Medical University of South Carolina

MEDICA

MUSC Theses and Dissertations

2015

Developing a Functional Measure Across the Continuum of Post-Acute Care

Chih-Ying (Cynthia) Li
Medical University of South Carolina

Follow this and additional works at: https://medica-musc.researchcommons.org/theses

Recommended Citation
Li, Chih-Ying (Cynthia), "Developing a Functional Measure Across the Continuum of Post-Acute Care" (2015). *MUSC Theses and Dissertations*. 467.
https://medica-musc.researchcommons.org/theses/467

This Dissertation is brought to you for free and open access by MEDICA. It has been accepted for inclusion in MUSC Theses and Dissertations by an authorized administrator of MEDICA. For more information, please contact medica@musc.edu.

DEVELOPING A FUNCTIONAL MEASURE ACROSS THE CONTINUUM OF POST-ACUTE CARE

BY

Chih-Ying (Cynthia) Li

A dissertation submitted to the faculty of the Medical University of South Carolina in partial fulfillment of the requirements for the degree
Doctor of Philosophy
In the College of Health Professions

© Chih-Ying (Cynthia) Li 2015 All rights reserved

DEVELOPING A FUNCTIONAL MEASURE ACROSS THE CONTINUUM OF POST-ACUTE CARE

BY

Chih-Ying (Cynthia) Li

Approved by:

Chair, Project Committee	Dr. Craig A. Velozo, PhD	08/06/2015 Date
Member, Project Committee	Dr. Kit N. Simpson, PhD	8/6/2015 Date
Member, Project Committee	Dr. Heather S. Bonilha, PhD	8/6/2015 Date
Member, Project Committee	Dr. Annie N. Simpson, PhD	8/6/2015 Date
Dean, College of Health Professions	Dr. Lisa Saladin, PhD	8/27/15 Date

Acknowledgements

I would like to thank my mentor, Dr. Craig A. Velozo, for his wonderful mentorship and guidance. I learned from him not only about how to become a critical-thinking researcher/ scientist but also benefited tremendously from his wonderful attitude towards life. He is a great role model and mentor for me and I greatly respect his academic professionalism.

I would like to thank my dissertation committees, Dr. Kit N. Simpson, for always checking on me to make sure I can efficiently complete data management when time was crunch; Dr. Heather Bonilha, for helping me to complete all the required courses efficiently when I first transferred to the MUSC; Dr. Annie N. Simpson, for always encouraging me, providing me warm support and solving statistical issues when I was troubled.

I also would like to thank Dr. Martin-Harris, for always giving me insightful suggestions and providing me helpful travel support so I can attend and present at the conferences. In addition, I would like to thank all the MUSC faculties and staffs, for their willingness and enthusiasm to provide me timely help whenever I had questions. MUSC provides such a resourceful and supportive environment that a student can ever dream of. I am eternally grateful that I can be part of the MUSC family and graduate from the MUSC.

I also would like to thank my previous college and graduate school mentor in Taiwan, Dr. Ay-woan Pan, who inspired me to pursue my PhD when I was a college student, and provided me the trust and support when I encountered the most difficult time in this journey.

I would like to thank my friends and colleagues in Boston, Gainesville, Charleston, Chicago, California Cleveland and Taiwan, without their support and love along my PhD journey, it is truly impossible for me to make it this far by myself.

Last but not least, I would like to thank my family, especially my parents, my sister and my brother, they are the strongest supporters to help me pursue my dream and get through all the difficult times in this journey, regardless of the challenges of long distance.

Thank you!

"*Success consists of going from failure to failure without loss of enthusiasm*"

- *Winston Churchill*

Abstract of Dissertation Presented to the
Doctor of Philosophy Program in Health and Rehabilitation Science
Medical University of South Carolina
In Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy

DEVELOPING A FUNCTIONAL MEASURE ACROSS
THE CONTINUUM OF POST-ACUTE CARE

By

Chih-Ying (Cynthia) Li

Chairperson: Craig A. Velozo, Ph.D., OTR/L
Committee: Kit N. Simpson, DrPH
Heather S. Bonilha, PhD, CCC-SLP
Annie N. Simpson, PhD

This dissertation proposed to establish a post-acute care continuum measurement system by creating an item bank that linked existing instruments. We linked two instruments measuring physical activities of daily living in the Veterans healthcare system, Functional Independence Measure which is used in Inpatient Rehabilitation Facilities and the Minimum Data Set which is used in the Community Living Centers. The objectives included: (a) creating an IRT-based item bank, (b) creating IRT-based short forms from the item bank, (c) comparing measurement precision of converted scores from varied FIM™ and MDS forms, and (d) comparing accuracy of the varied forms in generating functional related groups (FRG). We found measurement precision and accuracy decreased as the number of item decreased. FIM short forms (SFs) had similar precision and better accuracy than MDS SFs. The MDS_13-item form had acceptable precision and accuracy for generating FRGs, supporting developing a continuity measurement by linking existing instruments.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

Page

CHAPTER 1

CHAPTER 3

CHAPTER 5

ABBREVIATIONS

ADL	Activities of Daily Living
ADSA	Aging and Disability Services Administration
AHEAD	Asset and Health Dynamics Among the Oldest Old
AITC	Austin Information Technology Center
AM-PAC	Activity Measure for Post-acute Care
ANOVA	Analysis of Variance
BPRS	Brief Psychiatric Rating Scale
CAT	Computerized Adaptive Test
CARES	Cancer Rehabilitation Evaluation System
CARE tool	Continuity Assessment and Record Evaluation tool
CES-D	Center for Epidemiologic Studies Depression Scale
CFA	Confirmatory factor analysis
CFI	Comparative Fit Index
CGI	Clinical Global Impressions
CI	Confidence Interval
CINDRR	Center of Innovation on Disability and Rehabilitation Research
CLCs	Community Living Centers
CMGs	Case Mix Groups
CMS	Centers for Medicare and Medicaid Services
COIN	Center of Innovation
CORE	Center on Outcomes, Research and Education
CTT	Classical Test Theory
DIF	Differential Item Functioning
DRA	Deficit Reduction Act
DRG	Diagnosis-Related Group
DSH	Deliberate Self-Harm
DSHI	Deliberate Self-Harm Inventory
EAP	Expected a posteriori
EFA	Exploratory Factor Analysis
EORTC QLQ-C30	European Organization for Research and Treatment of Cancer Quality of

	Life Questionnaire-Core
FACIT-F	Functional Assessment of Chronic Illness Therapy-Fatigue Scale
FACT-G	Functional Assessment of Cancer Therapy (general version)
FIM	Functional Independence Measure
FLIC	Functional Living Index for Cancer
FM	Fibromyalgia
FRG	Functional-Related Group
FSOD	Function Status and Outcomes Dataset
GPCM	Generalized Partial Credit Model
GRM	Graded Response Model
HDI	Headache Disability Inventory
HHAs	Home Health Agencies
HIMQ	Headache Impact Questionnaire
HIT	Headache Impact Test
HRSD	Hamilton Rating Scale for Depression
ICC	Intraclass Correlation Coefficients
ICD-9 CM	International Classification of Diseases, 9th revision, Clinical Modification
IMMPACT	Initiative on Measurement, and Pain Assessment in Clinical Trials
IRB	Institutional Review Board
IRFs	Inpatient Rehabilitation Facilities
IRF-PAI	Inpatient Rehabilitation Facility Patient Assessment Instrument
IRT	Item Response Theory
ISAS	Inventory of Statements About Self Injury
ISR	ICD-10-Symptom Rating
Katz	Katz AOL Index
LORS	Levels of Rehabilitation Scale
LSU HSI	Louisiana State University Health Status Instruments
LTCHs	Long-Term Care Hospitals
MADRS	Montgomery Asberg Depression Rating Scale
MFIS	Modified Fatigue Impact Scale

MDD	Major Depressive Disorder
MDS	Minimum Data Set
MDS-PAC	Minimum Data Set-Post Acute Care
MIDAS	Migraine Disability Assessment Score
MMSE	Mini-Mental State Examination
MnSq	Mean Square Standardized Residuals
MedPAC	Medicare Payment Advisory Commission
MOA	Method of Administration
MS	Multiple Sclerosis
MSQ	Migraine Specific Questionnaire
MUSC	Medical University of South Carolina
HAQ-DI	Health Assessment Questionnaire Disability Index
NFSGVHS	Health Services Research and Development from North Florida/South Georgia Veterans Health System
NIH	National Institutes of Health
NHANES	National Health and Nutrition Examination Survey
QOL	Quality of Life
OASIS	Outcome and Assessment Information Set
OBRA 87	Omnibus Budget Reconciliation Act of 1987
OHSU	Oregon Health Sciences University
OLS	Ordinary Least Squares
OTS	Outpatient Therapy Services
PAC	Post-acute Care
PANSS	Positive and Negative Syndrome Scale
PCA	Principal Components Analysis
PCM	Partial Credit Model
PECS	Patient Evaluation and Conference System
PF10	Physical Functioning Scale (10-item)
PHQ	Patient Health Questionnaire
PPS	Prospective Payment System
PRO	Patient-Reported Outcomes

PROMIS	Patient Reported Outcomes Measurement Information System
Neuro-QOL	Quality of Life Outcomes in Neurological Disorders
RA	Rheumatoid Arthritis
RAI	Resident Assessment Instrument
RMSD	Root Mean Square Difference
RMSEA	Root Mean Square Error of Approximation
RTI	Research Triangle Institute
SD	Standardized Deviation
SE	Standard Error
SEM	Standard Error Of Measurement
SF	Short Form
SF-36	Short Form Health Survey (36 items)
SHI	Self-Harm Inventory
SHIF	Self-Harm Information Form
SIQTR	Self-Injury Questionnaire Treatment Related
SITBI	Self-Injurious Thoughts and Behaviours Interview
SLE	Systemic Lupus Erythematosus
SNF	Skilled Nursing Facility
SRMR	Standardized Root Mean Residuals
TBI	Traumatic Brain Injury
TICS	Telephone Interview for Cognitive Status
TLI	Tucker-Lewis Index
TSRQ	Treatment Self-Regulation Questionnaire
UDSmr	Uniform Data System for Medical Rehabilitation
UR	University of Rochester
VHA	Veteran’s Health Administration
WLSMV	Weighted Least Squares Means and Variance
ZSTD	Standardized Fit Statistics
1-PM	1-Parameter Model
2-PM	2-Parameter Model

CHAPTER ONE

INTRODUCTION

Background and Significance

A continuum of care across acute and post-acute services is an important and natural phenomenon in healthcare settings. Based on the varying ways in which diseases progress, patients need individualized trajectories of care across different facilities to obtain a variety of healthcare services that meet their needs. “A trajectory of care” is synonymous with the term “episode of care”, used in section 5008 of the Deficit Reduction Act (DRA) in 2005, meaning “the care a patient receives in order to treat a spell of illness associated with a hospitalization. A trajectory may include one or more settings” (Centers for Medicare and Medicaid Services [CMS], 2012); whereas “a spell of illness” covers, “all readmission and skilled nursing facility service use” based on Medicare's definition (Research Triangle Institute International [RTI], 2009). The US healthcare system provides a trajectory of care based on different recovery stages across acute and post-acute facilities, including acute hospitals, inpatient rehabilitation facilities (IRFs), skilled nursing facilities (SNFs; which is analogous to Community Living Centers [CLCs] in the Veterans’ healthcare system), home health agencies (HHAs), long-term care hospitals (LTCHs) and outpatient therapy services (OTS). Figure 1.1 demonstrates a general process of a trajectory of post-acute care based on 5.0% national sample of 2006 Medicare claims data. For instance, a person with acute stroke may proceed with a trajectory of care, which may include learning basic self-care skills in the IRFs or SNFs/CLCs; and maintaining a functional level in the chronic care setting (e.g., HHAs, OTS). If the stroke is minor, outpatient services may be necessary (e.g., OTS), but if the stroke is severe, then a long-term care facility may be required (e.g., LTCHs) (Figure 1.2). Based on a 5.0% national sample of 2006 Medicare claims data (RTI,

2009), over a third (35.2%; n=109,236) of all beneficiaries discharged from acute institutions continued to use at least one post-acute care (PAC), while almost 80% of this sample were discharged to either SNFs (41.1%) or HHAs (37.4%). Moreover, 52% of beneficiaries continue to use at least one additional service after receiving care at a first PAC site (RTI, 2009). In 2007, the Medicare Payment Advisory Commission (MedPAC) spent over $45 billion dollars on post-acute care for patients that had a stroke (RTI, 2009).

