of administration procedures (passive vs. active) or the rating scale structure, they also may indicate that the FMA-UE reflex items reflect a neurological mechanism that is different than that of the other

assessment items. A relatively simple spinal neural circuit governs the tendon tap muscle contraction, while a complex supraspinal and spinal neural network governs voluntary goal-directed movement [80]. Rather than a "hardwired" precursor to voluntary movement (as suggested by traditional motor control theory), stretch reflexes are important adjuncts of voluntary motor control [80, 82, 83] and the magnitude of the stretch reflex is amendable to operant training [86, 87]. Given that the FMA-UE is an assessment of voluntary movement ability, our results suggest the three reflex FMA-UE items diminish the construct validity of the instrument. As suggested by Gladstone et al. (2002), the reflex items may confound interpretation of the FMA-UE total score [55].

Item Hierarchy

Fugl-Meyer attributed the assessment structure and item choice, at least in part, to the earlier works of Twitchell (1951) [75] and Brunnstrom (1966) [76]. Twitchell described an "orderly progression of phenomena" characterizing the course of post-stroke UE motor recovery. Initial UE flaccidity was followed by emergence of gross flexor and extensor voluntary movement synergies, gradual separation of synergy movement patterns and finally distal fine motor coordination. Twitchell attributed this recovery sequence to neurophysiological mechanisms, i.e., the cortex reassuming its natural inhibition of primitive proprioceptive reflexes [75]. Brunnstrom, furthering Twitchell's work, documented specific movements that typified each stage of motor recovery thereby mapping the recovery progression. Fugl-Meyer appears to have, at least in part, applied a numerical rating scale to Brunnstrom's "Hemiplegia Classification and Progress Record" [76] to create the FMA-UE [15].

The Rasch item difficulty hierarchy may illustrate the pattern of post-stroke UE recovery. The keyform provides evidence that this hierarchy is replicated across different

individuals in this sample. This hierarchy however, does not follow the expected "stepwise" item arrangement of the FMA-UE. Contemporary motor control theoretical models suggest that UE movement is dynamically organized in response to the demands of a task rather than in a strict reflexive sequence [224]. Following stroke, UE muscle activation patterns and reaching kinematics for a task depend on contextual and mechanical factors such as target location [217, 218], characteristics of the environment [100] and external constraint [98]. Murray and Johnson (2004) found that healthy individuals scaled arm joint torques according to the number of joints involved in a task [106]. In addition, post-stroke UE motor control is also influenced by descending inputs from secondary non-cortical motor areas to axial musculature (likely spared from infarct) and inputs directly from primary cortical motor areas to forearm and hand muscles (likely damaged with infarct) [225]. Taken together, these studies suggest that for individuals with and without stroke, UE motor control is dependent on a dynamic interaction of neural, mechanical, and contextual factors. Our results suggest that the FMA-UE assessment items are arranged in a Rasch-derived difficulty order compatible with the task-specific complexity of motion. That is, arm movements are "easier "or "more challenging" based on the inherent demands of the movement. *Elbow flexion* may be "easier" than *wrist circumduction* because the movement is inherently less complex and the movement may be orchestrated via spared secondary motor areas. *Shoulder flexion to 180 degrees with the elbow extended* may be "difficult" because of the influence of gravity on a long lever arm.

Our task-specific interpretation of the item hierarchy may affect the interpretation of the FMA-UE score. Our results indicate that the FMA-UE aggregate score should not

be interpreted in a strict "proximal-to-distal" or "synergistic-to-isolated" manner. This is particularly important when the FMA-UE score is used to describe patient movement characteristics. Chae and colleagues (2003) [97] suggest that a low FMA-UE aggregate score indicated recovery of only proximal UE motor function, while higher FMA-UE aggregate score indicated recovery of both proximal and distal UE motor function. Chae's interpretation is consistent with Fugl-Meyer's proposed "stepwise" item order in which movements assumed to recover early in the process appear first on the assessment, while those to recover later in the process appear further along the assessment. Our results indicate that an individual with a low FMA-UE score is likely to exhibit some proximal arm motions characteristic of an abnormal synergy pattern (e.g., the "easy" items; *elbow flexion, shoulder adduction with internal rotation, scapular elevation*), but is unlikely to exhibit other proximal synergy-based movements (e.g., the "more difficult" items; *scapular retraction, shoulder external rotation, and forearm supination*). An individual with a low FMA-UE score is likely to exhibit some "easy" distal, non-synergistic, hand movements (e.g., *finger mass flexion, finger mass extension, and cylindrical grasp*). Participant groups formed using FMA-UE scores may be less homogenous than desired. Similarly, our item hierarchy, if assumed to illustrate the pattern of post-stroke UE recovery, shows that recovery does not proceed in a strict proximal-to-distal sequence as was traditionally endorsed by some therapists [226] and suggested by some therapy textbooks [227, 228].

Our results do not entirely contradict the original clinical observations of Twitchell and Brunnstrom. Indeed, the results provide at least partial support for their classic works. For example, both clinicians observed "flexion synergy" and "extension synergy"

movements to be among the first motions to return after stroke. Our results show that three of the six "flexor synergy" movements and two of the three "extension synergy" movements are among the least challenging items of the FMA-UE (*elbow flexion, shoulder abduction* and *scapular elevation*; and *forearm pronation* and *shoulder adduction with internal rotation*). Furthermore both clinicians described "movements deviating from synergy" occurring later in recovery. The Rasch hierarchy also partially supports this observation. Items requiring wrist motion are easier when the shoulder is neutral and the elbow is bent, than when the shoulder is flexed and the elbow is extended. For example *forearm pronation-supination, elbow at 90* and *wrist flexion-extension, elbow at 90* are moderately challenging items, while *pronation-supination, elbow extended* and *wrist flexion-extension, elbow extended* are among the hardest items.

In addition, Brunnstrom proposed separate hand and proximal limb evaluation scales. Several researchers [31, 34] have separated the hand items to focus their assessment on an area interest. It is feasible that different motor control constructs underlie arm versus hand movements given the different supraspinal areas likely influencing arm versus hand motorneurons [225]. Although the PCA rotated factor pattern provided some statistical support for this division, other results (e.g., Rasch fit statistics and PCA item correlations) support keeping the arm and hand voluntary movement items together as a unidimensional scale.

One surprising finding in the present study was that the *finger extension* item calibrated as an "easy" item. Studies have shown finger extension to be highly impaired after stroke [31, 229]. Our results may be explained in two ways. First, the vast majority of participants in our sample (90%) sustained mild or moderate stroke as defined by the

Orpington score. It is possible that our sample was not severely impaired and therefore did not exhibit impaired finger extension. Secondly, the FMA-UE defines finger mass extension as gross release of the mass flexion grasp [15]. *Finger extension*, as defined by the FMA-UE may actually reflect the natural tenodesis mechanism of grossly "extending" the fingers by flexing the wrist. It should be noted that some researchers have more rigorously evaluated finger extension by controlling for wrist position and defining finger extension as return to metacarpophalangeal joint neutral [31, 229]. One weakness of the FMA-UE is that there exist no standard administration guidelines [18, 56]. It is feasible that if the assessment were administered such that the wrist configuration during this item was controlled, the item would have a different difficulty level.

Limitations

There are a number of statistical and demographic limitations to this study that could limit the generalizability of our results. Traditional PCA was designed for continuous rather than ordinal data. Flora and Curran (2004) [230] suggest using factor analysis approaches designed specifically for multivariate categorical data when analyzing ordinal data (e.g., FMA-UE ratings). While these approaches are not yet common in the health care literature they should be considered for future studies. Additionally the generalizability of the findings is confined to sample characteristics of the study. Although our participants represented a range of stroke severity, the majority, 90%, of the sample experienced mild or moderate stroke. In order to generalize these findings across the broader range of stroke severity the study needs to be replicated with individuals with more severe motor impairment. Finally, the most critical limitation of,

the present study is that we used cross-sectional data to challenge Fugl-Meyer's "recovery" paradigm. We are currently replicating this study using longitudinal data.

Conclusion

In summary the Rasch analysis of the FMA-UE challenges the use of reflex items to measure UE movement and challenges the strict "definable stepwise" structure implied by the instrument. A better understanding of the post-stroke UE recovery progression and possible revision of the theoretical mechanisms underlying this recovery may be critical for clinical intervention. Instruments clearly reflecting the recovery pattern may more accurately identify the client's stage of recovery and inform the clinician of the appropriate treatment progression.

Table 3: Characteristics of the 512-person sample

Table 1: Characteristics of the 512 person sample

Years of Age (Mean ± SD)		69.8 ± 11.12
Gender	Males	242
	Females	270
Race	White	411
	African-American	85
	Other	16
Stroke Type	Ischemic	474
	Hemorrhagic	38
Stroke Location	Right Hemisphere	250
	Left Hemisphere	262
Days Since Stroke (Mean ± SD)		16.88 ± 31.23
Range of Days Since Stroke	Minimum	0
	Maximum	145
Stroke Severity	Minor (Orpington Score <3.2)	191 (37.3%)
	Moderate (3.2 <= Orpington <= 5.2)	269 (52.5%)
	Severe (Orpington > 5.2)	52 (10.2%)

Table 4: Item measures, fit statistics and score correlations for the full 33-item FMA-UE (n = 512)

FMA-UE Item	Measure	Error	Infit MnSq	Infit ZSTD	Outfit MnSq	Outfit ZSTD	Score Correlation
Normal reflex activity	3.40	0.11	1.34	3.8	1.10	0.4	.57
Wrist circumduction	1.55	0.09	0.79	-3.2	0.92	-0.3	.77
Hook grasp	1.26	0.09	1.29	3.9	1.08	0.5	.72
Shoulder flexion to 180°, elbow extended	1.20	0.09	0.70	-5.0	0.57	-2.7	.80
Spherical Grasp	1.15	0.09	0.96	-0.6	1.02	0.2	.77
Lateral prehension	1.05	0.09	0.96	-0.5	1.21	1.2	.76
Wrist flexion-extension, elbow extended	1.03	0.09	0.65	-5.8	0.57	-2.8	.82
Pronation-supination, elbow extended	0.98	0.09	0.57	-7.4	0.44	-4.1	.83
Wrist stable, elbow extended	0.93	0.09	0.96	-0.5	0.69	-2.0	.78
Movement with normal speed	0.88	0.09	1.10	1.4	0.93	-0.4	.76
Forearm supination	0.68	0.09	0.70	-4.8	0.73	-1.9	.84
Shoulder abduction to 90°, elbow extended	0.37	0.09	0.72	-4.3	0.53	-4.0	.83
Movement without dysmetria	0.36	0.09	1.37	4.7	1.21	1.4	.76
Shoulder external rotation	0.34	0.09	0.68	-5.0	0.63	-3.0	.85
Wrist stable, elbow at 90°	0.29	0.09	1.01	0.1	0.71	-2.3	.80
Wrist flexion-extension, elbow at 90°	0.24	0.09	0.58	-6.9	0.52	-4.3	.86
Palmar prehension	0.19	0.09	0.95	-0.7	0.83	-1.2	.81
Scapular retraction	0.16	0.09	0.89	-1.5	0.96	-0.3	.82
Pronation-supination, elbow at 90°	0.00	0.10	0.59	-6.5	0.48	-4.8	.86
Shoulder flexion to 90°, elbow extended	-0.03	0.10	0.78	-3.2	0.52	-4.4	.84
Hand to lumbar spine	-0.19	0.10	0.72	-4.0	0.62	-3.3	.85
Shoulder abduction	-0.32	0.10	0.79	-2.9	0.77	-1.8	.84
Elbow extension	-0.39	0.10	0.57	-6.6	0.53	-4.2	.87
Forearm pronation	-0.57	0.10	0.60	-5.9	0.64	-2.9	.86
Movement without tremor	-0.60	0.10	1.19	2.3	0.98	-0.1	.80
Cylindrical grasp	-0.74	0.10	0.88	-1.5	0.76	-1.7	.84

Table 4 continued

Finger mass extension	-0.86	0.10	0.70	-4.1	0.58	-3.3	.86
Scapular elevation	-0.97	0.11	0.93	-0.8	0.87	-0.8	.83
Finger mass flexion	-1.00	0.11	0.79	-2.7	0.70	-2.0	.85
Shoulder adduction with internal rotation	-1.09	0.11	0.65	-4.8	0.49	-3.8	.87
Elbow flexion	-1.23	0.11	0.67	-4.4	0.81	-1.1	.85
Triceps reflex	-3.66	0.13	5.20	9.9	9.90	9.9	.26
Biceps reflex	-4.40	0.14	3.51	9.9	9.90	8.2	.36

Table 5: Eigenvalues for the principal components analysis

Table 3: Eigenvalues for the Principle Components Analysis

Factor	Eigenvalue	Proportion of Variance	Cumulative Variance
1	22.48	0.68	0.68
2	1.60	0.05	0.73
3	1.31	0.04	0.77
4	1.04	0.03	0.80

Table 6: Item correlations with first principal component

Item Description	Component 1
Elbow Extension	0.91
Forearm Pronation	0.91
Pronation Supination with Elbow 90	0.91
Wrist Flexion Extension Elbow 90	0.90
Sh Flex to 90 Elbow Extended	0.90
Shoulder Adduction with IR	0.90
Finger Mass Extension	0.90
Hand to Lumbar Spine	0.89
Sh Abd to 90 Elbow Extended	0.88
Finger Mass Flexion	0.88
Elbow Flexion	0.88
Cylindrical Grasp	0.88
Shoulder Abduction	0.88
Shoulder External Rotation	0.88
Pronation Supination with Elbow Extended	0.87
Forearm Supination	0.86
Wrist Stable Elbow at 90	0.86
Scapular Elevation	0.85
Palmar Prehension	0.85
Wrist Flexion Extension Elbow Extended	0.85
Movement without tremor	0.84
Scapular Retraction	0.84
Sh Flex to 180 Elbow Extended	0.83
Wrist Stable Elbow Extended	0.83
Movement with normal speed	0.81
Wrist Circumduction	0.80
Lateral Prehension	0.80
Spherical Grasp	0.80
Movement without dysmetria	0.78
Hook Grasp	0.75
Biceps Reflex	0.24
Triceps Reflex	0.14
Normal Reflex Activity	0.53

Table 7: Rotated factor loadings (oblique rotation) for 4 principle components

Item Description	Factor 1	Factor 2	Factor 3	Factor 4
Scapular Elevation	0.78	0.04	0.10	0.15
Elbow Flexion	0.77	0.19	-0.02	0.15
Scapular Retraction	0.77	-0.08	0.25	0.10
Shoulder Abduction	0.75	0.04	0.18	0.11
Shoulder Adduction with IR	0.67	0.20	0.10	0.14
Shoulder External Rotation	0.61	0.06	0.30	0.13
Elbow Extension	0.61	0.21	0.19	0.13
Forearm Pronation	0.59	0.30	0.08	0.16
Movement without tremor	0.49	0.34	0.11	0.10
Hand to Lumbar Spine	0.46	0.26	0.23	0.17
Forearm Supination	0.40	0.31	0.23	0.15
Lateral Prehension	-0.04	0.77	0.15	0.09
Palmar Prehension	0.14	0.73	0.10	0.05
Spherical Grasp	0.09	0.71	0.10	0.06
Hook Grasp	0.01	0.69	0.18	0.01
Cylindrical Grasp	0.38	0.63	-0.03	0.07
Finger Mass Extension	0.37	0.61	-0.01	0.14
Finger Mass Flexion	0.37	0.55	0.04	0.13
Wrist Stable Elbow at 90	0.13	0.51	0.31	0.14
Movement with normal speed	0.21	0.48	0.25	0.04
Wrist Flexion Extension Elbow 90	0.21	0.46	0.33	0.13
Normal Reflex Activity	-0.13	-0.17	0.99	0.02
Sh Flex to 180 Elbow Extended	0.28	0.00	0.72	0.04
Pronation Supination with Elbow Extended	0.17	0.20	0.64	0.11
Triceps Reflex	-0.23	-0.10	-0.06	0.96
Biceps Reflex	0.00	-0.16	-0.11	0.92
Pronation Supination with Elbow 90 **	0.37	0.35	0.29	0.13
Movement without dysmetria **	0.32	0.37	0.18	0.09
Wrist Flexion Extension Elbow Extended	0.02	0.41*	0.54*	0.12
Wrist Stable Elbow Extended	-0.04	0.46*	0.52*	0.11
Wrist Circumduction	-0.05	0.47*	0.50*	0.08
Sh Abd to 90 Elbow Extended	0.45*	0.08	0.50*	0.05
Sh Flex to 90 Elbow Extended	0.44*	0.18	0.42*	0.09

Table 8: Item measures, fit statistics, and correlations for revised FMA-UE (n = 512, 30 Items)

FMA-UE Item	Measure	Error	Infit MnSq	Infit ZSTD	Outfit MnSq	Outfit ZSTD	Score Correlatio
Wrist circumduction	1.67	0.10	0.90	-1.5	0.96	-0.1	.77
Hook grasp	1.33	0.10	1.43	5.6	1.28	1.7	.73
Shoulder flexion to 180°, elbow extended	1.26	0.10	0.84	-2.5	0.80	-1.4	.80
Spherical Grasp	1.20	0.10	1.08	1.1	1.17	1.1	.78
Lateral prehension	1.08	0.10	1.10	1.5	1.30	2.0	.78
Wrist flexion-extension, elbow extended	1.06	0.10	0.77	-3.7	0.78	-1.6	.82
Pronation-supination, elbow extended	1.00	0.10	0.70	-4.8	0.63	-3.1	.83
Wrist stable, elbow extended	0.95	0.10	1.13	1.9	0.96	-0.2	.78
Movement with normal speed	0.89	0.10	1.25	3.4	1.13	1.0	.77
Forearm supination	0.65	0.10	0.82	-2.6	0.88	-1.0	.85
Shoulder abduction to 90°, elbow extended	0.28	0.10	0.86	-2.1	0.71	-2.7	.85
Movement without dysmetria	0.27	0.10	1.62	7.4	1.52	3.9	.77
Shoulder external rotation	0.24	0.10	0.83	-2.5	0.82	-1.6	.86
Wrist stable, elbow at 90°	0.18	0.10	1.20	2.7	0.96	-0.3	.82
Wrist flexion-extension, elbow at 90°	0.12	0.10	0.69	-4.8	0.66	-3.2	.88
Palmar prehension	0.06	0.10	1.10	1.3	1.01	0.1	.83
Scapular retraction	0.03	0.10	1.08	1.2	1.29	2.4	.84
Pronation-supination, elbow at 90°	-0.17	0.11	0.72	-4.2	0.67	-3.1	.88
Shoulder flexion to 90°, elbow extended	-0.21	0.11	0.94	-0.8	0.71	-2.7	.86
Hand to lumbar spine	-0.40	0.11	0.90	-1.4	0.85	-1.2	.87
Shoulder abduction	-0.56	0.11	0.99	-0.1	1.02	0.2	.87
Elbow extension	-0.64	0.11	0.70	-4.2	0.70	-2.4	.90
Forearm pronation	-0.87	0.11	0.76	-3.1	0.81	-1.4	.89
Movement without tremor	-0.91	0.11	1.51	5.4	1.36	2.2	.83
Cylindrical grasp	-1.10	0.12	1.09	1.1	0.94	-0.3	.87
Finger mass extension	-1.25	0.12	0.89	-1.3	0.76	-1.4	.89
Scapular elevation	-1.40	0.12	1.25	2.7	1.31	1.6	.86
Finger mass flexion	-1.44	0.12	1.04	0.4	1.03	0.2	.89
Shoulder adduction with internal rotation	-1.56	0.12	0.88	-1.4	0.74	-1.4	.90
Elbow flexion	-1.76	0.13	0.92	-0.8	1.18	0.8	.89

Table 9: Item difficulty hierarchy revised 30-item FMA-UE

FMA-UE item number	FMA-UE item description	Measure	Error
23	Wrist circumduction	1.67	0.10
26	Hook grasp	1.33	0.10
16	Shoulder flexion to 180°, elbow extended	1.26	0.10
30	Spherical grasp	1.20	0.10
27	Lateral prehension	1.08	0.10
22	Wrist flexion-extension, elbow extended	1.06	0.10
17	Pronation-supination, elbow extended	1.00	0.10
21	Wrist stable, elbow extended	0.95	0.10
33	Movement with normal speed	0.89	0.10
8	Forearm supination	0.65	0.10
15	Shoulder abduction to 90°, elbow extended	0.28	0.10
32	Movement without dysmetria	0.27	0.10
6	Shoulder external rotation	0.24	0.10
19	Wrist stable, elbow at 90°	0.18	0.10
20	Wrist flexion-extension, elbow at 90°	0.12	0.10
28	Palmar prehension	0.06	0.10
4	Scapular retraction	0.03	0.10
14	Pronation-supination, elbow at 90°	-0.17	0.11
13	Shoulder flexion to 90°, elbow extended	-0.21	0.11
12	Hand to lumbar spine	-0.40	0.11
5	Shoulder abduction	-0.56	0.11
10	Elbow extension	-0.64	0.11
11	Forearm pronation	-0.87	0.11
31	Movement without tremor	-0.91	0.11
29	Cylindrical grasp	-1.10	0.12
25	Finger mass extension	-1.25	0.12
3	Scapular elevation	-1.40	0.12
24	Finger mass flexion	-1.44	0.12
9	Shoulder adduction with internal rotation	-1.56	0.12
7	Elbow flexion	-1.76	0.13

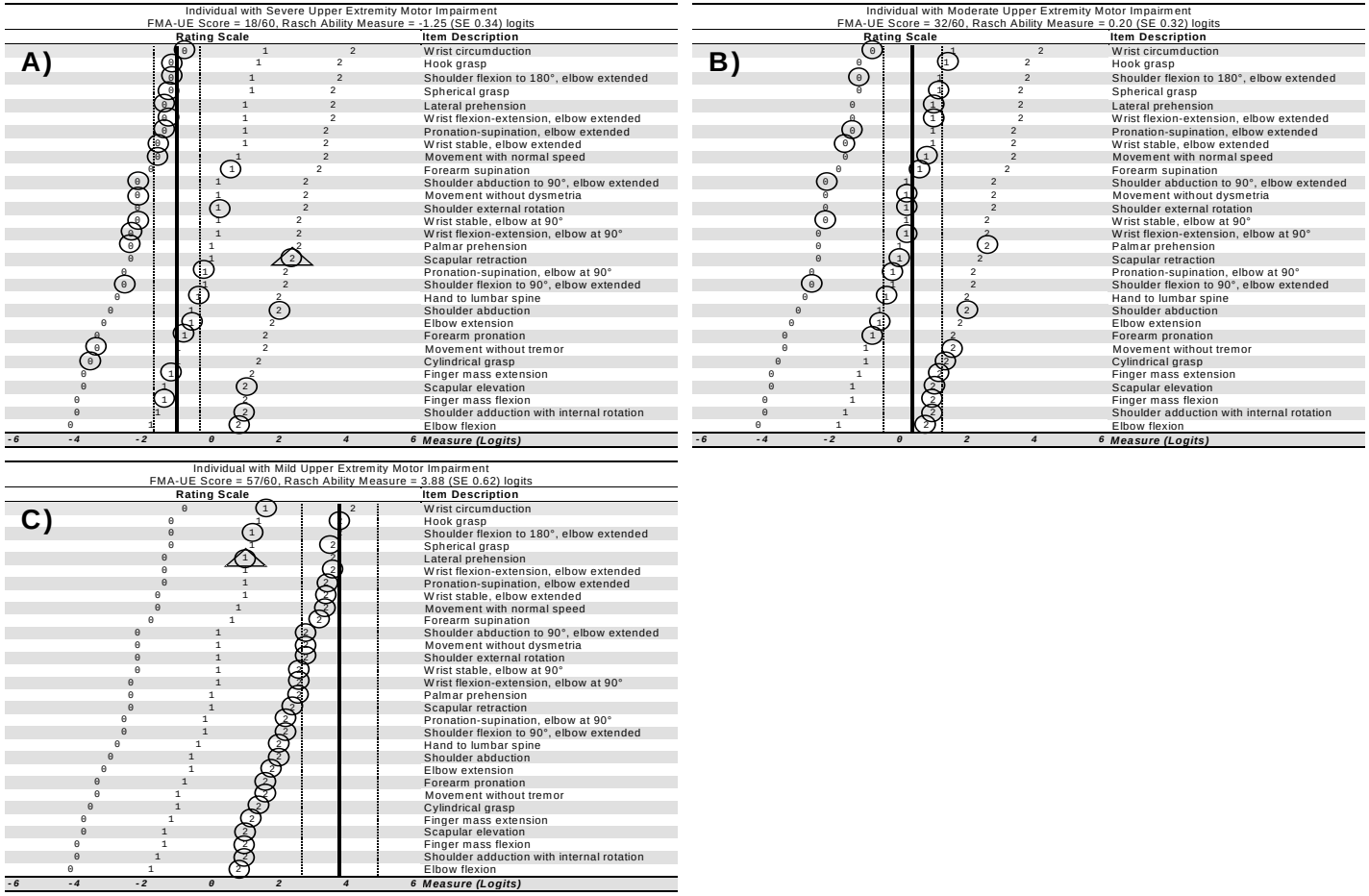

Figure 1: Keyform recovery maps for three individuals with stroke.

CHAPTER 3
LONGITUIDNAL STABILITY OF THE 30-ITEM FUGL-MEYER UPPER EXTREMITY ASSESSMENT ITEM DIFFICULTY HIERARCHY

The primary goal of many post-stroke upper extremity (UE) motor rehabilitation providers is to observe change in UE motor function attributable to the antecedent intervention. Assessing change requires repeated measurement of function before and after intervention. Furthermore many rehabilitation researchers and clinicians desire to understand the time course of post-stroke UE motor recovery. This also necessitates repeated measurement of function longitudinally over time. An assessment used to analyze clinical change and/or describe recovery must measure in a valid and consistent manner each time the assessment is administered [231].

The Fugl-Meyer Upper Extremity Assessment (FMA-UE) [15] is the most widely used clinical assessment of post-stroke UE motor function impairment [16]. It is often used to assess changes in UE motor function with rehabilitation. For example, the FMA-UE was the primary criterion for evaluating response to interventions such as rhythmic bilateral movement training [36, 37], neuromuscular stimulation [38], electromyographic triggered electrical stimulation [39, 40], Botox [41], home based exercise [30], community based exercise [42], robot aided therapy [43, 44], virtual reality [45], imagery [46], forced use [47], and modified constraint induced movement therapy [48-51]. Furthermore, the FMA-UE is often used in longitudinal characterizations of the post-stroke UE motor recovery process [2, 29, 232, 233]. When studied as a whole assessment, the FMA-UE has shown excellent psychometric properties in test-retest

situations [17, 18, 56]. However, there are no published analyses examining how individual FMA-UE items perform in test-retest situations.

Because the FMA-UE is often used as a longitudinal measure of UE motor function, it is critical to test whether the items forming this construct remain longitudinally stable. Item Response Theory, e.g., the Rasch measurement model, offers a method to examine assessment stability at the item level. Rasch analysis has been increasingly applied to health-care outcomes research to examine the stability of assessment items across testing occasions [161, 162, 182, 190].

In a previous cross-sectional study we used Rasch analysis and Principal Component Analysis to test the dimensionality and construct validity of the FMA-UE in a participant sample of 512 individuals 2 weeks to 3 months post-stroke. We found that three tendon-tap reflex items (*biceps reflex, triceps reflex, normal reflex activity*) were empirically disconnected from the other voluntary UE movement items. We removed the reflex items to form a modified 30-item FMA-UE.

The 30-item FMA-UE items were arranged along a hierarchical continuum from "least difficult" to "most difficult". This item hierarchy is a visual representation or map of the post-stroke UE recovery progression. This item difficulty hierarchy was consistent with contemporary motor control science suggesting that UE movements are more or less difficult depending on the task-specific complexities of the movement. The Rasch-derived item hierarchy was replicated across 98% of our sample with mild to severe UE motor impairment.

Ideally, the assessment should have the same item hierarchy each time it is administered. A stable item order assures that individuals with equivalent UE motor

ability will have equal probability of receiving a score on a given FMA-UE item irrespective of the occasion at which it was assessed. While we would expect that individual UE motor abilities would change over time, we would not expect the item hierarchy (i.e., the post-stroke UE recovery map) to change over time. An unstable item order may threaten the assessment's construct validity making it difficult to compare FMA-UE scores from one testing occasion to another testing occasion. It is possible that an assessment's item difficulty order may unexpectedly reorganize with a repeated test even if the overall consistency of the assessment as a whole remains high [180].

The purpose of the current study is to test whether the item difficulty hierarchy of the modified 30-item FMA-UE is equivalent at subsequent testing occasions. Specifically, we explored whether FMA-UE items display the same statistical characteristics, specifically item difficulties, in one sample of post-stroke individuals to whom the FMA-UE was administered at 2-weeks (T1) and 6 months (T2) post-stroke. Testing the longitudinal stability of the item hierarchy addresses the important issue of whether it is valid to make comparisons of repeated-measure FMA-UE scores.

Methods

Participants

A secondary analysis of FMA-UE scores from 459 individuals enrolled in the Kansas City Stroke Study [221] was conducted. Participants included in the Stroke Study met the following inclusion criteria; (1) stroke onset within 0 – 14 days prior to enrollment, (2) 18 years of age or older, (3) ischemic stroke as diagnosed by physician, (4) lived in the community prior to stroke onset, and (5) able to participate in baseline testing. The FMA-UE was administered to participants at admission to the study, i.e., 0 – 14 days post stroke (T1) and 6 months post stroke (T2). Patients were evaluated by staff,

all of who were health care professionals (e.g., nurses, physical therapists) and had undergone at least 2-weeks of training in the administration of the FMA-UE. For the purposes of our study we included only participants with FMA-UE scores at both T1 and T2. This resulted in exclusion of 82 participants from the larger database for a final sample size of 377 participants.

The 30-item Fugl-Meyer Upper Extremity Assessment

The modified 30-item FMA-UE, produced in a previous cross-sectional study, is identical to the traditional assessment except that the 3 tendon-tap reflex items (biceps reflex, triceps reflex, and normal reflex activity) present in the traditional 33-item FMA-UE [15] have been removed. Each of the 30-items are rated on a 3-point ordinal scale ("0" = unable to perform, "1" = partial performance, and "2" faultless performance). The items of the modified 30-item FMA-UE are arranged in a difficulty order consistent with contemporary motor control framework in which post-stroke UE recovery is viewed as a process of re-acquiring the ability to meet task-specific UE movement demands. Movements that were easier (i.e., less complex) for most people with stroke (shoulder adduction with internal rotation, elbow flexion, mass finger flexion) are located towards the easier end of the item hierarchy, and movements that were more challenging (i.e., more complex) for most people with stroke (wrist circumduction, lateral prehension, spherical grasp) are located towards the more difficult end of the item hierarchy. Ideally, item locations should be reproducible on different testing occasions.

Data Analysis

The analysis proceeded in four steps.

Step 1: Dimensionality

In order to create a legitimate measure, all items must contribute to the same construct, that is, the assessment must be unidimensional [143]. We ascertained that all items of the 30-item FMA-UE illustrated a single underlying trait using the WINSTEPS [234] software program. Rasch analysis produces infit and outfit statistics to determine the extent to which items contribute to a unidimensional construct. The infit statistic is most sensitive to ratings on items that are closely matched to participants' ability, the outfit statistic is most sensitive to ratings on items that are much easier or much harder than participants' ability [142]. Fit statistics are reported as mean square standardized residuals (MnSq) produced for each item of the instrument. MnSq represents observed variance divided by expected variance [172]. Consequently, the desired value of MnSq for an item is 1.0. We defined the acceptable criteria for unidimensionality as MnSq infit values < 1.7 associated with standardized Z values (ZSTD) < 2.0 as suggested by Wright and Linacre for assessments employing clinical observation [169]. High values (i.e., infit MnSq > 1.7, ZSTD > 2.0) indicate that scores are variant or erratic, suggesting that an item belongs to a construct that is different from that represented by the other items of the instrument, or that the item is being inaccurately scored. Low values indicate that the item is too predictable in its measurement. In this analysis we focused on items with high infit MnSq values, since this misfit represents a threat to construct validity.

Examining the point-measure correlation values of each item further dimensionality. This value is the correlation of an individual item with the other items of the assessment, omitting that item. The value indicates the extent to which the item contributes to measurement of the overall construct [235].

Step 2: Equivalence of item difficulties

To assess the longitudinal equivalence of the FMA-UE item structure, two procedures were followed. First, item difficulty parameters were derived with separate T1 and T2 Rasch analyses. Individual item measures were compared across the two testing occasions. Second, as suggested by Chang and Chan (1995) [231], an overall indication of the reproducibility of the item difficulty calibrations was obtained using an intraclass correlation coefficient (ICC) (model 2,1) [236].