In general, there are several challenges when providing the continuum of care from acute to post-acute settings. First, there are various ways for patients to initiate post-acute care due to the progress of certain illnesses and specific needs for services. While many patients start using post-acute care after being discharged from an acute hospital, this is not always the case; since patients may enter PAC facilities directly based on the nature of disease (e.g., fracture). Thus, the baseline for each patient to access the PAC could be varied, making it difficult to monitor patients' functional recovery after receiving each PAC. A second challenge in providing care along the acute and post-acute continuum is the difficulty in deciding which post-acute healthcare system contributes to the best treatment outcome. Because post-acute care varies significantly and is patient- and disease-specific, the services across PAC facilities are difficult to compare. For instance, people with exactly the same diagnoses or severity of illness may be referred to receive different PAC treatments based on a healthcare practitioner's personal recommendations, preferences or based on the availability of specific PAC facilities in the nearby area. A third challenge in providing care along this continuum is to determine a fair and standardized payment system across PAC facilities while differing payment metrics are used. For instance, acute hospitals use the Diagnosis-Related Group (DRG), IRFs use the Functional-Related Group (FRG), SNFs use lengths of time called benefit periods, HHA use a 60-day

episode based on functional measurement results, and outpatient facilities use G-codes as their payment systems. Thus, the challenges of various entry points into the healthcare systems, a diverse range of treatment provided, and varied benefit payment systems, make it difficult to standardize the measurements of patients' function across the continuum of post-acute care, and to monitor patient improvement and obtain fairness of healthcare insurance reimbursement across PAC.

Currently, the Medicare program requests that PAC facilities use patient assessment tools to measure medical, functional and cognitive information at admission and over the course of treatment (CMS, 2012). For instance, the required PAC site-specific patient assessment tools include the Inpatient Rehabilitation Facility Patient Assessment Instrument (IRF-PAI), i.e., the Functional Independence Measure (FIM™) with additional demographic data for IRFs, the Minimum Data Set (MDS) for SNFs/CLCs, and the Outcome and Assessment Information Set (OASIS) for HHAs (CMS, 2012) (Table 1.1). Since definitions and measurement scales of the items, data collection procedures, and data collection timeframes used across PAC facilities differ, CMS acknowledges that the data collected at different PAC facilities cannot be directly compared (CMS, 2012). To solve this issue, CMS has funded the development of the Continuity Assessment and Record Evaluation (CARE) standardized item set, a uniform patient assessment instrument designed to provide continuum care documentation across acute and post-acute facilities, including acute hospitals, IRFs, SNFs/CLCs, HHAs and LTCHs (CMS, 2012). The CARE item set uses the same measurement system across the PAC continuum, with the hope to generate comparable scores and standardize bill payment system and patient assessment data, including patients' functioning at admission and discharge, additional clinical information such as skin integrity and allergies/adverse drug reactions, the patient's demographic data, and

healthcare services that patients access (CMS, 2012). The CARE item set has a comprehensive item set and core item set as functional status quality metrics, including motor functional status (self-care and mobility) and cognitive functional status (memory, problem solving and communication) with a rating scale from one to six, representing complete dependence to complete independence (CMS, 2012).

Although the CARE item set seems promising for resolving the current issues, noticeable limitations still exist based on the tool's development, its feasibility and usefulness. First, the developmental procedures of the CARE item set cost considerable resources, including time, money and training. For instance, the CMS invested in a multi-year, multi-site CARE item set development project; thus, the costs for instrumental development are likely to represent only a fraction of the costs that will be incurred through implementation of the CARE item set across the range of PAC settings. Currently, CMS had spent more than 10 million dollars of developing and analyzing the CARE item set (CMS, 2011). Furthermore, data-collection software systems will require extensive modification or replacement, and instrument implementation will require extensive personnel training for assessment administration, which leading to additional burden for the healthcare practitioners. Increased measurement error is likely at the beginning of the implementation of the new instrument. Finally, there are already established reimbursement systems based on the existing instruments across PAC facilities, thus, the existing reimbursement systems will require significant restructuring. The reimbursement system for IRFs, for example, is currently based on the Functional-Related Group (FRG) measured by IRF-PAI, will have to be abandoned. The new reimbursement system of using the CARE item set will also need to be validated. Thus, it is expected to consume considerable time, effort, costs and resources before truly adopting the CARE item set into practice. Even with the aforementioned efforts, the CARE

item set still could not completely resolve the previously mentioned contextual challenges, such as various entry points into PAC and different PAC treatments provided across facilities; thus, the reliability and validity of using the CARE item set across facilities will require examinations to ensure the CARE item set will provide useful information to monitor patients' function and help obtain fair reimbursements across PAC facilitates.

While traditional psychometric methods support developing a single measurement system such as the CARE item set for all PAC venues, an alternative and practical solution is to use modern test theory, known as item response theory (IRT) and latent trait model, to link existing instruments across the PAC continuum. Traditional psychometric methods, known as classical test theory (CTT) or true score theory, are based on the following basic concept: observed score (X) = true score (T) + error (E). The development of the CARE item set is based on the concepts of CTT that measurement error will be diminished by using a single tool across the PAC continuum. However, it may also underestimate other error sources that could possibly occur from using a single tool across facilities such as the unfamiliarity of administering the new tool and the error attributable to this new single tool covering redundant or inappropriate items across settings and providing irrelevant information. On the other hand, the modern measurement methods based on the latent trait model provides a more cost-efficient approach to resolving the current issue by using existing instruments to generate a measurement common metric, with an assumption that allows for linking instruments when there is equivalence of the same latent trait. The latent trait model assumes that estimated scores of a respondent can be used to predict or explain test performance on the latent traits of the person (Hambleton, Swaminathan, Cook, Eignor & Gifford, 1978). Therefore, if the latent trait or person parameters measured across different instruments are assumed to be the same, then the IRT-based approach can co-calibrate

varied instruments to a common metric that measures the same latent trait. In other words, based on the latent trait model, the IRT approach could place different instruments on the same scale measuring the same latent construct and a score crosswalk could be generated among different instruments. A score crosswalk enables scores to be translatable across instruments. Furthermore, we assumed that the IRT-based approach could establish a linked instrument (item bank) with similar measurement precision compared to the CTT-based single-instrument. This hypothesis is based on the assumption that both approaches would generate instruments with similar levels of error, especially given the fact that the core function item set of the CARE item set has items that are similar to those of the FIM (Table 1.2).

The latent trait model, the foundation of IRT, is a measurement framework that we proposed to use to support the alternative solution of maintaining existing instruments in measuring people across the PAC continuum. The concept of linking is an initial attempt to consider subsets of items within existing instruments as tied to a single latent trait (Dorans, Pommerich, & Holland, 2010; Kolen & Brennan, 2004). Prior to perform linking, it is crucial to ensure that different instruments measure the same latent trait. In this study, “self-care physical/motor function” was considered as a single latent trait measured by both the FIM and the MDS (Table 1.2). Haley et al. (2011) successfully used an IRT test characteristic curve transformation method to link physical functioning items between the Activity Measure for Post-acute Care (AM-PAC) and the Quality of Life Outcomes in Neurological Disorders (Neuro-QOL) to produce a score conversion table between these two tests with a secondary sample who are community-dwelling adults (Haley et al., 2011). Velozo et al. (2007) and Wang et al. (2008a) also demonstrated and validated linked self-care physical/motor and cognitive items from the FIM and the MDS from a secondary Veterans dataset using Rasch modeling, to co-calibrate and

translate scores between instruments successfully. Thus, previous studies demonstrated successful evidences of linking instruments that measure the same latent trait to construct an item bank.

Item banking, allowing items from different instruments to represent a single latent trait, has great potential to improve health outcome assessments in rehabilitation (Bjorner, Chang, Thissen, & Reeve, 2007; Lai et al., 2011). Based on the latent trait theory, IRT-calibrated item banks can contain large numbers of items to illustrate a well-defined and unidimensional latent trait (Choi, Reise, Pilkonis, Hays & Cella, 2010). In addition, item banks have several advantages. First, item banking allows for automatic or immediate connection of measures across instruments since items across different instruments are co-calibrated altogether on the same continuum. Second, item banking allows for the development of shorter version for more efficient assessment, which could improve clinical use of the linked instruments. Lai et al. (2011) used the fatigue item bank through the National Institutes of Health (NIH) Patient Reported Outcomes Measurement Information System (PROMIS) to generate a computerized adaptive testing (CAT) and short form, showing that both CAT and the short form can measure more than 95% of the sample precisely with reliability greater than 0.9. An item bank composed of FIM and MDS can produce CATs and short forms, respectively, or collaboratively and each measure format generated from FIM and MDS (either separately or jointly) can demonstrate similar levels of measurement precision.

Short forms and CATs derived from the item bank could provide efficient and flexible measurement systems with less items compared to the original test, further decreasing assessment time and assessment burden for both the patients and the healthcare practitioners (Bjorner, Chang, Thissen, & Reeve, 2007; Choi et al., 2010; Ware, et al., 2005). For instance,

CAT assessment only needs as few as five polytomous items per domain in order to achieve high measurement precision (Bjorner, Chang, Thissen, & Reeve, 2007). In addition to significantly reducing the assessment burden, healthcare practitioners can choose the forms they prefer to use or the forms they are most familiar. For instance, therapists at the IRFs can use the FIM, short form FIM, or CAT FIM and the nurses at the SNFs/CLCs can use the MDS forms. The advantage to generate test forms from the item bank and further develop efficient test forms is to offer the opportunity for the practitioners to use already existing instruments in their clinical settings instead of learning how to use a new instrument. Thus, healthcare practitioners working at different facilities and having a different preference of instruments can still use their preferred instrument but the measurement results across settings and instruments will be comparable. It was hypothesized that no matter which form was used, different forms may generate comparable results. Furthermore, flexibility of the administration forms can also enhance implementation of the instruments developed from the item bank, thus further improving the feasibility and usefulness of the IRT-based test forms generated from the item bank and the crosswalk.

The purpose of this dissertation is to provide an alternative solution to use existing instruments to develop an item bank based on the IRT, and further establish and compare measurement precision and accuracy of shorter administration forms (i.e., short forms from the FIM™ and the MDS) across the continuum of PAC. This paper challenges the development of a uniform instrument across the PAC continuum based on the concept of CTT. While CTT is the theoretical base most commonly used for instrumental development and psychometric validation of instruments, the IRT provides a promising approach to calibrate all instruments on the same common metric across the care continuum among PAC facilities. In addition, an IRT-based item

bank can produce short versions of instruments such as short forms, providing more efficient and flexible assessment systems.

In summary, utilizing IRT-based concepts, such as latent trait model, and IRT-approaches, such as Rasch analysis, can create the state-of-art measurement systems of item banks and further develop short forms from existing instruments. Compared to using the CTT-based methods to develop a single instrument, linking instruments and developing different administration forms, IRT-based methods can decrease resources needed for instrumental development, minimize administration assessment burden for healthcare practitioners and patients, and provide comparable measurement for a fair reimbursement system for the healthcare policy makers, significantly contributing to the resolution of current measurement issues across PAC facilities.

CHAPTER TWO

LITERATURE REVIEW

Literature Review of the Problems, Research Design, and Methods

This chapter aimed to provide an overall review of current research using the methodologies of linking in healthcare. In education, scale equating and linking are crucial methodologies to generate comparable score across varied test forms and administration modes across time. High-stakes standardized academic examinations, such as the SAT™ and the ACT™ that determine college admission in the United States, using the linking and equating approaches to equate test performance of the test takers and further prevent cheating and maintain test fairness among the test takers (Dorans, 1999). The empirical applications of vertical (i.e., across time) or horizontal (i.e., across tests) linking and equating approaches have been evaluated and advanced by numerous published studies in the field of education for decades (Baker, 1993; Baker, & Al-karni, 1991; Dorans, Pommerich, & Holland, 2007; Kolen, & Brennan, 2004; Tate, 1999; von Davier, Holland, & Thayer, 2004; Wright, & Bell, 1984).

Compared to the field of education, the concepts of linking and equating are relatively sparse and underutilized in the field of health outcomes research, due to inherent testing contextual differences (e.g., more diverse and heterogeneous sample, smaller sample size, less items and commonly-used polytomous rating scales) (McHorney, & Cohen, 2000). One healthcare area that could benefit from linking is measuring patients across continuum of care. Linking measures across the continuum of care could advance healthcare services and functional assessments that would further benefit patients, healthcare practitioners and even healthcare policy makers. For instance, linking measures could address healthcare policy makers need for a fair healthcare reimbursements system for patients receiving healthcare across different post-

acute facilities which use different functional outcomes. Also linked measures would allow for healthcare practitioners to monitor a patient's functional changes across a continuum of care and communicate those findings to other healthcare professionals across facilities.

Literature Review for Classical Testing Theory (CTT)

Six published articles were found that used linking approaches based on traditional classical testing theory (CTT) methods in healthcare, in the professions of rehabilitation, psychiatry and aging (Table 2.1). Williams and colleagues (1997) initially published the first linking article by rescaling one instrument to the other based on expert panel determinations and observed relationships. The developed crosswalk was examined with Wilcoxon Rank Sum tests to compare differences between the Functional Independence Measure (FIM) scores and the Minimum Data Set (MDS)-derived scores (Pseudo-FIM). Williams and colleagues (1997) used ordinary least squares (OLS) linear regression to determine the percent of variance explained by the alternative subscale scores on the same population (patients who received rehabilitation). The results showed that intraclass correlation coefficients (ICC) between the FIM and Pseudo-FIM motor and cognitive subscales were both 0.8l and there were no significant differences ($p > 0.05$) of mean scores for five items (out of 12) between two scales (FIM and Pseudo-FIM). However, the mean scores of the remaining seven items were significantly different between FIM and Pseudo-FIM. The significant differences of mean scores of the seven items may be due to inherent errors within the instruments (Williams, Li, Fries, & Warren, 1997). Thus, this study showed mixed results and only partially supported the crosswalk between the FIM and the Pseudo-FIM.

Buchanan and colleagues (2003 & 2004) evaluated the planned prospective payment system (PPS) by substituting the Minimum Data Set-Post Acute Care (MDS-PAC) for the FIM

in the inpatient rehabilitation hospitals. The linking/translating score method used in this study included: (a) using telephone conferences between the two instrument development teams to identify potential problem translation areas, to refine both item and scoring for the functional status items, (b) realigning the seven scoring levels of the FIM, (c) incorporating ADL assist codes of the MDS, and (d) revising item-specific translation by adding supplemental items. The results showed that the mean score differences of the motor scales between FIM and the MDS-PAC translated were approximately 5 points in the 2003 study and 2.4 points in the 2004 study; the mean score differences of the cognitive scale were 0.01 point in the 2003 study and 0 point in the 2004 study.

In addition, Buchanan and colleagues (2004) found a 56% agreement of PPS classifications between FIM and MDS-PAC-to-FIM scores, and around 20% of the facilities had revenue shifts larger than 10% of the original cost with standardized deviation (SD) differences of $1,960, even though the mean payment between FIM and MDS-PAC-to-FIM was not significantly different. Based on the above results, Buchanan and colleagues (2004) concluded that the MDS-PAC should not be substituted for the FIM in determining the rehabilitation hospital PPS due to poor payment cell agreement and substantial revenue shifts, regardless of the positive findings of good item-level agreement between original and the translated scores.

Leucht and colleagues (2006) used equipercentile linking method to equate the Brief Psychiatric Rating Scale (BPRS)/Positive and Negative Syndrome Scale (PANSS) and compared the absolute change of the translated scores to the Clinical Global Impressions Ratings (CGI)-improvement and severity scores for patients with at least one psychiatric positive symptom. Leucht and colleagues (2006) found that correlations between various CGI and BPRS/PANSS/ PABPRS (PANSS-derived BPRS) scores for the whole sample at baseline and at weeks 1-6

ranged between 0.52 and 0.74, reflecting moderate to strong associations between the original and translated scores.

Fong and colleagues (2009) also used the equipercentile equating method (i.e., percentile equivalent equating) to link cut-point scores from a standard global cognitive function test (Mini-Mental State Examination; MMSE) to other tests (Telephone Interview for Cognitive Status; TICS; 30-item and 40-item versions) for community-dwelling elders. These investigators found the intraclass correlation coefficient for MMSE versus TICS-30 and TICS-40 was 0.80 (95% confidence limits of 0.78 to 0.83) and a cut-point category in MMSE and the corresponding cut-points for TICS-30 and TICS-40 both yield weighted k-values of 0.69, indicating substantial agreement exceeding chance. These findings support that the MMSE could be successfully linked to both TICS-30 and TICS-40.

In addition, Noonan and colleagues (2012) also used equipercentile equating and single-group design to develop a crosswalk and to cross-validate the crosswalk between the Modified Fatigue Impact Scale (MFIS) and the Patient Reported Outcome Measurement Information System (PROMIS) Fatigue Short Form (SF) at a follow-up time point for persons with Multiple Sclerosis (MS). Correlations between deviations (difference between projected and actual values) and fatigue level for the PROMIS Fatigue SF and MFIS were -0.31 and -0.30, respectively, indicating greater deviations of lower fatigue scores, meaning that the crosswalk is more accurate at higher than at lower levels of fatigue. In addition, the researchers found estimated sample means were impacted by sample size. When sample size is large, especially when sample size is 150 or greater, estimated sample means were much less varied.

In summary, for the six studies based on the CTT linking method, three studies positively supported linking approaches with two studies having successfully developed linked crosswalks

(Fong, et al., 2009; Leucht, et al., 2006), and one study positively supported the results of linking between two instruments under certain linking conditions (e.g., sample size larger than 150) (Noonan, et al., 2012). One study partially supports the concept of crosswalk between instruments by developing corresponding items conceptually between instruments and comparing their differences (William, Li, Fries, & Warren, 1997). The remaining two studies (both were from the same research team) concluded that the linking approach failed to replace original scores with the translated scores to adequately determine prospective payment (Buchanan, et al., 2003 & 2004).

While previous CTT-based linking articles demonstrated mixed findings of the linked crosswalks, it is important to recognize some major limitations of CTT methodologies regarding the linking result interpretations. The main and the most well recognized limitation of CTT methods is sample and test dependency, implying the inability of the CTT-based instruments to translate scores from one sample or one instrument to the other sample/instrument. Thus, due to sample and test dependency, the characteristics of the test are dependent on the sample from which those psychometrics were derived (McHorney, 2002; Thompson & Vacha-Haase, 2000), which could lead to limited generalizability of the findings. For instance, test dependency can result in inability to compare data using instruments with different numbers of items, types of rating scale and item difficulty levels, and the test performance across test takers may be dependent on a specific set of test items (McHorney & Cohen, 2000; McHorney, 2002). Consequently, an individual's score for a particular construct is dependent on the particular instrument. Thus, a test with easy items would generate higher scores and a test with more difficult items would generate lower scores, even when the ability level of the respondents is the same. Therefore scores between instruments cannot be comparable or translated.

Another critique is that the CTT-based linking approaches tend to simply use item-to-item matches conceptually based on expert panels, which would result in potential considerable error or bias (Haley, et al., 2011). Besides above described limitations of CTT-based methods, other factors could also potentially contribute to biased CTT-based linking results, or underestimate feasibility and usefulness of linking methodologies, such as inherent errors within instruments, item selection procedure, data collection procedure, instrumental administration process or reliability of the practitioners to administer the instruments.

Literature Review for Item Response Theory (IRT)

In both education and healthcare professions, another linking option is to use the modern testing theory, known as item response theory (IRT). The IRT approach avoids many limitations of CTT-based methods and offers a flexible and effective framework for linking scale scores based on its inherent linking nature. The IRT-based linking method is based on the fundamental assumption of the latent trait model, that different items measuring the same concept can be co-calibrated on a common underlying metric (Hambleton, Swaminathan, Cook, Eignor & Gifford, 1978; Ten Klooster, et al., 2013). Thus, unlike CTT-based methods, IRT linking methods have a "built-in" linking mechanism (Embretson, 1996; Orlando, Sherbourne, & Thissen, 2000), which can create conversion tables allowing a reliable score crosswalk among scales (Carmody et al., 2006; Orlando et al., 2000). One major advantage of IRT, in contrast to CTT, is that it is sample- and test- free, meaning that the obtained person/test parameter estimates are theoretically invariant regardless of the particular person/test used to estimate them (McHorney & Cohen, 2000). Thus, the person ability will be constant regardless of tests with different difficulty levels and different tests can generate comparable measures across tests.

We found 25 linking articles based on the IRT methodologies in the field of healthcare (Table 2.2). An increasing number of studies used IRT-based methods to link different patient-reported outcome measures. In rehabilitation, “physical function” is the most well-established domain that employed linking methodologies (Fisher, 1997; Fisher, Eubanks, & Marier, 1997; Fisher, Harvey, Taylor, Kilgore, & Kelly, 1995; Haley, et al., 2011; McHorney, 2002; McHorney & Cohen, 2000; Oude Voshaar, et al., 2014; Smith, & Taylor, 2004; Ten Klooster, et al., 2013; Velozo, Byers, Wang, & Joseph, 2007; Wang, Byers, & Velozo, 2008a). Besides physical function in rehabilitation, the earliest effort using IRT-based linking method was also found in the field of oncological, especially in the area of measuring quality of life (QOL) for patients with cancer (Chang, & Cella, 1997; Gonin, Lloyd, Cella, & Cray, 1996; Holzner, et al., 2006).

In addition, linked crosswalks based on the IRT methodologies have been applied in areas such as headache (Bjorner, Kosinski, & Ware, 2003), psychiatric symptoms (Leucht, et al., 2006), cancer (Holzner, et al., 2006), self-regulation (Masse, Allen, Wilson, & Williams, 2006), depression (Carmody, et al., 2006; Fischer, Tritt, Klapp, & Fliege, 2011; Fischer, Wahl, Fliege, Klapp, & Rose, 2012; Orlando, Sherbourne, & Thissen, 2000), asthma (Thissen, et al., 2011), pain (Askew, et al., 2013; Chen, Revicki, Lai, Cook, & Amtmann, 2009), spinal cord injury (Calhoun, et al., 2009; Slavin, Kisala, Jette & Tulsky, 2010), general quality of life (Haley, et al., 2011; Tulsky, et al., 2011), self-harm (Latimer, Covic, & Tennant, 2012) and fatigue (Lai, Cella, Yanez, & Stone, 2014; Noonan, et al., 2012).

In the area of physical function of rehabilitation, Fisher and colleagues published three articles applying IRT-based methods to link items across different instruments (Fisher, 1997; Fisher, Eubanks, & Marier, 1997; Fisher, Harvey, Taylor, Kilgore, & Kelly, 1995). Fisher and colleagues (1995) initiated the first study by developing a preliminary single rehabilitation-

measuring unit, *rehabit*, using a Rasch polytomous partial credit model to co-calibrate motor scales from two instruments, Functional Independence Measure (FIM), and Patient Evaluation and Conference System (PECS) for 54 patients with multiple neurological dysfunctions. This study (Fisher et al., 1995) showed the two calibrations between the FIM and the PECS correlates at 0.89, with an R^2 of 0.79, suggesting these two instruments were measuring the same construct, and their measures could be comparable. Subsequently, Fisher (1997) used pseudo-common item equating methods to calibrated similar but not identical items from four instruments, FIM, PECS, Katz AOL Index (Katz), and Levels of Rehabilitation Scale - III (LORS), derived from ten articles for five diagnostic groups of patients (brain injuries, neuromuscular, musculoskeletal, spinal cord and stroke). This study (Fisher, 1997) found the correlations among the four instruments and the seven pseudo-common items was 0.92 on average (an average p= 0.02), supporting quantitative stability of physical functioning as an independent construct across instruments and samples.

In a similar study, Fisher, Eubanks and Marier (1997) equated the physical functioning subscales based on a Rasch rating scale model of the Medical Outcomes Study Short Form 36 (SF36)'s 10-item physical functioning scale (PF10), and the 29-item Louisiana State University Health Status Instruments (LSU HSI) with a convenience sample of 285 patients in a public hospital general medicine clinic. The results showed that the two instruments had high correlations of item difficulty estimates (r = 0.95) and the paired-sample t-test between the PF10 and the LSU HSI is 0.95 (p= 0.34), indicating that the items from the two scales measure the same latent variable. In addition, the PF10 and the LSU HSI both fit to separated and merged Rasch rating scale models (Fisher, Eubanks, & Marier, 1997).

Smith and Taylor (2004) replicated Fisher and colleagues' (1995) study by using the same five diagnostic groups of patients, the same two instruments (FIM and PECS), and the same linking method (Rasch partial credit model) with a larger sample size of 500 patients on admission and at discharge to a free-standing rehabilitation hospital in early 1998. These investigators (Smith & Taylor, 2004) found that the correlation of the person measures between the FIM and PECS is 0.92 without counting measurement error, indicating that the common metric measures with equal-interval translation can be generated from either scale and are independent of the number of items and the rating scale structure in each instrument.

Similar to the early efforts of those linking studies in rehabilitation, three linking articles were found in oncological QOL clinical trial linked varied QOL instruments in 1996, 1997 and 2006. Gonin and colleagues (1996) initially used a Rasch rating scale model for 447 patients with cancer to equate scores of two QOL-questionnaires to demonstrate 'equatability' between the total scores of Functional Assessment of Cancer Therapy, general version (FACT-G, 7 items) and the Functional Living Index for Cancer (FLIC, 27 items); and the 'standard QOL scores' between the raw scores of the FACT-G and FLIC were also derived.

Follow-up, Chang and Cella (1997) extended findings of Gonin and colleagues' (1996) by linking five instruments using the same linking method (Rasch rating scale model) and comparing the total scores for 140 patients diagnosed of cancer of all types or HIV. The five instruments include the FACT-G, the Cancer Rehabilitation Evaluation System (CARES), the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire-Core (EORTC QLQ-C30), the Spitzer's Quality of Life-Index, and the Short Form Health Survey (SF-36). The results showed that the minimum value of Cronbach's alpha of all instruments was above 0.64, indicating acceptable internal consistency coefficient; and the item

reliabilities, such as person separations and the scale slopes of each scale, were similar. However, 0.64 may not be as good as expected. Chang and Cella (1997) found compatibility of five commonly used QOL measures and that each instrument retains different degrees of precision in relation to corresponding test-centered logits, still supported using the linking approach.