Step 3: Differential item functioning

A more robust comparison of T1 item difficulties to T2 item difficulties is accomplished through Differential Item Functioning (DIF) analysis. DIF is a statistical procedure commonly used to ascertain whether test-item measures are consistent irrespective of test-taker characteristics [181]. For the purposes of this study we used DIF to determine the consistency of item calibrations across testing occasions. The presence of DIF may indicate that an item measures a unique trait at one testing occasion when compared to another testing occasion or may signal problems with the administration of the assessment [182].

There are various approaches for examining DIF (see [183] for review). The Rasch model employs a relatively straightforward DIF procedure ("DIF=") in which each item's difficulty is directly compared between each time point using iterative t-tests [166]. The correct alpha value for this analysis is debated in the literature. Kothari and Haley (2003) [237] suggest an appropriate critical value of 1.96 to detect differences in item difficulty (α=0.05). This value does not protect against a Type I error (i.e., the probability of falsely detecting a significant difference) during multiple comparisons, but it does protect against a Type II error (i.e., the probability of overlooking a significant difference if one exists).

Small differences may occur by chance, therefore necessitating a corrected alpha value. We utilized a Bonferroni adjustment based on the 30 items in the test (0.05/30, α=0.0017). Using a technique similar to one recently used by Crane and Hart (2006) [238] we explored "small DIF" and "large DIF". Small DIF detects marginal item difficulty differences by comparing the probability of the difference to the unadjusted alpha level (α=0.05) and large DIF detects robust item difficulty differences by comparing the probability of the difference to the adjusted alpha level (α=0.0017).

DIF was visualized in a scatter plot of item difficulty values. A close fit of the item difficulty calibrations, within the 95% confidence interval measurement error boundaries, at each time point verifies item stability [239]. Items whose difficulty calibration logit values were more than 2 standard errors apart (falling outside the 95% confidence interval) were considered statistically different.

Step 4: Impact on assessment of arm motor function

The goal of examining the stability of the FMA-UE item difficulty order is to assure that the assessment is valid as a measure of UE motor ability changes over time. The presence of DIF does not always translate to poor measurement of person ability [189]. Therefore to empirically determine if the presence of DIF had any impact on the FMA-UE ability estimates, the assessment was adjusted for DIF by excluding potential items displaying DIF. UE motor ability was then measured at T1 and T2 using corrected and uncorrected assessments. The measurements obtained from the assessments were compared for similarity.

Results

Participants

Characteristics of the 377-person sample are presented in Table 10. The sample was 69.2 ± 11.2 years of age, 54% female and 80% white. The majority of the sample sustained ischemic stroke, with the location of the stroke equally distributed right versus left hemisphere. The sample primarily sustained minor (42%) or moderate (49%) stroke severity as defined by the Orpington Prognostic Scale [133].

Step 1: Dimensionality

Table 11 presents Rasch derived item infit statistics for the 30-item FMA-UE at each time point. All items in the assessment demonstrated acceptable fit statistics at T1, however hook grasp demonstrated an infit MnSq value > 1.70 with ZSTD > 2.0 at T2. Furthermore, all items demonstrated high point-measure correlations ranging from 0.76 – 0.91 at T1 to 0.75 - 0.88 at T2.

Step 2: Equivalence of item difficulties

The relationship between T1 item difficulty calibrations and T2 item difficulty calibrations is illustrated in Figure 2. A separate line represents the item difficulties at each testing occasion. The dotted line illustrates T1 item difficulties; the solid line illustrates T2 item difficulties. This figure demonstrates the relative invariance of the parameters. Despite some variations, the item ordering and item difficulty values are fairly consistent. Shoulder adduction with internal rotation, elbow flexion, and scapular elevation were among the easiest items at both T1 and T2. Wrist circumduction, hook grasp, and lateral prehension were among the most challenging items at both T1 and T2. The ICC across item difficulties at T1 and T2 was 0.95.

The reliability of comparing T1 and T2 item difficulties was assessed using an intraclass correlation coefficient (ICC) as suggested by Chang and Chan (1995) [231]. The present comparison yielded an ICC = 0.95 (model 2,1 [236]) indicating a high overall reliability between T1 item difficulties and T2 item difficulties.

Step 3: Differential item functioning T1 vs. T2

Table 12 presents DIF analysis of the 30-item FMA-UE item order. FMA-UE items are arranged in descending order, from more difficult items (top) to less difficult items (bottom). Rasch derived item difficulties for T1 and the standard error of the value are listed in the left two columns. Rasch derived item difficulties for T2 and the standard error of the value are listed in the next two columns. The DIF analysis compares the T1 and T2 measures of each item. The calculated t-value for each comparison and its associated probability level are listed in the right two columns. Two items: *shoulder flexion to 180, elbow extended* and *movement with normal speed*, show large DIF (using a p value corrected for multiple comparisons, $p<0.0017$). Six items: *hand to lumbar spine, shoulder abduction to 90 degrees with the elbow extended, pronation/supination with the elbow extended, wrist stable with elbow bent, movement without dysmetria and shoulder abduction*, show small DIF (using a p-value not corrected for multiple comparisons, $p<0.05$). Of these eight items, four (*shoulder flexion to 180, elbow extended, pronation/supination with the elbow extended, shoulder abduction to 90 degrees with the elbow extended, and hand to lumbar spine*) display lower T1 item difficulty values than T2 values. The remaining four items (*movement with normal speed, wrist stable with elbow bent, shoulder abduction, and movement without dysmetria*) display higher T1 item difficulty values than T2 values.

Figure 3 is a visual illustration of DIF. The item difficulty parameters of the assessment calculated at T1 (x-axis) are cross-plotted with the item difficulties of the assessment calculated at T2 (y-axis). The solid lines represent the 95% confidence interval for these values. The items outside the error bands are labeled. The two items furthest outside the error interval are shoulder flexion to 180 degrees with elbow straight and, movement with normal speed. Six items difficulties are on or slightly beyond the error interval (*hand to lumbar spine, shoulder abduction to 90 degrees with the elbow extended, pronation/supination with the elbow extended, wrist stable with elbow bent, movement without dysmetria and shoulder abduction*). Four items (*shoulder flexion to 180, elbow extended, pronation/supination with the elbow extended, shoulder abduction to 90 degrees with the elbow extended, and hand to lumbar spine*) are above the error band indicating that the item was less difficult at T1 compared to T2. Four items (*movement with normal speed, wrist stable with elbow bent, shoulder abduction, and movement without dysmetria*) are below the error band indicating that the items were more difficult at T1 than at T2.

Step 4: Impact of DIF on person ability measurement

To compare the possible impact of DIF on the assessment's validity as a measure of UE motor ability, the FMA-UE was adjusted by excluding items with DIF. T1 and T2 UE motor ability measures were calculated using three forms of the FMA-UE; (i) the full 30-item assessment, (ii) a 22-item assessment in which all DIF items (large and small) were removed, and (iii) a 28-item assessment in which only large DIF items were removed. Mean T1 and T2 person ability measures and their standard error of measurement (error bars) calculated using the three forms of the assessment are presented in figure 4. The pair of bars to the left illustrate the measures with no DIF items excluded

from the assessment, the bars in the middle illustrate the measures with all DIF items (eight items) excluded from the assessment, and the bars to the right illustrate the measures with only large DIF items (two items) excluded from the assessment. No significant differences in the measurement of participant's UE motor function are evident in the 30-item assessment compared to the 22-item and 28-item assessments neither at T1 nor at T2. The effect sizes for the change in UE motor function from T1 to T2 are virtually identical with and without DIF items (effect size = 0.52, 0.51, 0.51) for the three assessment versions.

Discussion

The purpose of this study was to test the equivalence of the FMA-UE item hierarchy across two testing occasions. We examined data from 377 post-stroke individuals to whom the FMA-UE was administered at two points during their recovery process. We tested whether FMA-UE items displayed invariant difficulty calibrations at each time point in the longitudinal study. Our data shows that the item difficulty order for eight items varies significantly, but that this DIF has no practical consequences for measuring UE motor function.

Rehabilitation researchers are often interested in whether an intervention is effective. To this end, researchers choose an appropriate assessment to score an outcome of interest and compare scores between and/or within groups. Regardless of the statistic used to assess effectiveness (e.g., comparison of group means, correlation test, etc.) it is implicitly assumed that the measurement properties of the assessment are identical each time the assessment is used [180].

In the present study, the majority of the FMA-UE items have equivalent T1 and T2 difficulty measures. However, the measures are not entirely identical. The scatter plot and

DIF statistical testing shows that eight item show significant DIF. Specifically, two items displayed difficulty estimates that were significantly different with a corrected alpha value (large DIF) at the two time points, and six other items displayed difficulty estimates that were significantly different with an uncorrected alpha value (small DIF) at the two time points.

Item instability may warn that an item is functioning in an unexpected way, i.e., the item measures a different trait on one testing occasion than another. Applied to the FMA-UE, item instability may reveal that the post-stroke UE motor ability recovery map illustrated by the item difficulty hierarchy at T1 is fundamentally different than the map illustrated by the item order at T2. Four items (*shoulder flexion to 180, elbow extended, pronation/supination with the elbow extended, shoulder abduction to 90 degrees with the elbow extended, and hand to lumbar spine*) displayed lower item difficulty values at T1 when compared to T2. That is, these items relocated from the "easy" end of the item difficulty hierarchy (at T1) towards the "more challenging" end of the item hierarchy (at T2). Four other items (*movement with normal speed, wrist stable with elbow at 90, movement without dysmetria, and shoulder abduction*) showed the opposite trend. That is, these items relocated from the "more challenging" end of the item difficulty hierarchy (at T1) towards the "easy" end of the item hierarchy (at T2).

According to item response theory, an individual's response to an assessment item depends on the individual's ability (with regards to the construct being assessed) and the item's difficulty. Test-takers with equal levels of UE motor ability should have the same probability of responding correctly to an assessment item irrespective of the time post-stroke at which the assessment is administered. A stable item (i.e., displaying no DIF)

maintains a consistent location on the post-stroke UE motor recovery "map" and therefore affords the same probability of correct response at each test situation. An unstable item (i.e., displaying DIF) does not maintain a constant location on the post-stroke UE motor recovery "map", and therefore does not afford the same probability of correct response at each test situation.

The occurrence of DIF among the FMA-UE items suggest that an individual tested at 0 – 14 days post-stroke is more likely to successfully perform *hand to lumbar spine, shoulder flexion to 90 with elbow extended, pronation-supination with elbow extended, and shoulder flexion to 180 with elbow extended* than an individual with equivalent UE motor ability tested at 6 months post-stroke. An individual tested at 0 – 14 days post-stroke is less likely to successfully perform *shoulder abduction, wrist stable with elbow bent, movement without dysmetria, and movement with normal speed* than an individual with equivalent UE motor ability tested at 6 months post-stroke. Of course individuals with more skill (i.e., greater UE motor ability) are more likely to successfully perform FMA-UE items when compared to individuals with less skill. However, the presence of DIF suggests that the assessment's underlying construct has shifted. This may threaten the assessment's construct validity.

Bingenheimer et al. (2005) [181] argues that measurement equivalence is necessary for valid interpretation of assessment scores. Specifically, an instrument should display item and scalar equivalence meaning that an assessment's items should assess the same trait each time it is given (item equivalence), and identical scores should always have the same interpretation (scalar equivalence). The items of an assessment are the operational definition of the trait the assessment is intended to measure, and the score quantifies how

much of the trait the test-taker possesses. If an item varies with regards to its meaning or the score varies with regards to its interpretation, the assessment is not an accurate measure of the trait. Bingenheimer argues that without convincing evidence of item and scalar equivalence, there is no basis for comparing scores on an assessment between or within groups.

In the healthcare literature DIF analyses have been useful to study item equivalence (i.e., to detect potential test item bias) among persons of different cultures and diagnostic groups. For example, using DIF analyses, Azocar and colleagues (2003) [185] showed that Latino-Americans were more likely to receive a higher score on some items of the Beck Depression Inventory than non Latino-Americans, potentially flagging a distinct Latino cultural interpretation of the items. Azocar recommended a reconsideration of the assessment's Spanish translation in order to facilitate cross-cultural comparisons of assessment scores. Tennant and colleagues (2004) [186] found that certain Functional Independence Measure self-care items displayed DIF. The researchers showed that culturally unique ways of bathing and grooming influenced the way the items were administered among European countries, To facilitate cross-country comparisons of scores the assessment scoring procedure was adjusted for those items. Roorda and colleagues (2004) [187] found that the hierarchical item ordering of a stair climbing assessment was different for amputees versus non-amputees. The researchers recommend using a shortened assessment, with DIF items removed, if the intention is to compare scores between the two diagnostic groups.

In the present study we used DIF analyses for a different purpose. Instead of testing the measurement equivalence of an assessment between cultures or diagnostic groups, we

examined the same group of people and compared the measurement equivalence of the assessment between testing occasions. The use of DIF analyses for this purpose has precedence in the literature. Tesio et al. (2003) [182] detected DIF among several Functional Independence Measure mobility items (transfer to a chair, locomotion, and stairs) when admission scores were compared to discharge scores. The items were "easy" for a group of orthopedic patients with mobility restrictions at admission (e.g., no lower extremity weightbearing due to orthopedic precautions). However, the items became "more difficult" for this group at discharge, after the mobility restrictions were removed. This counterintuitive result led the staff at this facility to examine the manner in which the assessment was scored. They learned that the admission score was based on a prediction of the patient's expected outcome, while the discharge score was based on actual observation of patient performance. In essence, the assessment was measuring a different construct at patient admission (e.g., ability of the staff to predict patient status) versus patient discharge.

Finlayson and Mallinson (2005) [161] applied DIF analysis to the primary functional outcome measure used in a large longitudinal study of aging and ADL performance in various healthcare settings. Rasch analysis of a 24-item ADL assessment was conducted at three time points over the 13-year study. Although DIF was detected among 4 items, the researchers were unable to discover a specific reason for it. Because statistically significantly differences in item difficulty do not always translate to practical differences in person ability scores [189], the researchers measured ADL ability using an assessment in which the 4 unstable items were removed and compared the results to measurement using the full assessment. The analysis suggested that DIF had only

minimal impact on the Rasch-derived ability measures over time or across settings leading the researchers to continue using the entire assessment.

How can DIF be interpreted in the present study? The presence of DIF may suggest that the FMA-UF is multidimensional [181]. It is possible that *shoulder flexion to 180 degrees with elbow extended* is a less demanding task for individuals earlier in the post-stroke recovery process than later in the process because spasticity and/or soft tissue shortening may occur over time. However, there does not seem to be an obvious trait shared by the DIF items that would lead to the hypothesize that the motor control construct represented by the items is conceptually different at 0-14 days post-stroke than at 6 months post-stroke. It is possible that our results could be attributed to potential differences in item administration at T1 compared to T2. Perhaps poor postural (e.g., decreased trunk control at T1 versus T2) and/or seating system (e.g., seated in a wheelchair at T1 versus a chair at T2) confounded the administration of the items. It is also possible that clients interpreted the item directions differently at each time point. These possibilities merit further investigation.

The presence of DIF can have significant impact on the validity of an assessment for measuring person ability. Linacre's (1994) [190] study provided a classic example of the impact of item instability on the measurement of patient change. Using Rasch analyses, Linacre found that when the 13 motor and 5 cognitive items of the Functional Independence Measure were combined into a single 18-item assessment, the admission and discharge item difficulty calibrations of the items were not equivalent. He suggests that the presence of variant item difficulty measures indicated that a single pattern of disability could not be identified when comparing patient admission scores to discharge

scores. That is, the meaning of the score was dependent on the time point (admission versus discharge) at which the assessment was administered. Linacre demonstrated that the statistical validity of the scale was improved if the motor and cognitive items were separated into sub-scales.

The relationship between the potential item bias detected in our study and the measurement of person ability was undertaken in a way similar to that of Finlayson and Mallinson (2005) [161] and recommended by Roznowski (1999) [189]. That is, we excluded DIF items, measured person-ability with the shortened assessment, and then compared the ability measures of the full scale to the shortened scale. There was essentially no difference in person ability measured by assessments with and without DIF items excluded. Furthermore, the effect sizes were identical for each assessment. The results suggest that the amount of DIF in the FMA-UE is negligible had no practical consequence on estimates of UE motor function. Therefore we suggest keeping all 30 items in the assessment for longitudinal measurement of UE motor ability.

Study Limitations

There are several limitations that might limit the generalizability of this study. First, although all items of the modified 30-item FMA-UE demonstrated acceptable fit statistics at T1, hook grasp was erratic at T2. The item displayed infit MnSq (1.82) and ZSTD (7.80) values which are beyond the acceptable ranges suggested by Wright and Linacre (1994) [169]. Smith (1996) [240] argues that Rasch analysis cannot proceed if its assumptions of unidimensionality are not met. High infit statistics, as displayed by hook grasp, may signal that the item is measuring a construct different than the construct measured by the other assessment items. We did not explore removing the item prior to proceeding with the DIF analysis in the present study, as would be a common procedure.

We did not do so for two reasons. The item, although erratic in the T2 analysis of this study, was not erratic in the T1 analysis or in the analysis we performed in a previous study with a larger sample size. Also, there is no motor control theory that would support a single prehension pattern such as hook grasp, belonging to a unique motor control construct.

The second limitation to this study is that we did not investigate whether other forms of DIF are present in the FMA-UE. For example, non-uniform DIF is a special type of item instability in which item difficulty characteristics change across clients of different ability levels. The Rasch method for detecting DIF used in this study has been method is criticized because of its inability to detect non-uniform DIF [183]. The presence of non-uniform DIF in the FMA-UE should be examined in future studies using appropriate methods for its detection.

Finally, in this study we examined the longitudinal stability of the 30-item FMA-UE rather than the traditional 33-item FMA-UE. Therefore our findings cannot be directly applied to the traditional 33-item FMA-UE. However, as we argued in a previous study, the 30-item FMA-UE is a unidimensional assessment with an item hierarchy consistent with contemporary motor control theory and evidence. The results of our previous study suggest that the three reflex items reflect a different motor control construct than the other items of the assessment. We purposely did not include the items in the present analysis to avoid DIF resulting from multi-dimensionality (i.e., DIF items do not consistently assess a single trait).

Conclusion

In test-retest situations it is important to examine, rather than assume, the longitudinal stability of test item difficulties. Although it is common to check the

construct validity of an assessment at the beginning of a longitudinal study, and check the reliability of the assessment at various points along the study, it is not common to check the stability of the assessment items [180]. Item stability is important since the items represent the construct an assessment intends to measure. Without assurance of item stability it is difficult to know whether the assessment measures the same trait across each testing occasion.

The present study illustrates a procedure for examining the longitudinal stability of assessment items. The presence of item instability may threaten repeated measures score comparisons and thus, this procedure should be a routine check of an assessment's measurement properties. DIF testing might also be a method useful to inform motor control theory. For example, testing item stability could yield valuable insights into the way motor control is theorized to recover over time.

Table 10: Sample characteristics (n = 377)

Years of Age (Mean ± SD)		69.2 ± 11.2
Gender	Males	174 (46%)
	Females	203 (54%)
Race	White	301 (80%)
	African-American	63 (17%)
	Other	13 (3%)
Stroke Type	Ischemic	355 (94%)
	Hemorrhagic	22 (6%)
Stroke Location	Right Hemisphere	173 (46%)
	Left Hemisphere	204 (54%)
Stroke Severity	Minor (Orpington Prognostic Score <3.2)	160 (42%)
	Moderate (3.2 <= Orpington <= 5.2)	184 (49%)
	Severe (Orpington > 5.2)	33 (9%)

Table 11: Rasch derived infit statistics T1 and T2 (n = 377)

	FMA-UE Item Description	T1 item MNSQ	T1 ZSTD	T1 rPM	T2 item MNSQ	T2 ZSTD	T2 rPM
Hard	Wrist circumduction	0.93	-0.8	0.79	0.93	-0.8	0.83
	Hook grasp	1.52	5.4	0.76	1.82 *	7.8 *	0.75
	Lateral prehension	1.18	2.0	0.80	1.39	4.1	0.79
	Spherical grasp	1.18	2.1	0.80	1.03	0.4	0.83
	Shoulder flexion to 180, elbow ext	0.82	-2.3	0.84	0.83	-2.1	0.84
	Wrist stable, elbow extended	1.10	1.2	0.80	1.04	0.5	0.84
	Wrist flexion-extension, elbow ext	0.76	-3.1	0.86	0.67	-4.3	0.87
	Movement with normal speed	1.27	2.9	0.80	0.85	-1.7	0.86
	Pronation-supination, elbow ext	0.64	-4.8	0.87	0.72	-3.6	0.87
	Forearm supination	0.77	-3.0	0.88	1.02	0.3	0.86
	Wrist stable, elbow at 90	1.11	1.2	0.84	1.04	0.5	0.84
	Movement without dysmetria	1.58	5.7	0.80	1.40	3.7	0.81
	Wrist flexion-extension, elbow at 90	0.72	-3.5	0.89	0.69	-3.7	0.88
	Shoulder external rotation	0.88	-1.4	0.88	0.81	-2.0	0.88
	Scapular retraction	1.09	1.0	0.86	1.19	1.9	0.84
	Pronation-supination elbow at 90	0.68	-4.0	0.89	0.72	-3.0	0.88
	Shoulder abduction to 90, elbow ext	0.77	-2.8	0.88	0.80	-2.3	0.86
	Palmar prehension	1.16	1.7	0.85	1.05	0.5	0.85
	Shoulder flexion to 90, elbow ext	0.86	-1.6	0.88	0.81	-2.0	0.86
	Shoulder abduction	1.00	0.1	0.88	1.05	0.5	0.85
	Elbow extension	0.76	-2.6	0.90	0.87	-1.2	0.87
	Forearm pronation	0.77	-2.6	0.90	1.15	1.3	0.83
	Hand to lumbar spine	0.83	-1.7	0.89	1.14	1.3	0.84
	Movement without tremor	1.62	5.3	0.83	1.32	2.9	0.82
	Finger mass extension	0.99	-0.1	0.89	0.88	-1.1	0.86
	Cylindrical grasp	1.06	0.6	0.88	0.91	-0.7	0.84
	Finger mass flexion	1.12	1.0	0.89	0.85	-1.3	0.86
	Scapular elevation	1.10	0.9	0.89	1.21	1.6	0.83
	Elbow flexion	0.94	-0.5	0.90	1.08	0.7	0.82
Easy	Sh adduction with internal rotation	0.79	-1.9	0.91	1.06	0.5	0.84

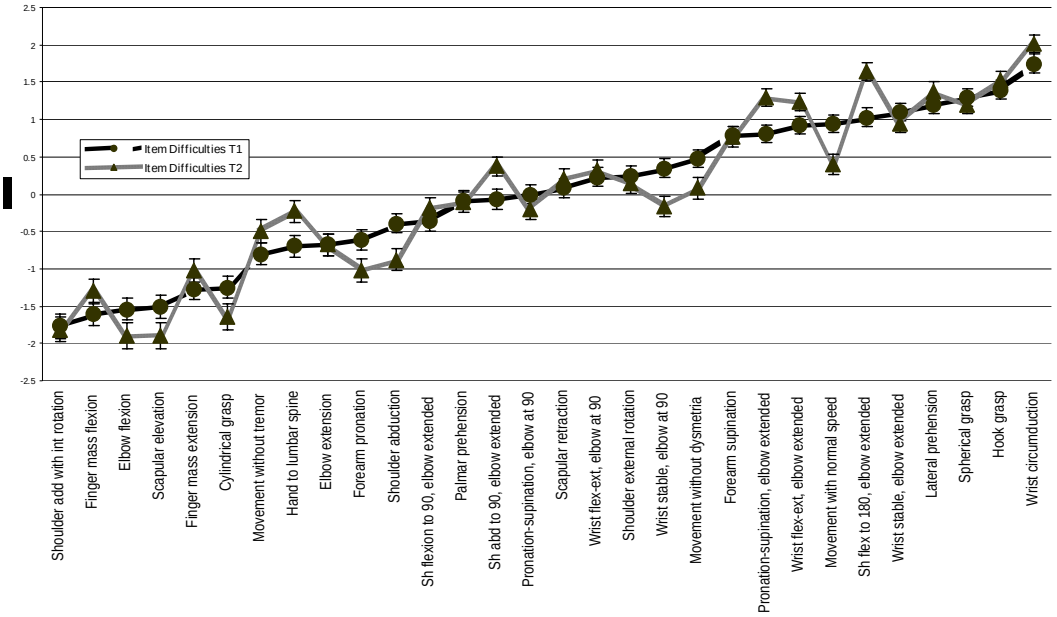

Figure 2: item difficulty measures of T1 (0 – 14 days post-stroke) and T2 (6 months post-stroke)

Table 12: Differential item functioning analysis (n= 377, df = 594, ** p<0.0017, * p<0.05)

	FMA-UE Item Description	T1 item difficulty (Logits)	T1 error	T2 item difficulty (Logits)	T2 error	Calculated t-value	Probability	
Hard	Wrist circumduction	1.76	0.12	1.97	0.12	-1.24	0.2155	
▲	Hook grasp	1.41	0.12	1.49	0.12	-0.48	0.6301	
	Lateral prehension	1.21	0.12	1.35	0.12	-0.80	0.4247	
	Spherical grasp	1.31	0.12	1.17	0.12	0.83	0.4054	
	Shoulder flexion to 180, elbow extended	1.04	0.12	1.61	0.12	-3.36	0.0008	**
	Wrist stable, elbow extended	1.11	0.12	0.92	0.13	1.07	0.2831	
	Wrist flexion-extension, elbow extended	0.93	0.12	1.20	0.12	-1.53	0.1259	
	Movement with normal speed	0.96	0.12	0.38	0.13	3.26	0.0012	**
	Pronation-supination, elbow extended	0.82	0.12	1.26	0.12	-2.56	0.0106	*
	Forearm supination	0.80	0.12	0.75	0.13	0.30	0.7654	
	Wrist stable, elbow at 90	0.36	0.13	-0.18	0.14	2.89	0.0040	*
	Movement without dysmetria	0.48	0.12	0.06	0.14	2.29	0.0226	*
	Wrist flexion-extension, elbow at 90	0.23	0.13	0.31	0.13	-0.45	0.6545	
	Shoulder external rotation	0.25	0.13	0.14	0.13	0.61	0.5445	
	Scapular retraction	0.09	0.13	0.19	0.13	-0.56	0.5753	
	Pronation-supination elbow at 90	0.00	0.13	-0.22	0.14	1.19	0.2363	
	Shoulder abduction to 90, elbow extended	-0.06	0.13	0.35	0.13	-2.23	0.0262	*
	Palmar prehension	-0.08	0.13	-0.13	0.14	0.24	0.8126	
	Shoulder flexion to 90, elbow extended	-0.36	0.13	-0.20	0.14	-0.80	0.4247	
	Shoulder abduction	-0.39	0.13	-0.88	0.15	2.40	0.0167	*
	Elbow extension	-0.69	0.14	-0.68	0.15	-0.06	0.9557	
	Forearm pronation	-0.61	0.14	-1.02	0.15	1.96	0.0499	
	Hand to lumbar spine	-0.71	0.14	-0.24	0.14	-2.36	0.0186	*
	Movement without tremor	-0.80	0.14	-0.51	0.14	-1.48	0.1396	
	Finger mass extension	-1.28	0.15	-1.02	0.15	-1.22	0.2213	
	Cylindrical grasp	-1.25	0.15	-1.62	0.16	1.67	0.0958	
	Finger mass flexion	-1.62	0.15	-1.28	0.16	-1.51	0.1315	
	Scapular elevation	-1.52	0.15	-1.88	0.17	1.55	0.1223	
	Elbow flexion	-1.55	0.15	-1.88	0.17	1.44	0.1493	
Easy	Shoulder adduction with internal rotation	-1.79	0.16	-1.79	0.17	0.02	0.9848	

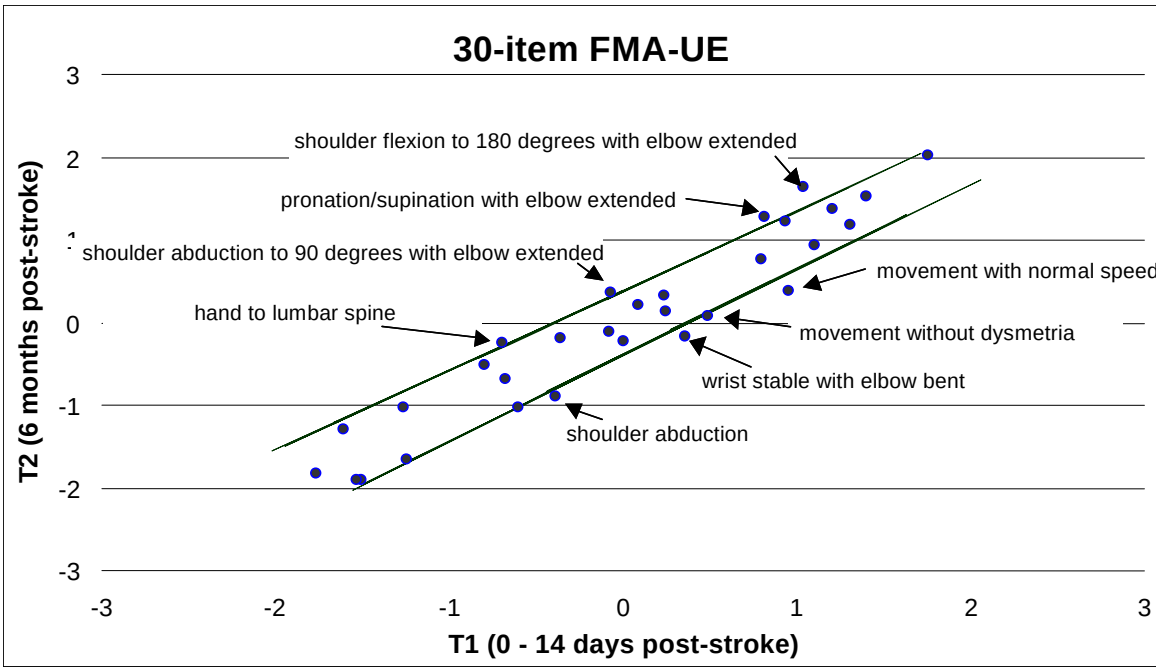

Figure 3: scatter plot of 30-item FMA-UE item difficulties at T1 versus T2

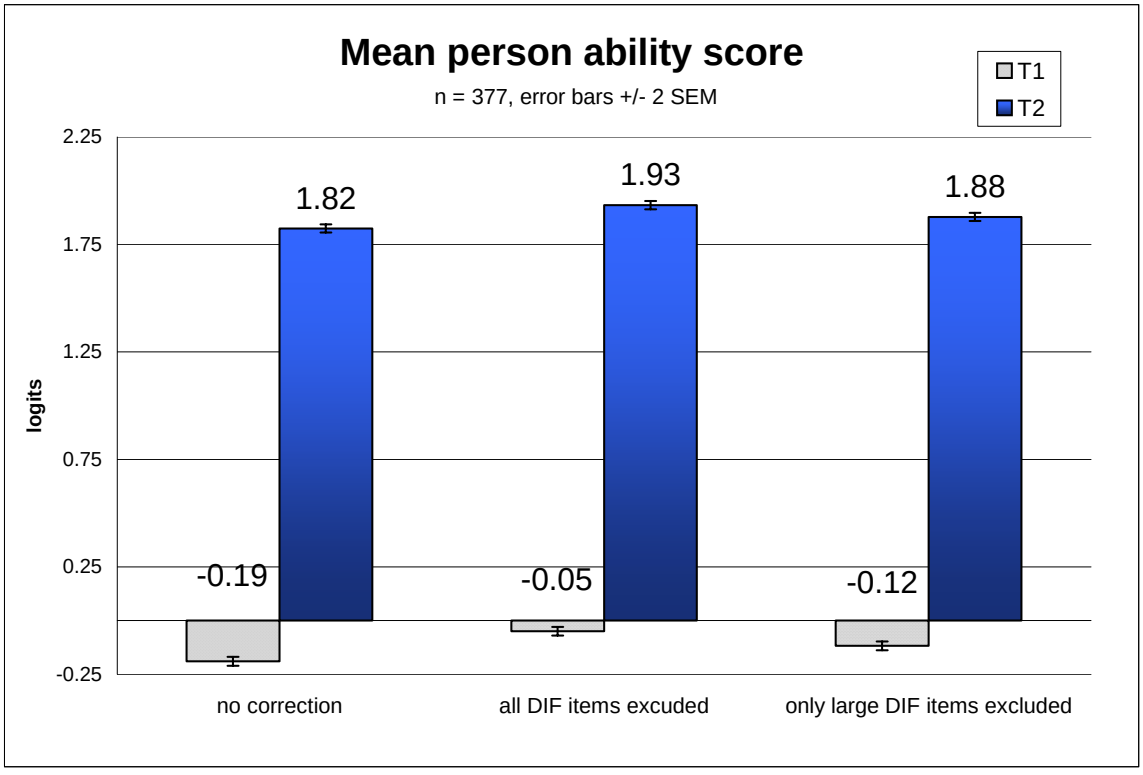

Figure 4: Comparison of mean person ability with and without DIF adjustment

CHAPTER 4
LINKING THE FUGL-MEYER TO THE WOLF MOTOR FUNCTION TEST: ASSESSING THE CONTINUUM OF IMPAIRMENT TO FUNCTION

Each year over 700,000 individuals in the United States have a stroke [1] with three-fourths of these individuals experiencing some degree of upper extremity (UE) paralysis [2]. Although UE motor function improves rapidly within the first 3-months post-stroke [3], it rarely recovers completely. At four years post-stroke 85% of survivors report residual arm or hand motor impairment severe enough to interfere with performance of daily tasks [4]. Clearly, there is an ongoing need to aim post-stroke rehabilitation interventions at reducing UE motor impairment. Furthermore, rehabilitation researchers must assure that interventions effective for reducing impairment also improve the client's ability to use his/her arm in "real life" daily activities.