Finally, Holzner and colleagues (2006) applied both classical test theory and the Rasch measurement model to investigate the equivalence of the EORTC QLQ-C30 and the FACT-G, the two most widely used oncological QOL instruments, for the patients with cancer in Germany. Holzner and colleagues (2006) found that the physical, emotional and functional/role domains of the FACT-G and EORTC were equitable with good internal consistency (ranging from 0.75 ~ 0.89) and acceptable correlation between corresponding subscales (range of r: 0.60 ~ 0.77). But for the social domain, serious discrepancies between the corresponding subscales were detected with very low correlation of 0.09 and therefore social subscales were not qualified for linking. This implied that prior to conducting linking, it is essential to ensure that the two instruments measure the same construct and have acceptable correlations.

Other researchers carried out studies with the aims to develop and validate linking approaches that allow instruments to be equivalent. In the more well-established domain of physical self-care functioning, an additional six published articles were found (McHorney, 2002; Haley, et al., 2011; Qude, et al., 2014; Ten Klooster, et al., 2013; Velozo, Byers, Wang, & Joseph, 2007; Wang, Byers, & Velozo, 2008a). McHorney (2002) linked three modules of functional status items in the Asset and Health Dynamics Among the Oldest Old (AHEAD) study with 4655 elders aged 70 years old or older, and found the six common Activity of Daily Living (ADL) items constructing a single dominant dimension, accounting for 48% of the variations. Both sets of items were successfully linked to the common items, allowing all items

to be placed on the same underlying ability measure. McHorney (2002) used a 2-parameter (P) model given the results showing that the 2-P model fits the data better compared to the 1-P model because the 2-P model has better flexibility allowing item difficulty and item discrimination to be different. Velozo and colleagues (2007) applied the 1-P, IRT model, the Rasch model, to calibrate items on a common scale between FIM and the Minimum Data Set (MDS) using secondary Veterans data of 236 patients from four facilities. The results showed good internal consistency of the combined FIM-MDS item pool (Cronbach alpha = 0.94), with 21 of the 26 items showing acceptable fit statistics. In addition, good correlations of raw scores and measures were found between the FIM and the MDS (r= -0.81 and 0.78, respectively). Wang and colleagues (2008a) further replicated Velozo et al. (2007)'s study with larger sample size, including 654 Veterans as the calibration sample, and 1476 Veterans as the validation sample, to determine the accuracy and applicability of the crosswalk based on the function-related groups (FRGs) classifications at three levels: (1) individual patient, (2) classification system, and (3) facilities. The results demonstrated a fair to substantial strength of agreement between FRGs classifications generated from the MDS-derived FIM and actual FIM scores, with the mean differences within 1.3 and 0.1 points for the motor and cognition scales, respectively. However, individual equivalence was relatively low with only 35 ~ 67% of the translated scores within 5 points of the FIM actual scores, which was slightly worse than the previous studies by Buchanna and colleagues (45.3 ~ 50.3%) (2003 & 2004).

Haley and colleagues (2011) linked the physical functioning items from two instruments, Activity Measure for Post-Acute Care (AM-PAC) and Quality of Life Outcomes in Neurological Disorders (Neuro-QOL), using IRT-based generalized partial credit model methods (Stocking-Lord method) with two samples: 1041 post-acute patients and 549 community-dwelling adults.

The results supported the use of a nonequivalent sampling design to link two instruments of different item difficulty levels by using common items. The authors (Haley, et al., 2011) provided a score conversion table and suggested that a future prospective study should ask participants to respond to both instruments in order to replicate and validate the crosswalk generated from this study.

Two linking articles were published by Netherland researchers. Ten Klooster and colleagues (2013) developed and evaluated a crosswalk between scores on the PF-10 and Health Assessment Questionnaire disability index (HAQ-DI) in patients with rheumatoid arthritis (RA), with 532 patients as the baseline developmental sample and 276 patients as the validation sample of Dutch descent. The result showed that the agreements between predicted and observed scores from the Rasch-based crosswalk in the cross-validation sample had high intra-class correlation coefficients (ICCs) (95% CI) for both HAQ-DI (0.72 to 0.81) and the PF-10 (0.75 to 0.82), respectively (Ten Klooster, et al., 2013).

Qude and colleagues (2014) replicated Klooster and colleagues' (2013) study by developing and evaluating the crosswalk between PF-10 and HAQ-DI with a larger and more diverse sample, including rheumatoid arthritis (RA; n=29,020), fibromyalgia (FM; n=3,776) and systemic lupus erythematosus (SLE; n=1,609) who participated in the National Data Bank for Rheumatic Diseases. The results found that the ICCs between predicted and actual scores ranged from 0.70–0.78, indicating that the crosswalk was sufficiently reliable for group-level use across diagnostic subgroups (Qude, et al., 2014). In addition, the mean difference between observed and expected scores was close to zero in US patients with RA (Qude, et al., 2014).

In summary, the linking studies in the domain of physical self-care function were advanced across almost 20 years, and demonstrates fairly consistent results that support (a)

physical self-care can be treated as a single latent trait, allowing for the use of a linking approach in this domain, and (b) most studies showed acceptable to good ICCs between the original and the translated scores, implying feasibility and validity of the crosswalk, which could possibly be used in clinical healthcare, especially given the similar results from several replicated studies.

Besides the domains of physical self-care functioning in rehabilitation and QOL in oncology, four articles were found using IRT-based methods to equate instruments in the domain of depression for clinical trials (Carmody, et al., 2006; Fischer, Wahl, Fliege, Klapp, & Rose, 2012; Fischer, Tritt, Klapp, & Fliege, 2011; Orlando, Sherboune, & Thissen, 2000). Orlando, Sherboune and Thissen (2000) used an IRT summed scores approach, a similar method as common person equating but with a focus mainly on translating summed scores between instruments, to calibrate a modified 23-item version of the Center for Epidemiologic Studies Depression Scale (CES-D) to the standard scale of 20-teim CES-D for 1120 patients with depression. The study compared the classification rates of respondents at the 18-month as depressed using both the 20 CES-D items (cut score of 16) and the 23-item scale (corresponding cut score of 20); and the result showed that nearly 95% of the sample were classified in the same way regardless of which criterion was used, indicating that this linking method can successfully generate comparable scores and result in similar classification results (Orlando, Sherboune, & Thissen, 2000).

Carmody and colleagues (2006) used Samejima's graded IRT model based on Orlando et al. (2000)'s procedures to equate total scores for each pair of scales, and estimate item parameters for each item of each instrument. The three instruments included the Hamilton Rating Scale for Depression-17 (HRSD17; items=17), the Hamilton Rating Scale for Depression-6 (HRSD6; items=6), and the Montgomery Asberg Depression Rating Scale (MADRS; items=10).

The research team (Carmody, et al., 2006) used first sample for calibration of 233 outpatients with depression who were highly treatment resistant and the second sample for validation of 985 outpatients with nonpsychotic major depressive disorder (MDD). The results demonstrated that three instruments had high correlations ranging from 0.86 to 0.89 for the first sample and 0.91 to 0.94 for the second sample, with moderate to high internal consistency (0.78 to 0.92) and moderate item-total correlation (0.50 to 0.78) (Carmody, et al., 2006).

Fischer, Tritt, Klapp and Fliege (2011) used a general response partial credit model to link the ICD-10-Symptom Rating (ISR) depression scale, the Patient Health Questionnaire (PHQ) depression scales (PHQ-9) and PHQ-2 (only first two items of PHQ-9) with 2258 inpatients and outpatients of a psychosomatic clinic as a construction sample and 2259 as a validation sample in Germany. The results showed that the first eigenvalue is 6.99, substantially greater than the second eigenvalue (which is 1.00), and accounts for 54% of the total variance, indicating unidimensionality. The authors also found the predicted scores provided by the conversion tables are similar to the observed scores in a validation sample, given that the converted PHQ-9 and the ISR scores contain about 66% (mean ± 1 SD) and 95% (mean ± 2 SD) of the means of the actual scores (Fischer, Tritt, Klapp, & Fliege, 2011).

Fischer, Wahl, Fliege, Klapp, & Rose (2012) replicated Fischer, et al. (2011)'s study to evaluate the validity of the conversion table between PHQ and ICD-10-Symptom Rating (ISR) by comparing treatment outcomes with 1066 patients with some types of mental and/or behavioral disorders from two psychosomatic clinics in Germany using generalized partial credit model. The results showed no difference in variance between the original PHQ-9 scores and the PHQ-9 scores transformed from ISR scores ($p= 0.76$), but a significant difference in means ($p= 0.04$, effect size = 0.03), with original PHQ-9 scores being slightly higher than ISR scores that

were transformed to PHQ-9 scores (11.09 vs. 10.90). The correlation between original PHQ-9 summary scores and transformed PHQ-9 sum scores was 0.82 ($p < 0.001$) (Fischer, Wahl, Fliege, Klapp, & Rose, 2012).

In addition, the Patient-Reported Outcomes Measurement Information System (PROMIS), an initiative sponsored by the National Institutes of Health (NIH) (Cella, et al., 2007), developed the Patient-Reported Outcomes (PRO) Rosetta Stone (PROsetta Stone®) project to develop and apply linking or equating methods between the PROMIS measures and related "legacy" instruments. Thus, the range of PRO assessment options could be expanded based on the concept of using a common and standardized metric (Choi, et al., 2012). The PROsetta Stone project identifies and applies appropriate linking methods, thus, the scores on a range of PRO instruments can be used as standardized T-score metrics linking to the PROMIS (Choi, et al., 2012). Three articles were found with such attempts (Askew, et al., 2013; Lia, Cella, Yanez, & Stone, 2014; Thissen, et al., 2011).

Thissen and colleagues (2011) used Samejima's graded IRT model and Expected a posteriori (EAP) with a method called calibrated projection to calibrate the PedsQLTM Asthma Symptoms Scale 3.0 asthma module to obtain scores comparable with those of the PROMIS pediatric asthma impact scale (PAIS) with approximately 300 children, age 8–17. Calibrated projection is a method using a full-information factor analytic approach to link without a need for two instruments to measure a single construct (Carle, et al., 2011). Thissen and colleagues (2011) found that the estimated correlation between theta 1 (the underlying construct measured by the PAIS) with theta 2 (underlying construct measured by the PedsQL^{TM}) was 0.96 and the likelihood ratio test for the difference in fit rejected the unidimensional model, indicating the PAIS exhibited strong convergent validity with the PedsQL Asthma Symptoms Scale, and

weaker relations with the other five scales (Treatment, Worry, and Communication Scales, and the DISABKIDS Asthma Impact and Worry Scales). The results showed that only one of the legacy scales was linked to the metric of the PAIS, while the other five scales appeared to measure constructs different from the PAIS.

Askew and colleagues (2013) used a two-parameter logistic graded response model to develop a crosswalk table to transform Brief Pain Inventory pain interference scale (BPI-PI) scores to PROMIS-PI short form (PROMIS-PI SF) scores for the multiple sclerosis (MS) patients, with 369 patients as a developmental calibration sample and 360 patients as a validation sample. The results showed that the mean difference between observed and cross-walked T scores was 0.51 (SD = 3.9) in the calibration sample and -1.47 (SD = 4.2) in the validation sample; and that root mean square difference (RMSD) estimates ranged from 0.01 to 0.06, indicating that the crosswalk table produced very similar observed and cross-walked scores across subgroups in the validation sample (Askew, et al., 2013).

Lia, Cella, Yanez and Stone (2014) used the Stocking-Lord calibration and fixed-parameter calibration to develop linked crosswalk tables to enable the direct comparison of fatigue scores from the three most widely used fatigue instruments, including PROMIS-Fatigue with Functional Assessment of Chronic Illness Therapy-Fatigue Scale (FACIT-F), SF-36 and Neuro-QOL, to the same metric in order to facilitate fatigue outcomes interpretations. The Stocking-Lord linking method belongs to characteristic curve method that uses separate calibration instead of concurrent calibration. The factor analysis confirmed the assumption of unidimensionality of the combined three scales and the correlations between instruments are high ($r \geq 0.88$), while the T-score discrepancies (Stocking-Lord minus fixed-parameter) ranged from -

0.30 to 1.10 with a mean of 0.06 (SD =.01), and only one participant had a discrepancy greater than 1 T-score unit (0.1 SD), supporting the score comparability between three instruments.

Additional areas such as headache, pain, self-regulation and self-harm were also found using the linking methods to facilitate score comparisons. Bjorner, Kosinski, & Ware (2003) used a generalized partial credit model (GPCM) to develop and assess the calibration of IRT-based scores on the Headache Impact Test (HIT) into the metrics of the traditional headache scales, including Migraine Specific Questionnaire (MSQ), Headache Disability Inventory (HDI), Headache Impact Questionnaire (HIMQ), Migraine Disability Assessment Score (MIDAS) using telephone interview data (n=1016) and internet data (n=1103) from general population surveys of recent headache sufferers. The results showed ICC's of calibrated HIT and the observed traditional scores were between 0.80 and 0.94 and the relative validity analyses showed the maximum mean difference between the observed and expected scores was 1.7 points on a 0–100 scale, supporting that the IRT approach could achieve comparability of new and widely-used scales (Bjorner, Kosinski, & Ware, 2003).