The Fugl-Meyer Upper Extremity Assessment (FMA-UE) is the most widely used measure of post-stroke UE motor impairment [16]. The FMA-UE is the primary assessment for evaluating the success of post-stroke rehabilitation interventions such as rhythmic bilateral movement training [36, 37], neuromuscular stimulation [38], electromyographic triggered electrical stimulation [39, 40], Botox [41], home based exercise [30], community based exercise [42], robot aided therapy [43, 44], virtual reality [45], imagery [46], forced use [47], and modified constraint induced movement therapy [48-51]. However, one of the primary problems facing users of the FMA-UE is the difficulty determining the functional relevance of its impairment-level score [24, 51]. That is, its score has no obvious functional interpretation. What does it mean that a client

has shown a 10-point improvement in his/her FMA-UE score? Is this client better able to accomplish activities of daily living as a result of the intervention? Although the 10-point FMA-UE gain may be "statistically significant," what is the "clinical significance" of this change [136]?

Conceptually, there are two approaches for determining what constitutes "significant" change: distribution and anchor (or criterion) based approaches [61, 138]. "Distribution-based" methods define a "significant" change score as one that exceeds a pre-defined level of error [24, 138, 241, 242]. However, the results of distribution-based approaches depend on the characteristics of the sample. Therefore it is unlikely that there exists a single change score that defines "significant" change across all samples and contexts. On the other hand "anchor-based" methods relate a change score to a functionally relevant criterion pre-determined by client or clinician. This criterion "anchors" the meaning of the change. Both approaches have been applied to interpretation of FMA-UE change scores [24, 42, 242].

Feyes et al. (1998) [135] compared 50 individuals with stroke who received sensorimotor stimulation to the affected limb to 50 individuals with stroke who did not receive the intervention over the first year post-stroke. The FMA-UE was administered at 6 and 12 months. The researchers found that the FMA-UE change score was approximately 10% greater for those who received intervention compared to those who did not receive intervention. The researchers suggested the magnitude of the difference between the mean change scores represented the distinction between more impaired and less impaired stages of recovery, and therefore was a "clinically relevant" difference. Van der Lee et. al. (1999) [47] found no literature guiding the specification of a FMA-UE

"minimal clinically important difference", but noted that a 10% change in the overall FMA-UE score was of a magnitude that made measurement error unlikely [24]. Using Van der Lee's criteria, McCombe-Waller and Whitall (2005) [139] interpreted their 8% gain in FMA-UE score as "approaching clinical relevance". Pang et al. (2006) [42] took a criterion-based rather than distribution based approach. In this study the researchers simultaneously administered the FMA-UE and a client-satisfaction survey to 63 participants in a community based exercise program. Following the intervention, treatment effect sizes (assessed by the FMA-UE) ranged from 0.71 – 0.81, and participants reported increased ability to perform simple functional tasks such as "now able to switch on/off lights". Pang concluded that the treatment effects appear to be "clinically significant" from the patient's perspective.

Item Response Theory offers another approach to anchor-based interpretation of the FMA-UE change score. IRT provides the theory and method for linking the FMA-UE to a more functionally oriented assessment, thereby embedding functional criterion into the FMA-UE structure [143, 191, 192]. Linking is a method that re-calibrates separate assessments onto a common measurement scale. The theory behind this process has been termed "scale free measurement."

Scale free measurement means that what one measures is independent of the instrument used to make the measurement. Applied to post-stroke UE rehabilitation, scale-free measurement theorizes that an abstract trait (e.g., UE motor ability) can be measured by an "infinite" number of items representing the trait. Any two assessments are simply sub-samples of this item pool and therefore can be mathematically linked.

Linking establishes the relationship between different assessments therefore allowing direct comparison of items and scores [167].

The Wolf Motor Function Test (WMFT) [26, 201] is a logical choice of assessments to link with the FMA-UE. The World Health Organization's International Classification of Functioning and Disability (ICF) [62] provides a conceptual framework to define UE motor behaviors characterizing a specific health domain and provides a framework for measuring UE motor behaviors across health domains [199]. According to the ICF, body function includes "control and coordination of single and multiple joint movements." The FMA-UE is an assessment of UE body function impairment [14]. The first seven items of the WMFT may also represent this domain because these items require simple limb movements without functional endpoint (e.g., *place forearm on table, place hand on table*) and appear to be logical extensions of FMA-UE items. For example, the FMA-UE requires an individual to perform "shoulder abduction to 90 degrees with elbow extended" and "flex shoulder to 90 degrees with elbow extended". The WMFT requires an individual to "extend the elbow to the side by sliding it across a table top" and "lift the hand from a table to place it on a box placed at shoulder height." The items appear to require similar UE motor ability. The ICF's activity performance domain includes "reaching, grasping, and lifting". The last eight WMFT items represent this domain because the items require performance of functional tasks (e.g., lift soda can, lift a pencil) all of which require reaching to, grasping and lifting an object. These items can be viewed as logical extensions of the FMA-UE "hand" items. For example, the FMA-UE requires an individual to perform "cylindrical grasp" (i.e., maintain the shape the hand around a small can against minimal resistance), and "palmar prehension" (i.e.,

maintain one's hold on a pencil against minimal resistance). The WMFT items are functional applications of these hand positions requiring the individual to "lift a soda can" and "lift a pencil." Again, the FMA-UE and WMFT items appear to require similar UE motor ability. Therefore, it is reasonable to postulate that the two assessments measure a common underlying UE motor ability trait.

If the FMA-UE and WMFT are both measures of UE motor ability they can be linked to a common scale thereby creating a single item hierarchy comprised of impairment-level FMA-UE items and more functional-level WMFT items. Linking would provide a more comprehensive range of behaviors evidencing post-stroke UE motor recovery. IRT, specifically, Rasch analysis places person measures and item difficulty on the same linear continuum. Therefore, a FMA-UE score (measure) can be connected to the likelihood of successfully achieving (or not achieving) particular items on the assessment. Furthermore, if the impairment-level items of the FMA-UE are linked to the functional-level items of the WMFT, the combined assessment allows the score (measure) to be connected to the likelihood of successfully performing non-functional and/or functional tasks

The overall purpose of this study is to link the FMA-UE and WMFT using Rasch analysis. To do so, we will test the hypothesis that the FMA-UE and the WMFT measure the same UE motor ability construct. We will also test the hypothesis that there is a "threshold" value at which the impairment-level FMA-UE score transitions to the more functional-level WMFT score.

Methods

Participants

Secondary analysis of data from 100 persons enrolled in a randomized clinical trial of therapeutic exercise was performed [220]. Persons with stroke were included in the exercise study if they met the following inclusion criteria; (1) stroke within 30 to 50 days, (2) ability to ambulate 25 ft independently, (3) mild to moderate stroke deficits defined by a total Fugl-Meyer score of 27-90 (upper and lower extremities), (4) Orpington Prognostic Scale score of 2.0 to 5.2, (5) palpable wrist extension on the involved side, and (6) Folstein Mini-Mental Status exam score > 16 [220]. The FMA-UE and WMFT were administered to the sample prior to and immediately following the three month, in-home, progressive exercise intervention targeting balance, flexibility and strength (the WMFT was known as the Emory motor assessment [200] at the time of the exercise study). The assessments were administered according to a standardized procedure by trained staff. Because the purpose of the present study was ultimately to explore the usefulness of linking the two assessments for interpreting change scores, we included only participants who had completed both assessments pre (T1) and post (T2) intervention. This resulted in the exclusion of 9 individuals who were lost to follow up, or failed to complete one assessment. Thus the final sample size for the present study was 91 participants.

Instruments

The WMFT is comprised of two scales, the Functional Abilities Scale and performance time scale. The Functional Abilities Scale is an observational rating scale in which the test administrator rates the patient's quality of movement on a 6-point ordinal rating scale ("0" patient does not attempt to use more affected arm for task to "5" affected

arm is used for the task and its movement appears normal). Performance time is the number of seconds, recorded with a hand-held stopwatch, required by the client to complete each of the 15 WMFT tasks. If the client is unable to complete a task within two minutes, the task is considered unachievable and a score of 120 seconds is given. The final time score reported is the median time required for all timed tasks as per WMFT protocol [201]. A lower score (faster performance) indicates more UE motor ability. The WMFT has been shown to have high interrater reliability (intraclass correlation coefficients, ≥ 0.88), internal consistency (Cronbach $\alpha \geq 0.86$), and test-retest reliability ($r \geq 0.90$) [26, 201]. The Functional Abilities Scale rating scale data was not available, so for the purposes of the present study we used the WMFT performance time data.

To facilitate comparison of the FMA-UE and WMFT with Rasch analysis, the WMFT performance time data was re-coded into a rating scale. The overall intention of the procedure used to recode the data was to capture the connection between performance time (seconds) and UE motor ability thereby preserving as much information contained in the data as possible. Participants' performance time data was divided into performance categories, which were ordered to reflect increasing amounts of UE motor ability (slower time = less ability). Each category was assigned a rating scale rank (lower value = less ability). Category boundaries were established using a deductive process in which cut-off times were checked against criteria, and if the criterion was not achieved, other cut-off times were selected. This procedure was continued until the criteria were achieved [243-246].

The intent of the criteria was to insure that the rating-scale selected preserved the measurement characteristics of the original time data. The following criteria is based on those outlined by Linacre (2002) [247] and Hands (1999) [244]:

1. At least 10 observations per rating scale category for all items to assure the stability of item difficulty and person ability parameters.
2. Monotonically increasing item step calibrations to assure that increasing performance on each item is reflected as increasing progression thorough the rating scale steps.
3. Person separation index value > 2.0. A person separation index describes how well the instrument differentiates individuals in the sample. It is desirable to have a high person separation index (persons "spread out" along the measurement scale) because the intention of the WMFT is to discriminate between persons of various levels of UE motor ability.
4. A strong correlation between estimates of UE motor ability measured with the traditional WMFT (summed time to perform each item) and the Rasch person measures (logits) that were derived using the converted-WMFT data. This step assures that the score conversion captures the information originally obtained using the WMFT performance times.

Data Analysis

Once the WMFT data was recoded, a six-stage methodology was employed. We tested the WMFT for unidimensionality (step 1) and longitudinal stability (step 2). Once the validity of the WMFT item content and structure was established, we combined the FMA-UE and WMFT into a common item pool and concurrently calibrated item and person parameters (step 3). The newly created FMA-UE + WMFT item hierarchy was tested for unidimensionality (step 4) and longitudinal stability (step 5). Finally, we explored interpretation of FMA-UE change scores using the linked assessment (step 6).

Step 1: Dimensionality

In order to create a legitimate measure, all items must contribute to the same construct [143]. The first step towards testing the dimensionality of the combined FMA-UE + WMFT item pool was to ascertain whether each assessment functioned as a unidimensional measure when examined independently. Once unidimensionality of the

separate assessments was established, the items were pooled and the dimensionality of the combined assessment was examined.

In previous studies, we established the unidimensionality and longitudinal stability of a modified 30-item FMA-UE. Therefore the initial focus of the present study was to establish the unidimensionality and longitudinal stability of the WMFT.

Principal Components Analysis (PCA)

A principal components analysis (PCA) was performed (SPSS v. 11) on categorized WMFT data collected at two time points: prior to the exercise intervention (T1) and following the exercise intervention (T2). The intent of the PCA was to reorganize the multivariate data (participants' WMFT ordinal item scores) into a limited number of components (factors) so that each component captured a substantial amount of the overall variance within the dataset. That is, the purpose of the PCA was to detect "obvious" factors within the WMFT. The WMFT is comprised of two types of UE motor tasks; single/multiple joint motions (items 1 – 7) and functional tasks (items 8 – 15) [26]. To date, there are no published studies that have examined whether the items represent a single trait or separate constructs. We hypothesized that the assessment items illustrated a single dimension: "UE motor ability". To test the hypothesis we examined PCA eigenvalues and factor loading statistics. We employed the Kaiser rule [157] as the criteria for retaining principal components, i.e., only the components with an eigenvalue > 1.0 were retained. We expected one eigenvalue would explain the majority of the variance in the data and considered a component to be a unique dimension if it contributed > 10% of explanatory variance [161, 248]. Furthermore we expected that the majority of WMFT assessment items would correlate with a single principal component as evidenced by factor loadings on the first component > 0.40 [219].

Rasch analysis

A Rasch partial credit analyses was applied to the WMFT data using the Winsteps software program [234]. The partial credit analysis allows items to have unique, rather than consistent, response structures. The partial credit model was considered the most appropriate for the data because the rating scale was uniquely defined for each item based on item-specific cut-off times.

Dimensionality of the data was further examined with Rasch-derived fit statistics. Under the Rasch model, each person with a given level of UE motor ability is expected to do well (score higher) on relatively easy tasks, and have more difficulty (progressively lower scores) on tasks of increasing difficulty. The model expectations are assessed through item goodness-of-fit statistics. Fit statistics are reported as the mean square (MnSq) of the item residuals, where a residual is the difference between a person's actual response to an assessment item and the response expected by the Rasch model. Consequently the MnSq statistic has an ideal value of 1.0 with departures from this value indicating potential departures from unidimensionality. Fit statistics are also reported in a standardized form (ZSTD), which is the transformation of the mean square of the residuals to a statistic with a z-distribution. Two types of fit statistics are calculated with Rasch analysis. The infit statistic is weighted according to the variance in response patterns (i.e., the pattern of ratings that the sample exhibited on a given item) on items with difficulties similar to the ability of the sample. The outfit statistic is an unweighted value based on the sum of the mean squares of the residuals and therefore is sensitive to items that depart from model expectations when the item is much harder or easier than a person's ability level (i.e., outliers). The acceptable criteria for unidimensionality depends on the intended purpose of the measure and the degree of rigor desired. For

surveys using ordinal rating scales, Wright and Linacre [169] suggest reasonable ranges of MnSq fit values between 0.5 and 1.7 associated with standardized z-values (ZSTD) less than 2.0 for clinical observations. This study focused on items with high infit statistics, i.e., those items that depart from model expectations with regards to person ability, because they pose the greatest threat to unidimensionality.

Item correlations

Two types of item correlations were calculated using the Winsteps program, Cronbach's alpha and point measure correlations. Cronbach's alpha is an index of internal consistency or homogeneity of an assessment [249]. When all items of an assessment measure a single attribute the items correlate with one another, i.e., the assessment is internally consistent. Nunnally (1978) suggests Cronbach's alpha values > 0.70 to evidence satisfactory internal consistency [250]. Point-measure item correlations (rPM) are the correlation of an item with the other items of the assessment, omitting that item. The value indicates the extent to which an item contributes to measurement of the assessment's overall construct [235]. Point-measure correlations were calculated in the same manner as point-biserial correlations, except that Rasch analysis replaced total raw scores with Rasch-derived parameters [165]. The rPM coefficients were interpreted in the traditional manner; coefficients below 0.50 suggested poor reliability, coefficients from 0.50 – 0.75 suggested moderate reliability, and values greater than 0.75 suggested good reliability [236].

Step 2: Longitudinal Stability

When an assessment is used to measure change, the longitudinal stability of the item hierarchy should be examined to determine whether it is valid to make comparisons of repeated-measures [180]. In other words, it is important to assure that all items

measure the same trait at each testing occasion. A robust longitudinal investigation of item difficulties is accomplished through Differential Item Functioning (DIF) analysis. DIF is a statistical procedure in which item difficulty estimates are examined, one at a time, to ascertain whether items display the same statistical characteristics irrespective of test-taker characteristics (e.g., gender, culture) or testing occasion. The presence of DIF may indicate multidimensionality or problems with the administration of the assessment [182].

Following a procedure that we had previously applied to the FMA-UE, we performed a DIF analysis of the WMFT with Winsteps [234]. The analysis calculated the differences in WMFT item difficulty estimates between T1 and T2. The significance of the differences was tested with iterative paired t-tests. We made a Bonferroni correction for multiple comparisons based on the number of items in the WMFT, thus setting the alpha value at 0.05/15 = 0.0033.

Step 3: Linking and Concurrent Calibration of the FMA-UE + WMFT Item Pool

The goal of this study was to develop a general UE motor ability scale, and specifically to relate the location of item content across the two assessments. After establishing the dimensionality and longitudinal stability of each assessment separately, we proceeded with linking the assessments by merging the FMA-UE and WMFT items into a common item pool. Fisher (1997) suggested that concurrent calibration of instruments with two types of rating scales is best accomplished by analyzing the data using two Rasch models simultaneously [197]. Following these guidelines, we concurrently applied a Rasch rating scale analysis (Groups = 1) to the FMA-UE items, and a Rasch partial credit analyses (Groups=0) to the WMFT items, combining the items of both instruments into a single analysis. Applying the two Rasch models to the data

allowed the items of the FMA-UE to share one rating scale structure, and the items of the WMFT to share another. The concurrent analysis simultaneously estimated item difficulty and person ability parameters in the overall subject and item pool so that the two assessments were linked to the same measurement scale.

Step 4: Dimensionality of the Common Item Pool

We tested the hypothesis that the item pool represented a single trait, "UE motor ability", using the same procedures described above. We applied a traditional PCA using the same PCA criteria (eigenvalues > 1.0, unique dimensions defined by contributing > 10% of explanatory variance [161, 248], and factor loadings > 0.40 [157]) as described above. In addition we conducted a factor rotation of the principal components to further explore our hypothesis. Factor rotation is a mathematical technique in which the multi-dimensional axes of initial PCA factor loadings (the number of axes matches the number of retained factors) are rotated to model the data. The rotation re-expressed the factors as a less complex structure to facilitate interpretation [164]. This procedure produced a factor matrix with factor loading values (equivalent to regression coefficients) indicating the "weight" given to the item when interpreting the factor [157]. We considered factor loadings meaningful if the absolute value exceeded 0.40 [219]. Finally, we conduced a Rasch analysis and examined Rasch-derived fit statistics with the same criteria for acceptable infit statistics (MnSq < 1.7 with ZSTD < 2.0 [169]) and examined item correlations.

Step 5: Longitudinal Stability of the Common Item Pool

We examined DIF statistics to test whether all items of the overall UE motor ability scale were longitudinally stable with regards to their item difficulty calibrations using a

corrected alpha value for multiple comparisons based on the number of items within the combined item pool (0.05/45 = 0.0011).

Step 6: Interpretation of Change Scores

Once the dimensionality and longitudinal stability of the item hierarchy derived with concurrent calibration was established we addressed the final hypothesis for this study; that there exists a "threshold" FMA-UE score at which changes in a FMA-UE score translate to changes in the WMFT score. To accomplish this, the combined assessment was plotted onto an item map. The item map displayed the items of each assessment along a common measurement metric. The FMA-UE + WMFT item map also displayed the relationship between the assessments with regards to the overlap in the range of the trait measured by each assessment. The boundaries of the range were determined by examining the extreme category estimates for the easiest and most difficult items of each assessment [148, 251]. A category estimate is the average ability measure for the individuals receiving either the lowest rating ("0" for the FMA-UE or "1" for the WMFT) on the easiest item, or the highest rating ("2" for the FMA-UE or "4" for the WMFT) on the hardest items. The Rasch analysis was used to determine these values for each assessment, and then the values were plotted onto the item map.

Functional threshold

The point at which the functional items of the WMFT (i.e., items 8 – 15) overlap the FMA-UE was defined as the "functional threshold". The functional threshold is the zone in which individuals have enough UE motor ability to be able to successfully perform the more functional items of the WMFT. The lower boundary of this zone was defined as the logit value corresponding to the lowest category estimate of the easiest WMFT functional item.

Functional interpretation of a FMA-UE change score

FMA-UE person ability measures were calculated in separate T1 and T2 analyses with FMA-UE item difficulties anchored to the values derived from the concurrent calibration. To explore interpretation of FMA-UE change scores, person ability measures at each time point were overlaid onto the item map. The item map displayed the functional threshold. Thus, the overlay of person ability measures onto this map visually illustrated the relationship between item difficulties and person abilities. This allowed for comparison of change scores between individuals with more severe versus less severe stroke with regards to the likelihood that individuals were able to perform UE movements (i.e., FMA-UE items) and/or simple UE functional tasks (i.e., WMFT items), i.e., gain enough UE motor ability to move past the "functional threshold".

Results

Participants

Characteristics of the 91-person sample are presented in Table 13. Participants were 69.4 ± 10.3 years of age with slightly more males than females represented in the sample. A majority of the participants were Caucasian. Participants were 76.1 ± 27.1 days post stroke at admission to the study with an approximately equal distribution of right and left hemispheric involvement. The majority of the sample had a minor (54.9%) or moderate (39.6%) stroke as defined by the Orpington Prognostic Score [252].

Conversion of WMFT Performance Time Data to Performance Categories

A rating scale should meaningfully separate participants into ability levels. The greater the number of rating scale categories created, the greater the potential for increasing the measurement precision of the scale which leads to finer discrimination between individuals [244]. However, if too many rating scale categories are created, it is

likely some ratings will not be well represented, i.e., the width of a particular rating category may be too narrow to correspond to a distinct participant ability level [253].

We compared two ways of creating rating scale performance categories from time data based on examples cited in the literature. We first divided each item into 5 performance categories ("1" = very severely impaired performance" to "5" = near normal performance) with cutoffs at the 20th, 40th, 60th, and 80th percentiles. Percentiles were calculated separately for each item. This percentile division was employed by Hands et al. (2001) [243] to convert children's motor performance data obtained from serial assessments each with distinct units of measurement (e.g., time to walk a distance, number of skips) to a common rating scale for Rasch analysis. We also recoded the WMFT time data into a 4-point performance category rating scale based on quartile divisions (25th, 50th, and 75th percentiles). Percentiles were calculated separately for each item. Williams (2005) [246] and Reeves (2004) [245] categorized continuous data (walking speed) and interval data (financial assets) into rating-scale performance categories with quartile divisions for Rasch analysis.

Rasch analysis was applied to each dataset and results were compared. We considered four criteria outlined by Linacre (2002) [247] and Hands (1999) [244] to determine which performance category rating scale was optimally functioning. Results from two items, *forearm to box* and *turn key*, were chosen to represent our reasoning process. These results are presented in Table 14. Results from the other 13 items are similar.

Criteria 1: at least 10 observations per rating scale category for all items. Table 14 presents two examples of the rating scale statistics, comparing results from the 4-point

scale (left side of the table) to results from the 5-point scale (right side of the table). The number of observations per rating scale category for *forearm to box* ranges from 20 – 23 for the 4-point scale, and 5 – 25 for the 5-point scale. The number of observations per category for *turn key* ranges from 17 – 31 for the 4-point scale, and 10 – 26 for the 5-point scale. The 5-point scale does not have the recommended number of observations per category in the rating scale category "1" for the item *forearm to box*, and barely meets the criteria in the rating scale category "2" for the item *turn key*. Inspection of other items (data not presented) reveals that 15/15 items met Linacre's criteria when the 4-point scale was employed, but only 4/15 items met Linacre's criteria when the 5-point scale was employed. Among the 11/15 items not meeting Linacre's criteria with the 5-point scale, the rating of either "1" or "2" (sometimes both) had less than 10 observations per category. This means that these rating scale categories, representing the least amount of UE motor ability (most severely impaired), were not well represented by the 5-point rating scale solution. As the inclusion criteria for the exercise study was mild to moderate stroke deficits defined by a total Fugl-Meyer score of 27-90 (upper and lower extremities), and Orpington Prognostic Scale score of 2.0 to 5.2 (mild to moderate impairment), poor representation of individuals with severe stroke in the sample is reflected in the lack of observations for the lower categories. The 4-point rating scale appears to meet the first criteria better than the 5-point scale because it has a wider range of cut off boundaries for the lower ability ratings.

Criteria 2: monotonically increasing item step calibrations. An item step calibration (also termed "item step threshold" [165]) can be conceptualized as the boundary between adjacent rating scale categories, i.e., the boundary between "1 = severely impaired motor

performance" and "2 = impaired motor performance". Table 14 presents the step thresholds for the two representative items. Step thresholds calculated with the 4-point rating scale increase monotonically for both items shown in the table. For example, under the 4-point scale the item *forearm to box* exhibits an increasing progression of thresholds (–0.44, 0.00, 0.44) as does the item turn key (-0.97, 0.11, 0.86). Under the 5-point scale the item *forearm to box* exhibits increasing steps from –2.92, -0.03, 1.11 to 1.84. However, the item *turn key* exhibits a disordered step threshold. The first step is –0.85 but the second step decreases to –1.07 before increasing to 0.65 and 1.26. All other items (data not presented) exhibit monotonically increasing step thresholds under the 4-point scale. *Turn key* is the only item with disordered thresholds under the 5-point scale.

Criteria 3: person separation index value > 2.0. The person separation index for the 4-point scale were 2.93 (reliability = 0.90), and for the 5-point scale was 3.78 (reliability = 0.93). The person separation values indicate that 4-point scale split the sample into 4 statistically distinct strata (as per the equation: strata = (4Gp + 1) / 3 where Gp is the person separation index value [254]), and the 5-point scale split the sample into 5 strata.

Criteria 4: strong correlation between estimates of UE motor ability measured with the traditional WMFT (summed time to perform each item) and the Rasch analysis (logits). Prior to calculating the correlation between the Rasch-derived ability estimates and the original performance time data, we examined the performance time data for outliers to identify any extreme scores that may distort the statistical analyses. An outlier was defined as a performance time greater than or equal to three standard deviations beyond the mean performance time for that item [236]. Because the outliers were actual true data (i.e., were not aberrant) all outliers were retained, and adjusted to one decimal

point outside the closest non-outlying score [246]. The correlation between the sum of each participant's performance time and the Rasch-derived ability measure showed an inverse relationship as expected because higher performance time means less UE motor ability and visa versa. Using the 4-point ordinal scale the absolute value of the correlation between WMFT time scores and Rasch-derived person ability measures was 0.77 indicating that the 4-point ordinal scale was strongly correlated with the time scores. Using the 5-point ordinal scale, this value dropped slightly to 0.71 but still reflected a strong correlation between the measures.

The intention of the conversion of the WMFT continuous time data to ordinal data was to preserve as much of the information contained in the continuous performance time data as possible with an optimal and effective rating scale. To that end, we compared two rating scales against quality control criteria. Both scales appeared to satisfy the majority of the criteria. However, the 4-point rating scale was chosen since it satisfied all of the criteria especially with regard to criteria 1.

Step 1: Dimensionality of the WMFT

Table 15 presents the results from the WMFT PCA at T1 (left) and T2 (right). According to the Kaiser rule [157], the analysis retained two factors with eigenvalues greater than one. Each analysis showed the first eigenvalue to large (8.28 T1, 8.54 T2) indicating that a single factor explains the majority of the variance in the data, 55% and 59% respectively. The second eigenvalue is small (1.54 T1, 1.12 T2) meaning that a second factor contributes little (10% and 7%) explanation of the data.

Table 16 presents the unrotated principal component factor pattern for each of the two retained factors at T1 (left) and T2 (right). In an unrotated factor pattern matrix, factor loadings are similar to correlation coefficients and are considered meaningful if the

value is > 0.40 [157]. At each time point, all items correlate strongly with one factor. First factor loadings range from 0.65 – 0.83 at T1 and 0.64 – 0.87 at T2.

Table 17 presents Rasch derived Wolf Motor Function Test item measures, fit statistics, and point measure correlations. Infit statistics for all items at each time point are below the cutoff suggested by Linacre and Wright [247] (MnSq values are < 1.7 with associated ZSTD values < 2.0).

The Cronbach's alpha values for the assessment were $\alpha = 0.94$ at T1, and $\alpha = 0.95$ at T2 suggesting the items have strong inter-correlations and the assessment is internally consistent. Item point measure correlations (rPM) (table 17) are strong; ranging from 0.63 – 0.78 at T1 and 0.64 – 0.80 at T2 suggesting that the items work together to measure a common trait.

In summary, the dimensionality tests (PCA, Rasch infit statistics, Cronbach's alpha and item point measure correlations) suggest that the WMFT is a unidimensional assessment.

Table 18 presents the Wolf Motor Function Test item difficulty hierarchy with items arranged in T1 difficulty order so that easy items are at the bottom and more challenging items at the top. Item difficulties spread only 0.50 logits at T1 (0.25 logits to –0.25 logits) and 0.22 logits at T2 (0.09 logits to –0.13 logits). In general, items representing simple limb movements without functional endpoint (e.g., *hand to table and forearm to table*) are easier than items representing functional activities requiring fine hand movements (e.g., *turn key and lift pencil*).

Step 2: Longitudinal stability of the WMFT

Table 19 presents the Wolf Motor Function Test Differential Item Function Analysis. From left to right, the table presents the T1 and T2 item difficulty measures and

standard error of the measure followed by t-test results. The alpha value for this analysis was based on a Bonferroni correction for 15 comparisons, thus the level of significance was determined to be $p<0.003$. No items display statistically significant differences in item difficulty, suggesting that the item difficulties remain consistent at each testing occasion.

In summary, the dimensionality and differential item functioning analyses suggest that the WMFT functions as a unidimensional, longitudinally stable assessment. The modified 30-item FMA-UE has been previously shown to have the similar measurement properties (see Chapter 3).

Step 3: Linking and Concurrent Calibration of the FMA-UE + WMFT

The overall goal of this study was to develop a general UE motor ability scale by combing the FMA-UE with the WMFT. To accomplish this, we ran a co-calibrated Rasch analysis, with the items for both instruments

Step 4: Dimensionality of the FMA-UE + WMFT Item Pool

A PCA was conducted on T1 and T2 measures derived from FMA-UE and WMFT items from the co-calibrated analysis (SPSS v. 11). Table 20 presents the eigenvalues of the FMA-UE + WMFT linked assessment PCA. Figure 5 presents the scree plots of the FMA-UE + WMFT linked assessment PCA. Using the Kaiser rule [157] (retain factors with eigenvalues > 1) the PCA retained 8 factors at T1 and 10 factors at T2. At each testing occasion the first eigenvalue was large (19.06 T1; 17.56 T2), the second eigenvalue was much smaller (3.21 T1; 3.29 T2), and subsequent eigenvalues were progressively smaller. The first factors accounted for 42% and 39% of the variance in the data at T1 and T2, respectively, while the second factors only contributed an additional 7% of explanatory variance at each time point.

Because the Kaiser rule suggested retaining so many factors, Cattell's scree test was employed [157, 158]. The scree plot (T1 to the left, T2 to the right of Figure 5) is a graph of all eigenvalues in their decreasing order. The plot looks like the side of a mountain, and "scree" refers to the debris fallen from a mountain and lying at its base. According to the scree test, factors are retained if their data points occur prior to the point at which the "mountain" ends and the "debris" begins. Eigenvalues of "scree" are trivial because of measurement error, only those that markedly deviate from a linear trend are valuable to the analysis [157]. According to this rule, 2 or possibly 3 factors should be retained in this analysis (See Figure 5).

Although we hypothesized that the item pool represents a single construct, "UE motor ability", it is possible that the item pool represents two constructs; one represented primarily by FMA-UE items another by WMFT items. The FMA-UE is comprised of a majority of items requiring the client to perform UE movements in free space, that is, achieve a limb configuration without a specific target endpoint or interaction with object. In contrast, the WMFT items require the client to perform a specific task or move to a specific target. Furthermore the WMFT places a speed demand on the client's performance (e.g., "move as fast as you can") whereas the FMA-UE places no such demand on the client. Therefore, to explore if the items split into assessment-specific dimensions, we retained two factors and performed a factor rotation (varimax) to facilitate factor interpretation.

Table 21 presents the FMA-UE + WMFT item pool rotated factor matrix. The items from the FMA-UE and WMFT are listed in the left column, FMA-UE items above, and WMFT below. Factor loadings for two factors at T1 are presented first (middle

columns), and factor loadings for two factors at T2 are presented next (right side columns). Strong factor loadings (values > 0.40 [157]) are shaded in grey. In general, FMA-UE items tend to load on factor 1 while WMFT items tend to load on factor 2. Seventy-three percent (22/20) FMA-UE items solely load onto factor 1 at T1, and 86 percent (26/30) at T2. Fifty-three percent of the WMFT items load solely on factor 2 at T1 and 100 percent at T2.

Two items in T1, scapular elevation, and movement without dysmetria, fail to load strongly onto any factor (items marked with a single asterisk). Twelve items at T1 and 4 items at T2 "cross-load" onto more than one factor (marked with a double asterisk). Cross-loaded items inter-correlate with the factors and are generally considered un-interpretable within the framework of one factor [255]. For example, the items shoulder flexion to 90 with elbow extended, and lift soda can load onto factor 1 (0.66, 0.43) and factor 2 (0.45, 0.68) at T1 and the items shoulder flexion to 180 with elbow extended and movement with normal speed load onto factor 1 (0.60, 0.71) and factor 2 (0.45, 0.42) at T2.

In summary, while the majority of the PCA variance for the combined FMA-UE and WMFA is accounted by a single factor, there are some indications that the two instruments may be measuring distinct UE motor ability constructs (see discussion of multidimensionality in the discussion section below). For purposes of this study, we chose to continue the analysis with the entire item pool.

Table 22 presents the FMA-UE + WMFT item measure, fit statistics, and point measure correlations. Only 1 item in the item pool (FMA-UE movement without dysmetria) showed infit values that are beyond the acceptable ranges described by Wright

and Linacre (1994) [169] at T1 and approach unacceptability at T2 (MnSq >1.7 with associated ZSTD values > 2.0, are shaded in grey). The results suggest that all remaining items fit the unidimensional expectations of the Rasch model at each time point.

Person separation was 5.02 T1 and 4.69 at T2 with person reliability at 0.96 at each time point, indicating that the sample was split into 4 statistically distinct strata [254].

The internal consistency of the pooled items within the overall UE motor ability scale was strong (Cronbach's $\alpha = 0.96$) at both time points suggesting the items measure a similar underlying construct and the item pool is internally valid [249]. Table 22 presents correlations. The item point-measure item correlation values (rPM) at T1 ranged from 0.22 – 0.76 and at T2 ranged from 0.19 – 0.79. Four items at T1 and 7 items at T2, all FMA-UE items, demonstrated poor rPM values (below 5.0). These values are shaded in grey in the columns labeled "rPM" in table 22. While a subset of the FMA-UE items showed poor rPM values, a majority of the pooled items (91 percent (41/45) at T1 and 84 percent (38/45) at T2) showed acceptable values suggesting that the majority of the items contribute to the measurement of a single overall construct of UE motor ability.

Step 5: Longitudinal Stability of the FMA-UE + WMFT Item Pool

Figure 6 presents the FMA-UE + WMFT differential item functioning scatter plot. The item difficulty parameters of the item pool calculated at T1 (x-axis) are cross-plotted with the item difficulties of the item pool calculated at T2 (y-axis). The solid lines represent the 95% confidence interval for these values. The items outside the error bands are labeled. Three of the 45 items fall outside the error interval (forearm supination, wrist stable with elbow extended, wrist stable with elbow bent). The items are below the error band indicating that the items were unexpectedly more difficult for the sample at T1 than at T2. However, no T1 to T2 comparisons of item difficulty were statistically significant

at a significance level corrected for multiple comparisons ($p<0.0011$) suggesting that the three items falling outside the error interval in figure 6 may do so because of chance (i.e., type I error, $p<0.05$).

In summary, the dimensional and longitudinal stability analysis results show that while the FMA-UE + WMFT combined item pool exhibits some degree of multidimensionality the results tend to support viewing the item pool as a unidimensional set of items representing a single underlying trait. Both the traditional PCA and PCA of the standardized residuals suggest the WMFT items measure a somewhat different trait than the FMA-UE. However, a single factor explained the vast majority of the variance in the data in both of the PCA analyses. Although one FMA-UE item demonstrated poor infit statistics at one time point (movement without dysmetria), this item did not misfit at T2, nor did it misfit in our previous Rasch analyses of the FMA-UE with a larger dataset. The rPM statistics show that 11 FMA-UE items display poor item measures. The DIF statistics suggest that three FMA-UE items display instability of item difficulties at the two testing occasions, however, in a previous analysis with a larger sample only one of these items (wrist stable with elbow bent) not stable. Finally, the Rasch analysis showed that all of the pooled items had acceptable infit statistics with strong item point measure correlations).

Step 6: Interpretation of Change Scores

The FMA-UE + WMFT items were combined into an overall UE motor ability measure. The item difficulty parameters were concurrently calibrated so that range of content of the two assessments could be directly compared. Figure 7 displays the FMA-UE + WMFT linked item map. FMA-UE items are located to the left of the diagram, WMFT are located to the right of the diagram. The center scale represents UE motor

ability, measured in logits, ordered from less ability (bottom of the diagram) to more ability (top of the diagram). The logit scale is an equal interval scale. Note that each item is located along the UE motor ability metric. Each item's location on the metric corresponds to the average UE motor ability required for an individual to have a 50% probability of successfully accomplishing that motor behavior. For example, an individual has a 50% probability of performing the WMFT item, hand to lumbar spine if he/she has –0.5 logits of UE motor ability, while 1.3 logits of UE motor ability is required for a person to have a 50% probability of successfully performing the FMA-UE item, picking up a pencil.

The FMA-UE + WMFT item map also illustrates the range of UE motor ability measured by each assessment. The shaded boxes surrounding each item set represent this range. The FMA-UE items span a wider range overall and are more evenly distributed along the measurement scale than the WMFT items. The low end of the range (bottom of the colored box) is calculated as the mean UE motor ability score for the individuals who received the lowest rating ("0" for the FMA-UE and "1" for the WMFT) on the easiest item on each assessment. The high end of the range (top of the colored box) is calculated as the mean UE motor ability score of the individuals who received the highest rating ("2" for the FMA-UE and "4" for the WMFT) on the hardest item on each assessment. The FMA-UE covers a range of UE motor ability content 4.80 logits wide with the lowest average measure for the easiest FMA-UE item (elbow flexion) being –2.69 logits while the highest average measure for the hardest item (shoulder flexion to 180 with elbow extended) being 2.11 logits. The WMFT covers a range of UE motor ability content 3.22 logits wide with the lowest average measure for the easiest WMFT item (hand to table)

being –0.98 logits while the highest average measure for the hardest WMFT item (turn key) being 2.24 logits. The upper half of the FMA-UE range overlaps with the full range of the WMFT items.

Person measures have a direct relationship to the item difficulties of the combined FMA-UE – WMFT instruments as illustrated by the person UE motor ability distribution curve that has been overlaid onto Figure 7. The peak of the distribution curve corresponds to the mean person-ability measure of the sample at T1. Items at located at an individual's ability measure reflect items that the individual has a 50% probability of being able to achieve. For example, a person with an UE motor ability measure of -1.0 logits has a 50% probability of being able to pronate and supinate his forearm with elbow bent. He likely can accomplish shoulder abduction and scapular elevation (-1.0 logits), but not is not likely to be able to place hand to lumbar spine (-0.5 logits), achieve hook grasp (0.5 logits) or lift a pencil (1.3 logits). A person with 1.3 logits of UE motor ability is likely to be able to perform all of these motor behaviors. An individual's change score can also be directly connected to a description of the UE motor behaviors that he can or cannot perform. For example, while a person with 0.0 logits of UE motor ability can perform palmar prehension (hold a pencil between the tips of his thumb and index finger against slight resistance) (0.0 logits), he is unlikely to be able to lift the pencil (1.3 logits). If this individual responds to an UE rehabilitation intervention as evidenced by a 1 logit increase in his UE motor ability score (from 0.0-1.0 logits), this he has likely gained the ability to pronate and supinate forearm with elbow extended (0.9 logits) and place hand to box (1.0 logits), but still cannot lift the pencil (1.3 logits). A further 0.05 logit increase

in his/her UE motor ability would make it likely that he/she can lift the pencil (1.3 logits) and turn a key (1.3 logits).

Functional threshold

We would expect that depending on the individual's initial UE motor ability we should see different interpretations of FMA-UE change scores. We hypothesized that there exists a "threshold" zone of UE motor ability corresponding to the point at which a FMA-UE score can be directly connected to the likelihood that the individual can successfully perform the WMFT functional items (items 8 – 15). This threshold zone is defined as the point at which the functional items of the WMFT overlap the FMA-UE and is represented by the shaded rectangular area of the UE motor ability logit ruler in Figure 7. The lower boundary of this zone was determined by calculating the lowest category estimate of the easiest WMFT item. The easiest functional WMFT item was stack checkers. Its lowest category estimate (i.e., the mean person score of the individuals who received a "1" on this item) was calculated by WINSTEPS as 0.45 logits. In essence, this overlapping area reflects the amount of UE motor ability (represented by the FMA-UE items) needed to perform simple functional UE motor tasks (represented by the WMFT).

Functional interpretation of a FMA-UE change score

Figure 8 presents FMA-UE person ability measures overlaid onto the FMA-UE + WMFT item map to illustrate how the "threshold" zone aids in interpretation of FMA-UE change scores. Figure 8 presents the UE motor ability logit ruler running vertically through the center of the diagram. The shaded box near the top of the UE motor ability logit ruler is the "functional threshold" (corresponding to the shaded area of the ruler in

Figure 7), that is, the range of UE motor ability measured by the WMFT functional items 8 - 15.

Rasch derived FMA-UE motor ability measures (i.e., person ability measures in logits) for the 91-person sample, are overlaid onto this item map. T1 FMA-UE ability measures are displayed to the left of the ruler, and T2 FMA-UE ability measures are displayed to the right of the ruler. The T1 and T2 measures of two individuals are highlighted. The reader's attention is drawn to individuals #49 and #29 whose identification codes are marked with a box. These individuals will be compared because while each individual demonstrates similar FMA-UE change scores, the meaning of the FMA-UE change score differs depending on the individual's T1 measure.

At T1 individual #49T1 (boxed identification code) demonstrated a FMA-UE ability measure of –2.39 logits. At T2 this individual demonstrated a FMA-UE ability measure of –0.55 logits (#49T2). Individual #49 experienced a 1.84-logit gain in UE motor ability; however, his/her T2 ability measure is not in the "functional threshold" zone (shaded area). Because his/her T2 ability measure does not match the difficulty of WMFT functional items, he/she is unlikely to be able to successfully perform these WMFT functional items. Therefore it is unlikely that he/she experienced functional benefits from the rehabilitation intervention.

In comparison, individual #29 (boxed identification code) experienced a similar FMA-UE change score but with a different functional interpretation. At T1 this individual #29 demonstrated a FMA-UE ability measure –0.64 logits, and at T2 a FMA-UE ability measure of 1.16 logits. Similarly to individual #49, he/she demonstrated a 1.80-logit gain in UE motor ability, however, by T2, individual #29 gained enough UE

motor ability to move into the "threshold" zone. His/her T2 UE motor ability matches the difficulty of WMFT functional items making it likely that he/she can successfully perform these "functional" items. In contrast to individual #49, individual #29 appears to have experienced functional benefits from the rehabilitation intervention.

In summary, although individuals #49 and #29 experienced nearly identical FMA-UE change scores, the individual beginning the intervention with more UE motor ability appeared to have experienced greater functional benefits of the intervention.

Discussion

The purpose of this study was to link the FMA-UE and WMFT using Rasch analysis. The results suggest that the two instruments can be conceptualized as measuring a single trait "UE motor ability." It was feasible to link the two assessments to a common measurement scale. While others [42] have related the FMA-UE score to external functional criterion, this study demonstrates a method to directly embed a functional criterion within the FMA-UE by linking the FMA-UE to the WMFT. This method is advantageous because it affords a connection between the FMA-UE score an its functional interpretation. Linking an impairment-level motor assessment with a more functionally-oriented motor assessment offers a novel approach to exploring two important issues facing rehabilitation scientists; 1) interpreting impairment-level change scores in terms of more functional-level gains, and 2) understanding the relationship between stroke-related UE motor impairment and activity performance.

Linking as a Method to Explore "Significant" Change

The linked item map provides an opportunity to consider the interaction between changes in individual ability and functional responses to specific items. The map shows that FMA-UE change scores for individuals at the lower end of the scale (those with less

ability) will have different meaning than change scores for individuals in the middle or at the top of the scale (those with more ability). Individuals with less ability will require larger increases in the FMA-UE score before functional gains are realized, and individuals with more ability will realize functional gains with smaller increases in the FMA-UE score. While traditional methods for defining responsiveness (i.e., "distribution-based approaches" seek to define clinically significant change in terms of a single change score value [24, 138, 241, 242], the FMA-UE + WMFT item map demonstrates that a single value will not adequately capture the "significance" of change for all individuals.

Linking as a Method to Measure Body Function and Activity Performance

The item map may be useful for testing specific hypotheses regarding measurement of impairment-level and/or activity-level motor behaviors. For example, the creators of the WMFT hypothesized that the assessment is most appropriate for individuals with mild and moderate stroke related UE motor impairment [26, 200]. Our results not only support this hypothesis but also point to a specific "quantity" of UE motor ability that is necessary to be successful with WMFT items. Viewing FMA-UE change scores (Rasch-derived person ability measures) with regards to the "threshold" marked by the UE motor ability necessary to successfully perform functional WMFT items illustrates that individuals with less post-stroke UE motor ability will require larger gains in motor ability before gaining functional motor skills when compared to those individuals with more post-stroke UE motor ability. Pang (2006) [42] found that high baseline FMA-UE scores experienced greater ability to perform WMFT items when compared to those with lower baseline FMA-UE scores. The FMA-UE + WMFT item map offers a direct way to

examine the relationship between stroke-related UE motor impairment and performance of specific functional activities.

Linking as a Way to Conceptualize the Motor Rehabilitation Treatment Process

The Rasch-derived item difficulty hierarchy may illustrate the pattern of post-stroke recovery. Specifically, the hierarchy illustrates a progression of the reacquisition of single joint and multiple joint movements (FMA-UE items) to reacquisition of simple functional reaching, grasping and lifting tasks (WMFT items). This hierarchy is consistent with contemporary motor control expectations that arm movements are "easier "or "more challenging" based on the inherent task-specific neural, mechanical and contextual complexities of the movement [122, 126, 127, 256, 257] (see Chapter 2).

Contemporary motor control research suggests that task-specific increases in neural, mechanical and/or contextual requirements concomitantly increase the computational demands on the motor system [112]. This concept may explain why WMFT items are more difficult than FMA-UE items. First, the WMFT items require interaction with an object (e.g., table, box, soda can etc.). Interacting with an object not only increases the mechanical load on the limb but also increases the perceptual processing requirements of the task, e.g., attention to the characteristics of the object [126], feedforward adjustments of posture [127] and increased reliance on feedback to correct movement errors [70]. Second, WMFT items are considered "serial" rather than "discrete" movement tasks. A discrete movement is a "one step" motion with definable beginning and end points, whereas a serial task is comprised of an ordered series of discrete movements [66]. A serial task may be more difficult than a discrete task because the motor program must blend successive movement components to accomplish the goal [122]. Third, the WMFT items place a speed requirement on the individual's

performance (“do this task as fast as you can”). An individual must make additional adaptations to his/her motor plan to meet this externally imposed demand on the information processing system [258].

The map may assist in conceptualizing a progression of intervention activities and treatment goals for the UE motor rehabilitation process. For example, the item map reveals that in general, FMA-UE movement items are less difficult than WMFT functional items. This may imply that simple UE movements without functional endpoint may be appropriate treatment activities and/or treatment goals for those with more severe post-stroke UE motor impairment or for those earlier in the recovery process. These individuals may benefit from motor experiences in which the motor control system is challenged to attain/maintain various arm postures. Individuals with less severe post-stroke UE motor impairment (i.e., those with more UE motor ability) may benefit from more functionally-oriented interventions and treatment goals. These individuals may benefit from treatment aimed at challenging the motor system to interact with objects, perform serial tasks and move with varying speeds.

Limitations of this Study: Conversion of Performance Time data to Ordinal Data

It is possible that the precision of the WMFT performance time data was compromised during the conversion to the ordinal rating scale, although the correlations between the scores produced by timed performance data and ordinal scales was $r = 0.77$. Conversion of continuous data to ordinal/categorical data is not uncommon in the research literature because it is sometimes advantageous to group patients based on an interval measurement with cut-off scores. For example, Hands (2001) [243] desired to create an assessment of children’s motor development comprised of milestone behaviors. These behaviors (i.e., the “items” on the new assessment) included skipping, running, and

hopping, however each item was scored with a unique unit of measurement (i.e., number of skips in a set time, time to run a set distance, number of hops for a set distance). To facilitate the comparison of the behaviors she created categories for each item, using percentile divisions of the item scores across participants, so that each child's item score (number of skips, running velocity etc.) fell into one of the categories. Each category was assigned an ordinal rank, which became the item score on the new assessment.

Similarly, Williams (2005) [246] sought to create an assessment to measure high level mobility of individuals with brain injury comprised of behaviors measured as counts and continuous data (e.g., number of stairs ascended/descended in a period of time, gait velocity over uneven surfaces etc.). In a procedure similar to Hands, he created categories for each item using percentile divisions of item scores across participants. The categories were assigned ordinal ranks, and Rasch analysis was applied to explore the psychometrics of the new assessment.

We followed the examples of Hands and Williams in converting the WMFT time data to performance categories. To assure that our "rating scale" was effective at capturing client UE motor ability, we compared my results at each stage of the process against quality-control criteria specified by Linacre (2002) [247] and Hands (1999) [244]. While this approach met these criteria, there is the possibility of at least some loss of the precision and psychometric properties of the timed scale when converting it to a rating scale.

Limitations of this Study: Sample Characteristics

The low categories (representing "less UE motor ability") were not well represented in the sample that, by design, only included individuals with mild and moderate stroke. This limitation was noticed when checking the number of observations

per category in the low categories of the 5-point scale. The present study would benefit from re-analysis with a sample including persons of more severe UE motor impairment. It would also be interesting to determine whether similar findings would be obtained if I used the WMFT Functional Abilities Scale rating scale data rather than the converted time data. Unfortunately, the rating scale was not available the datasets used for this study.

Limitations of this Study: Possible Multidimensionality of the Item Pool

It may at first appear contradictory that the PCA results suggested the FMA-UE + WMFT item pool exhibited multidimensionality; while the Rasch fit statistics suggested the item pool is unidimensional. IRT measurement models assume unidimensionality, and thus the co-calibration process is invalid if this assumption is violated [144]. Muraki and colleagues (2000) [167] note the difficulties meeting the unidimensionality assumption when IRT models are applied to performance assessments. Murakai argues that because performance assessments, by their very nature, require examinees to complete tasks that demonstrate their ability to apply their skills to a "real life" situation, the data gathered from the assessment is influenced by "construct-irrelevant" variance such as environmental effects, practice effects and examinee fatigue. According to the researchers, unidimensionality is violated to some extent in every dataset, and even more in data from performance assessments. However, according to Murakai, as long as the violations are not "extreme" IRT models appear to be robust. McHorney (2002) [248] suggested that if greater than 20% of the variance in the data is accounted for by the first factor in a PCA, unidimensionality is established. Finlayson and Mallinson (2005) [161] suggested that a factor must contribute a minimum of 10% of explanatory variance to be considered a separate factor in modeling the data. Our results (table 20) support the

unidimensionality of the item pool according to these suggestions. The first factor explained 42% of the variance in the data at T1 and 39% of the variance in the data at T2. The second factor contributed 7% of additional explanatory variance at each testing occasion, and subsequent factors contributed less than 7%. These findings appear to indicate "noise" in the data, but do not appear to support the presence of distinct multidimensional constructs.

Because it is likely that performance assessment date is multidimensional to some extent, various multidimensional IRT models have been proposed but are not widely used [167]. The findings from this study suggest that these models should be considered for future studies.

Limitations of this Study: Sample Size

The use of a 91 person sample size for this study may be criticized because of the commonly referenced "rule of thumb" specifying that a minimum of 200 subjects is necessary for Rasch analysis [203]. In Rasch measurement, the researcher seeks to obtain stable person, item and rating scale estimates. A stable parameter is one that has low standard error. In a Monte Carlo simulation study, Wang (2005) [208] found unstable item parameters at either the higher or lower end of the item difficulty hierarchy with sample sizes of 100 - 2000 individuals and an assessment with 10 to 60 items. Item parameters were stable with 100 - 150 participants and longer assessments with 30 – 40 items. Wang's study suggests that the present sample size of 91 participants and assessment lengths provide stable item parameter estimation by limiting the error with which the parameters are associated. The findings of this study should be empirically confirmed by conducting the study with larger sample sizes.

Conclusion

The use of IRT measurement methods to link existing motor function assessments is being increasingly applied to healthcare outcomes research [148, 191, 197, 198]. The conceptual basis of linking existing assessments is that items, although from different assessments, share common underlying motor behaviors, and thus are appropriate to be measured on a common metric [148, 192]. In this study we have shown that the FMA-UE and WMFT items require similar UE motor behaviors and represent a single UE motor ability construct. Linking assessments allows scores from one assessment to directly translate to a score on another assessment [197]. This method has been useful for tracking recovery across the continuum of care [148] and "extending" an assessment that had previously exhibited a ceiling effect [198]. We have suggested that linking the FMA-UE to the WMFT improves the functional interpretability of the FMA-UE change score, and may be a useful platform to explore recovery of impairment-level UE motor behaviors and functional performance. Specifically, the linked item map can show how changes in motor impairments, often the target of therapy, can translate to improvements at the functional level. Calibrating the FMA-UE and WMFT together onto the same measurement scale may be one of example of, as Fisher (1997) stated, "taking full advantage of scale-free measurement's possibilities" [191].

Table 13: Sample characteristics (n = 91)

Table 13: Characteristics of the 91 person sample		
Years of Age (Mean ± SD)		69.4 ± 10.3
Gender	Female	43
	Male	48
Race	White	74
	African-American	16
	Other	1
Stroke Type	Ischemic	82
	Hemorrhagic	9
Stroke Location	Right Hemisphere	44
	Left Hemisphere	47
Days Since Stroke (Mean ± SD)		76.1 ± 27.1
Stroke Severity	Minor (Orpington Prognostic Score <3.2)	50 (54.9%)
	Moderate (3.2 <= Orpington <= 5.2)	36 (39.6%)
	Severe (Orpington > 5.2)	2 (2.2%)
	Orpington score missing	3

Table 14: Example of rating scale statistics

	4-point scale			5-point scale		
Item description	Rating Scale category	Number of observations per category	Step Threshold	Rating Scale category	Number of observations per category	Step Threshold
Forearm to box	1	21	none	1	5	none
	2	22	-0.44	2	20	-2.92
	3	23	0.00	3	25	-0.03
	4	20	0.44	4	25	1.11
				5	25	1.84
Turn key	1	31	none	1	16	none
	2	19	-0.97	2	10	-0.85
	3	17	0.11	3	26	-1.07
	4	19	0.86	4	24	0.65
				5	24	1.26

Table 15: Principal components analysis eigenvalues of the Wolf Motor Function Test

	T1		T2	
Factor	Eigenvalue	Proportion of Variance	Eigenvalue	Proportion of Variance
1	8.28	55%	8.54	59%
2	1.54	10%	1.12	07%

Table 16: Principal component factor loadings

	T1	**T2**
Item Description	Factor 1	Factor 1
Forearm to table (side)	0.758	0.690
Forearm to box (side)	0.705	0.772
Extend elbow (side)	0.724	0.790
Extend elbow, push weight	0.650	0.701
Hand to table	0.821	0.867
Hand to box	0.816	0.816
Reach and retrieve	0.693	0.734
Lift soda can	0.831	0.831
Lift pencil	0.790	0.728
Lift paperclip	0.737	0.720
Stack checkers	0.755	0.750
Flip cards	0.684	0.735
Turn key	0.726	0.740
Fold towel	0.664	0.643
Lift basket	0.757	0.771

Table 17: Wolf Motor Function Test item measures, fit statistics, and point measure correlations.

	T1				T2			
WMFT Item Description	Difficulty measure (SE)	Infit MnSq	Infit ZSTD	rPM	Difficulty measure (SE)	Infit MnSq	Infit ZSTD	rPM
Forearm to table	-0.17 (0.14)	0.97	-0.2	0.71	-0.16(0.14)	1.24	1.6	0.67
Forearm to box	-0.09 (0.14)	1.14	1.0	0.69	-0.30(0.14)	0.94	-0.4	0.73
Extend elbow to side	-0.11 (0.14)	1.07	0.5	0.70	-0.08 (0.14)	0.89	0.7	0.74
Extend elbow, push weight	-0.11 (0.14)	1.32	2.1	0.63	-0.15 (0.14)	1.18	1.2	0.68
Hand to table	-0.25 (0.14)	0.77	-1.6	0.78	-0.13 (0.14)	0.63	-2.8	0.80
Hand to box	-0.06 (0.14)	0.79	-1.5	0.76	0.07 (0.14)	0.81	-1.4	0.76
Reach and retrieve	-0.15 (0.14)	1.17	1.2	0.67	-0.16 (0.14)	1.09	0.6	0.68
Lift soda can	0.10 (0.14)	0.69	-2.3	0.76	-0.03 (0.14)	0.75	-1.8	0.77
Lift pencil	0.24 (0.13)	0.83	-1.2	0.73	0.10 (0.14)	1.09	0.7	0.70
Lift paperclip	0.06 (0.14)	1.01	0.2	0.69	0.10 (0.14)	1.11	0.8	0.69
Stack checkers	0.02 (0.14)	0.96	-0.2	0.71	0.07 (0.14)	1.01	0.1	0.69
Flip cards	0.13 (0.14)	1.18	1.2	0.66	0.03 (0.14)	1.06	0.5	0.71
Turn key	0.24 (0.13)	1.04	0.3	0.67	0.10 (0.14)	1.04	0.4	0.71
Fold towel	0.12 (0.14)	1.23	1.5	0.65	0.10 (0.14)	1.37	2.3	0.64
Lift basket	0.03 (0.14)	0.96	-0.2	0.70	0.19 (0.14)	0.93	-0.4	0.71

Table 18: Wolf Motor Function Test item difficulty hierarchy

	Item Description	T1 Difficulty (logits)	SE	T2 Difficulty (logits)	SE
hard	Turn key	0.25	0.14	0.09	0.14
▲	Lift pencil	0.23	0.14	0.09	0.14
	Flip cards	0.10	0.14	0.03	0.14
	Fold towel	0.12	0.14	0.09	0.14
	Lift soda can	0.12	0.14	-0.03	0.14
	Lift paperclip	0.06	0.14	0.09	0.14
	Stack checkers	0.02	0.14	0.07	0.14
	Lift basket	0.00	0.14	0.18	0.14
	Hand to box	-0.07	0.14	0.07	0.14
	Forearm to box (side)	-0.09	0.14	-0.03	0.14
	Extend elbow (side)	-0.11	0.14	-0.07	0.14
	Extend elbow, push weight	-0.11	0.14	-0.13	0.14
	Reach and retrieve	-0.13	0.14	-0.15	0.14
	Forearm to table (side)	-0.13	0.14	-0.15	0.14
easy	Hand to table	-0.25	0.14	-0.13	0.14

Table 19: Wolf Motor Function Test differential item function analysis

	T1		T2		t		
	T1 item difficulty	SE	T2 item difficulty	SE	calculated t value	df	probability
Forearm to table	-0.13	0.14	-0.15	0.14	0.08	170	0.9385
Forearm to box	-0.09	0.14	-0.03	0.14	-0.32	170	0.7499
Extend elbow to side	-0.11	0.14	-0.07	0.14	-0.22	170	0.8262
Extend elbow push weight	-0.11	0.14	-0.13	0.14	0.08	170	0.9393
Hand to table	-0.25	0.14	-0.13	0.14	-0.61	170	0.5426
Hand to box	-0.07	0.14	0.07	0.14	-0.71	170	0.4763
Reach and retrieve	-0.13	0.14	-0.15	0.14	0.08	170	0.9385
Lift soda can	0.12	0.14	-0.03	0.14	0.76	170	0.4505
Lift pencil	0.23	0.14	0.09	0.14	0.75	170	0.4516
Lift paperclip	0.06	0.14	0.09	0.14	-0.13	170	0.8987
Stack checkers	0.02	0.14	0.07	0.14	-0.22	170	0.8224
Flip cards	0.10	0.14	0.03	0.14	0.36	170	0.7168
Turn key	0.25	0.14	0.09	0.14	0.85	170	0.3951
Fold towel	0.12	0.14	0.09	0.14	0.17	170	0.8682
Lift basket	0.00	0.14	0.18	0.14	-0.91	170	0.3625

Table 20: Eigenvalues of the FMA-UE + WMFT linked assessment

	T1			T2		
Factor	Eigenvalue	Proportion of Variance	Cumulative	Eigenvalue	Proportion of Variance	Cumulative
1	19.06	42%	42%	17.56	39%	39%
2	3.21	07%	49%	3.29	07%	46%
3	2.29	05%	55%	2.41	05%	51%
4	1.52	03%	58%	1.72	04%	56%
5	1.37	03%	61%	1.61	04%	59%
6	1.35	09%	64%	1.42	03%	62%
7	1.26	03%	67%	1.40	03%	65%
8	1.12	02%	69%	1.26	03%	68%
9	na	na	na	1.17	03%	71%
10	na	na	na	1.03	02%	73%

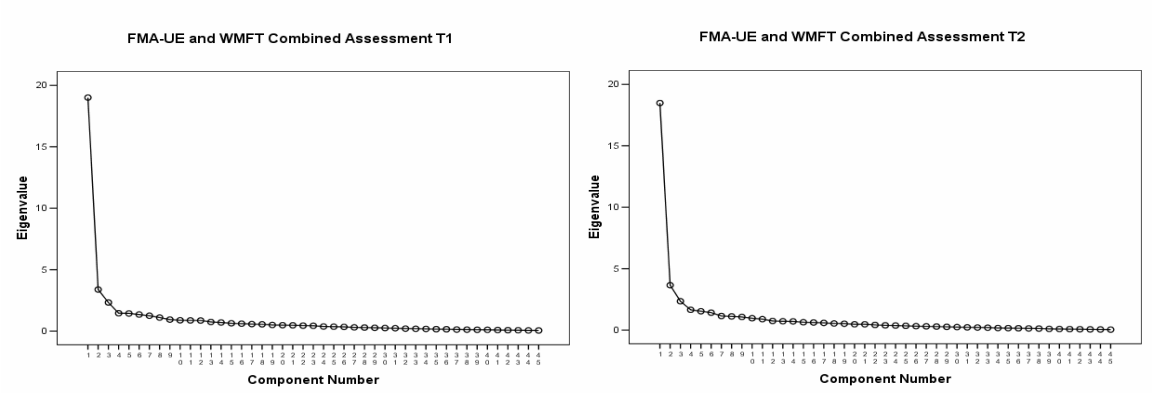

Figure 5: Scree plots of the FMA-UE + WMFT linked assessment

Table 21: FMA-UE + WMFT item pool rotated factor matrix

		T1 Rotated Factor Matrix		T2 Rotated Factor Matrix	
		Factor 1	Factor 2	Factor 1	Factor 2
Fugl-Meyer Upper Extremity Assessment Items	Scapular elevation	*0.20	*0.35	0.71	0.26
	Scapular retraction	0.12	0.44	0.54	0.11
	Shoulder abduction	0.29	0.56	0.68	0.23
	Shoulder external rotation	0.58	0.28	0.68	0.25
	Elbow flexion	0.57	0.26	0.53	-0.02
	Forearm supination	0.63	0.02	0.52	0.34
	Shoulder adduction, int. rotation	0.59	0.23	0.57	0.16
	Elbow extension	0.55	0.29	0.64	0.16
	Forearm pronation	0.72	0.32	0.61	0.24
	Hand to lumbar spine	0.49	0.37	0.63	0.15
	Shoulder flex. to 90, elbow ext.	**0.66	**0.45	**0.72	**0.41
	Pron-sup, elbow at 90	0.65	0.13	0.68	0.28
	Shoulder abd. to 90, elbow ext	**0.55	**0.44	0.72	0.29
	Shoulder flex to 180, elbow ext	**0.57	**0.40	**0.60	**0.45
	Pronation-supination, elbow ext	0.67	0.31	0.71	0.27
	Wrist stable, elbow at 90	0.63	0.33	0.59	0.29
	Wrist flex-ext, elbow at 90	0.62	0.29	0.66	0.21
	Wrist stable, elbow ext	0.70	0.38	0.71	0.38
	Wrist flex-ext, elbow ext	**0.49	**0.41	0.53	0.17
	Wrist circumduction	0.71	0.29	0.60	0.27
	Finger mass flexion	0.76	0.15	0.64	0.36
	Finger mass extension	0.77	0.18	0.64	0.31
	Hook grasp	0.70	0.19	**0.58	**0.43
	Lateral prehension	0.76	0.14	0.60	0.32
	Palmar prehension	0.70	0.17	0.56	0.37
	Cylindrical grasp	0.61	0.27	0.49	0.23
	Spherical grasp	0.74	0.05	0.54	0.23
	Movement without tremor	0.48	0.26	0.36	0.14
	Movement without dysmetria	*0.20	*0.03	0.30	-0.07
	Movement with normal speed	0.70	0.37	**0.71	**0.42
Wolf Motor Function Test Items	Forearm to table	0.16	0.77	0.14	0.68
	Forearm to box	0.11	0.79	0.15	0.77
	Extend elbow side	0.13	0.76	0.25	0.75
	Extend elbow push weight	0.09	0.72	0.10	0.71
	Hand to table	0.31	0.79	0.32	0.81
	Hand to box	0.28	0.80	0.24	0.78
	Reach and retrieve	0.06	0.76	0.19	0.70
	Lift soda can	**0.43	**0.68	0.35	0.75
	Lift pencil	**0.55	**0.53	0.33	0.65
	Lift paperclip	**0.56	**0.48	0.25	0.67
	Stack checkers	**0.49	**0.53	0.27	0.69
	Flip cards	0.62	0.38	0.34	0.66
	Turn key	**0.48	**0.49	0.22	0.71
	Fold towel	0.33	0.52	0.19	0.62
	Lift basket	**0.48	**0.58	0.23	0.74

grey shadings indicate strong factor loadings (value > 0.40), ** indicated cross-loadings,
* indicate a failure to load strongly on either factor, n = 91, varimax rotation