Masse, Allen, Wilson, & Williams (2006) used the partial credit model to compare test scores from two 8-item self-regulation scales retrieved from the Treatment Self-Regulation Questionnaire (TSRQ) with 627 firefighters aimed at improving dietary and physical activity behaviors from Oregon Health Sciences University (OHSU) and 355 adult smokers in a tobacco dependence treatment and diet intervention study from the University of Rochester (UR) using the common items as an anchor for the linking. The results showed that the principal component analysis indicated that the eight items assigned to OHSU and UR explained 40.3 and 41.6% of the total variance, respectively; and the two, eight-item TSRQ scales can be linked if they have at least four items in common (Masse, Allen, Wilson, & Williams, 2006). Masse and colleagues

(2006) found that scale reliability was reduced when fewer overlapping items were in the scales (e.g., reliability is 0.81 for 15 overlapping items and the reliability is 0.64 when there are eight overlapping items).

Chen, Revicki, Lai, Cook, & Amtmann (2009) used two approaches, common item non-equivalent group design and separately calibrated with Samejima's graded response model, to simultaneously calibrate pain items onto a common scale from two independent surveys, Initiative on Measurement, and Pain Assessment in Clinical Trials (IMMPACT) Pain Modules (n=148) and Center on Outcomes, Research and Education (CORE) Survey (n=400). The results showed the two linking approaches produce similar linking results but that the simultaneous IRT calibration method produced more stable item parameters across independent samples than separated calibration (i.e., separated calibration produced extreme item parameter estimates as high as 16.16 and 37.0). The correlations between the IRT scores of the two approaches was 0.999 for the IMMPACT and CORE samples, meaning the two calibration approaches produced very similar item characteristics (Chen, et al., 2009).

Latimer, Covic and Tennant (2012) used Rasch analysis to co-calibrate six deliberate self-harm (DSH) instruments, Self-Injury Questionnaire Treatment Related (SIQTR), Self-Injurious Thoughts and Behaviours Interview (SITBI), Deliberate Self-Harm Inventory (DSHI), Inventory of Statements About Self Injury (ISAS), Self-Harm Information Form (SHIF), Self-Harm Inventory (SHI), to develop a common measurement metric for 568 Australians aged 18-30 years old in Australia. The results had three co-calibrations with Cronbach's alpha ranging from 0.690 to 0.827 and different scales occupied different ranges on the hierarchy of DSH (prevalence estimates ranging from 47.7 to 77.1%), meaning scales with different difficulty

levels can still be co-calibrated. This study provides a raw score conversion table and the hierarchy of DSH behaviors from six DSH scales (Latimer, Covic, & Tennant, 2012).

In summary, varied IRT linking methods were used in the previous studies, including Rasch partial credit, Rasch rating scale model, IRT summed scores approach, Samejima's graded response IRT model, general response/generalized partial credit model, two-parameter logistic graded response model, common person equating and Stocking-Lord calibration, compared to qualitative and conceptual linking or equipercentile methods used in the CTT studies. Table 2.2 demonstrates a summary of each article that used IRT-based linking methodologies in different domains of healthcare in the order of time.

The majority of the crosswalk validation studies supported score translatability between instruments with acceptable agreement using statistics such as intraclass correlation coefficients (ICC) or Cohen's effect size at group-level comparison (Askew, at al., 2013; Bjorner, Kosinski, & Ware, 2003; Holzner, et al., 2006; Qude, et al., 2014; Ten Klooster, et al., 2013; Wang , Byers, & Velozo, 2008a). For instance, Orlando and colleagues (2000) examined the validity of the cut score generated from the sum-score translation method by comparing depression classification rates of respondents at the 18-month using both the original and the translated scores, and found nearly 95% of the sample are classified in the same categories. Ten Klooster and colleagues (2013) found different IRT models can generate reliable crosswalks between observed and translated scores with similar agreement of ICC ranging from 0.72 to 0.82. Qude and colleagues (2014) found that the crosswalk between instruments could produce reliable score conversions at the diagnostic-subgroup level in a cross-cultural setting.

While most studies showed successful linking results using IRT at the group-level, it is noticeable that linking may not work as reliably as expected at the individual-level (Askew, at al.,

2013; Holzner, et al., 2006; Ten Klooster, et al., 2013; Wang , Byers, & Velozo, 2008a). For instance, Holzner and colleagues (2006) found that the confidence intervals of translated scores for individual subjects were very large, thus the limited precision of individual scores are likely to lead to unreliable measures of individual differences. Wang and colleagues (2008a) found that only 37 ~ 67% of the translated scores were within 5 points of the actual scores at individual-level comparison. Fischer and colleagues (2011) found that individual scores comparison is imprecise due to substantial statistical spread. Askew and colleagues (2013) recommended that individual scores derived from crosswalks should be used for the group-level analysis instead of using in clinical care given the additional source of inherent errors. In addition, Ten Koolster and colleagues (2014) found substantial discrepancies in agreement within individual patients. Thus, we expected that linking approach would produce better accuracy at group-level classification.

Methodological Issues Related to Linking

Chen and colleagues (2009) stated that when conducting linking, it is important to recognize the strategies in sampling and linking procedures (Haly et al., 2011). Dorans (2007) suggested three types of sampling procedures in linking, including sampling the same people, collecting the same test items, or a combination of both; and two types of linking procedures, one is to put all items in the same pool and co-calibrate the items, while the other is to use the common items to calibrate different instruments (Haly et al., 2011). In addition, three different approaches can be used to link scores from different instruments, including equating, scale alignment and prediction (Dorans, 2007). Noonan and colleagues (2012) compared these three linking methods and proclaimed that the more restrictive the approach used, the closer the link between scores. The most restrictive linking method is equating with five required assumptions:

equal construct, equal reliability, population invariance, equity and symmetrical of the linked instruments.

Consequently, several potential concerns needed to be addressed when conducting linking to ensure minimizing potential errors and maximizing reliability and validity of the final linking product. Based on the literature, the factors potentially influencing the linking results include sample size, source of items, number of items, breadth and depth of measurement, item difficulty, type of rating scale, scaling method, and psychometric rigor of the linked instruments (Chen, et al., 2009; Doran, 2007; Lia, Cella, Yanez, & Stone, 2014).

For instance, Fisher (1997) examined several studies with sample size ranging from 53 to 30,000 subjects, and along with Cook et al.'s (2007) study, these researchers stated that it is necessary to have sample sizes of 300 or more for linking health outcome measures when using IRT methodologies (i.e., Graded Response Model (GRM), the Partial Credit Model (PCM), and the Generalized Partial Credit Model (GPCM)) with the empirical evidence showing that the averaged R square values within in the sample size of 150 was 0.91 and for all other sample size from 150, 200, 300, 400, 500, 750, 1000, 1500 to 2000, the averaged R square values increased to 0.92. But in general, there is no interaction effect between model and sample size. Fischer and colleagues (2012) found inherent psychometric properties did not significantly change the results of transformed sum scores, but could lead to significantly different F values and effect sizes due to the increased main effects and interaction (Fischer, et al., 2012).

Several linking studies controlled for the pre-existing errors by removing invalid subjects or items before conducting linking procedures using a developmental sample (Latimer, Covic, & Tennant, 2012; Velozo, Byers, Wang, & Joseph, 2007). Some studies examined internal consistency of the instruments or conducted total score correlation between instruments prior to

executing linking procedures to ensure that there was a similar construct measured across instruments (Carmody, et al., 2006; Holzner, et al., 2006).

IRT Models

Thus, it is critical to choose an appropriate linking method and fulfill corresponding assumptions in order to use the linking strategy successfully. However, when considering linking strategies, multiple IRT-based linking strategies are available (Embretson, 1996; Orlando, Sherbourne, & Thissen, 2000; McHorney & Cohen, 2000). Accordingly, when using IRT-based analysis, one should take into account the different model assumptions, and the final model choice should be selected based on several different aspects, such as dimensionality, or the discrimination equality of the items (Embretson, 1996; Orlando, Sherbourne, & Thissen, 2000; Ten Klooster, et al., 2013).

In general, every IRT model needs to consider three item parameters: item discrimination (a parameter), item difficulty (b parameter), and guessing (c parameter). While the 1-parameter model (1-P; assumes that the data have no discrimination differences and guessing) and 2-parameter model (2-P; assumes that the data have no guessing) are most commonly used in healthcare because guessing parameter is not a crucial concern as it is in education. It may be challenge to determine whether 1-P or 2-P is the best model to apply since each model has its own specific strengths and limitations.

For instance, 1-P holds the strictest assumptions which are not easy to be fulfilled by real observations, but it is the easiest model to interpret both the results and its implication. Thus, a 1-P-based instrument may be more meaningful and easier for the practitioners to use. While a 2-P may fit better with the real observations with more flexibility compared to 1-P, it is more difficult to interpret the 2-P-based results. One of the major limitations of the 2-P was that 2-P

could adjust item discrimination to improve the data-model fit, so fit statistics from 2-P are lacking the confirmatory function as those in 1-P due to the fact that 1-P identifies the ideal model in advance.

However, when comparing the results statistically generated from 1-P and 2-P methods, there was a high correlation (nearly 99% in certain scenarios) of person measures between these two models (Hambleton, 1989). Ten Klooster and colleagues (2013) also found that different IRT models (i.e., 1-P model, 2-P model (Generalized Partial Credit Model; GPCM) and 3-P model (multidimensional GPCM model)) produced similar linking products even though the fundamental model assumptions are inherently different. Thus, it could simply be considered that 1-P and 2-P have “methodological differences”.

Although using the 2-P extension may improve model fit, a 2-P-based linking approach is less straightforward compared to a 1-P method, because the observed sum score is no longer a sufficient statistic for the trait level estimation and resulting crosswalk contains a second source of statistical error (Ten Klooster, et al., 2013). A more conservative way is to report the results from both models (1-P and 2-P) and to examine if any differences of the results exist between models.

The Rasch model, belonging to the 1-P family, has the major advantage of the capability to generate a more straightforward crosswalk that is more robust against statistical error than the 2-P family. Since all items are equally discriminating and each observed total score is associated with only one latent trait (theta) score in the Rasch model (Andrich, 2004; Bond, & Fox, 2007; Ten Klooster, et al., 2013). In addition, the Rasch model is the only IRT model that allows translating one-to-one from the IRT score (measure score, logit) to the summed scores (raw score), thus a linear raw-measure score conversion can be automatically generated (Orlando,

Sherbourne, & Thissen, 2000). Due to its straightforward linking characteristic and simplicity in result interpretation and application, the Rasch model was selected as a fundamental basis to link instruments in this dissertation.

Creating an Item Bank

While there is considerable evidence to support translating scores between instruments, the findings have been limited to translating scores between two or more instruments. An important implication not addressed by the literature is that the statistical findings that support translating scores across instruments also support combining existing instruments into an item bank. The authors know of no study that has combined translatable instruments (existing instruments) into a single item bank and further create short forms.

The proposed study will combine two features of (a) the previous linking studies by combining existing instruments (co-calibration) to create and item bank, and (b) the del Toro and colleagues (2011) approach to develop short forms from the item bank and further validate their accuracy and precision. In contrast to the previous linking studies, this dissertation focused on the psychometric development of the item bank instead of simply developing a score conversion table. This dissertation also compared the precision of different test forms such as the item banks, short forms with different numbers of items.

Studies are needed to compare the psychometrics of different test forms derived from the item bank using existing instruments. Few studies (n=3) in healthcare using IRT models to address the comparisons of different test forms (Choi, Reise, Pilkonis, Hays, & Cella, 2010; Lai, et al., 2011, Bojner, 2014). Regarding of different test forms such as short forms and CATs besides item bank, Choi and colleagues (2010) found that short forms and CATs produced highly correlated scores compared with full-bank scores, and dynamic short form (using a two-step

process including a screening question to select one of two short forms) generate measures that have comparable to CATs. Lai and colleagues (2011) found CATs in general had better precisions than short forms but all three short forms (4, 8, 12 items) showed good precision for more than 95% of the sample (individuals with fatigue) with a reliability greater than 0.9. Boiner and colleagues (2014) found that no statistically or clinically significant differences in score levels in different methods of administration among two non-overlapping parallel 8-item forms from three PROMIS domains (physical function, fatigue, and depression).

When considering measurement precision, since the item bank has all the items of the combined instruments, the item bank was expected to have the highest precision. While the CATs were expected to be able to approximate the precision of the item bank, recent studies surprisingly demonstrated that well-developed short forms could approximate the precision of the item bank and the CATs (Bjorner, et al., 2014; Choi, Reise, Pilkonis, Hays, & Cella, 2010; Fries, Cella, Rose, Krishnan, & Bruce, 2009; Lai, et al., 2011). This dissertation investigated precision and accuracy of different test forms generated from the item bank.