Table 22: FMA-UE + WMFT item measure, fit statistics, and point measure correlations

Item Description	T1 Item Difficulty	Error	Infit MNSQ	ZSTD	rPM	T2 Item Difficulty	Error	Infit MNSQ	ZSTD	rPM
FMA-UE shoulder flexion to 180, elbow extended	1.40	0.19	0.97	-0.1	0.67	1.29	0.18	0.85	-1.2	0.72
WMFT turn key	1.29	0.13	1.00	0.0	0.67	1.41	0.13	1.09	0.7	0.67
WMFT lift pencil	1.29	0.13	0.70	-2.3	0.75	1.41	0.13	1.02	0.2	0.70
WMFT flip cards	1.19	0.13	0.93	-0.5	0.71	1.36	0.13	0.98	-0.1	0.71
WMFT fold towel	1.18	0.13	1.29	2.0	0.61	1.41	0.13	1.36	2.3	0.60
WMFT lift soda can	1.16	0.13	0.70	-2.4	0.76	1.30	0.13	0.75	-1.9	0.77
WMFT lift paperclip	1.13	0.13	0.84	1.1	0.72	1.41	0.13	1.11	0.8	0.67
WMFT lift basket	1.10	0.13	0.87	-0.9	0.72	1.50	0.13	0.93	-0.5	0.69
WMFT stack checkers	1.09	0.13	0.90	-0.7	0.72	1.39	0.13	0.99	0.0	0.69
FMA-UE wrist circumduction	1.07	0.19	0.70	-2.4	0.71	1.08	0.19	0.74	-2.1	0.61
WMFT hand to box	1.02	0.13	0.88	-0.8	0.73	1.39	0.13	0.90	-0.7	0.74
WMFT forearm to box	1.00	0.13	1.28	1.9	0.62	1.30	0.13	1.07	0.5	0.67
WMFT extend elbow to side	0.98	0.13	1.28	1.9	0.62	1.26	0.13	0.94	-0.4	0.72
WMFT extend elbow, push weight	0.98	0.13	1.44	2.8	0.57	1.20	0.13	1.37	2.4	0.61
WMFT reach and retrieve	0.95	0.14	1.39	2.5	0.57	1.18	0.14	1.11	0.8	0.64
WMFT forearm to table	0.93	0.14	1.15	1.1	0.64	1.19	0.13	1.32	2.1	0.60
FMA-UE pronation-supination, elbow extended	0.89	0.19	0.98	-0.1	0.67	0.70	0.19	0.99	0.0	0.67
WMFT hand to table	0.85	0.13	0.87	-0.9	0.75	1.21	0.13	0.68	-2.5	0.79
FMA-UE wrist flexion-extension, elbow extended	0.82	0.19	0.72	-2.2	0.63	0.56	0.19	0.90	-0.8	0.49
FMA-UE hook grasp	0.53	0.19	1.41	2.8	0.64	0.24	0.19	1.30	2.1	0.69
FMA-UE spherical grasp	0.49	0.19	0.99	0.0	0.59	0.01	0.20	1.03	0.3	0.54
FMA-UE lateral prehension	0.31	0.19	1.08	0.6	0.65	0.74	0.19	0.78	-1.7	0.63
FMA-UE shoulder abduction to 90, elbow extended	0.24	0.19	0.97	-0.2	0.67	0.35	0.19	1.00	0.1	0.67
FMA-UE forearm supination	0.20	0.19	0.86	-1.0	0.47	-0.57	0.21	0.87	-0.8	0.60
FMA-UE movement with normal speed	0.17	0.19	0.87	-1.0	0.75	-0.03	0.20	0.90	-0.7	0.77
FMA-UE wrist stable, elbow extended	0.13	0.19	1.29	2.1	0.75	-0.70	0.22	1.19	1.2	0.74
FMA-UE palmar prehension	-0.02	0.19	0.94	-0.4	0.62	-0.31	0.20	0.83	-1.3	0.64
FMA-UE shoulder external rotation	-0.29	0.20	0.77	-1.9	0.63	-0.52	0.21	0.84	-1.1	0.64

Table 22 continued

FMA-UE movement without dysmetria	-0.29	0.20	1.75	4.7	0.22	-0.35	0.20	1.65	3.9	0.19
FMA-UE scapular retraction	-0.41	0.20	0.98	-0.1	0.39	-0.39	0.21	1.06	0.5	0.45
FMA-UE hand to lumbar spine	-0.45	0.20	0.84	-1.2	0.60	-0.19	0.20	0.89	-0.8	0.52
FMA-UE shoulder flexion to 90, elbow extended	-0.57	0.20	1.16	1.2	0.77	-0.48	0.21	1.10	0.7	0.76
FMA-UE wrist flexion-extension, elbow at 90	-0.65	0.20	0.69	-2.5	0.64	-0.85	0.22	0.78	-1.4	0.58
FMA-UE wrist stable, elbow at 90	-0.82	0.21	1.42	2.7	0.69	-2.01	0.28	1.25	1.1	0.63
FMA-UE elbow extension	-0.95	0.21	0.77	-1.7	0.57	-1.21	0.24	1.01	0.1	0.55
FMA-UE pronation-supination, elbow at 90	-1.00	0.21	0.87	-0.9	0.56	-1.21	0.24	0.83	-1.0	0.65
FMA-UE cylindrical grasp	-1.14	0.22	0.91	-0.5	0.63	-1.78	0.27	1.08	0.4	0.47
FMA-UE movement without tremor	-1.34	0.23	1.24	1.5	0.53	-1.10	0.23	1.33	1.9	0.38
FMA-UE finger mass flexion	-1.44	0.23	0.77	-1.5	0.63	-1.05	0.23	0.63	-2.5	0.67
FMA-UE shoulder abduction	-1.72	0.24	0.86	-0.8	0.59	-1.21	0.24	0.83	-1.0	0.62
FMA-UE scapular elevation	-1.78	0.24	1.55	2.8	0.38	-1.10	0.23	0.89	-0.6	0.65
FMA-UE finger mass extension	-1.84	0.25	0.72	-1.7	0.67	-1.93	0.28	0.69	-1.6	0.64
FMA-UE shoulder adduction with internal rotation	-1.90	0.25	0.89	-0.6	0.56	-2.17	0.30	1.11	0.5	0.48
FMA-UE forearm pronation	-2.16	0.27	0.62	-2.2	0.74	-2.36	0.31	0.79	-0.9	0.59
FMA-UE elbow flexion	-3.66	0.39	0.51	-1.9	0.63	-3.39	0.42	0.96	0.0	0.35

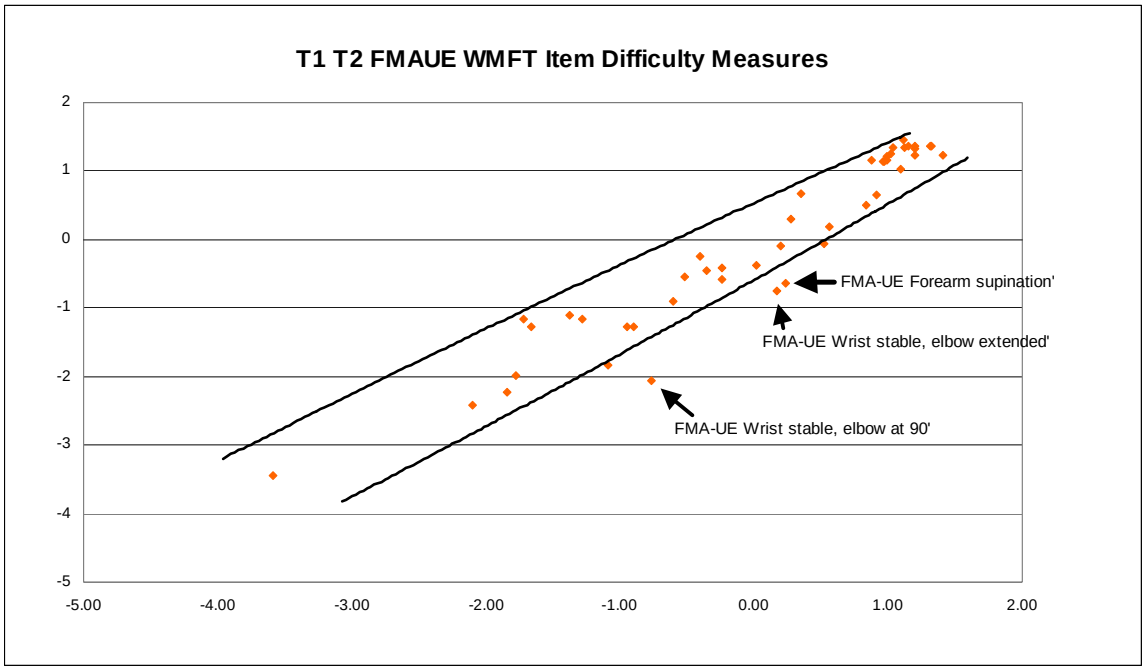

Figure 6: FMA-UE + WMFT differential item functioning scatter plot

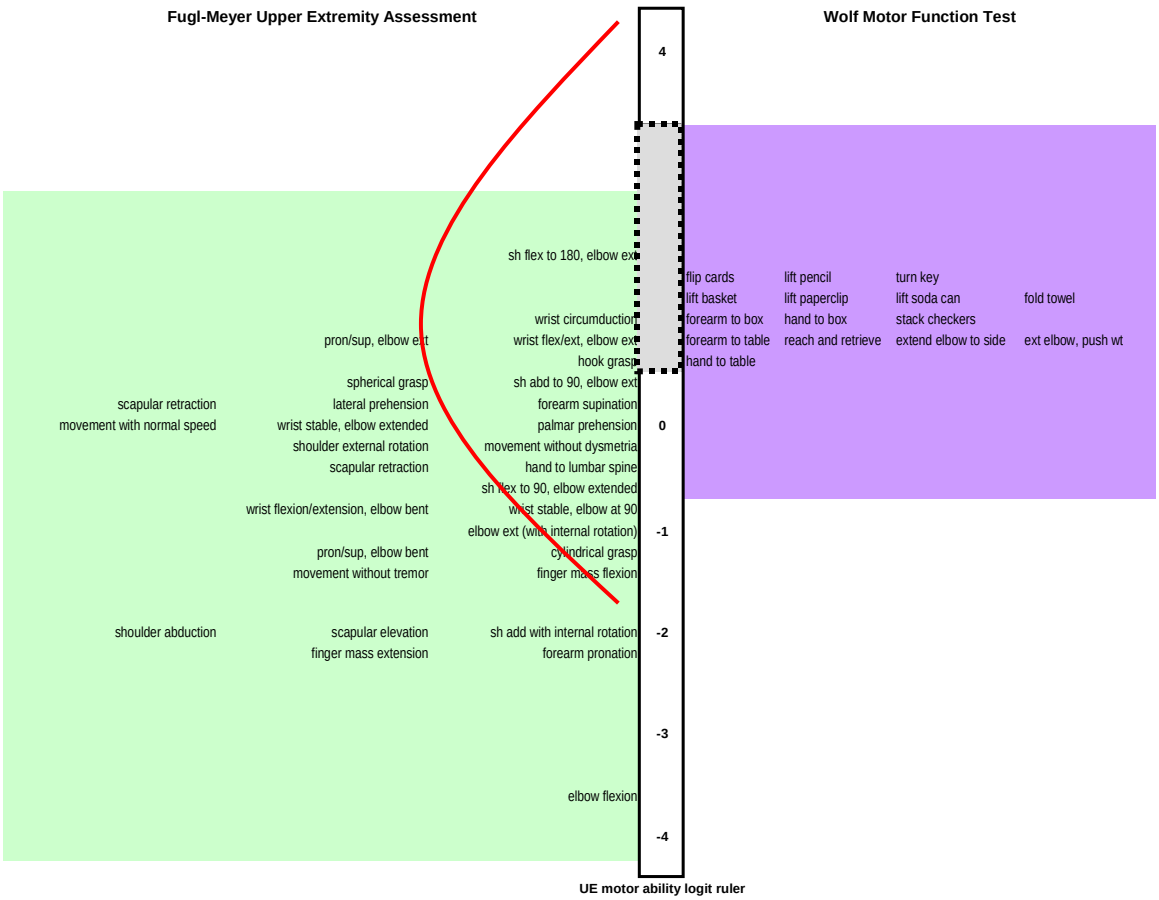

Figure 7: FMA-UE + WMFT linked item map

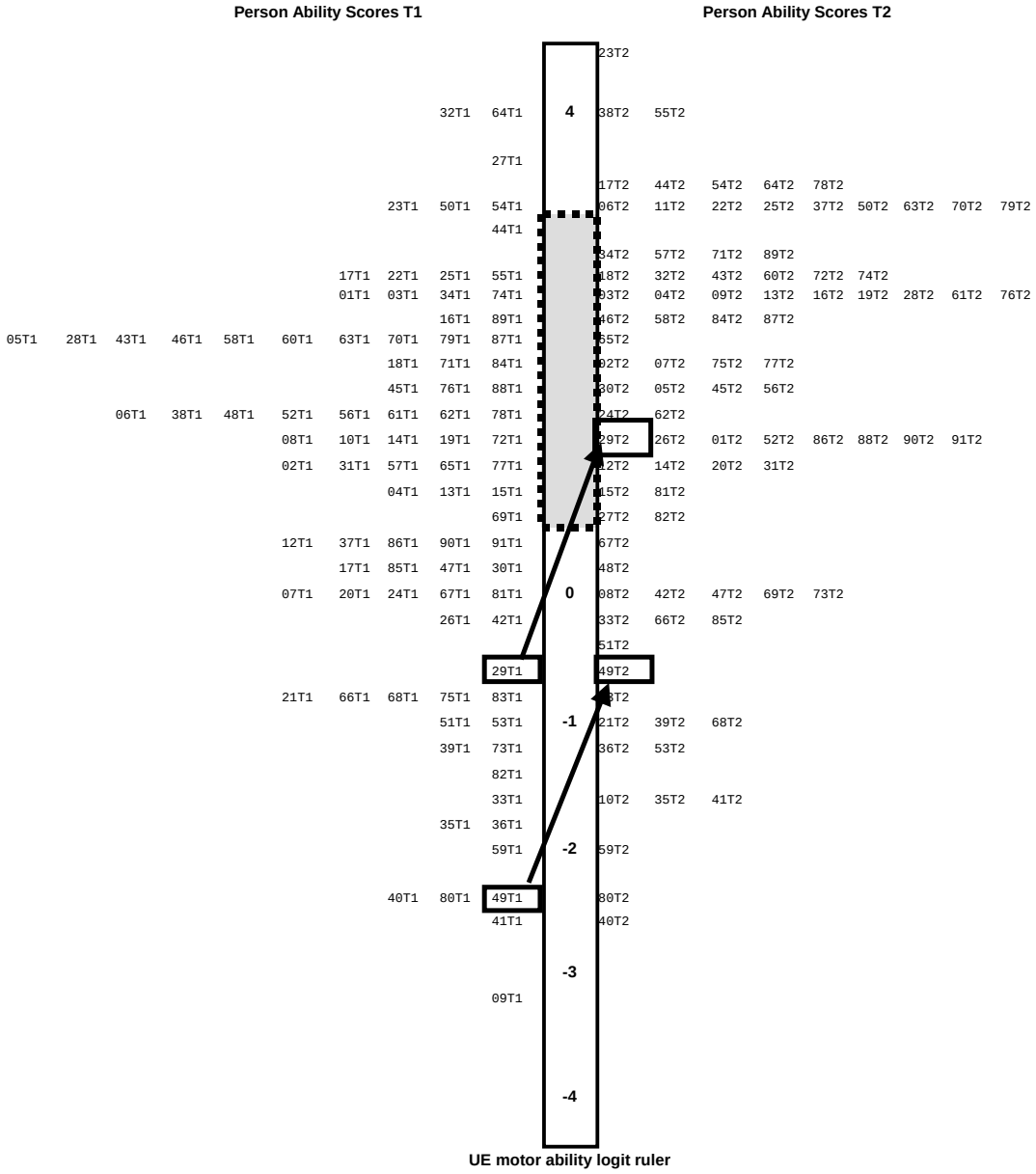

Figure 8: FMA-UE person ability measures overlaid onto the FMA-UE + WMFT item map

CHAPTER 5
A COMPARISON OF THE RASCH MODEL AND A TWO PARAMERTER ITEM RESPONSE MODEL TO DERIVE ITEM AND PERSON PARAMETERS

Item Response Theory (IRT) is a family of mathematical models each designed to describe the relationship between an individual's ability level and his/her pattern of responses to assessment items. All IRT models calculate one ability parameter for each individual, however the models vary with regards to the number of item parameters used to describe ability. The least complex, the one-parameter logistic models (1PL), express individual ability as a function of item difficulty. The Rasch model is typically considered a 1PL IRT model. The two-parameter logistic models (2PL) express individual ability as a function of item difficulty and item discrimination, and the three-parameter logistic models (3PL) express individual ability as a function of item difficulty, item discrimination and guessing [144]. The merits and pitfalls of each model are debated, sometimes passionately, in the literature [160, 209, 211, 259].

The question of which IRT model to select for any analysis is complex because one must first carefully consider how the assumptions of the IRT model match one's a-priori hypotheses regarding the characteristics of the given population (i.e., the data) [209, 212]. Models with additional parameters will more flexibility fit a greater variety of data [260]. The Rasch model is unique among IRT models because it assumes all items are equally discriminating whereas the 2PL and 3PL models accommodate varying discriminations. Item discrimination describes how well a test item differentiates between test-takers of higher and lower ability [144]. Items with high discrimination are thought to provide

more information about a test-taker than items with low discrimination [261]. Traub [262] stated that the a-priori assumption that an assessment's items all have equal discriminative ability "flies in the face of common sense and a wealth of empirical evidence accumulated over the last 80 years." Indeed, studies have shown that item discriminations vary greatly among items measuring activities of daily living [159], cognition [263], and self-esteem [261].

The evidence suggests that item discrimination values vary among items; therefore it is reasonable to assume that this is also true of FMA-UE items. In choosing the Rasch model for this dissertation, we did not account for the likelihood that FMA-UE item discriminations vary and so it is possible that we have inappropriately applied an IRT model to the dissertation data. Model misfit (i.e., choosing the wrong IRT model) is thought to negatively effect the estimation of person and item parameters and may invalidate an analysis [209, 260, 264].

The purpose of the present study is to test whether the choice of IRT model affected the estimation of FMA-UE item difficulty and person ability parameters as accomplished in the previous chapters of this dissertation. We will address this purpose by testing three hypotheses: (1) the 2PL is a better fit to the data than the Rasch model, (2) item discriminations vary among FMA-UE items, and (3) the choice of IRT model affects estimates of item difficulty and person ability. The hypotheses will be tested by comparing 2PL-derived item and person parameters to Rasch-derived item and person parameters. We will compare the item and person parameters derived from two datasets: the 512-person dataset (D512) utilized in the first study of the dissertation, and the 377 person dataset (D377) used as the post-test data in the second study of this dissertation.

The reason that the two datasets were chosen was to examine the person and item parameters influencing our conclusions regarding the cross sectional and longitudinal measurement properties of the FMA-UE.

Methods

Participants

Two separate 2PL analyses were conducted. The first analysis was applied to the pooled dataset described in study #1 of this dissertation. The sample consisted of 512 individuals. Please refer to study #1 for a more complete description of the sample. For the purposes of the present study I will refer to this dataset as D512.

The second analysis was applied to the sample who were administered the FMA-UE at the second time point (T2), 6 months post stroke, in the Kansas City Stroke study. This sample consisted of 377 individuals and is described in study #2 of this dissertation. For the purposes of the present study I will refer to this dataset as D377.

Dimensionality

Prior to conducting any IRT analysis a necessary first step is to determine if the data meet the unidimensional assumptions common to all IRT models. Because we have previously conducted analyses of unidimensionality (e.g., principal components analysis) of the 30-item FMA-UE we will not repeat the analyses here.

Preparing the data for 2PL analysis

The software for 2PL analysis (Multilog [265]) does not recognize “0” as an item score, instead treating the value as missing data. Therefore we recoded all FMA-UE data from both time points from a 0- 2 rating scale to a 1-3 rating scale. The data was recoded in SPSS (v. 11).

Model fit

The fit of an IRT model to the data was examined with a likelihood ratio test, which assumes that data are better modeled with more mathematical parameters. We followed a procedure outlined by McHorney and Kirisci to perform the likelihood ratio test [213, 248]. If the likelihood for a model with fewer parameters is not significantly different than the likelihood for a model with more parameters, then the model with fewer parameters is chosen. The likelihood ratio is a chi-square statistic and the significance of the ratio test was set at $p<0.01$.

2PL analysis

FMA-UE data were fit to a 2-parameter graded response model using the Multilog for Windows software program [265] to calculate item difficulty and item discrimination parameters. The graded response model is useful for analyzing responses in multiple ordered categories [266]. The data were subsequently used to create maximum likelihood estimates (MLE) of person ability, also conducted with Multilog.

Comparison of parameter estimates

We compared three parameters estimated with each model.

Item discrimination

The Rasch model software calculates an item discrimination index [148]. However, in contrast to the 2-parameter IRT model this index does not enter into person ability estimates. The index describes whether items are close to or depart from the ideal discrimination value of "1". In other words, the index reveals how consistent items are with the unidimensional expectations of the Rasch model. Low values indicate that an item fails to discriminate as well as other items. We compared item discrimination values as calculated by Rasch analysis to those calculated by the 2PL analysis.

Person ability and item difficulty

We compared estimates of person ability and item difficulty obtained with the 2PL analysis to the estimates of person ability and item difficulty that we had previously obtained with Rasch analysis. We utilized Pearson product moment correlations to determine the correlations between the estimates.

Multilog calculates parameters in a unit of measurement termed "probits" while Winsteps (Rasch analysis) calculates parameters in "logits." In general, 1 logit = 1.7 probits [234]. To facilitate comparison of the item difficulty hierarchies obtained with each model, we converted logit values to probit values using the above conversion ratio.

Results

Model Fit

Multilog was used to calculate the "negative two times log likelihood" values for the 1PL and 2PL logistic models. These values were 7645.6 and 9960.8 for the 1PL and 2PL models respectively. The likelihood ratio was 2315.2, the difference between the values. The 1PL analysis calculated 30 parameters (30 items in the assessment), the 2PL analysis calculated 60 parameters. The df for this analysis is equal to the difference between these parameters and therefore df = 30. The likelihood ratio is distributed as a chi square statistic, thus using the chi square table to calculate the critical value; the likelihood ratio of 2315.2 is significant at $p<0.01$ (the critical chi square value is equal to 50.89 at the 0.01 level). The result suggests the 2PL model fit these data better than the 1PL model.

Item Discrimination D512

The comparisons of Rasch-derived item discrimination values and those obtained from the 2PL analysis in the first dataset, D512, are presented in table 23. Item

discriminations calculated by Multilog ranged from 2.26 to 4.92. Item discriminations calculated by Winsteps ranged from 0.35 to 1.38. In general, the two models reveal the same information. That is, each item response model identified the same items as most and least discriminating. Specifically, wrist flexion-extension with the elbow at 90 degrees, shoulder flexion to 90 degrees with the elbow extended, wrist flexion-extension with elbow extended and pronation-supination with the elbow at 90 degrees are among the most discriminating items at in each dataset as identified by each model. Movement without dysmetria, hook grasp, and lateral prehension are among the least discriminating items. The item discrimination values calculated by each model at T1 were strongly correlated; $r = 0.88$.

D377 Item Discrimination

The comparisons of Rasch-derived item discrimination values and those obtained from the 2PL analysis in the second dataset, D377, are presented in Table 24. Item discriminations calculated by Multilog ranged from 6.13 to 2.66. Item discriminations calculated by Winsteps ranged from 1.31 to 0.50. As in the time 1 analysis above, the two models yield identical information. That is, the items the 2PL model identified as the most discriminating were the same items Rasch model identified as most discriminating. The items the 2PL identified as least discriminating were the same items the Rasch model identified as least discriminating. The item discrimination values calculated by each model were strongly correlated; $r = 0.89$. The item discrimination pattern for the 377-subject analysis is similar to that generated with the 512-subject analysis.

D512 Item Difficulty

Item difficulty parameters calculated by the 2PL analysis in the first dataset, D512, are compared to those obtained by Rasch analysis in table 25. The 2PL results are presented on the left side of the table, the Rasch results on the right side of the table. Moving from left to right, item descriptions are in the first column, the next two columns present “beta 1” and “beta 2” values. These values are item difficulty estimates at each adjacent response category boundary. Beta 1 is the boundary between “0” unable to perform and “1” partially able to perform, beta 2 is the boundary between “1” partially able to perform and “2” near normal performance. The fifth and sixth columns from the left report the average of the betas values and the average standard error of the difficulty measures SEM). The betas were averaged so as to more easily compare 2PL item difficulty estimates with the Rasch item difficulty estimates that are reported as an average value. The two columns to the right report the Rasch-derived item difficulty values with its SEM. 2PL values are reported in probits, Rasch values are reported in logits.

The item difficulty hierarchies are directly compared in figure 9. Item descriptions are located along the x-axis, item difficulty values (probits) are reported on the y-axis. Each item’s difficulty value as calculated by the 2PL analysis is shown as a circle, while its item difficulty value as calculated by Rasch is shown as a square. Rasch estimates have been converted to probits by multiplying the logit value (sixth column from the left in table 25) by 0.59. Error bars are reported as ± 2 SEM. Figure 9 shows that, in general, the hierarchies are identical. The values for 22 of the 30 items are identical within error. Eight items show different difficulty values. Linacre reports that the conversion ratio (1

logit = 1.7 probits) is an “approximate” relationship and is not expected to yield identical values [165].

Figure 10 presents a scatter plot of the 2PL item difficulty estimates (x-axis) plotted against Rasch item difficulty estimates (y-axis). The Pearson product moment correlation between the difficulty estimates is strong; r = 0.99.

D377 Item Difficulty

Item difficulty parameters calculated by each model in the second dataset, D377, are compared in table 26. As described above, the 2PL results are presented on the left side of the table, the Rasch results on the right side of the table. The item difficulty hierarchies are directly compared in figure 14. Item descriptions are located along the x-axis, item difficulty values (probits) are reported on the y-axis. Each item’s difficulty value as calculated by the 2PL analysis is shown as a circle, while its item difficulty value as calculated by Rasch is shown as a square. Rasch estimates have been converted to probits by multiplying the logit value (sixth column from the left in table 26) by 0.59. Error bars are reported as ± 2 SEM.

Figure 11 is a comparison of item hierarchies generated by Multilog and Winsteps in the D377 dataset. The figure shows that, in general, the hierarchies are identical. The values for 22 of the 30 (73%) items are identical within error.

Figure 12 presents a scatter plot of the 2PL item difficulty estimates (x-axis) plotted against Rasch item difficulty estimates (y-axis). The Pearson product moment correlation between the difficulty estimates is strong; r = 0.98.

D512 Person Ability

Figure 13 presents a scatter plot of person-ability measures derived with the 2PL analysis (x-axis) plotted against person-ability measures derived with the Rasch analysis (y-axis) for D512. The Pearson product moment correlation between the person-ability estimates is strong; r = 0.99.

Figure 14 presents a cross plot of person-ability measures derived with the 2PL analysis (x-axis) plotted against person-ability measures derived with the Rasch analysis (y-axis) for D377. The Pearson product moment correlation between the person-ability estimates is strong; r = 0.99.

Discussion

The purpose of this study was to test whether the choice of IRT model affected the estimation of FMA-UE item difficulty and person ability parameters. The results show that, although the 2PL model fit the data better than a 1PL model and FMA-UE item discrimination values varied, item difficulty and person ability estimates generated by the two IRT models were essentially identical in each dataset.

Which IRT Model, 2PL or Rasch, Best Fits the FMA-UE Data?

The likelihood ratio was calculated to determine which model, 2PL or 1PL, was the best "fit" to the data. This procedure is often utilized as a first step in IRT analysis [213, 248, 261, 267, 268]. The chi-square test revealed that the 2PL model better fits the FMA-UE data than the 1PL model. This result is not surprising. Adding parameters to a mathematical model increases its flexibility for describing variations in assessment data [144]. From this result it is inferred that FMA-UE items differ in their discriminability and that the more complex 2PL was necessary to adequately describe FMA-UE item characteristics [261].

Within Each Dataset, Do FMA-UE Item Discriminations Vary?

The results show that FMA-UE item discrimination values vary within each dataset. D512 FMA-UE item discriminations ranged from 4.92 to 2.26 probits and D377 item discriminations ranged from 6.13 to 2.66 probits, a 1.17% and 1.30% difference respectively (calculated as [maximum-minimum]/minimum). These results are consistent with others who have shown activities of daily living, cognitive, and personality items differ in their discriminability. McHorney and Cohen (2000) [159] showed a 10% variation in discrimination values among 71 ADL and IADL item (discriminations ranged from 0.31 to 3.46 probits). Teresi et al (1995) [263] showed an 8% variation in 50 items measuring cognition (discriminations ranged from 0.51 5.00 to probits), and Gray-Little et al. (1997) [261] showed a 1.32% variation in item discriminations among 30 items measuring self-esteem (discriminations ranged from 2.70 to 1.16 probits).

The Rasch model is criticized because it assumes equal item discrimination, constraining discriminations to a value of "1" when calculating item difficulty and person ability parameters [165]. Cook and colleagues (2003) [160] argued that the Rasch assumption does not "conform to reality". Indeed, the results of the present study show that FMA-UE items are not equally discriminating. Item discrimination indices calculated by Winsteps support this finding, indicating that D512 item discriminations ranged from 0.35 to 1.38 and D377 discriminations ranged from 0.50 to 1.31, a 2.94% and 1.62% variation respectively.

Perhaps the important question is not "do FMA-UE item discriminations vary?" The results of both the 2PL and Rasch analyses show that they do. Instead, the important question is, given that FMA-UE item discriminations vary, "does the choice of IRT

model affect the measurement properties of the FMA-UE with regards to estimation of item difficulty and person ability?"

Does the Choice of IRT Model Affect Either Item or Person Parameters?

The results show that within both datasets, the item hierarchies estimated by the 2PL and Rasch models are consistent and the item difficulty values (measured in probits) are nearly identical. Moreover, person ability measures were strongly, nearly perfectly correlated. The results show that the item and person parameters estimated by the Rasch model were robust to unequal FMA-UE item discriminations.

The results are consistent with Teresi et al. (1995) [263] who fit both a 1PL and 2PL model to her analysis of cognitive assessment among individuals of different ethnic groups and levels of education. Teresi showed an 8% variation in item discriminations, yet each model was robust with regards to item and person parameter estimation. Penfield (2004) [264] in simulation studies, found that when the data had small levels of variation in item discrimination about the value of 1.0 (as calculated by Rasch analysis) the person and item parameters were not impacted by the choice of Rasch partial credit model or 2PL graded response model. However, when the variation was large for all items of the scale, the person and item parameters became increasingly biased, meaning that it was difficult to reliably estimate the parameters. It is possible that the FMA-UE items vary so little with regards to their discrimination that the estimation of person and item parameters is not impacted.

What Does Item Discrimination Mean?

Item discrimination is conceptually related to item bias [269]. A non-biased item measures the same underlying attribute across subgroups. Item discrimination describes how well a test item differentiates between subgroups of test-takers: those with higher

ability and those with lower ability [144]. Item discrimination is not an index of item bias, however items with different magnitudes of discrimination relate to the underlying attribute differently and therefore might measure different constructs in examinees of different ability levels [144, 263].