This dissertation differs from the PROMIS approach in that two existing instruments were combined into an item bank without changing the original root or rating scales of the original items. Typically, PROMIS studies go through an extensive process to create an item bank by modifying existing items so that all items have the same root and rating scale. Combining existing instruments instead of modifying current items to construct the item bank has the advantage for the researchers and the clinicians to use sets of items from the instruments in their original item structure (e.g., if clinicians are used to using a particular instrument, they can select items from that instrument) instead of imposing them to use a new instrument or modified items.

In addition to comparing the precision of the varied short forms and the item bank, a critical question is about how well these short forms can perform in real-life applications. For instance, the FIM is used by the CMS in an algorithm to derive FRGs for the PPS. Thus, it is important to know whether the converted score and also the short forms derived from a "function" item bank (e.g., combining ADL instruments) generated comparable FRG classifications to those derived from the original FIM. If using different test forms (here meaning different instruments, and also different lengths of the instruments) can generate comparable FRG classifications when measuring patients' function, then the usefulness of short forms can be further established. If the measurement precision and accuracy among different test forms are similar, then no matter which test form the clinical practitioners choose to use, they can obtain equivalent results.

In summary, linking can enhance meaningful score comparison, facilitate interpretation of scores across studies or populations, and may be useful for measuring longitudinal effects or monitoring continuous functional changes. In addition, generating shorter version of the instrument from the linked item bank could facilitate feasibility of the linked instrument. The present study may be a precursor to using IRT-based linking strategies to co-calibrate different instruments (e.g., depression or pain measures) into an item bank based on selected item and person parameters. Developing an item bank of existing instruments further facilitated the generation of a variety of administration test forms, could provide a viable alternative to mandating that all rehabilitation facilities use existing instruments, allowing healthcare facilities to continue using current instruments and avoid the training and costs associated with adopting a new measurement system. By validating the precision and accuracy of different test forms, the findings of this dissertation will facilitate generating state of art healthcare measurement across the continuum of care measurement.

CHAPTER THREE

METHODOLOGIES

Hypotheses, Research Designs Measurement and Statistical Approaches

3.1 Specific Aims and Hypotheses

The overall purpose of this dissertation is to utilize an IRT measurement model to establish the best item bank (self-care physical function) using the existing instruments, Functional Independence Measure (FIM) and the Minimum Data Set (MDS) accuracy across the continuum of post-acute care (PAC), and also to develop flexible administration formats (4-item and 8-item short forms) and validate their measurement precision and accuracy.

The fundamental theoretical basis to link FIM and MDS is the latent trait model, assuming the same construct measured across instruments can be equivalently compared (Hambleton, Swaminathan, Cook, Eignor & Gifford, 1978). After a single item bank was developed by linking FIM and MDS, there were two specific phases in this dissertation, including phase 1, to build the state-of-art instruments, including full item bank, 4-item and 8-item short forms (Aims I and II) and phase 2, to validate precision and accuracy of the varied instruments (Aims III and IV) (Figure 3.1). A detailed study procedure of both phases is illustrated in Figure 3.1. Specific aims for this dissertation were described as follows:

Specific Aim I: Create a FIM-MDS item bank measuring daily motor that meets Item Response Theory (IRT) model requirements

Hypothesis of Aim I: This aim did not have hypothesis. However, prior to proceed to Aim I, the operational hypothesis is that, based on the latent trait model, the FIM and MDS measure the same latent trait (daily motor**)**; therefore, the instruments can be linked.

Specific Aim II: **Generate IRT-based 4-item and 8-item short forms from the item bank**

This aim did not have hypothesis. But this aim assumed that once the item bank meets the IRT requirements, for instance, the criteria of unidimensionality, then IRT-based short forms could be established.

Specific Aim III: **Compare measurement precision of the IRT-based short forms and the MDS converted score to the original FIM scores**

Hypothesis of Aim III: The 4- and 8-item short forms created from the previous Aims and the MDS converted scores have similar measurement precision compared to the original measure.

Specific Aim IV: Assess the accuracy of the IRT-based short forms and the MDS converted scores in classifying Veterans into Function Related Groups (FRGs) compared to the data collected from the original FIM (treating as a standard)

Hypothesis of Aim IV: The 4- and 8-item short forms and the MDS converted scores will categorize Veterans into the same FRGs levels that are categorized using the original FIM score.

3.2 Data Source

Data were retrieved from existing databases maintained by the Austin Information Technology Center (AITC) in Texas. The FIM and MDS data reside in two separate databases at the AITC. FIM data are contained in the Function Status and Outcomes Dataset (FSOD) (10N), and MDS data are maintained in the dataset for the Office of the Assistant Deputy under Secretary for Health at the Patient Care Services (10P4).[1]

Demographic variables such as age, gender, ethnicity and marital status were retrieved from the FSOD. Clinical and administrative variables were retrieved from the FSOD and MDS, including the impairment classification system of International Classification of Diseases, 9th revision, Clinical Modification (ICD-9 CM), the duration between dates for admission and discharge assessments of the FIM and the MDS. The two datasets were merged based on the scramble social security number for each Veteran at the Center of Innovation (COIN) on Disability and Rehabilitation Research (CINDRR). We only obtained de-identified data and analyzed the data at the Medical University of South Carolina (MUSC). This dissertation is part of the larger research project funded by the Department of Veterans Affairs, Health Services Research and Development from North Florida/South Georgia Veterans Health System (NFSGVHS) CINDRR. The Institutional Review Board for Human Research (IRB) at the NFSGVHS, UF and MUSC approved this study protocol prior to executing any study analysis.

3.3 Study Design

This dissertation used retrospective, secondary, national Veterans data and IRT common person equating method to link and validate a crosswalk between the FIM and MDS. We chose

[1] This study is part of the funded grant entitled "Item Banking across the Continuum of Care" funded by the Department of Veterans Affairs, Health Services Research and Development. Thus the method session is largely overlapped with the contents in the grant written by the original Principle Investigator, Dr. Craig A. Velozo. When Dr. Velozo took a position at the Medical University of South Carolina and a WOC at the Ralph Johnson VA Medical Center in Charleston, SC, Dr. Sergio Romero at the CINDRR became the PI.

to use common person equating method because the dataset had the same individual responded to both instruments (Dorans, 2007). In contrast to using raw score methodologies, We used Rach analysis, one-parameter IRT model, to create interval measures, an essential requirement for the most basic arithmetic operations, and also to create sample-free item calibrations, thus allowing the creation of FIM-MDS short forms (SFs)[1].

Based on the IRT assumptions, FIM-MDS item bank and the generated short forms would retain their item calibration structure for any sample from a population. Thus, the item bank created from this study provide a critical connection across two important continuums of health care measures, the FIM used at the inpatient rehabilitation facilities (IRFs) and the MDS used at the Community Living Centers (CLCs) in the Veterans healthcare services system.

3.4 Participants

To minimize the potential functional status change in Veterans between FIM and MDS assessments, only respondents from the Veterans AITC system who completed both the FIM and the MDS assessments within seven days or less were selected for analysis. We decided on seven days because FIM is required to be re-assessed every week and the MDS is required to be re-assessed within 14 days. This inclusion criterion included the patients who had rapid transition between the IRFs and CLCs.

A total number of 3000 Veterans were stratified randomized into two samples for phases 1 and 2 to represent the diversity of diagnoses. The first sample of 500 Veterans was used for Aims I and II; and the second sample of 2500 Veterans was used for Aims III and IV. First sample (N=500) was used to create a FIM-MDS item bank that meets IRT requirements, and generate IRT-based 4-item and 8-item short forms (SFs) from the item bank (Aims I and II). The

second sample (N=2500) was used to compare precision and accuracy of the IRT-based SFs, MDS converted score and the original FIM measure (Aims III and IV).

3.5 Clinical Measures

The Veteran's Health Administration (VHA) system incorporated components of the Uniform Data System for Medical Rehabilitation (UDSmr), the most widely used clinical database for assessing inpatient rehabilitation facilities (IRFs) outcomes, into the VHA Functional Status and Outcomes Database (FSOD) (Fiedler, & Granger, 1997; Granger, & Hamilton, 1993). [1]

The VHA Directive 2000-016 requires every VHA medical center to assess functional status of every Veteran patient who has new stroke, lower extremity amputee, and orthopedic impairment; thus the rehabilitation outcomes of these patients could be tracked in the FSOD (VHA Directive 2000-016, 2002). All clinical raters who submit data to the AITC need to complete training and credentials on FIM data collection to achieve 80% agreement through the UDSmr FIM Credentialing Examination. The practitioners who administered the MDS also need to complete required training before executing MDS assessment.

Self-care motor, as recognized as the Activity of Daily Living (ADL), was be represented by 13 items from the FIM (in the FSOD) and 13 items from the MDS (Table 3.1). Both the physical ADL items (total N=26) were included in the analysis. The FIM items were administered in inpatient rehabilitation facilities (IRFs) settings while the MDS items were administered in the Community Living Centers (CLCs) (also known as skilled nursing facilities, SNFs).

The FIM has 18 items measuring disability from basic activities of daily living to global activities, representing the core functional status measure of the FSOD. The FSOD is administered by clinicians and is used to produce IRF quarterly reports that provide the determinations of the Function Related Groups (FRG), the most common basis for development of quality indicators in rehabilitation. In this dissertation, we used 13 items from the FIM motor subscale to create the item bank.

The 13 FIM motor items have a 7-point rating scale (1 total assist, 2 maximal assist, 3 moderate assist, 4 minimal assist, 5 supervision, modified independence-device, 7 complete independence-no device), and 12 of 13 MDS motor items have two ratings scales: self-performance (0 independent, 1 supervision, 2 limited assistance, 3 extensive assistance, 4 total dependence, 8 activity did not occur) and support provided (0 no setup or physical help, 1 setup help only, 2 one person physical assist, 3 more than two physical assist, 8 activity did not occur over the last 7 days). Three items in the MDS have rating scales that differ from above (0-4, and 8; 0, 2, 3, and 4) (Table 3.1).

While the IRFs use the FIM as the gold standard for measuring functional outcomes, the Minimum Data Set (MDS) of the Resident Assessment Instrument (RAI), is the gold standard used for monitoring similar functional outcomes in CLCs. The Omnibus Budget Reconciliation Act of 1987 (OBRA 87) federally mandated that all CLCs in the United States report the MDS for Medicare prospective payment reimbursement (Rantz, 1999). CLCs play a critical role for providing the context and tracking the healthcare status for elderly Veterans. Specifically, the VHA is the largest single provider of skilled nursing home care in the U.S., with 133 community living centers (Tsan, et al., 2008) and at least 1.5 million skilled nursing facility residents

participating in the Medicare or Medicaid programs nationwide (Jones, Dwyer, Bercovitz, & Strahan, 2009).

The MDS has 284 items assessing the cognitive, behavioral, functional and medical status of individuals residing in the skilled nursing facility (Morris, 1990), which was later renamed as Community Living Centers (CLC). Lawton et al. (1998) concluded that the items used in the MDS reflected important indicators of the physical and cognitive status of CLC residents and, thus, could be used to determine quality of care. The nurses in charge of each unit monitor assessment processes of the MDS along with relevant information provided by licensed nursing assistants, social workers, activities staffs, and medical staff (Lawton, et al., 1998). The MDS is assessed at patient admission to the skilled nursing facility, subsequently each quarter (approximately every 92 days), and/or when there is a relevant change in the patient's condition (Lawton, et al., 1998).

Previous research has provided evidence that both the FIM and the MDS have adequate reliability and validity. For the FIM, Stineman and colleagues (1996) identified the factor structure of the FIM with motor and cognitive dimensions across 20 impairment categories with 93,829 rehabilitation inpatients. Internal reliability for the FIM subscales ranged from 0.86 to 0.97 and exceeded the minimum criterion for discriminate validity (Stineman, et al., 1996). In a meta-analysis of 11 studies, the median inter-rater reliability for the total FIM was 0.95 and the test-retest reliability of the FIM was 0.95 (Ottenbacher, Hsu, Granger, & Fiedler, 1996). Rasch analysis, a 1- parameter IRT model, supported and indicated that the FIM had two constructs: motor and cognitive dimensions (Linacre, Heinemann, Wright, Granger, & Hamilton, 1994).

For the MDS, early studies showed that MDS items had excellent reliability with interclass correlations of 0.7 or higher in both the physical and cognition functioning domains (Hawes, et al., 1995). Sixty-three percent of the MDS items achieved reliability coefficients of 0.6 or higher and 89% achieved 0.4 or higher. The MDS cognitive scale corresponded closely with the Mini-Mental State Examination (MMSE), nursing judgments of disorientation, and clinical neurological diagnoses of Alzheimer's disease and other dementias (Morris, et al., 1994). The seven MDS cognitive items (short term memory, long term memory, decision making and four categories of memory recall) had an internal reliability of 0.83 to 0.88 (Morris, et al., 1994). The MDS assesses two unidimensional constructs, physical and cognition functioning (Wang, Byers, & Velozo, 2008a). In this dissertation, we only used 13 items from the MDS.