McHorney and Cohen (2000) [159] suggest that poorly discriminating items inefficiently measure and yield less information than other items because factors other than ability will be affecting the persons response to the item. Gray-Little et al. (1997) [261] found that the poorly discriminating self-esteem items were the same items that the factor analysis revealed to be less correlated with the primary factor in describing the unidimensional underlying construct. Masters (1988) [269] argued that in some cases items with the highest discrimination values were biased. He argued that item discrimination interacts with dimensionality of the assessment, i.e., highly discriminating items are biased in that they do not measure the intended construct underlying the assessment. Masters makes the point that a researcher must carefully inspect item discrimination values to learn what the values reveal about the data rather than just accepting discrimination indices at face value, as the basis for making a subsequent judgment about the items usefulness. Masters' viewpoint touches upon the motivational differences that separate the Rasch from other IRT models.

The underlying conflict between the Rasch model and other IRT models has to do with views of how data provide information [211]. In short: should a measurement model explain data, or should a measurement model disclose information perhaps hidden in data? Andrich (2004) [211] frames this controversy using the language of mathematician-philosopher Thomas Kuhn as a clash of "incompatible paradigms".

Kuhn stated that the purpose of measurement is to disclose anomalies in data that cannot be explained by the theory underlying the measurement model with which the data is analyzed. With this approach one is able to ask questions as to why data does not fit a given theory thereby potentially learning new information that contributes to advancement of the theory. Kuhn's view is in stark contrast to traditional measurement theory that asserts data should fit a model so as to support the theory underlying the model. This issue is notably evident in controversies surrounding the Rasch model [211].

According to non-Rasch IRT models, data is best interpreted according to a mathematical model found to explain the largest amount of variance in the data. Models are tried and abandoned depending on how well the model explains the data. The Rasch model has the inverse assumption. That is, data should be viewed with the assumption that it either fits or does not fit the model. If it does not fit, then, rather than abandoning the model in favor of another, the data is examined for "anomalies" [211]. According to Andrich [211], the process of examining misfitting data (either items or persons) "uncovers" information that would have been hidden otherwise. New information may lead to advances in theory. From the perspective of the Rasch model, fitting a model to the data might mask important features in the data that might provide clues that inform researchers where to look so as to expand or revise an existing theory.

The overarching purpose of this dissertation was to examine the measurement properties of the FMA-UE. In the three studies preceding the present study we assumed an item-oriented viewpoint from which to investigate (1) the validity of the FMA-UE item content, (2) the reproducibility of its item structure, and (3) interpretation of its score. Throughout the studies we have explored the FMA-UE item-level construct

validity, which has served to focus attention on traditional and contemporary motor control theories. The intention of the dissertation has been to challenge and enhance the measurement properties of the FMA-UE by testing and increasing its match to contemporary theory. The Rasch model has been a valuable method for exploring the interdependence of measurement and theory. The present study demonstrates the robustness of the Rasch model for accomplishing the purpose of this dissertation.

Conclusion

Choosing the correct IRT model for a given analysis is said to be complex because the researcher must carefully consider how the assumptions of the IRT model describe the characteristics of the given population (i.e., the data) [209, 212]. Models with additional parameters, i.e., the 2PL or 3PL models, will more flexibility fit a greater variety of data [260]. The Rasch model assumes all assessment items are equally discriminating. In this study we have shown that the FMA-UE items are not equally discriminating. Yet, despite this violation of Rasch model assumptions, Rasch-derived FMA-UE item and person parameters are robust. The choice of the most parsimonious IRT model, the Rasch model, to conduct analyses for this dissertation is supported.

Table 23: Comparison of item discrimination values calculated by Multilog to those calculated by Winsteps in the first dataset

D512 Item Discrimination		
Item Description	**Multilog estimated Item Discrimination**	**Winsteps estimated item discrimination**
Elbow extension	4.92	1.24
Sh flexion to 90, elbow extended	4.86	1.21
Pronation-supination, elbow extended	4.57	1.38
Pronation-supination, elbow at 90	4.56	1.29
Wrist flex-ext, elbow at 90	4.35	1.31
Forearm pronation	4.30	1.21
Sh abd to 90, elbow extended	4.30	1.23
Shoulder adduction with internal rotation	4.05	1.08
Wrist stable, elbow extended	3.93	0.99
Shoulder external rotation	3.90	1.11
Wrist stable, elbow at 90	3.88	0.96
Sh flex to 180, elbow extended	3.82	1.21
Hand to lumbar spine	3.78	1.10
Finger mass extension	3.76	1.12
Wrist flex-ext, elbow extended	3.73	1.25
Elbow flexion	3.69	1.00
Forearm supination	3.67	1.08
Finger mass flexion	3.49	0.92
Shoulder abduction	3.41	0.98
Cylindrical grasp	3.26	0.98
Wrist circumduction	3.21	1.11
Movement with normal speed	3.14	0.79
Scapular elevation	3.08	0.74
Scapular retraction	3.07	0.76
Movement without tremor	2.92	0.68
Palmar prehension	2.91	0.92
Spherical grasp	2.68	0.81
Hook grasp	2.61	0.56
Lateral prehension	2.56	0.77
Movement without dysmetria	2.26	0.35

Table 24: Comparison of item discrimination values calculated by Multilog to those calculated by Winsteps in the second dataset

D377 Item Discrimination		
Item Description	Multilog estimated Item Discrimination	Winsteps estimated item discrimination
Wrist flex-ext, elbow at 90	6.13	1.31
Sh flexion to 90, elbow extended	5.57	1.29
Wrist stable, elbow extended	5.29	1.14
Wrist flex-ext, elbow extended	5.13	1.31
Wrist stable, elbow at 90	5.13	1.16
Pronation-supination, elbow at 90	5.07	1.26
Sh abd to 90, elbow extended	5.01	1.27
Pronation-supination, elbow extended	4.90	1.30
Movement with normal speed	4.75	1.25
Shoulder external rotation	4.71	1.12
Finger mass flexion	4.45	1.11
Finger mass extension	4.36	1.19
Elbow extension	4.29	1.02
Cylindrical grasp	4.21	1.17
Sh flex to 180, elbow extended	4.01	1.10
Forearm supination	3.76	0.79
Wrist circumduction	3.69	1.06
Hand to lumbar spine	3.66	0.98
Shoulder adduction with internal rotation	3.52	0.96
Palmar prehension	3.47	0.92
Shoulder abduction	3.32	0.94
Movement without tremor	3.29	0.86
Forearm pronation	3.28	0.88
Elbow flexion	3.24	0.85
Scapular elevation	3.21	0.81
Spherical grasp	3.16	0.85
Scapular retraction	3.09	0.69
Movement without dysmetria	2.82	0.71
Hook grasp	2.69	0.30
Lateral prehension	2.66	0.50

Table 25: D512 comparison of item parameters derived with the 2PL and Rasch models

D512	**2PL (probits)**				**Rasch Model (logits)**	
Item Description	β1	β2	Ave. β	Ave. SEM	Item difficulty	SEM
Scapular elevation	-0.89	0.15	-0.37	0.07	-1.40	0.12
Scapular retraction	-0.71	0.76	0.03	0.08	0.03	0.10
Shoulder abduction	-0.64	0.45	-0.10	0.05	-0.56	0.11
Shoulder external rotation	-0.55	0.79	0.12	0.07	0.24	0.10
Elbow flexion	-0.93	0.05	-0.44	0.08	-1.76	0.13
Forearm supination	-0.49	0.98	0.25	0.07	0.65	0.10
Shoulder add with internal rotation	-0.95	0.16	-0.40	0.07	-1.56	0.12
Elbow extension	-0.74	0.48	-0.13	0.07	-0.64	0.11
Forearm pronation	-0.61	0.32	-0.15	0.07	-0.87	0.11
Hand to lumbar spine	-0.55	0.47	-0.04	0.06	-0.40	0.11
Sh flexion to 90, elbow extended	-0.16	0.39	0.12	0.05	-0.21	0.11
Pronation-supination, elbow at 90	-0.46	0.56	0.05	0.05	-0.17	0.11
Sh abd to 90, elbow extended	-0.19	0.64	0.23	0.05	0.28	0.10
Sh flex to 180, elbow extended	0.04	0.98	0.51	0.05	1.26	0.10
Pronation-supination, elbow ext	-0.04	0.91	0.44	0.05	1.00	0.10
Wrist stable, elbow at 90	0.01	0.41	0.21	0.05	0.18	0.10
Wrist flex-ext, elbow at 90	-0.49	0.70	0.11	0.05	0.12	0.10
Wrist stable, elbow extended	0.21	0.67	0.44	0.05	0.95	0.10
Wrist flex-ext, elbow extended	-0.13	0.66	0.27	0.06	1.06	0.10
Wrist circumduction	0.13	1.18	0.66	0.07	1.67	0.10
Finger mass flexion	-1.00	0.20	-0.40	0.07	-1.44	0.12
Finger mass extension	-0.77	0.20	-0.29	0.07	-1.25	0.12
Hook grasp	0.18	0.88	0.53	0.07	1.33	0.10
Lateral prehension	-0.13	1.03	0.45	0.07	1.08	0.10
Palmar prehension	-0.34	0.57	0.12	0.07	0.06	0.10
Cylindrical grasp	-0.60	0.16	-0.22	0.07	-1.10	0.12
Spherical grasp	-0.14	1.11	0.49	0.07	1.20	0.10
Movement without tremor	-0.32	0.08	-0.12	0.08	-0.91	0.11
Movement without dysmetria	-0.25	0.62	0.19	0.08	0.27	0.10
Movement with normal speed	0.10	0.73	0.42	0.06	0.89	0.10

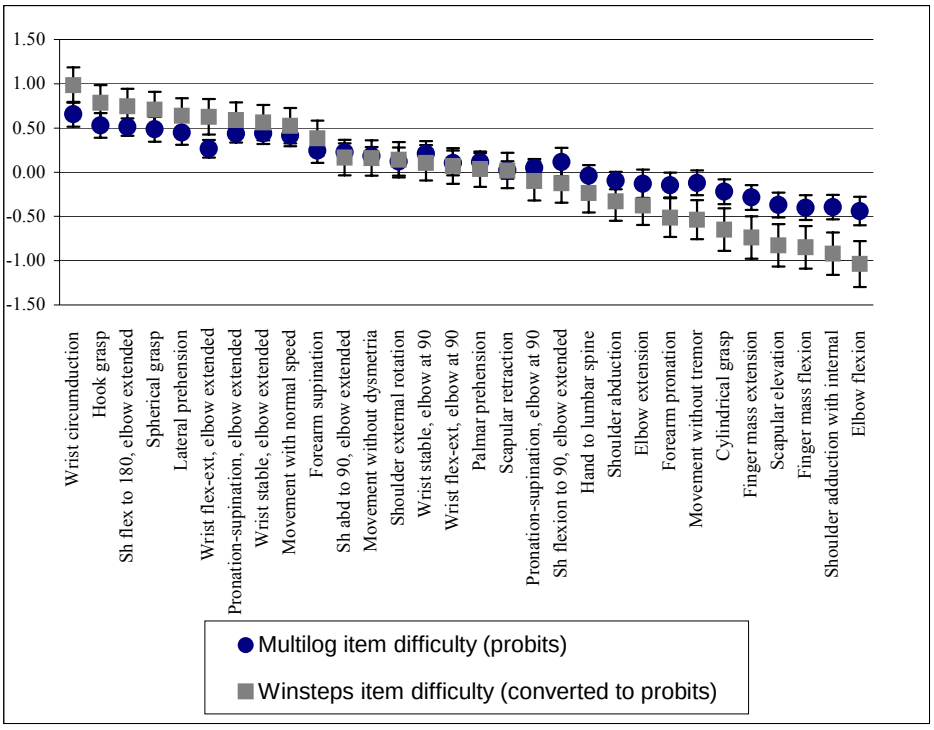

Figure 9: D512 comparison of item hierarchies generated by Multilog (probits) to the item hierarchy generated by Winsteps (logits converted to probits by multiplying by 0.59). Error bars ± 2 SEM

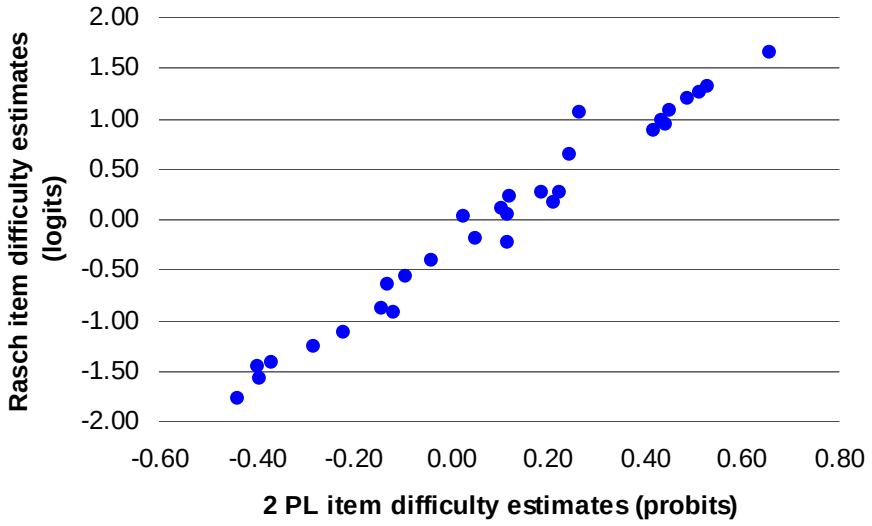

Figure 10: D512 scatter plot of item difficulty estimates (n = 512, r = 0.99)

Table 26: D377 comparison of item parameters derived with the 2PL and Rasch models

D377	**2PL (probits)**				**Rasch Model (logits)**	
Item Description	β1	β2	Ave. β	Ave. SE	Item difficulty	SE
Scapular elevation	-1.39	-0.18	-0.79	0.11	-1.90	0.17
Scapular retraction	-0.90	0.46	-0.22	0.09	0.21	0.14
Shoulder abduction	-0.97	0.08	-0.45	0.10	-0.88	0.15
Shoulder external rotation	-0.79	0.47	-0.16	0.08	0.15	0.14
Elbow flexion	-1.28	-0.21	-0.75	0.10	-1.90	0.17
Forearm supination	-0.86	0.71	-0.08	0.09	0.77	0.13
Shoulder add with internal rotation	-1.27	-0.16	-0.72	0.09	-1.81	0.17
Elbow extension	-1.01	0.23	-0.39	0.08	-0.68	0.15
Forearm pronation	-0.86	-0.02	-0.44	0.09	-1.02	0.16
Hand to lumbar spine	-0.51	0.12	-0.20	0.08	-0.23	0.14
Sh flexion to 90, elbow extended	-0.33	0.16	-0.09	0.06	-0.19	0.14
Pronation-supination, elbow at 90	-0.61	0.29	-0.16	0.07	-0.21	0.14
Shoulder abd to 90, elbow ext	-0.28	0.35	0.04	0.06	0.37	0.13
Sh flexion to 180, elbow extended	-0.19	0.83	0.32	0.06	1.65	0.12
Pronation-supination, elbow ext	-0.23	0.71	0.24	0.06	1.29	0.12
Wrist stable, elbow at 90	-0.23	0.10	-0.07	0.08	-0.17	0.14
Wrist flex-ext, elbow at 90	-0.48	0.45	-0.02	0.07	0.33	0.13
Wrist stable, elbow extended	0.03	0.42	0.23	0.06	0.95	0.13
Wrist flex-ext, elbow extended	-0.29	0.72	0.22	0.07	1.23	0.12
Wrist circumduction	-0.05	0.92	0.44	0.07	2.02	0.12
Finger mass flexion	-1.11	0.03	-0.54	0.08	-1.29	0.16
Finger mass extension	-0.77	-0.04	-0.41	0.07	-1.02	0.16
Hook grasp	0.08	0.47	0.28	0.80	1.53	0.12
Lateral prehension	-0.37	0.77	0.20	0.09	1.38	0.12
Palmar prehension	-0.72	0.29	-0.22	0.08	-0.11	0.14
Cylindrical grasp	-0.88	-0.27	-0.58	0.08	-1.64	0.17
Spherical grasp	-0.39	0.71	0.16	0.08	1.20	0.12
Movement without tremor	-0.43	-0.03	-0.23	0.09	-0.50	0.15
Movement without dysmetria	-0.45	0.24	-0.11	0.11	0.08	0.14
Movement with normal speed	-0.21	0.33	0.06	0.07	0.40	0.13

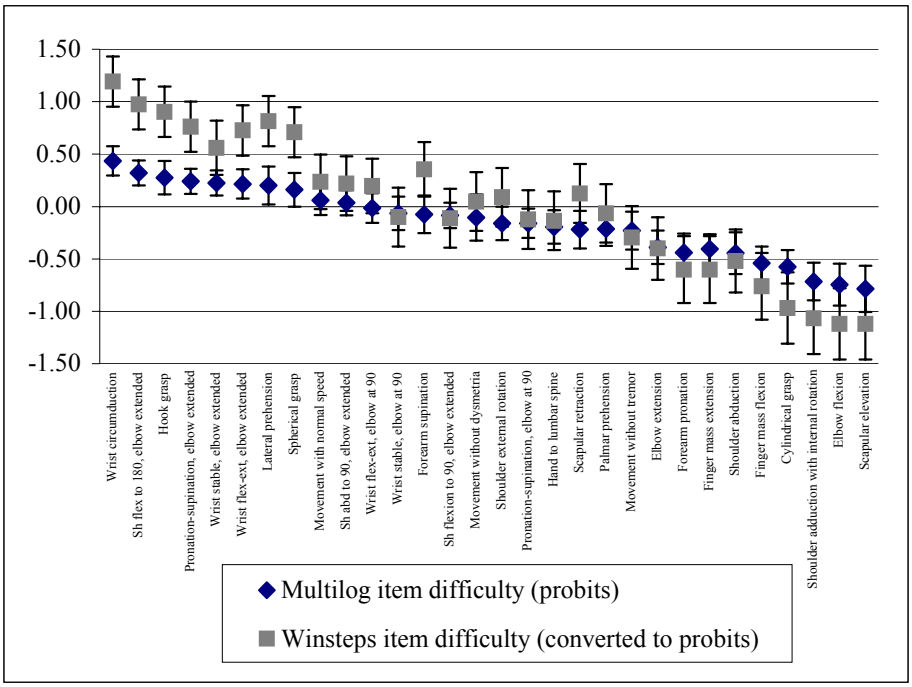

Figure 11: D377 comparison of item hierarchies generated by Multilog (probits) to the item hierarchy generated by Winsteps (logits converted to probits by multiplying by 0.59). Error bars ± 2 SEM

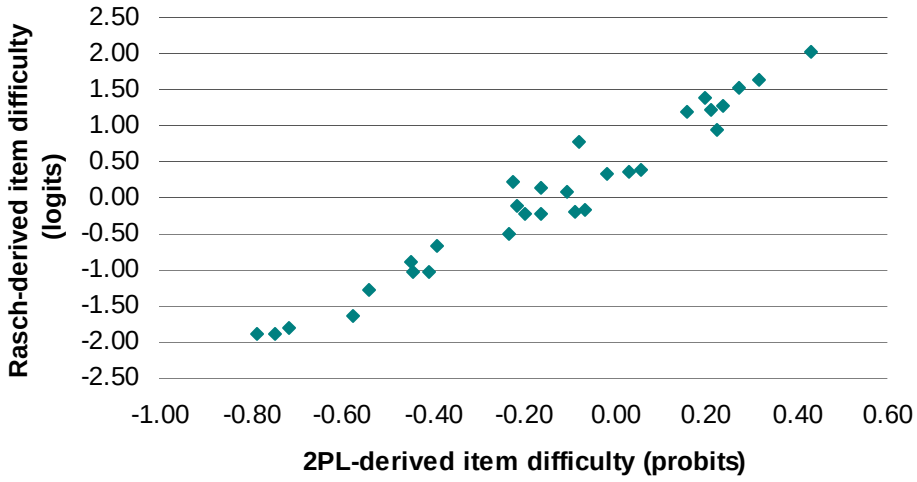

Figure 12: D377 scatter plot of item difficulty estimates (n = 377, r = 0.98)

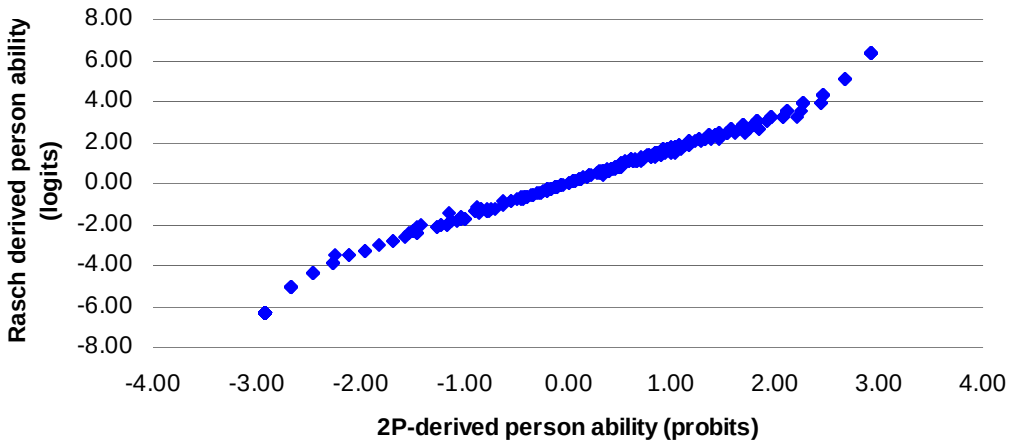

Figure 13: D512 scatter plot of person ability estimates (n = 512, r = 0.99)

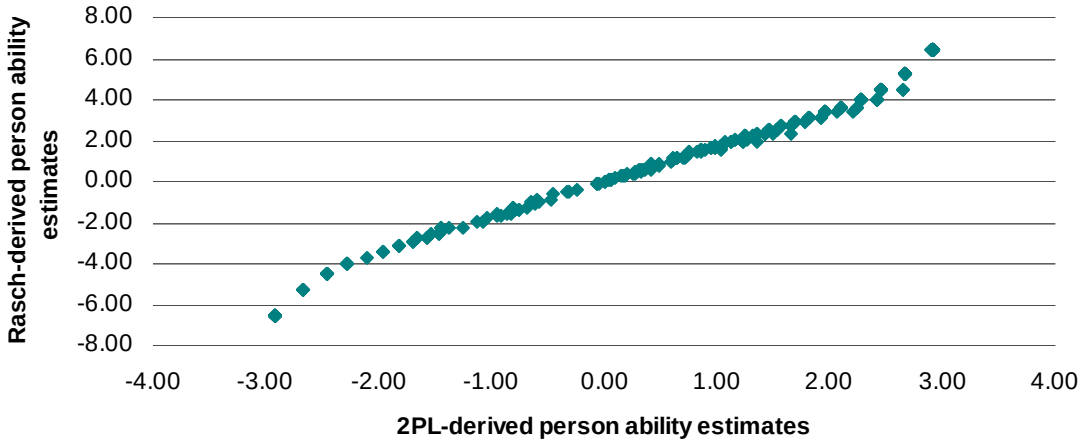

Figure 14: D377 scatter plot of person ability estimates (n = 377, r = 0.99)

CHAPTER 6
SUMMARY AND CONCLUSION

The FMA-UE is considered the "gold standard" and is the most widely used clinical assessment of post-stroke UE motor impairment. However its item-level measurement properties have not been well studied with contemporary measurement models, e.g., item response theory. A review of the literature identified several potential deficiencies in FMA-UE item level content and structure; (1) all items might not be consistent with current motor control theoretical expectations of post-stroke impairment and recovery, (2) the reproducibility of the construct "recovery" illustrated by its item hierarchy had not been established across subjects of various ability levels, nor within subjects over time, and (3) in its traditional format, the FMA-UE score had no clear functional meaning. These possible deficiencies were addressed in this project. The overall purpose of this dissertation was to examine, and potentially enhance, the measurement properties of the FMA-UE. Item Response Theory, i.e., the Rasch measurement model, was used to investigate (1) the validity of the FMA-UE item content, (2) the reproducibility of its item structure, and (3) interpretation of its score. The purpose of this dissertation was addressed in four studies.

The aim of study 1 was to test the validity of the traditional post-stroke UE motor control theory that underlies the FMA-UE by examining its item-level dimensionality and item difficulty hierarchy. Principal Components Analysis and Rasch analysis was applied to the data from 512 individuals 1-6 months post-stroke. The results indicated that the items of the FMA-UE contribute to the measurement of multiple constructs. Specifically,

the three reflex items were empirically disconnected from the other items of the assessment. The reflex items were removed, creating a modified 30-item FMA-UE, a unidimensional assessment of post-stroke upper extremity motor ability. Rasch analysis revealed that the FMA-UE item difficulty order was not consistent with the item order proposed by Fugl-Meyer, but instead was consistent with contemporary motor control theoretical views that post-stroke upper extremity movements are "easier" or "more challenging" based on task-specific contextual requirements of the movement. The Rasch-derived 30-item FMA-UE difficulty hierarchy remained consistent, independent of person ability, across 98% of the 512-person sample.

The aim of study 2 was to test the longitudinal reproducibility of the 30-item FMA-UE item structure. Rasch analysis was used to examine data from 377 individuals with stroke to whom the FMA-UE was administered at two points during the recovery process, 1-3 months post-stroke, and 6 months post-stroke. Using Differential Item Functioning procedure, the item difficultly measures at each time point were compared. Ideally, item difficulties should be invariant each time an assessment is administered to assure that the score is reproducible and comparable. It was found that the item difficulty order varies slightly, but that this variance has no practical consequences for measuring upper extremity motor ability, suggesting that the 30-item FMA-UE item structure is reproducible in test-retest situations.

The aim of study 3 was to test a method for enhancing the interpretability of a FMA-UE change score by linking it to the Wolf Motor Function Test (WMFT) an upper extremity motor ability assessment with functionally oriented items. Using Rasch analysis it was determined that a single trait is measured by both assessments. It was

feasible to link the two assessments to a common measurement scale. This method afforded an obvious connection between the FMA-UE score an its functional interpretation, thereby increasing the interpretability of the score.

Because Rasch analysis was the primary method used throughout these three studies, the aim of study 4 was to compare the results obtained with the Rasch model to those obtained with another item response theory model to assure that the choice of statistical method was valid. The results showed that the results obtained with a two-parameter item response theory mathematical model are virtually identical to those obtained by the Rasch analysis. This finding suggests that the results obtained using Rasch analysis in the dissertation are valid.

This project is significant for two reasons. First, the FMA-UE item content and structure represents traditional assumptions regarding the sequence of post-stroke upper extremity recovery. Rasch analysis provided insights about the dimensionality and difficulties of FMA-UE items thereby challenging traditional assumptions and providing insight into the construct of recovery consistent with contemporary motor control theory. Second, linking the FMA-UE to the WMFT has produced a item difficulty hierarchy that might illustrate the pattern of post-stroke recovery, progressing from reacquisition of single joint and multiple joint movements (FMA-UE items) to reacquisition of simple functional reaching, grasping and lifting tasks (WMFT items). The combined item difficulty hierarchy shows that FMA-UE change scores for individuals at the lower end of the scale (those with less upper extremity motor ability) will have different meaning than FMA-UE change scores for individuals in the middle or at the top of the scale (those with more upper extremity motor ability). Individuals with less ability will require larger

increases in the FMA-UE score before functional gains are realized, and individuals with more ability will realize functional gains with smaller increases in the FMA-UE score. The combined item hierarchy also may be the basis for exploring hypotheses regarding post-stroke UE motor impairment and the functional consequences of impairment and/or assisting rehabilitation professionals in designing the sequence for therapeutic interventions to reduce impairment and improve functional ability.

The overarching purpose of this dissertation and the aims of each project were consistent with that of an overall emphasis within the field of rehabilitation research to assure that clinical assessment tools are optimal for measuring targeted outcomes [58-60] and translate to functionally interpretable measurements of change [61]. An emphasis on evidence-based practice in post-stroke motor rehabilitation has amplified the need for valid, reliable, and functionally relevant outcome measures to evaluate the effects of intervention. This dissertation has examined and improved the measurement properties of the FMA-UE, a long-standing and important post-stroke rehabilitation measure.

APPENDIX A
RASCH ANALYSIS CONTROL FILES

Control File for Project 1 (Chapter 2)

- Title="Fugl-Meyer Upper Extremity Motor Assessment"
- ;512 person pooled dataset
- ;reflex items removed
- IDFILE=*
- 1
- 2
- 18
- *
- CODES=012
- NAME1=1
- NAMELEN=3
- NI=33
- ITEM1=5
- TABLES=11111111111111111111
- DISCRIM=YES
- &END
- 1 Biceps reflex
- 2 Triceps reflex
- 3 Scapular elevation
- 4 Scapular retraction
- 5 Shoulder abduction
- 6 Shoulder external rotation
- 7 Elbow flexion
- 8 Forearm supination
- 9 Shoulder adduction with internal rotation
- 10 Elbow extension
- 11 Forearm pronation
- 12 Hand to lumbar spine
- 13 Sh flexion to 90, elbow extended
- 14 Pronation-supination, elbow at 90
- 15 Sh abd to 90, elbow extended
- 16 Sh flex to 180, elbow extended
- 17 Pronation-supination, elbow extended
- 18 Normal reflex activity
- 19 Wrist stable, elbow at 90

- 20 Wrist flex-ext, elbow at 90
- 21 Wrist stable, elbow extended
- 22 Wrist flex-ext, elbow extended
- 23 Wrist circumduction
- 24 Finger mass flexion
- 25 Finger mass extension
- 26 Hook grasp
- 27 Lateral prehension
- 28 Palmar prehension
- 29 Cylindrical grasp
- 30 Spherical grasp
- 31 Movement without tremor
- 32 Movement without dysmetria
- 33 Movement with normal speed
- END NAMES

Control File for Project 2, T1 (Chapter 3)

- Title=”T1 30-item Fugl-Meyer Upper Extremity Motor Assessment, n= 377"
- ;KC stroke study T1 data but only those with T2 data only n = 377
- CODES=012
- NAME1=1
- NAMELEN=3
- NI=33
- ITEM1=5
- IDFILE=*
- 1
- 2
- 18
- *
- TABLES=11111111111111111111
- DISCRIM=YES
- &END
- 1 Biceps reflex
- 2 Triceps reflex
- 3 Scapular elevation
- 4 Scapular retraction
- 5 Shoulder abduction
- 6 Shoulder external rotation
- 7 Elbow flexion
- 8 Forearm supination
- 9 Shoulder adduction with internal rotation
- 10 Elbow extension
- 11 Forearm pronation

- 12 Hand to lumbar spine
- 13 Sh flexion to 90, elbow extended
- 14 Pronation-supination, elbow at 90
- 15 Sh abd to 90, elbow extended
- 16 Sh flex to 180, elbow extended
- 17 Pronation-supination, elbow extended
- 18 Normal reflex activity
- 19 Wrist stable, elbow at 90
- 20 Wrist flex-ext, elbow at 90
- 21 Wrist stable, elbow extended
- 22 Wrist flex-ext, elbow extended
- 23 Wrist circumduction
- 24 Finger mass flexion
- 25 Finger mass extension
- 26 Hook grasp
- 27 Lateral prehension
- 28 Palmar prehension
- 29 Cylindrical grasp
- 30 Spherical grasp
- 31 Movement without tremor
- 32 Movement without dysmetria
- 33 Movement with normal speed
- END NAMES

Control File for Project 2, T2 (Chapter 3)

- Title="T2 30-item Fugl-Meyer Upper Extremity Motor Assessment, n = 377"
- ;KC stroke study T2 data but only those with T1 data only, n = 377
- CODES=012
- NAME1=1
- NAMELEN=3
- NI=33
- ITEM1=5
- IDFILE=*
- 1
- 2
- 18
- *
- TABLES=111111111111111111111
- DISCRIM=YES
- &END
- 1 Biceps reflex
- 2 Triceps reflex

- 3 Scapular elevation
- 4 Scapular retraction
- 5 Shoulder abduction
- 6 Shoulder external rotation
- 7 Elbow flexion
- 8 Forearm supination
- 9 Shoulder adduction with internal rotation
- 10 Elbow extension
- 11 Forearm pronation
- 12 Hand to lumbar spine
- 13 Sh flexion to 90, elbow extended
- 14 Pronation-supination, elbow at 90
- 15 Sh abd to 90, elbow extended
- 16 Sh flex to 180, elbow extended
- 17 Pronation-supination, elbow extended
- 18 Normal reflex activity
- 19 Wrist stable, elbow at 90
- 20 Wrist flex-ext, elbow at 90
- 21 Wrist stable, elbow extended
- 22 Wrist flex-ext, elbow extended
- 23 Wrist circumduction
- 24 Finger mass flexion
- 25 Finger mass extension
- 26 Hook grasp
- 27 Lateral prehension
- 28 Palmar prehension
- 29 Cylindrical grasp
- 30 Spherical grasp
- 31 Movement without tremor
- 32 Movement without dysmetria
- 33 Movement with normal speed
- END NAMES

WMFT Control File for Project 3, T1 (Chapter 4)

- Title="T1 WMFT n= 91 quartiles"
- CODES=1234
- ISGroups=0
- NAME1=1
- NAMELEN=4
- NI=15
- ITEM1=6
- ;DIF=$S3W2

- ;DIF analysis with classification in column 34 with a width of 2 characters (T1 or T2)
- TABLES=111111111111111111111
- &END
- 1 Forearm to table
- 2 Forearm to box
- 3 Extend elbow to side
- 4 Extend elbow push weight
- 5 Hand to table
- 6 Hand to box
- 7 Reach and retrieve
- 8 Lift soda can
- 9 Lift pencil
- 10 Lift paperclip
- 11 Stack checkers
- 12 Flip cards
- 13 Turn key
- 14 Fold towel
- 15 Lift basket
- END NAMES

WMFT Control File for Project 3, T2 (Chapter 4)

- Title="WMFT Pepper T2 n= 91 quartiles"
- ;Pepper study T2 n = 91, timed data made into ordinal with quartiles
- CODES=1234
- ISGroups=0
- NAME1=1
- NAMELEN=4
- NI=15
- ITEM1=6
- ;DIF=$S3W2
- ;DIF analysis with classification in column 34 with a width of 2 charachters (T1 or T2)
- TABLES=111111111111111111111
- &END
- 1 Forearm to table
- 2 Forearm to box
- 3 Extend elbow to side
- 4 Extend elbow push weight
- 5 Hand to table
- 6 Hand to box
- 7 Reach and retrieve

- 8 Lift soda can
- 9 Lift pencil
- 10 Lift paperclip
- 11 Stack checkers
- 12 Flip cards
- 13 Turn key
- 14 Fold towel
- 15 Lift basket
- END NAMES

FMA-UE + WMFT Control File for Project 3, T1 (Chapter 4)

- Title="FMA-UE + WMFT T1 Combined n= 91"
- ;Pepper study combined analysis of 91 FMA and WMFT data T1
- CODES= 01234
- NAME1=1
- NAMELEN=4
- NI=48
- ITEM1=6
- ISGROUPS=11111111111111111111111111111111000000000000000
- ;DIF=$S3W2
- ;DIF analysis with classification in column 34 with a width of 2 characters (T1 or T2)
- TABLES=11111111111111111111
- IDFILE=*
- 1
- 2
- 18
- *
- &END
- 1 FMA-UE Biceps reflex
- 2 FMA-UE Triceps reflex
- 3 FMA-UE Scapular elevation
- 4 FMA-UE Scapular retraction
- 5 FMA-UE Shoulder abduction
- 6 FMA-UE Shoulder external rotation
- 7 FMA-UE Elbow flexion
- 8 FMA-UE Forearm supination
- 9 FMA-UE Shoulder adduction with internal rotation
- 10 FMA-UE Elbow extension
- 11 FMA-UE Forearm pronation
- 12 FMA-UE Hand to lumbar spine
- 13 FMA-UE Sh flexion to 90, elbow extended

- 14 FMA-UE Pronation-supination, elbow at 90
- 15 FMA-UE Sh abd to 90, elbow extended
- 16 FMA-UE Sh flex to 180, elbow extended
- 17 FMA-UE Pronation-supination, elbow extended
- 18 FMA-UE Normal reflex activity
- 19 FMA-UE Wrist stable, elbow at 90
- 20 FMA-UE Wrist flex-ext, elbow at 90
- 21 FMA-UE Wrist stable, elbow extended
- 22 FMA-UE Wrist flex-ext, elbow extended
- 23 FMA-UE Wrist circumduction
- 24 FMA-UE Finger mass flexion
- 25 FMA-UE Finger mass extension
- 26 FMA-UE Hook grasp
- 27 FMA-UE Lateral prehension
- 28 FMA-UE Palmar prehension
- 29 FMA-UE Cylindrical grasp
- 30 FMA-UE Spherical grasp
- 31 FMA-UE Movement without tremor
- 32 FMA-UE Movement without dysmetria
- 33 FMA-UE Movement with normal speed
- 34 WMFT Forearm to table
- 35 WMFT Forearm to box
- 36 WMFT Extend elbow to side
- 37 WMFT Extend elbow push weight
- 38 WMFT Hand to table
- 39 WMFT Hand to box
- 40 WMFT Reach and retrieve
- 41 WMFT Lift soda can
- 42 WMFT Lift pencil
- 43 WMFT Lift paperclip
- 44 WMFT Stack checkers
- 45 WMFT Flip cards
- 46 Turn key
- 47 WMFT Fold towel
- 48 WMFT Lift basket
- END NAMES

FMA-UE + WMFT Control File for Project 3, T2 (Chapter 4)

- Title="FMA-UE + WMFT T2 Combined n= 91"
- ;Pepper study combined analysis of 91 FMA and WMFT data T2
- CODES= 01234
- NAME1=1

- NAMELEN=4
- NI=48
- ITEM1=6
- ISGROUPS=11111111111111111111111111111111000000000000000
- ;DIF=$S3W2
- ;DIF analysis with classification in column 34 with a width of 2 characters (T1 or T2)
- TABLES=11111111111111111111
- IDFILE=*
- 1
- 2
- 18
- *
- &END
- 1 FMA-UE Biceps reflex
- 2 FMA-UE Triceps reflex
- 3 FMA-UE Scapular elevation
- 4 FMA-UE Scapular retraction
- 5 FMA-UE Shoulder abduction
- 6 FMA-UE Shoulder external rotation
- 7 FMA-UE Elbow flexion
- 8 FMA-UE Forearm supination
- 9 FMA-UE Shoulder adduction with internal rotation
- 10 FMA-UE Elbow extension
- 11 FMA-UE Forearm pronation
- 12 FMA-UE Hand to lumbar spine
- 13 FMA-UE Sh flexion to 90, elbow extended
- 14 FMA-UE Pronation-supination, elbow at 90
- 15 FMA-UE Sh abd to 90, elbow extended
- 16 FMA-UE Sh flex to 180, elbow extended
- 17 FMA-UE Pronation-supination, elbow extended
- 18 FMA-UE Normal reflex activity
- 19 FMA-UE Wrist stable, elbow at 90
- 20 FMA-UE Wrist flex-ext, elbow at 90
- 21 FMA-UE Wrist stable, elbow extended
- 22 FMA-UE Wrist flex-ext, elbow extended
- 23 FMA-UE Wrist circumduction
- 24 FMA-UE Finger mass flexion
- 25 FMA-UE Finger mass extension
- 26 FMA-UE Hook grasp
- 27 FMA-UE Lateral prehension
- 28 FMA-UE Palmar prehension
- 29 FMA-UE Cylindrical grasp
- 30 FMA-UE Spherical grasp

- 31 FMA-UE Movement without tremor
- 32 FMA-UE Movement without dysmetria
- 33 FMA-UE Movement with normal speed
- 34 WMFT Forearm to table
- 35 WMFT Forearm to box
- 36 WMFT Extend elbow to side
- 37 WMFT Extend elbow push weight
- 38 WMFT Hand to table
- 39 WMFT Hand to box
- 40 WMFT Reach and retrieve
- 41 WMFT Lift soda can
- 42 WMFT Lift pencil
- 43 WMFT Lift paperclip
- 44 WMFT Stack checkers
- 45 WMFT Flip cards
- 46 WMFT Turn key
- 47 WMFT Fold towel
- 48 WMFT Lift basket
- END NAMES

APPENDIX B
SAS AND MULTILOG CONTROL FILES

SAS PCA Control File for FMA-UE Data Project 1 (Chapter 2)

- data;
- input i1 5 i2 6 i3 7 i4 8 i5 9 i6 10 i7 11 i8 12 i9 13 i10 14 i11 15 i12 16 i13 17
- i14 18 i15 19 i16 20 i17 21 i18 22 i19 23 i20 24 i21 25 i22 26 i23 27 i24 28 i25 29
- i26 30 i27 31 i28 32 i29 33 i30 34 i31 35 i32 36 i33 37;
- label
- i1 = 'Biceps Reflex'
- i2 = 'Triceps Reflex'
- i3 = 'Scapular Elevation'
- i4 = 'Scapular Retraction'
- i5 = 'Shoulder Abduction'
- i6 = 'Shoulder External Rotation'
- i7 = 'Elbow Flexion'
- i8 = 'Forearm Supination'
- i9 = 'Shoulder Adduction with IR'
- i10 = 'Elbow Extension'
- i11 = 'Forearm Pronation'
- i12 = 'Hand to Lumbar Spine'
- i13 = 'Sh Flex to 90 Elbow Extended'
- i14 = 'Pronation Supination with Elbow 90'
- i15 = 'Sh Abd to 90 Elbow Extended'
- i16 = 'Sh Flex to 180 Elbow Extended'
- i17 = 'Pronation Supination with Elbow Extended'
- i18 = 'Normal Reflex Activity'
- i19 = 'Wrist Stable Elbow at 90'
- i20 = 'Wrist Flexion Extension Elbow 90'
- i21 = 'Wrist Stable Elbow Extended'
- i22 = 'Wrist Flexion Extension Elbow Extended'
- i23 = 'Wrist Circumduction'
- i24 = 'Finger Mass Flexion'
- i25 = 'Finger Mass Extension'
- i26 = 'Hook Grasp'
- i27 = 'Lateral Prehension'
- i28 = 'Palmar Prehension'
- i29 = 'Cylindrical Grasp'
- i30 = 'Spherical Grasp'

- i31 = 'Movement without tremor'
- i32 = 'Movement without dysmetria'
- i33 = 'Movement with normal speed'
- ;
- datalines;
- 001 22222222222221211021210210022102 1
- 002 22000000000000000000000000000000
- 003 22222222222222222222212201212222
- 004 22000000000000000000000000000000
- 005 22111111111211111000001100010111
- (insert data here)
- proc factor rotate=obvarimax reorder;
- run;
- quit;

SAS PCA Control File for WMFT Data Project 3 (Chapter 4)

- data;
- input i1 6 i2 7 i3 8 i4 9 i5 10 i6 11 i7 12 i8 13 i9 14 i10 15 i11 16 i12 17 i13 18
- i14 19 i15 20;
- label
- i1 = 'Forearm to table'
- i2 = 'Forearm to box'
- i3 = 'Extend elbow'
- i4 = 'Ext elbow push weight'
- i5 = 'Hand to table'
- i6 = 'Hand to box'
- i7 = 'Reach and retrieve'
- i8 = 'Lift soda can'
- i9 = 'Lift pencil'
- i10 = 'Lift paperclip'
- i11 = 'Stack checkers'
- i12 = 'Flip cards'
- i13 = 'Turn key'
- i14 = 'Fold towel'
- i15 = 'Lift basket'
- ;
- datalines;
- 02T1 443333344342412
- 03T1 433443232222231
- 04T1 342433443431332
- 06T1 222331211232233
- (insert data here)

- proc factor rotate=obvarimax reorder;
- run;
- quit;

SAS PCA Control File for FMA-UE + WMFT Data Project 3 (Chapter 4)

- data;
- input i1 6 i2 7 i3 8 i4 9 i5 10 i6 11 i7 12 i8 13 i9 14 i10 15 i11 16 i12 17 i13 18
- i14 19 i15 20 i16 21 i17 22 i18 23 i19 24 i20 25 i21 26 i22 27 i23 28 i24 29 i25 30
- i26 31 i27 32 i28 33 i29 34 i30 35 i31 36 i32 37 i33 38 i34 39 i35 40 i36 41 i37 42
- i38 43 i39 44 i40 45 i41 46 i42 47 i43 48 i44 49 i45 50;
- label
- i1 = 'FMA-UE Scapular elevation'
- i2 = 'FMA-UE Scapular retraction'
- i3 = 'FMA-UE Shoulder abduction'
- i4 = 'FMA-UE Shoulder external rotation'
- i5 = 'FMA-UE Elbow flexion'
- i6 = 'FMA-UE Forearm supination'
- i7 = 'FMA-UE Shoulder adduction with internal rotation'
- i8 = 'FMA-UE Elbow extension'
- i9 = 'FMA-UE Forearm pronation'
- i10 = 'FMA-UE Hand to lumbar spine'
- i11 = 'FMA-UE Sh flexion to 90, elbow extended'
- i12 = 'FMA-UE Pronation-supination, elbow at 90'
- i13 = 'FMA-UE Sh abd to 90, elbow extended'
- i14 = 'FMA-UE Sh flex to 180, elbow extended'
- i15 = 'FMA-UE Pronation-supination, elbow extended'
- i16 = 'FMA-UE Wrist stable, elbow at 90'
- i17 = 'FMA-UE Wrist flex-ext, elbow at 90'
- i18 = 'FMA-UE Wrist stable, elbow extended'
- i19 = 'FMA-UE Wrist flex-ext, elbow extended'
- i20 = 'FMA-UE Wrist circumduction'
- i21 = 'FMA-UE Finger mass flexion'
- i22 = 'FMA-UE Finger mass extension'
- i23 = 'FMA-UE Hook grasp'
- i24 = 'FMA-UE Lateral prehension'
- i25 = 'FMA-UE Palmar prehension'
- i26 = 'FMA-UE Cylindrical grasp'
- i27 = 'FMA-UE Spherical grasp'
- i28 = 'FMA-UE Movement without tremor'
- i29 = 'FMA-UE Movement without dysmetria'

- i30 = 'FMA-UE Movement with normal speed'
- i31 = 'WMFT Forearm to table'
- i32 = 'WMFT Forearm to box'
- i33 = 'WMFT Extend elbo'
- i34 = 'WMFT Ext elbow push weight'
- i35 = 'WMFT Hand to table'
- i36 = 'WMFT Hand to box'
- i37 = 'WMFT Reach and retrieve'
- i38 = 'WMFT Lift soda can'
- i39 = 'WMFT Lift pencil'
- i40 = 'WMFT Lift paperclip'
- i41 = 'WMFT Stack checkers'
- i42 = 'WMFT Flip cards'
- i43 = 'WMFT Turn key'
- i44 = 'WMFT Fold towel'
- i45 = 'WMFT Lift basket'
- ;
- datalines;
- 01T2 22211122211121211022121121222122244413424434231 3
- 02T2 22211121222222212021211222112222234344434323211 3
- 03T2 22222221222122222222221222112112233222211332131 2
- (insert data here)
- proc factor nfact= 2 rotate=obvarimax reorder;
- run;
- quit;

Multilog Control File for D512 (Chapter 5)

- MULTILOG for Windows 7.00.2327.2
- Created on: 7 June 2006, 13:22:57
- >PROBLEM RANDOM,
- INDIVIDUAL,
- DATA = '512FMAT1Recod.DAT',
- NITEMS = 30,
- NGROUPS = 1,
- NEXAMINEES = 512,
- NCHARS = 3;
- >TEST ALL,
- GRADED,
- NC = (3(0)30);
- >END ;
- 3
- 123

- 1111111111111111111111111111
- 2222222222222222222222222222
- 3333333333333333333333333333
- (33A1)

Multilog Control File for D377 (Chapter 5)

- MULTILOG for Windows 7.00.2327.2
- Created on: 10 June 2006, 13:37:43
- >PROBLEM RANDOM,
- INDIVIDUAL,
- DATA = 'FMAN377T2.DAT',
- NITEMS = 30,
- NGROUPS = 1,
- NEXAMINEES = 377,
- NCHARS = 3;
- >TEST ALL,
- GRADED,
- NC = (3(0)30);
- >END ;
- 3
- 123
- 1111111111111111111111111111
- 2222222222222222222222222222
- 3333333333333333333333333333
- (33A1)

LIST OF REFERENCES

1. American-Heart-Association, 2005 Heart Disease and Stroke Statistics - 2005 Update. 2005, American Heart Association: Dallas, Texas.

2. Kwakkel, G, BJ Kollen, J van der Grond, and AJ Prevo, Probability of regaining dexterity in the flaccid upper limb: impact of severity of paresis and time since onset in acute stroke. Stroke, 2003. 34(9): p. 2181-6.

3. Duncan, PW, LB Goldstein, RD Horner, PB Landsman, GP Samsa, and DB Matchar, Similar motor recovery of upper and lower extremities after stroke. Stroke, 1994. 25(6): p. 1181-8.

4. Broeks, JG, GJ Lankhorst, K Rumping, and AJ Prevo, The long-term outcome of arm function after stroke: results of a follow-up study. Disabil Rehabil, 1999. 21(8): p. 357-64.

5. Nakayama, H, HS Jorgensen, HO Raaschou, and TS Olsen, Recovery of upper extremity function in stroke patients: the Copenhagen Stroke Study. Arch Phys Med Rehabil, 1994. 75(4): p. 394-8.

6. Nudo, R, Adaptive plasticity in motor cortex: implications for rehabilitation after brain injury. J Rehabil Med, 2003. 41: p. 7-10.

7. Lum, PS, CG Burgar, PC Shor, M Majmundar, and M Van der Loos, Robot-assisted movement training compared with conventional therapy techniques for the rehabilitation of upper-limb motor function after stroke. Arch Phys Med Rehabil, 2002. 83(7): p. 952-9.

8. Winstein, CJ, JP Miller, S Blanton, E Taub, G Uswatte, D Morris, D Nichols, and S Wolf, Methods for a multisite randomized trial to investigate the effect of constraint-induced movement therapy in improving upper extremity function among adults recovering from a cerebrovascular stroke. Neurorehabil Neural Repair, 2003. 17(3): p. 137-52.

9. Michaelsen, SM, R Dannenbaum, and MF Levin, Task-specific training with trunk restraint on arm recovery in stroke: randomized control trial. Stroke, 2006. 37(1): p. 186-92.

10. Howle, J, Neuro-Developmental Treatment Approach: Theoretical Foundations and Principles of Clinical Practice. 2002, Laguana Beach, CA: Neuro-Developmental Treatment Association.

11. Maher, CG, C Sherrington, M Elkins, RD Herbert, and AM Moseley, Challenges for evidence-based physical therapy: accessing and interpreting high-quality evidence on therapy. Phys Ther, 2004. 84(7): p. 644-54.

12. Coster, W, International conference on evidence-based practice: A collaborative effort of the American Occupational Therapy Association, the American Occupational Therapy Foundation, and the Agency for Healthcare Research and Quality. Am J Occup Ther, 2005. 59(3): p. 356-8.

13. Whitall, J, Stroke rehabilitation research: time to answer more specific questions? Neurorehabil Neural Repair, 2004. 18(1): p. 3-8.

14. Salter, K, JW Jutai, R Teasell, NC Foley, and J Bitensky, Issues for selection of outcome measures in stroke rehabilitation: ICF Body Functions. Disabil Rehabil, 2005. 27(4): p. 191-207.

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

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

17. Duncan, PW, M Propst, and SG Nelson, Reliability of the Fugl-Meyer assessment of sensorimotor recovery following cerebrovascular accident. Phys Ther, 1983. 63(10): p. 1606-10.

18. Sanford, J, J Moreland, LR Swanson, PW Stratford, and C Gowland, Reliability of the Fugl-Meyer assessment for testing motor performance in patients following stroke. Phys Ther, 1993. 73(7): p. 447-454.

19. Wood-Dauphinee, SL, JI Williams, and SH Shapiro, Examining outcome measures in a clinical study of stroke. Stroke, 1990. 21(5): p. 731-9.

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

21. Filiatrault, J, AB Arsenault, E Dutil, and D Bourbonnais, Motor function and activities of daily living assessments: a study of three tests for persons with hemiplegia. Am J Occup Ther, 1991. 45(9): p. 806-10.

22. Katz, RT, GP Rovai, C Brait, and WZ Rymer, Objective quantification of spastic hypertonia: correlation with clinical findings. Arch Phys Med Rehabil, 1992. 73(4): p. 339-47.

23. Gowland, C, P Stratford, M Ward, J Moreland, W Torresin, S Van Hullenaar, J Sanford, S Barreca, B Vanspall, and N Plews, Measuring physical impairment and disability with the Chedoke-McMaster Stroke Assessment. Stroke, 1993. 24(1): p. 58-63.

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

25. Wang, CH, CL Hsieh, MH Dai, CH Chen, and YF Lai, Inter-rater reliability and validity of the stroke rehabilitation assessment of movement (STREAM) instrument. J Rehabil Med, 2002. 34(1): p. 20-4.

26. Wolf, SL, PA Catlin, M Ellis, AL Archer, B Morgan, and A Piacentino, Assessing the Wolf Motor Function Test as outcome measure for research in patients after stroke. Stroke, 2001. 32(7): p. 1635-9.

27. Duncan, PW, SM Lai, and J Keighley, Defining post-stroke recovery: implications for design and interpretation of drug trials. Neuropharmacology, 2000. 39(5): p. 835-41.

28. Boissy, P, D Bourbonnais, C Kaegi, D Gravel, and BA Arsenault, Characterization of global synkineses during hand grip in hemiparetic patients. Arch Phys Med Rehabil, 1997. 78(10): p. 1117-24.

29. Desrosiers, J, L Noreau, A Rochette, D Bourbonnais, G Bravo, and A Bourget, Predictors of long-term participation after stroke. Disabil Rehabil, 2006. 28(4): p. 221-30.

30. Duncan, PW, LG Richards, D Wallace, J Stoker-Yates, P Pohl, C Luchies, A Ogle, and S Studenski, A randomized, controlled pilot study of a home-based exercise program for individuals with mild and moderate stroke. Stroke, 1998. 29(10): p. 2055-60.

31. Raghavan, P, E Petra, JW Krakauer, and AM Gordon, Patterns of impairment in digit independence after subcortical stroke. J Neurophysiol, 2006. 95(1): p. 369-78.

32. Harris-Love, ML, S McCombe Waller, and J Whitall, Exploiting interlimb coupling to improve paretic arm reaching performance in people with chronic stroke. Arch Phys Med Rehabil, 2005. 86(11): p. 2131-7.

33. Levin, MF, RW Selles, MH Verheul, and OG Meijer, Deficits in the coordination of agonist and antagonist muscles in stroke patients: implications for normal motor control. Brain Res, 2000. 853(2): p. 352-69.

34. Michaelsen, SM, S Jacobs, A Roby-Brami, and MF Levin, Compensation for distal impairments of grasping in adults with hemiparesis. Exp Brain Res, 2004. 157(2): p. 162-73.

35. Levin, MF, Interjoint coordination during pointing movements is disrupted in spastic hemiparesis. Brain, 1996. 119: p. 281-93.

36. Stinear, JW and WD Byblow, Rhythmic bilateral movement training modulates corticomotor excitability and enhances upper limb motricity poststroke: a pilot study. J Clin Neurophysiol, 2004. 21(2): p. 124-31.

37. Luft, AR, S McCombe-Waller, J Whitall, LW Forrester, R Macko, JD Sorkin, JB Schulz, AP Goldberg, and DF Hanley, Repetitive bilateral arm training and motor cortex activation in chronic stroke: a randomized controlled trial. JAMA, 2004. 292(15): p. 1853-61.

38. Chae, J, F Bethoux, T Bohine, L Dobos, T Davis, and A Friedl, Neuromuscular stimulation for upper extremity motor and functional recovery in acute hemiplegia. Stroke, 1998. 29(5): p. 975-9.

39. Francisco, G, J Chae, H Chawla, S Kirshblum, R Zorowitz, G Lewis, and S Pang, Electromyogram-triggered neuromuscular stimulation for improving the arm function of acute stroke survivors: a randomized pilot study. Arch Phys Med Rehabil, 1998. 79(5): p. 570-5.

40. Page, SJ and P Levine, Back from the brink: electromyography-triggered stimulation combined with modified constraint-induced movement therapy in chronic stroke. Arch Phys Med Rehabil, 2006. 87(1): p. 27-31.

41. Viriyavejakul, A, R Vachalathiti, and N Poungvarin, Botulinum treatment for post-stroke spasticity: low dose regime. J Med Assoc Thai, 1998. 81(6): p. 413-22.

42. Pang, MY, JE Harris, and JJ Eng, A community-based upper-extremity group exercise program improves motor function and performance of functional activities in chronic stroke: a randomized controlled trial. Arch Phys Med Rehabil, 2006. 87(1): p. 1-9.

43. Krebs, HI, BT Volpe, ML Aisen, and N Hogan, Increasing productivity and quality of care: robot-aided neuro-rehabilitation. J Rehabil Res Dev, 2000. 37(6): p. 639-52.

44. Finley, MA, SE Fasoli, L Dipietro, J Ohlhoff, L Macclellan, C Meister, J Whitall, R Macko, CT Bever, Jr., HI Krebs, and N Hogan, Short-duration robotic therapy in stroke patients with severe upper-limb motor impairment. J Rehabil Res Dev, 2005. 42(5): p. 683-92.

45. Piron, L, F Cenni, P Tonin, and M Dam, Virtual Reality as an assessment tool for arm motor deficits after brain lesions. Stud Health Technol Inform, 2001. 81: p. 386-92.

46. Page, S, Imagery improves upper extremity motor function in chronic stroke patients: a pilot study. The Occupational Therapy Journal of Research, 2000. 23(3): p. 200-215.

47. van der Lee, JH, RC Wagenaar, GJ Lankhorst, TW Vogelaar, WL Deville, and LM Bouter, Forced use of the upper extremity in chronic stroke patients: results from a single-blind randomized clinical trial. Stroke, 1999. 30(11): p. 2369-75.

48. Page, SJ, SA Sisto, P Levine, MV Johnston, and M Hughes, Modified constraint induced therapy: a randomized feasibility and efficacy study. J Rehabil Res Dev, 2001. 38(5): p. 583-90.

49. Page, SJ, S Sisto, MV Johnston, and P Levine, Modified constraint-induced therapy after subacute stroke: a preliminary study. Neurorehabil Neural Repair, 2002. 16(3): p. 290-5.

50. Page, SJ, S Sisto, MV Johnston, P Levine, and M Hughes, Modified constraint-induced therapy in subacute stroke: a case report. Arch Phys Med Rehabil, 2002. 83(2): p. 286-90.

51. Bonifer, NM, KM Anderson, and DB Arciniegas, Constraint-induced movement therapy after stroke: efficacy for patients with minimal upper-extremity motor ability. Arch Phys Med Rehabil, 2005. 86(9): p. 1867-73.

52. Heart and Stroke Foundation of Ontario. Management of the Post Stroke Arm and Hand: Treatment Recommendations of the 2001 Consensus Panel., http://209.5.25.171/ClientImages/1/PostStrokeArmandHandFinal2002.pdf., Accessed May 3, 2006.

53. Hambleton, RK, H Swaminathan, and HJ Rogers, Fundamentals of Item Response Theory. 1991, Newbury Park, CA: Sage Publications.

54. De Weerdt, W and MA Harrison, Measuring recovery of arm-hand function in stroke patients: a comparison of the Brunnstrom-Fugl-Meyer test and Action Research Arm test. Physiother Can, 1985. 37: p. 65-70.

55. Gladstone, DJ, CJ Danells, and SE Black, The Fugl-Meyer Assessment of motor recovery after stroke: a critical review of its measurement properties. Neurorehabil Neural Repair, 2002. 16(3): p. 232-40.

56. Platz, T, C Pinkowski, F van Wijck, IH Kim, P di Bella, and G Johnson, Reliability and validity of arm function assessment with standardized guidelines for the Fugl-Meyer Test, Action Research Arm Test and Box and Block Test: a multicentre study. Clin Rehabil, 2005. 19(4): p. 404-11.

57. Berglund, K and AR Fugl-Meyer, Upper extremity function in hemiplegia. A cross-validation study of two assessment methods. Scand J Rehabil Med, 1986. 18(4): p. 155-7.

58. Wolf, S and JA Kleim. Top 10 Take Home Messages. in The III Step Conference: Linking Movement Science and Intervention. 2005. Salt Lake City, UT.

59. Winstein, C, Why is the functional independence measure used to identify some rehabilitation needs in stroke survivors when there are better tools? Physiother Res Int, 2004. 9(4): p. 182-4.

60. Research plan for the national center for medical rehabilitation research. 1993, National Institutes of Health.

61. Haley, SM and MA Fragala-Pinkham, Interpreting change scores of tests and measures used in physical therapy. Phys Ther, 2006. 86(5): p. 735-43.

62. International Classification of Functioning (ICF), Disability and Health. 2001, Geneva: World Health Organization.

63. Jette, AM, Toward a common language for function, disability, and health. Phys Ther, 2006. 86(5): p. 726-734.

64. Salter, K, JW Jutai, R Teasell, NC Foley, J Bitensky, and M Bayley, Issues for selection of outcome measures in stroke rehabilitation: ICF Activity. Disabil Rehabil, 2005. 27(6): p. 315-40.

65. Salter, K, JW Jutai, R Teasell, NC Foley, J Bitensky, and M Bayley, Issues for selection of outcome measures in stroke rehabilitation: ICF Participation. Disabil Rehabil, 2005. 27(9): p. 507-28.

66. Shumway-Cook, A and MH Woollacott, Motor Control: Theory and Practical Applications. 2nd ed. 2001, Philadelphia: Lippincott Williams & Wilkins.

67. McDowell, I and C Newell, The Theoretical and Technical Foundations of Health Measurement, in Measuring Health: A Guide to Rating Scales and Questionnaires, I. McDowell and C. Newell, Editors. 1996, New York:University Press: p. 10-46.

68. Barreca, SR, PW Stratford, LM Masters, CL Lambert, and J Griffiths, Comparing two versions of the Chedoke Arm and Hand Activity Inventory with the Action Research Arm Test. Phys Ther, 2006. 86(2): p. 245-53.

69. Jackson, JH, Selected Writings. 1958, New York: Basic Books.

70. Schmidt, RA and TD Lee, Motor Control and Learning: a Behavioral Emphasis. 3rd ed. 1999, Champaign, IL: Human Kinetics.

71. Clower, WT, Early contributions to the reflex chain hypothesis. J Hist Neurosci, 1998. 7(1): p. 32-42.

72. Mathiowetz, V and JB Haugen, Motor behavior research: implications for therapeutic approaches to central nervous system dysfunction. Am J Occup Ther, 1994. 48(8): p. 733-45.

73. Bobath, B, Observations on adult hemiplegia and suggestions for treatment. Physiotherapy, 1960. 46: p. 5-14.

74. Kabat, H and M Knott, Proprioceptive facilitation techniques for treatment of paralysis. Phys Ther Rev, 1953. 33(2): p. 53-64.

75. Twitchell, TE, The restoration of motor function following hemiplegia in man. Brain, 1951. 74: p. 443-480.

76. Brunnstrom, S, Motor testing procedures in hemiplegia: based on sequential recovery stages. Phys Ther, 1966. 46(4): p. 357-75.

77. Brunnstrom, S, Movement Therapy in Hemiplegia: a Neurophysiological Approach. 1970, New York: Harper & Row.

78. Carr, JH and RB Shepherd, Movement Science: Foundations for Physical Therapy in Rehabilitation. 2nd ed. 2000, Gaithersburg, Md.: Aspen Publishers.

79. Krakauer, JW, Motor learning: its relevance to stroke recovery and neurorehabilitation. Curr Opin Neurol, 2006. 19(1): p. 84-90.

80. Zehr, EP, Neural control of rhythmic human movement: the common core hypothesis. Exerc Sport Sci Rev, 2005. 33(1): p. 54-60.

81. Taub, E, Movement in nonhuman primates deprived of somatosensory feedback. Exerc Sport Sci Rev, 1976. 4: p. 335-74.

82. Levin, MF and AG Feldman, The role of stretch reflex threshold regulation in normal and impaired motor control. Brain Res, 1994. 657(1-2): p. 23-30.

83. Wolpaw, JR and AM Tennissen, Activity-dependent spinal cord plasticity in health and disease. Annu Rev Neurosci, 2001. 24: p. 807-43.

84. Nakazawa, K, SI Yamamoto, T Ohtsuki, H Yano, and T Fukunaga, Neural control: novel evaluation of stretch reflex sensitivity. Acta Physiol Scand, 2001. 172(4): p. 257-68.

85. Ogiso, K, JM McBride, T Finni, and PV Komi, Stretch-reflex mechanical response to varying types of previous muscle activities. Journal of Electromyography and Kinesiology, 2002. 12(1): p. 27-36.

86. Wolpaw, JR, Adaptive plasticity in the spinal stretch reflex: an accessible substrate of memory? Cell Mol Neurobiol, 1985. 5(1-2): p. 147-65.

87. Wolf, SL and RL Segal, Reducing human biceps brachii spinal stretch reflex magnitude. J Neurophysiol, 1996. 75(4): p. 1637-1646.

88. Burne, JA, VL Carleton, and NJ O'Dwyer, The spasticity paradox: movement disorder or disorder of resting limbs? J Neurol Neurosurg Psychiatry, 2005. 76(1): p. 47-54.

89. Lindmark, B and E Hamrin, Evaluation of functional capacity after stroke as a basis for active intervention. Validation of a modified chart for motor capacity assessment. Scand J Rehabil Med, 1988. 20(3): p. 111-5.