Studies suggest that the cognitive scale of the FIM and the MDS, respectively, are not as sensitive as the motor scale. For instance, Davidoff, Roth, Haughton, and Ardner (1990) failed to find a significant relationship between the cognitive subscale of the FIM and a comprehensive neuropsychological battery for patients with spinal cord injury discharged from acute rehabilitation. In addition, the cognitive construct of the MDS is not as effective as the FIM's motor scale in stratifying the functional level of CLC residents (Wang, Byers, & Velozo, 2008). Thus, in this dissertation, we only linked motor items from the FIM (n=13) and the MDS (n=13).

3.6 Statistical Software and Data Management

Microsoft Access was used for merging data and matching data. SAS version 9.4 was used to manage data and conduct descriptive/inferential analysis (SAS Institute; Carry, NC, USA). Winsteps version 3.57.2 was used for Rasch analysis, including fits statistics, rating scale diagnoses, monotonicity and person strata (Linacre, 2014). To ensure we used consistent model

across all the analyses, we also use Winsteps to identify Differential Item Functioning (DIF) items and obtain person measure errors to draw total test error plots (Linacre, 2014). Mplus version 7.1 was used for factor analysis and residual correlation matrix (Muthén, & Muthén, 2014). For all statistical analyses, the selected level of significance was set at 0.05.

3.7 Data Analyses

Descriptive statistics was performed for the two subsamples (N=500 and N=2500), such as age, gender, ethnicity, diagnoses, marital status, days between administrations of FIM and MDS, FIM/MDS raw scores and measure scores. Each aim in this dissertation has its own specific plans of statistical analysis, listed as follows:

Aim I: Create a FIM-MDS Item Bank that Meets Item Response Theory (IRT) Model Requirements

We conducted the IRT and related psychometric analyses based on the PROMIS instrumental developing and maintaining procedures for item bank. The purpose of Aim I was to develop an IRT-based item bank. Thus, the item bank needs to fulfill the IRT models assumptions, including unidimensionality, local independence and monotonicity.

3.7.1 Unidimensionality

Unidimensionality is a principal requirement of the IRT model, representing a scale measures only one construct and the single construct accounts for all item covariance (Tennant, & Pallant, 2006). We used both the fit statistics and the factor analysis to determine if the proposed self-care motor item bank is “essentially” unidimensional that meets with the following required standards of unidimensionality.

3.7.1.1 Rasch Fit Statistics

Rasch fit statistics is an index to measure the difference between the estimated scores of the Rasch model and the observed scores (Bond & Fox, 2007; Wu & Adams, 2013). MnSq (mean square standardized residuals), representing observed variance divided by expected variance, was used to assess the extent of unidimensional level of each item. A low MnSq value (e.g., <0.9) implies that an item fails to discriminate respondents with different levels of ability or that item is redundant. While a high MnSq value (e.g., >1.1) implies that scores are variant or erratic, indicating that item does not belong to the same continuum as the other items or that the item is probably misinterpreted. Items with high MnSq values represent a threat to validity and were given greater consideration. For clinical scales, Wright and Linacre (1994) suggested a reasonable range of MnSq fit values being within 0.5 to 1.7, along with associated standardized fit statistics (ZSTD) values between ±2.0.

It is important to note that fit statistics alone are not sufficient to be used as assessing the dimensionality of an instrument (Smith, 2002). The more appropriate approach is to consider together both the results from fit statistics and factor analyses.

3.7.1.2 Confirmatory factor analysis (CFA)

The CFA identifies the number and nature of the underlying latent factors with the prior assumption that all items load on the same/one factor based on unidimensional model. A polychoric correlations matrix was analyzed using a weighted least squares estimator with four model fit indices, including the comparative fit index (CFI, > 0.95), Tucker-Lewis Index (TLI, > 0.95), Root Mean Square Error of Approximation (RMSEA,

< 0.06) and standardized root mean residuals (SRMR, < 0.08) (Hu, & Bentler,1996). The factor loadings and average absolute residual correlations were used to confirm the factor structure.

3.7.1.3 Principal Components Analysis (PCA) of Rasch residuals

The Rasch residual PCA was used to assess if there were meaningful structures of residuals after extracting the primary Rasch dimension. First contrast in the Rasch residual PCA represents the first PCA component in the correlation matrix of the residuals after extracting the Rasch dimension (Linacre, 2004, 2010 & 2012). Linacre (2004, 2010 & 2012) suggests that unidimensionality of an instrument is supported when the Rasch dimension explains more than 40% variance of the data, the first contrast of the Rasch residual explains less than 5% variance of the data, and the eigenvalue of the first contrast is less than or equal to 2.0.

3.7.2 Local Independence

Local independence means the response to any item is unrelated to the response to any other item, which can be identified by the residual correlation matrix produced by the factor analyses with Mplus. High residual correlation was an indication of local dependence and the cut-off point of 0.2 from PRIMIS standard manual was used (PROMIS®, 2014). In other words, items with residual correlations above 0.2 were flagged as violating local independence (Reeve, et al., 2007b).

However, local dependence could be a particular challenge in this study because it is reasonable to maintain as many as possible items from the FIM and the MDS in the final item bank with which clinicians are familiar (e.g., FIM in IRFs and MDS in CLCs). Consequently, it

is likely that the final item bank may include items that are locally dependent (e.g., eating item from the FIM and eating item from the MDS). Thus, Reeve and colleagues' approach (Reeve, et al., 2007b) of retaining locally dependent items was used to maintain the quality of preserving items, but marking them as "enemies" preventing locally dependent items from being administered to any individual. This procedure allowed us to create a "FIM" short form and a "MDS" short form generated from the item bank, allowing clinicians to use the items with which they are most familiar with (e.g., FIM or MDS) but are not locally dependent.

3.7.3 Monotonicity

Monotonicity signifies that the average ability estimates for all persons in the sample who choses that particular response category increase as the numbers in the rating scale increases. In other words, the probability of endorsing a rating scale response indicative of better function should increase as person ability increases. If the predicted order is reversed, meaning this item "violates" monotonicity. The monotonous pattern of category logit measure was examined by the ordered pattern of the rating scale response from the Winsteps Rasch diagnostic summary table outputs.

3.7.4 Differential Item Functioning (DIF)

DIF item means that individuals with the same level of ability do not have the same probability of endorsing a particular item due to the fact that they are belonging to different groups (e.g., male, female). For instance, diagnostic DIF item (i.e., stroke, traumatic brain injury, and lower extremity amputee patients) for the FIM and the MDS could be the communication items because respondents with similar cognitive abilities are likely to show different levels of communication abilities (i.e., respondents with left hemisphere stroke would possibly

demonstrate more deficits than those with orthopedic damage on the communication item) due to different diagnoses (i.e., left hemispheric stroke versus orthopedic damage). Winsteps Rasch-Welch (logistic regression) t-test was used to examine differential item functioning (DIF) items for Veterans under or over 65 years old (Linacre, 2014). The items are identified as a moderate to large DIF item if the DIF contrast ≥ 0.64 logits at significant level of $p>0.05$; and identified as a slight to moderate DIF item if the DIF contrast ≥ 0.43 logits at significant level of $p>0.05$ (Zwick, Thayer, & Lewis, 1999).

Aim II: Generate IRT-based Short Forms and Computer Adaptive Tests from The FIM-MDS Item Bank

We recognized that varied ways could be used to construct short forms (del Toro, et. al., 20111; PROMIS®, 2014; Yu, et al., 2011). Since there are no definitive studies showing one method is superior over another, we used the short form development procedures based on the simplest model, Rasch model, by del Toro and colleagues' (2011).

3.7.5 Short Form Development

We eliminated any items with high residual correlation to construct the short form used in this dissertation. To ensure that each patient responded consistently to both instruments before developing a valid item bank, we also eliminated Veterans with person measures that fell outside of the 95% confidence interval error identity line. We used del Toro and colleagues' (2011) Boston naming short form procedures, including: (a) excluding items with high residual correlations > ±0.2 to minimize item redundancy, (b) creating intervals with 2 standard errors apart starting at the item with mean item difficulty level (logit=0) to cover a full spectrum of

item difficulty, and (c) choosing the items with item discrimination closest to 1 to best fit the Rasch model.

We anchored the FIM and the MDS items to the item bank using the co-calibrated item difficulties and item step thresholds prior to developing the short forms. The short form analysis was then anchored on the co-calibrated item difficulties and step thresholds. Two final short forms were constructed. The 4-item short form and the 8-item short form generated from the item bank, FIM, and MDS. Each short form consisted of items spread across difficulty levels, and item discrimination values that were close to 1.

Aim III: Compare Measurement Precision of the Varied IRT-Based Short Forms and the MDS Converted Scores to the Original FIM Measures

An independent validation dataset of 2500 participants was used to compare the precision of the varied IRT-based short forms and the MDS (n=13) converted scores. The ability estimate based on the original FIM was considered as the "gold standard."

Six new administration forms (short forms from the FIM, the MDS and the item bank) were generated. A series of analyses were conducted to compare the measurement properties across different administration forms: 1) original FIM (13 items), 2) 4-item FIM short-form, 3) 8-item FIM short-form, 4) original MDS (13 items), 5) 4-item MDS short-form, and 6) 8-item MDS short-form for measuring self-care motor.

The ability estimates and associated standard error (SE) from different administration forms were obtained. It is assumed that each respondent answered identically in the full administration of the item bank and also each administration form (original, 4- and 8-item short-

forms). We defined “bias” as the difference in the ability estimate associated between the standard and an administrative form.

3.7.6 Person- and Item-level Psychometrics Comparisons

Person- and item-level psychometrics of each test form were reported, including: person ability (Mean ± SD), minimum and maximum of person measure, item difficulty (Mean ± SD), minimum and maximum of item difficulty, percentage of persons with maximum person measure, and percentage of persons with minimum person measure.

Significant ceiling/floor effects were identified when more than 5% of the sample had the maximum/minimum person measures. We also calculated the correlations between the full-length test forms (i.e., item bank, FIM_13 and MDS_13) and the corresponding 4- and 8-item SFs (i.e., item bank_8 items, item bank_4 items, FIM_8 items, FIM_4 items, MDS_8 items, and MDS_4 items).

3.7.7 Precision Comparisons

For each test form (original test form and generated short forms), we compared their measurement precision based on three approaches:

(a) Comparing person strata calculated from the person separation index of Rasch analysis. Person separation Index from Rasch analysis was used to determine the number of person ability strata (clinical group differences; distinguishable person ability levels) with the formula of (person separation index*4+1)/3 (Andrich, 1982).

(b) Generating the standard error of measurement (SEM) plot for each test form based on Rasch model. Gibbons and colleagues (2014) suggested using a cut-off value of SEM as 0.3 to

represent a reliability level of 0.90 for a scale with 12 items. The SEM values were presented graphically over the challenge level of test items in order to investigate how much the scale attains measurement precision across the challenge level of the scale.

(c) Calculating 95% confidence interval (CI) of the person measure standard error (SE) between the full-length administration form (i.e., item bank, FIM_13 and MDS_13) and the corresponding 4- and 8-item SFs.

Aim IV: Assess Measurement Accuracy of the IRT-Based Short Forms and Item Bank in Classifying Veterans into Function Related Groups (FRGs)

The Functional Independence Measure–Function Related Groups (FRGs) classification system was developed by Stineman and colleagues (1994, 1995 & 1997). We used the FRG classification system to examine whether the IRT-based short forms, the MDS_13 converted scores could classify the same patient into the same or a similar classification group compared to that derived from the original FIM measure.

The Centers for Medicare & Medicaid Services (CMS) uses Case Mix Groups (CMGs), a form of FRGs, as a basis for the IRF prospective payment system (PPS) (Stineman, 1995). The FRG algorithm uses the FIM motor (13 items) and the FIM cognitive (5 items), along with patient's age at admission to the IRF to predict the costs of treating Medicare patients (Figure 3.2; for the Rehabilitation Impairment Classification – RIC for stroke). Based on an impairment (i.e., stroke or lower extremity amputation), patients were classified into one of 20 impairment categories. Note that each category has a specific FRG model. Figures 3.3 – 3.5 showed the FRG algorithms for lower extremity amputation, knee replacement and hip replacement. Patients assigned to different FRGs are expected to have different rehabilitation outcomes and total costs

of care. Thus, the FRGs classification system provided a pragmatic accuracy examination of the newly generated measures (i.e., short forms) when comparing with the original FIM scores.