90. Welmer, AK, LW Holmqvist, and DK Sommerfeld, Hemiplegic limb synergies in stroke patients. Am J Phys Med Rehabil, 2006. 85(2): p. 112-9.

91. Beer, RF, JD Given, and JP Dewald, Task-dependent weakness at the elbow in patients with hemiparesis. Arch Phys Med Rehabil, 1999. 80(7): p. 766-72.

92. Dewald, JP and RF Beer, Abnormal joint torque patterns in the paretic upper limb of subjects with hemiparesis. Muscle Nerve, 2001. 24(2): p. 273-83.

93. Lum, PS, CG Burgar, and PC Shor, Evidence for strength imbalances as a significant contributor to abnormal synergies in hemiparetic subjects. Muscle Nerve, 2003. 27(2): p. 211-21.

94. Beer, RF, JP Dewald, ML Dawson, and WZ Rymer, Target-dependent differences between free and constrained arm movements in chronic hemiparesis. Exp Brain Res, 2004. 156(4): p. 458-70.

95. Kamper, DG, AN McKenna-Cole, LE Kahn, and DJ Reinkensmeyer, Alterations in reaching after stroke and their relation to movement direction and impairment severity. Arch Phys Med Rehabil, 2002. 83(5): p. 702-7.

96. Trombly, CA, Deficits of reaching in subjects with left hemiparesis: a pilot study. Am J Occup Ther, 1992. 46(10): p. 887-97.

97. Chae, J, I Labatia, and G Yang, Upper limb motor function in hemiparesis: concurrent validity of the Arm Motor Ability test. Am J Phys Med Rehabil, 2003. 82(1): p. 1-8.

98. Michaelsen, SM, A Luta, A Roby-Brami, and MF Levin, Effect of trunk restraint on the recovery of reaching movements in hemiparetic patients. Stroke, 2001. 32(8): p. 1875-83.

99. Latash, ML, Bernstein’s traditions in movement studies, in Progress in Motor Control, M.L. Latash, Editor. 1998, Champaign:Human Kinetics.

100. Wu, C, CA Trombly, K Lin, and L Tickle-Degnen, A kinematic study of contextual effects on reaching performance in persons with and without stroke: influences of object availability. Arch Phys Med Rehabil, 2000. 81(1): p. 95-101.

101. Wu, C, CA Trombly, K Lin, and L Tickle-Degnen, Effects of object affordances on reaching performance in persons with and without cerebrovascular accident. Am J Occup Ther, 1998. 52(6): p. 447-56.

102. Wu, CY, CA Trombly, and KC Lin, The relationship between occupational form and occupational performance: a kinematic perspective. Am J Occup Ther, 1994. 48(8): p. 679-87; discussion 688.

103. Ma, HI and CA Trombly, The comparison of motor performance between part and whole tasks in elderly persons. Am J Occup Ther, 2001. 55(1): p. 62-7.

104. Ma, HI, CA Trombly, and C Robinson-Podolski, The effect of context on skill acquisition and transfer. Am J Occup Ther, 1999. 53(2): p. 138-44.

105. Jaric, S and ML Latash, Learning a motor task involving obstacles by a multi-joint, redundant limb: two synergies within one movement. J Electromyogr Kinesiol, 1998. 8(3): p. 169-76.

106. Murray, IA and GR Johnson, A study of the external forces and moments at the shoulder and elbow while performing every day tasks. Clin Biomech, 2004. 19(6): p. 586-94.

107. Latash, ML and JJ Nicholas, Motor control research in rehabilitation medicine. Disabil Rehabil, 1996. 18(6): p. 293-9.

108. Latash, ML and JG Anson, What are "normal movements" in atypical populations? Behav Brain Sci, 1996. 19(1): p. 55-106.

109. Mercier, C, AM Bertrand, and D Bourbonnais, Comparison of strength measurements under single-joint and multi-joint conditions in hemiparetic individuals. Clin Rehabil, 2005. 19(5): p. 523-30.

110. Zackowski, KM, AW Dromerick, SA Sahrmann, WT Thach, and AJ Bastian, How do strength, sensation, spasticity and joint individuation relate to the reaching deficits of people with chronic hemiparesis? Brain, 2004.

111. McCrea, PH, JJ Eng, and AJ Hodgson, Saturated muscle activation contributes to compensatory reaching strategies after stroke. J Neurophysiol, 2005. 94(5): p. 2999-3008.

112. Scott, SH and KE Norman, Computational approaches to motor control and their potential role for interpreting motor dysfunction. Curr Opin Neurol, 2003. 16(6): p. 693-8.

113. Shelton, FN and MJ Reding, Effect of lesion location on upper limb motor recovery after stroke. Stroke, 2001. 32(1): p. 107-12.

114. Gowland, C, H deBruin, JV Basmajian, N Plews, and I Burcea, Agonist and antagonist activity during voluntary upper-limb movement in patients with stroke. Phys Ther, 1992. 72(9): p. 624-33.

115. Mercier, C, AM Bertrand, and D Bourbonnais, Differences in the magnitude and direction of forces during a submaximal matching task in hemiparetic subjects. Exp Brain Res, 2004. 157(1): p. 32-42.

116. Cirstea, MC and MF Levin, Compensatory strategies for reaching in stroke. Brain, 2000. 123: p. 940-53.

117. Taub, E, G Uswatte, and T Elbert, New treatments in neurorehabilitation founded on basic research. Nat Rev Neurosci, 2002. 3(3): p. 228-36.

118. Fisher, BE and KJ Sullivan, Activity-dependent factors affecting poststroke functional outcomes. Top Stroke Rehabil, 2001. 8(3): p. 31-44.

119. Taub, E and G Uswatte, Constraint-induced movement therapy: bridging from the primate laboratory to the stroke rehabilitation laboratory. J Rehabil Med, 2003(41 Suppl): p. 34-40.

120. Davies, PM, Steps to follow: a guide to the treatment of adult hemiplegia based on the concept of K. and B. Bobath. 1984, New York: Springer-Verlag.

121. Mulder, T, W Zijlstra, and A Geurts, Assessment of motor recovery and decline. Gait Posture, 2002. 16(2): p. 198-210.

122. Rohrer, B, S Fasoli, HI Krebs, R Hughes, B Volpe, WR Frontera, J Stein, and N Hogan, Movement smoothness changes during stroke recovery. J. Neurosci., 2002. 22(18): p. 8297-8304.

123. Ellis, MD, BG Holubar, AM Acosta, RF Beer, and JP Dewald, Modifiability of abnormal isometric elbow and shoulder joint torque coupling after stroke. Muscle Nerve, 2005. 32(2): p. 170-8.

124. Michaelsen, SM and MF Levin, Short-term effects of practice with trunk restraint on reaching movements in patients with chronic stroke: a controlled trial. Stroke, 2004. 35(8): p. 1914-19.

125. Thielman, GT, CM Dean, and AM Gentile, Rehabilitation of reaching after stroke: Task-related training versus progressive resistive exercise ,. Archives of Physical Medicine and Rehabilitation, 2004. 85(10): p. 1613-1618.

126. Gentilucci, M, E Daprati, M Gangitano, MC Saetti, and I Toni, On orienting the hand to reach and grasp an object. Neuroreport, 1996. 7(2): p. 589-92.

127. Zattara, M and S Bouisset, Posturo-kinetic organisation during the early phase of voluntary upper limb movement. 1. Normal subjects. J Neurol Neurosurg Psychiatry, 1988. 51(7): p. 956-65.

128. Roth, E, A Heinemann, L Lovell, R Harvey, J McGuire, and S Diaz, Impairment and disability: Their relation during stroke rehabilitation. Archives of Physical Medicine and Rehabilitation, 1998. 79(3): p. 329-335.

129. Mahoney, FI and DW Barthel, Functional evaluation: the Barthel Index. Md State Med J, 1965. 14: p. 61-5.

130. Winstein, CJ, DK Rose, SM Tan, R Lewthwaite, HC Chui, and SP Azen, A randomized controlled comparison of upper-extremity rehabilitation strategies in acute stroke: A pilot study of immediate and long-term outcomes. Arch Phys Med Rehabil, 2004. 85(4): p. 620-8.

131. Shelton, FD, BT Volpe, and M Reding, Motor impairment as a predictor of functional recovery and guide to rehabilitation treatment after stroke. Neurorehabil Neural Repair, 2001. 15(3): p. 229-37.

132. Mercier, C and D Bourbonnais, Relative shoulder flexor and handgrip strength is related to upper limb function after stroke. Clin Rehabil, 2004. 18(2): p. 215-21.

133. Lai, SM, PW Duncan, and J Keighley, Prediction of functional outcome after stroke: comparison of the Orpington Prognostic Scale and the NIH Stroke Scale. Stroke, 1998. 29(9): p. 1838-42.

134. Wilson, DJ, LL Baker, and JA Craddock, Functional test for the hemiparetic upper extremity. Am J Occup Ther, 1984. 38(3): p. 159-64.

135. Feys, HM, WJ De Weerdt, BE Selz, GA Cox Steck, R Spichiger, LE Vereeck, KD Putman, and GA Van Hoydonck, Effect of a therapeutic intervention for the hemiplegic upper limb in the acute phase after stroke: a single-blind, randomized, controlled multicenter trial. Stroke, 1998. 29(4): p. 785-92.

136. Wade, D, 'But is the difference clinically significant?' Clin Rehabil, 2005. 19(4): p. 349-53.

137. Testa, MA, Interpretation of quality-of-life outcomes: issues that affect magnitude and meaning. Med Care, 2000. 38(9 Suppl): p. II166-74.

138. Terwee, CB, FW Dekker, WM Wiersinga, MF Prummel, and PM Bossuyt, On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. Qual Life Res, 2003. 12(4): p. 349-62.

139. McCombe Waller, S and J Whitall, Hand dominance and side of stroke affect rehabilitation in chronic stroke. Clin Rehabil, 2005. 19(5): p. 544-51.

140. Duff, MJ, LB Okun, and G Veneziano, Trialogue on the number of fundamental constants. Journal of High Energy Physics, 2002. 3(23).

141. Wright, B, Definition of objective measurement, http://www.rasch.org/define.htm, Accessed May 27, 2005.

142. Bond, TG and CM Fox, Applying the Rasch Model: Fundamental Measurement in the Human Sciences. 2001, Mahwah, N.J.: Erlbaum.

143. Velozo, CA, G Kielhofner, and JS Lai, The use of Rasch analysis to produce scale-free measurement of functional ability. Am J Occup Ther, 1999. 53(1): p. 83-90.

144. Crocker, LM and J Algina, Introduction to Classical and Modern Test Theory. 1986, New York: Holt Rinehart and Winston.

145. White, LJ and CA Velozo, The use of Rasch measurement to improve the Oswestry classification scheme. Arch Phys Med Rehabil, 2002. 83(6): p. 822-31.

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

147. Sheehan, TJ, S DuBrava, J Fifield, S Reisine, and L DeChello, Rate of change in functional limitations for patients with rheumatoid arthritis: effects of sex, age, and duration of illness. J Rheumatol, 2004. 31(7): p. 1286-92.

148. Coster, WJ, SM Haley, PL Andres, LH Ludlow, TL Bond, and PS Ni, Refining the conceptual basis for rehabilitation outcome measurement: personal care and instrumental activities domain. Med Care, 2004. 42(1 Suppl): p. I62-72.

149. Avery, LM, DJ Russell, PS Raina, SD Walter, and PL Rosenbaum, 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. 84(5): p. 697-705.

150. Wright, BD, JM Linacre, RM Smith, AW Heinemann, and CV Granger, FIM measurement properties and Rasch model details. Scand J Rehabil Med, 1997. 29(4): p. 267-72.

151. Velozo, CA, LC Magalhaes, AW Pan, and P Leiter, Functional scale discrimination at admission and discharge: Rasch analysis of the Level of Rehabilitation Scale-III. Arch Phys Med Rehabil, 1995. 76(8): p. 705-12.

152. Fisher, AG, The assessment of IADL motor skills: an application of many-faceted Rasch analysis. Am J Occup Ther, 1993. 47(4): p. 319-29.

153. Smith, EV, Jr., Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. J Appl Meas, 2002. 3(2): p. 205-31.

154. Linacre, JM, Detecting multidimensionality: which residual data-type works best? J Outcome Meas, 1998. 2(3): p. 266-83.

155. Hattie, J, Methodology review: assessing unidimensionality of tests and items. Applied Psychological Measurement, 1985. 9(2): p. 139-164.

156. Smith, EV, Jr. and BD Johnson, Attention Deficit Hyperactivity Disorder: scaling and standard setting using Rasch measurement. J Appl Meas, 2000. 1(1): p. 3-24.

157. Lattin, J, JD Carroll, and P Green, Analyzing Multivariate Date. 2003, Pacific Grove, CA: Thomson Brooks-Cole.

158. Hong, S, SK Mitchell, and RA Harshman, Bootstrap scree tests: A Monte Carlo simulation and applications to published data. Br J Math Stat Psychol, 2006. 59(Pt 1): p. 35-57.

159. McHorney, CA and AS Cohen, Equating health status measures with item response theory: illustrations with functional status items. Med Care, 2000. 38(9 Suppl): p. II43-59.

160. Cook, KF, PO Monahan, and CA McHorney, Delicate balance between theory and practice: health status assessment and item response theory. Med Care, 2003. 41(5): p. 571-4.

161. Finlayson, M, T Mallinson, and VM Barbosa, Activities of daily living (ADL) and instrumental activities of daily living (IADL) items were stable over time in a longitudinal study on aging. J Clin Epidemiol, 2005. 58(4): p. 338-49.

162. Haley, SM, WJ Coster, PL Andres, LH Ludlow, P Ni, TL Bond, SJ Sinclair, and AM Jette, Activity outcome measurement for postacute care. Med Care, 2004. 42(1 Suppl): p. I49-61.

163. Humphreys, LG and RG Montanelli, An investigation of the parallel analysis criterion for determining the number of common factors. Multivariate Behav. Res., 1975. 10: p. 193–206.

164. Norman, GR and DL Streiner, Biostatistics: the Bare Essentials. 2000, Hamilton, Ont.: B.C. Deker Inc.

165. Linacre, JM, A Users Guide to Winsteps and Ministeps Rasch-Model Computer Programs. 2006, Chicago: Winsteps.com.

166. Wright, B and GN Masters, Rating Scale Analysis. 1982, Chicago, IL: MESA press.

167. Muraki, E, CM Hombo, and YW Lee, Equating and linking of performance assessments. Appl Psych Meas, 2000. 24(4): p. 325-337.

168. Hart, T, J Whyte, S Millis, RK Bode, J Malec, RN Richardson, and F Hammond, Dimensions of Disordered Attention in Traumatic Brain Injury: Further Validation of the Moss Attention Rating Scale. Arch Phys Med Rehabil, 2006. 87(5): p. 647-655.

169. Wright, BD and JM Linacre, Reasonable mean-square fit values, http://www.rasch.org/rmt/rmt83b.htm, Accessed May 15, 2006.

170. Ryall, NH, SB Eyres, VC Neumann, BB Bhakta, and A Tennant, Is the Rivermead Mobility Index appropriate to measure mobility in lower limb amputees? Disabil Rehabil, 2003. 25(3): p. 143-53.

171. Linacre, JM, Test validity and Rasch measurement: construct, content etc. Rasch Measurement Transactions, 2004. 18(1): p. 970-971.

172. Wright, BD and MH Stone, Best Test Design. 1979, Chicago: Mesa Press.

173. Kornetti, DL, SL Fritz, YP Chiu, KE Light, and CA Velozo, Rating scale analysis of the Berg Balance Scale. Arch Phys Med Rehabil, 2004. 85(7): p. 1128-35.

174. Russell, DJ, LM Avery, PL Rosenbaum, PS Raina, SD Walter, and RJ Palisano, Improved scaling of the gross motor function measure for children with cerebral palsy: evidence of reliability and validity. Phys Ther, 2000. 80(9): p. 873-85.

175. Haley, SM, LH Ludlow, and WJ Coster, Pediatric Evaluation of Disability Inventory: clinical interpretation of summary scores using Rasch rating scale methodology. Phys Med Rehabil Clin N Am, 1993. 4: p. 529-40.

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

177. Sabari, JS, AL Lim, CA Velozo, L Lehman, O Kieran, and JS Lai, 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. 86(8): p. 1609-15.

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

179. Chiu, YP, SL Fritz, KE Light, and CA Velozo, Use of item response analysis to investigate measurement properties and clinical validity of data for the dynamic gait index. Phys Ther, 2006. 86(6): p. 778-87.

180. Gluck, J and A Indurkhya, Assessing changes in the longitudinal salience of items within constructs. J Adolescent Res, 2001. 16(2): p. 169-187.

181. Bingenheimer, JB, SW Raudenbush, T Leventhal, and J Brooks-Gunn, Measurement equivalence and differential item functioning in family psychology. J Fam Psychol, 2005. 19(3): p. 441-55.

182. Tesio, L, Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. Journal of Rehabilitation Medicine, 2003. 35(3): p. 105-115.

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

184. Teresi, JA, M Kleinman, and K Ocepek-Welikson, Modern psychometric methods for detection of differential item functioning: application to cognitive assessment measures. Stat Med, 2000. 19(11-12): p. 1651-83.

185. Azocar, F, P Arean, J Miranda, and RF Munoz, Differential item functioning in a Spanish translation of the Beck Depression Inventory. J Clin Psychol, 2001. 57(3): p. 355-65.

186. Tennant, A, M Penta, L Tesio, G Grimby, JL Thonnard, A Slade, G Lawton, A Simone, J Carter, A Lundgren-Nilsson, M Tripolski, H Ring, F Biering-Sorensen, C Marincek, H Burger, and S Phillips, Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model: the PRO-ESOR project. Med Care, 2004. 42(1 Suppl): p. I37-48.

187. Roorda, LD, ME Roebroeck, T van Tilburg, GJ Lankhorst, and LM Bouter, Measuring activity limitations in climbing stairs: development of a hierarchical scale for patients with lower-extremity disorders living at home. Arch Phys Med Rehabil, 2004. 85(6): p. 967-71.

188. Dallmeijer, AJ, J Dekker, LD Roorda, DL Knol, B van Baalen, V de Groot, VP Schepers, and GJ Lankhorst, Differential item functioning of the Functional Independence Measure in higher performing neurological patients. J Rehabil Med, 2005. 37(6): p. 346-52.

189. Roznowski, M and J Reith, Examining the Measurement Quality of Tests Containing Differentially Functioning Items: do biased items result in poor measurement? Educational and Psychological Measurement, 1999. 59(2): p. 248-269.

190. Linacre, JM, AW Heinemann, BD Wright, CV Granger, and BB Hamilton, The structure and stability of the Functional Independence Measure. Arch Phys Med Rehabil, 1994. 75(2): p. 127-32.

191. Fisher, WP, Jr., Physical disability construct convergence across instruments: towards a universal metric. J Outcome Meas, 1997. 1(2): p. 87-113.

192. Fisher, WP, RF Harvey, P Taylor, KM Kilgore, and CK Kelly, Rehabits - a common language of functional assessment. Arch Phys Med Rehabil, 1995. 76(2): p. 113-122.

193. Wolfe, EW, Equating and item banking with the Rasch model. J Appl Meas, 2000. 1(4): p. 409-434.

194. Crane, PK, G van Belle, and EB Larson, Test bias in a cognitive test: differential item functioning in the CASI. Stat Med, 2004. 23(2): p. 241-56.

195. Hanson, BA and AA Beguin, Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. Appl Psych Meas, 2002. 26(1): p. 3-24.

196. Kim, SH and AS Cohen, A comparison of linking and concurrent calibration under item response theory. Appl Psych Meas, 1998. 22(2): p. 131-143.

197. Fisher, WP, Jr., RL Eubanks, and RL Marier, Equating the MOS SF36 and the LSU HSI Physical Functioning Scales. J Outcome Meas, 1997. 1(4): p. 329-62.

198. Segal, ME, AW Heinemann, RR Schall, and BD Wright, Rasch analysis of a brief physical ability scale for long-term outcomes of stroke. Physical Medicine and Rehabilitation: State of the Art Review, 1997. 11(2): p. 385-96.

199. Latham, N and SM Haley, Measuring functional outcomes across postacute care: current challenges and future direction. Critical Reviews in Physical and Rehabilitation Medicine, 2003. 15(2): p. 83-98.

200. Wolf, SL and SA Binder-MacLeod, Electromyographic biofeedback applications to the hemiplegic patient. Changes in upper extremity neuromuscular and functional status. Phys Ther, 1983. 63(9): p. 1393-403.

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

202. Linacre, JM, Sample size and item calibration stability. Rasch Measurement Transactions, 1994. 7(4): p. 328.

203. Hambleton, RK, Educational Measurement, in Principles and Selected Applications in Item Response Theory: 3rd Ed., R.L. Lynn, Editor. 1989, New York:American Council on Education/McMillan Publishing.

204. Embretson, S and SP Reise, Item Response Theory for Psychologists. 2000, Mahwah, New Jersey: Lawrence Erlbaum Associates.

205. Reeve, BB and P Fayers, Applying item response theory modeling for evaluating questionnaire item and scale properties, in Assessing Quality of Life in Clinical Trials Methods and Practice: 2nd Ed., P. Fayers and R.D. Hays, Editors. 2005, p. 55-73.

206. Lai, JS, D Cella, CH Chang, RK Bode, and AW Heinemann, Item banking to improve, shorten and computerize self-reported fatigue: an illustration of steps to create a core item bank from the FACIT-Fatigue Scale. Qual Life Res, 2003. 12(5): p. 485-501.

207. Holman, R, CA Glas, and RJ de Haan, Power analysis in randomized clinical trials based on item response theory. Control Clin Trials, 2003. 24(4): p. 390-410.

208. Wang, WC and CT Chen, Item parameter recovery, standard error estimates, and fit statistics of the Winsteps program for the family of Rasch models. Educational and Psychological Measurement, 2005. 65(3): p. 376-404.

209. McHorney, CA and PO Monahan, Postscript: Applications of Rasch analysis in health care. Med Care, 2004. 42(1 Suppl): p. I73-8.

210. .Masters, GN, Undesirable item discrimination. Rasch Measurement Transactions, 1993. 7(2): p. 289.

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

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

213. Kirisci, L, RE Tarter, M Vanyukov, C Martin, A Mezzich, and S Brown, Application of item response theory to quantify substance use disorder severity. Addictive Behaviors, 2006. 31(6): p. 1035-1049.

214. Wright, BD and JM Linacre, Observations are always ordinal; measurements, however, must be interval. Arch Phys Med Rehabil, 1989. 70(12): p. 857-60.

215. Tate, DG, The state of rehabilitation research: art or science? Arch Phys Med Rehabil, 2006. 87(2): p. 160-6.

216. Murray, IA and GR Johnson, A study of the external forces and moments at the shoulder and elbow while performing every day tasks. Clin Biomech (Bristol, Avon), 2004. 19(6): p. 586-94.

217. Buchanan, TS, DP Almdale, JL Lewis, and WZ Rymer, Characteristics of synergic relations during isometric contractions of human elbow muscles. J Neurophysiol, 1986. 56(5): p. 1225-1241.

218. Beer, RF, JP Dewald, and WZ Rymer, Deficits in the coordination of multijoint arm movements in patients with hemiparesis: evidence for disturbed control of limb dynamics. Exp Brain Res, 2000. 131(3): p. 305-19.

219. Floyd, FJ and KF Widaman, Factor analysis in the development and refinement of clinical assessment instruments. Psychol Assessment, 1995. 7(3): p. 286-299.

220. Duncan, PW, S Studenski, LG Richards, S Gollub, S-M Lai, D Reker, S Perera, J Yates, V Koch, S Rigler, and D Johnson, Randomized clinical trial of therapeutic exercise in subacute stroke. Stroke, 2003. 34(9): p. 2173-2180.

221. Patel, AT, PW Duncan, SM Lai, and S Studenski, The relation between impairments and functional outcomes poststroke. Arch Phys Med Rehabil, 2000. 81(10): p. 1357-63.

222. Luft, AR, S Waller, L Forrester, GV Smith, J Whitall, RF Macko, JB Schulz, and DF Hanley, Lesion location alters brain activation in chronically impaired stroke survivors. Neuroimage, 2004. 21(3): p. 924-35.

223. Wright, BD and GN Masters, Rating scale analysis. 1982, Chicago: Mesa Press. xi, 206.

224. Gordon, J, ed. Assumptions underlying physical therapy intervention: theoretical and historical perspectives. 2nd ed. Movement science: foundations for physical therapy in rehabilitation, ed. J.H. Carr and R.B. Shepherd. 2000, Aspen Publishers: Gaithersburg, Md. xiii, 220.

225. Colebatch, JG, JC Rothwell, BL Day, PD Thompson, and CD Marsden, Cortical outflow to proximal arm muscles in man. Brain, 1990. 113 (Pt 6): p. 1843-56.

226. Gordon, C, Assumptions underlying physical therapy interventions, theoretical and historical perspectives, in Movement science: Foundations for Physical Therapy in Rehabilitation, 2nd ed, J.H. Carr and R.B. Shepherd, Editors. 2000, Gaithersburg, Md:Aspen Publishers.

227. Warren, M, Strategies for sensory and neuromotor remediation, in Occupational Therapy: Overcoming Human Performance Deficits, C. Christiansen and C. Baum, Editors. 1991, Thorofare NJ:Slack: p. 633-664.

228. Pedretti, LW, Occupational Therapy: Practice Skills for Physical Dysfunction. 4th ed. 1996, St. Louis: Mosby. xv, 876.

229. Kamper, DG, RL Harvey, S Suresh, and WZ Rymer, Relative contributions of neural mechanisms versus muscle mechanics in promoting finger extension deficits following stroke. Muscle Nerve, 2003. 28(3): p. 309-18.

230. Flora, DB and PJ Curran, An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. Psychol Methods, 2004. 9(4): p. 466-91.

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

232. Duncan, P, LB Goldstein, RD Horner, PB Landsman, GP Samsa, and DB Matchar, Similar motor recovery of upper and lower extremities after stroke. Stroke, 1994. 25(6): p. 1181-8.

233. Feys, H, J Hetebrij, G Wilms, R Dom, and W De Weerdt, Predicting arm recovery following stroke: value of site of lesion. Acta Neurol Scand, 2000. 102(6): p. 371-7.

234. Linacre, JM, WINSTEPS. 2004, Winsteps.com: Chicago.

235. Smith, RM, Polytomous Mean-Square Fit Statistics, http://www.rasch.org/rmt/rmt103a.htm, Accessed October 10, 2005.

236. Portney, L and M Watkins, Foundations of Clinical Research, 2nd edition. 2000, Upper Saddle River, NJ.: Prentice Hall Health.

237. Kothari, DH, SM Haley, KM Gill-Body, and HM Dumas, 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. 83(9): p. 776-85.

238. Crane, PK, DL Hart, LE Gibbons, and KF Cook, A 37-item shoulder functional status item pool had negligible differential item functioning. Journal of Clinical Epidemiology, 2006. 59(5): p. 478-484.

239. Wright, BD, Time 1 to Time 2 Comparison. Rasch Measurement Transactions, 1996. 10(1): p. 478-9.

240. Smith, RM, A comparison of methods for determining dimensionality in Rasch measurement. Structural Equation Modeling, 1996. 3: p. 25-40.

241. Wallace, D, PW Duncan, and SM Lai, 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. 55(9): p. 922-8.

242. Rabadi, MH and FM Rabadi, Comparison of the Action Research Arm Test and the Fugl-Meyer Assessment as measures of upper-extremity motor weakness after stroke. Archives of Physical Medicine and Rehabilitation, 2006. 87(7): p. 962-66.

243. Hands, B and D Larkin, Using the Rasch measurement model to investigate the construct of motor ability in young children. J Appl Meas, 2001. 2(2): p. 101-20.

244. Hands, B, B Sheridan, and D Larkin, Creating performance categories from continuous motor skill data using a Rasch measurement model. J Outcome Meas, 1999. 3(3): p. 216-32.

245. Reeves, TC and EW Ford, Strategic management and performance differences: nonprofit versus for-profit health organizations. Health Care Manage Rev, 2004. 29(4): p. 298-308.

246. Williams, GP, V Robertson, KM Greenwood, PA Goldie, and ME Morris, The high-level mobility assessment tool (HiMAT) for traumatic brain injury. Part 2: content validity and discriminability. Brain Inj, 2005. 19(10): p. 833-43.

247. Linacre, JM, Understanding Rasch measurement: optimizing rating scale category effectiveness. Journal of Applied Measurement, 2002. 3(1): p. 85-106.

248. McHorney, CA, Use of item response theory to link 3 modules of functional status items from the Asset and Health Dynamics Among the Oldest Old study. Arch Phys Med Rehabil, 2002. 83(3): p. 383-94.

249. Bland, JM and DG Altman, Cronbach's alpha. British Medical Journal, 1997. 314(7080): p. 572.

250. Nunnally, JC, Psychometric Theory. 2nd ed. 1978, New York: McGraw-Hill.

251. Jette, AM, SM Haley, and P Ni, Comparison of functional status tools used in post-acute care. Health Care Financ Rev, 2003. 24(3): p. 13-24.

252. Lai, S-M, PW Duncan, and J Keighley, Prediction of Functional Outcome After Stroke : Comparison of the Orpington Prognostic Scale and the NIH Stroke Scale. Stroke, 1998. 29(9): p. 1838-1842.

253. Stone, MH, Substantive scale construction. J Appl Meas, 2003. 4(3): p. 282-297.

254. Smith, EV, Jr., Evidence for the reliability of measures and validity of measure interpretation: a Rasch measurement perspective. J Appl Meas, 2001. 2(3): p. 281-311.

255. Costello, AB and JW Osborne, Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis, http://pareonline.net/pdf/v10n7.pdf, Accessed May 30, 2006.

256. Graham, KM, KD Moore, DW Cabel, PL Gribble, P Cisek, and SH Scott, Kinematics and kinetics of multijoint reaching in nonhuman primates. J Neurophysiol, 2003. 89(5): p. 2667-77.

257. Galloway, JC and GF Koshland, General coordination of shoulder, elbow and wrist dynamics during multijoint arm movements. Exp Brain Res, 2002. 142(2): p. 163-80.

258. Fitts, PM, The information capacity of the human motor system in controlling the amplitude of movement. 1954. J Exp Psychol Gen, 1992. 121(3): p. 262-9.

259. Linacre, JM, What is item response theory? A tentative taxonomy. Rasch Measurement Transactions, 2004. 17(2): p. 926-927.

260. May, K and TS Jackson, IRT item parameters and the reliability and validity of pretest, posttest, and gain scores. International Journal of Testing, 2005. 5(1): p. 63-73.

261. Gray-Little, B, VSL Williams, and TD Hancock, An item response theory analysis of the Rosenberg self-esteem scale. Pers Soc Psychol B, 1997. 23(5): p. 443-451.

262. Traub, RE, A priori considerations in choosing an item response model, in Applications of Item Response Theory, R.K. Hambleton, Editor. 1983, Vancouver, Canada:Educational and Research Institute of British Columbia.

263. Teresi, JA, RR Golden, P Cross, B Gurland, M Kleinman, and D Wilder, Item bias in cognitive screening measures: comparisons of elderly white, Afro-American, Hispanic and high and low education subgroups. J Clin Epidemiol, 1995. 48(4): p. 473-83.

264. Penfield, RD, The impact of model misfit on partial credit model parameter estimates. J Appl Meas, 2004. 5(2): p. 115-28.

265. Thissen, D, Multilog for Windows. 2003, Scientific Software International: Lincolnwood, IL.

266. Samejima, F, Graded response model, in Handbook of modern item response theory, W. van der Linden and R. Hambleton, Editors. 1997, New York, NY:Springer: p. 85–100.

267. Marino, RJ and DE Graves, Metric properties of the ASIA motor score: subscales improve correlation with functional activities. Archives of Physical Medicine and Rehabilitation, 2004. 85(11): p. 1804-1810.

268. Krueger, RF and MS Finger, Using item response theory to understand comorbidity among anxiety and unipolar mood disorders. Psychol Assess, 2001. 13(1): p. 140-51.

269. Masters, GN, Item discrimination: When more is worse. J Educ Meas, 1988. 25(1): p. 15-29.

BIOGRAPHICAL SKETCH

Michelle Woodbury graduated from Fredonia High school, Fredonia, NY, in 1982. From 1982 to 1986 she attended Elizabethtown College, Elizabethtown, PA, graduating with a Bachelor of Science degree in occupational therapy. She worked as an occupational therapist in both inpatient and outpatient neurorehabilitation settings in New York, Florida and South Carolina full time from 1986 to 1998. From 1993 to 1995 she attended Columbia International University, Columbia, SC, graduating with a Masters of Arts degree in intercultural studies. She joined the faculty of the Medical University of South Carolina in Charleston, SC, in 1998 and taught neurorehabilitation occupational therapy courses until 2002. From 2002 to 2006 she attended the University of Florida, Gainesville, FL, graduating with a PhD in rehabilitation science and a concentration in movement science.