To assess the accuracy between administration forms in classifying Veterans into FRGs, we used weighted kappa to examine agreement strength for the stroke, knee replacement, and hip replacement FRG calculations. We used kappa and McNemar's test to provide a 2x2 table for the lower extremity amputation FRG calculation due to its dichotomous FRG classification algorithm. A weighted kappa statistic for categorical data ranging from 0.21 to 0.40 demonstrates a fair strength of observer agreement, from 0.41 to 0.60 represents a moderate strength of agreement, and from 0.61 to 0.80 indicates a substantial strength of agreement (Landis & Koch, 1977). McNemar's statistics was used to test whether any association existed between classification results. The McNemar test is a test on a 2x2 classification table to test the difference between paired proportions. A value of 0.05 was used as cutoff significance in this study. Kappa statistics was used to quantify the strength of association; a kappa statistic ranging from 0.21 to 0.40 indicating a fair strength of agreement, 0.41 to 0.60 indicating a moderate strength of agreement, and 0.61 to 0.80 indicating a substantial strength of agreement (Landis & Koch, 1977).

Since the variability of the data could significantly bias the kappa classification results, we examined the percentage of agreement in each diagnostic group instead of simply relying on weighted kappa results. Finally, we also calculated a two-way mixed method Intraclass Correlation Coefficient (ICC) between FRGa (FRG generated from the actual FIM score) and FRGc (FRG generated from the converted FIM score) for all test forms across the four diagnostic groups. However, ICC also had similar limitation as the kappa results.

3.8 Final Products Generated for Each Specific Aim

The end product for each specific aim was described as follows: For Aim I, a final item set, the motor item bank, was generated after the items meet the IRT-based criteria, including unidimensionality, model fit, monotonicity and local independence, and also the criteria of differential item functioning. For Aim II, the IRT-based short forms were established, including: FIM_4-item short form, FIM_8-item short form, MDS_4-item short form, MDS_4-item short form, Item Bank_4-item short form, and Item Bank_4-item short form. For Aim III, the test error plots were generated and the person strata were calculated for each administration form. For Aim IV, the percentage of individuals classified into the same, one FRG category apart (±1 level) and two FRG categories apart (±2 levels) were calculated. The strength of agreement between the original and the converted scores, as well as the ICC was presented. A summary table of each specific Aim with corresponding hypotheses, statistical methods and final expected products was demonstrated in Table 3.2.

3.9 Strengths and Limitations of the Methods Used in this Study

In order to recognize the advantages and limitations of the methods used in this dissertation, a comparison was made with three other study designs, using the dataset of (a) the National Health and Nutrition Examination Survey (NHANES), (b) the Medicare Data, and (c) a prospective study using a single tool at different facilities (Table 3.3). Both the NHANES and the Medicare datasets are national retrospective datasets. While the NHANES is a cross-sectional database containing serial national survey data since 1960 on the health and nutritional status of community-dwelling individuals in the United States (NHANES, 2014), the Medicare dataset is administrative data with CMS separated Medicare billing data from different healthcare

providers, such as inpatient hospitals, Medicare Part B providers and skilled nursing homes. The prospective study is a hypothesized study that aims to collect data for the same patient using both the single instrument (i.e., CARE item set) and the existing instruments (i.e., FIM and MDS) and compared the differences of the measurement results. To the authors' knowledge, currently the CMS funded researchers are conducting a prospective study; however, we have not found any published articles, therefore, we did not have any evidence to support or against our hypothesis that whether using one single instrument would generate the same or different errors as using existing instruments.

The advantages and the limitations of each study design is addressed based on the following features: sampling frame, characteristics of the dataset, required resources, internal validity, external validity and miscellaneous factors that may contribute to secondary variance or errors which may influence the study results (Table 3.3). The advantage of the proposed study design includes large sample size, less resourced needed (also time and cost) in terms of data collection and better internal reliability compared to the prospective study. In addition, the two instruments (FIM and MDS) are actual tests developed independently and are extensively used in current IRFs and CLCs compared to the NHANES study design. An advantage of the proposed retrospective study versus a prospective study is that both the patients the practitioners were blind to the study purposes when their data were collected, which contributes to better internal validity.

An additional advantage is that this dissertation used the data collected for clinical and administrative reporting purposes in real life, implying the real-life applicability, for instance, the data used in the present study may include the error encountered in real-life practice and could reflect the real-life scenario in the Veterans healthcare system.

The limitation of the proposed study include the homogeneity of the Veterans' dataset leading to decreased external validity (generalizability) because the sample is restricted to the Veterans population instead of the general population. For instance, the Veterans dataset had a characteristic that the vast majority of the respondents were male compared to the general population. In addition, even though we only included the same patient who took the FIM and MDS within 6 days, to avoid possible functional changes between being assessed by the two instruments, however, it is possible that the patients' functional status may change over these 6 days, which could possibly produce undesirable secondary variance on the outcome variable such as responding to the two instruments inconsistently. However, Wang and colleagues (2008a) found that decreased the days between two instruments administrated (e.g., decreased to 3 days) still produced similar results as 7 days. We decided to use the common scenario, which was to use a discharge FIM from IRFs and the MDS on admission to CLCs.

In summary, the study design of this dissertation has several advantages in terms of sampling frame, required resources and internal validity compared to the other three study types. However, the Medicare project may have comparable advantages and limitations and the CMS-funded prospective study may have better generalizability even though the prospective study would require much more additional cost and time to be completed (Table 3.3).

3.10 Conclusion and Implications

This study aims to link the FIM (13 items) and the MDS (13 items) motor items of the same person based on common person equating methods using the IRT Rasch model, and to validate the measurement precision of different administration forms (4-item and 8-item short form generated from the item bank). We assumed that the linking tools could provide comparable

precision and also accuracy when classifying patients into FRGs compared to using a single instrument twice within the same period of time.

The proposed study intended to develop the state-of-art motor measure across the continuum of post-acute care (PAC) for the Veteran population. In this dissertation, we specifically focused on the transition from acute to IRF to CLC (SNF) settings. In addition, we generated multiple IRT-based administration forms to reduce patients' and healthcare practitioners' assessment burden while at the same time maximizing measurement precision with sufficient breath that the item bank provides.

This dissertation challenged the current efforts to develop a single instrument across PAC and represents the potential for considerable cost savings by maintaining existing instruments and reimbursement systems (i.e., it would be unnecessary to develop the new instrument and also to unnecessary to train practitioners on new instruments). Future studies can apply the same methodologies in the extended dataset for different research areas. For instance, using the Medicare dataset to compare the total cost between using the linking tool and a single tool, in terms of FRGs classification results. In addition, future studies could link additional instruments used currently across PAC, such as MDS (used in the SNFs) and Outcome and Assessment Information Set (OASIS) (used in the Home Health Agencies; HHAs). Additionally, the same study design and methodologies could be used with different population (e.g., depression) and for different instruments (e.g., varied fear of falling scales), to replicate and validate the study design and results. In summary, this dissertation could provide meaningful and practical applications in the field of healthcare measurement.

CHAPTER FOUR (Manuscript_1)

Continuum of Care Assessment across Post-Acute Care in Veterans: Linking Existing Instruments to Develop an Activity of Daily Living Item Bank

Abstract

Objective: This paper aimed to develop and examine dimensionality and item-level psychometric properties of an item bank measuring Activities of Daily Living (ADL) physical function in the continuum of post-acute care settings.

Design: An item response theory-based common person equating method was used with the retrospective data. Factor analyses, fit statistics and principal component analysis of Rasch residuals were used to examine dimensionality, model fit, local independence and monotonicity. Differential item functioning (DIF) was used to determine DIF items.

Setting: Inpatient rehabilitation facilities and community living centers in the Veterans healthcare system.

Participants: 371 Veterans completed both instruments within 6 days from October 2008 to September 2010.

Interventions: NA

Main Outcome Measure(s): Pooled item responses from the Functional Independence Measure (FIM) and the Minimum Data Set (MDS)

Results: The FIM-MDS item bank demonstrated good internal consistency (Cronbach alpha= 0.98), met three criteria for the rating scale diagnoses (e.g., monotonicity) and three of the four model fit statistics (unidimensionality: CFI/TLI=0.98, RMSEA=0.14, and SRMR=0.07). One item (MDS walk in corridor) had residual correlation ≥ 0.2, violating local independence.

Principal component analysis of Rasch residuals showed that the item bank explained 94.2% variance. The item bank covered the range of theta from -1.50 to 1.26 (item), -3.57 to 4.21 (person) with person strata of 6.3. One item (MDS bowel control) (3.8%) had slight to moderate DIF across age groups, with a DIF contrast from Winsteps larger than 0.43 ($p<0.05$).

Conclusions: The findings indicated the ADL physical function item bank constructed from FIM and MDS items measured a single latent trait with overall acceptable item-level psychometric properties, suggesting it is an appropriate source for developing efficient test forms such as short forms and computerized adaptive tests.

Keywords: continuity of patient care, activities of daily living, Veterans

Introduction

Based on the nature of disease progress, patients may need healthcare services in a variety of post-acute care (PAC) to meet with their evolving needs. The term “trajectory of care” has been coined to describe healthcare services that a patient receives during their recovery process. “A trajectory of care” is synonymous with the term “episode of care”, used in section 5008 of the Deficit Reduction Act (DRA) in 2005, meaning “*the care a patient receives in order to treat a spell of illness associated with a hospitalization. A trajectory may include one or more settings*” (Centers for Medicare and Medicaid Services; CMS, 2012), whereas “*a spell of illness*” covers “*all readmission and skilled nursing facility service use*” based on Medicare's definition (Research Triangle Institute International (RTI), 2009). A trajectory of PAC is provided across varied facilities, such as inpatient rehabilitation facilities (IRFs), skilled nursing facilities (SNFs; known as community living centers, CLCs, in the Veterans healthcare system), home health agencies (HHAs), long-term care hospital (LTCH) and outpatient therapy services (OTS). Based on a five percent national sample of 2006 Medicare claims data, over a third (35.2%, n=109,236) of all beneficiaries discharged from acute facilities transited to at least one type of PAC facility

(RTI, 2009). In addition, 52 percent of this group of beneficiaries went on to use at least one additional PAC service after the first PAC site (RTI, 2009). In 2007, the Medicare Payment Advisory Commission (MedPAC) spent over $45 billion dollars on PAC (RTI, 2009). Based on its high utilization rate and cost, PAC plays an important role for patients, healthcare practitioners and policy makers.

One major challenges resulting from the continuum of post-acute care is to assess and monitor the function of patients as they transfer across different facilities. The main reason this challenge exists is that different instruments are used across the PAC continuum. For instance, the required PAC site-specific patient assessment tools for different settings include the Inpatient Rehabilitation Facility Patient Assessment Instrument (IRF-PAI) (i.e., the Functional Independence Measure (FIM) with additional demographic data such as age and gender) for the IRFs and the Minimum Data Set (MDS) for the SNFs/CLCs.

The use of different instruments across the PAC results in two major issues: 1) patient care is interrupted since functional scores are not easily translated from one facility to the next and 2) it is difficult to establish a fair reimbursement system when different facilities base their reimbursement on different functional scores. Two potential solutions could possibly solve the above-mentioned challenges. The traditional psychometric method, known as Classical Test Theory (CTT) or true score theory, supports the development of a single measurement system for all PAC venues. This is based on the concept that using a single instrument could potentially decrease measurement errors and thus further increase test reliability. However, the development and implementation of a single measurement system has significant drawback in terms of the considerable costs and challenges in implementing a new tool (e.g., modifying electronic medical records, re-training therapists on a new assessment). These barriers resulted in the CMS

terminating the implementation of the MDS-PAC, as a uniform PAC outcomes measure in 2000 (Wang, Byers, & Velozo, 2008a).

An alternative solution, which avoids the aforementioned drawbacks, is to use modern test theory, such as item response theory (IRT)/latent trait model, to link existing instruments and translate scores from different instruments across the PAC continuum. That is, all facilities could continue to use their existing instruments since a conversion system would be created to translate measures across existing instruments. The IRT methods accomplish this by assuming that an equivalent construct can be co-calibrated across different instruments, and the estimated scores of a respondent can be used to predict or explain test performance based on the latent traits of a person (Hambleton, Swaminathan, Cook, Eignor & Gifford, 1978). We hypothesized that the IRT methods can be used to combine existing measures into a single item bank that measures a single latent trait with measurement precision similar to that of using a single instrument.

An initial demonstration of the latent trait model that would support using existing instruments to measure equivalent construct across the PAC continuum is to determine whether the items on different instruments can be linked (Dorans, Pommerich, & Holland, 2010; Haley, et al., 2011; Kolen & Brennan, 2004; Velozo, Byers, Wang, & Joseph, 2007; Wang, Byers, & Velozo, 2008a). This study aimed to establish a FIM-MDS item bank that provides acceptable IRT psychometrics based on the assumption that the FIM and MDS measures a single latent trait, activity of daily living (ADL).

Methods

Participants

Data for the study were extracted from the existing databases maintained by the Veterans Austin Information Technology Center (AITC). The FIM and the MDS data resided in two

separate databases at the AITC. FIM data were contained in the Function Status and Outcomes Dataset (FSOD), and the MDS data were maintained in the dataset for the Office of the Assistant Deputy under Secretary for Health at the Patient Care Services. These two datasets were merged by patient identifiers and these data were then de-identified at the COIN (Center of Innovation): Center of Innovation on Disability and Rehabilitation Research (CINDRR); North Florida/South Georgia and Tampa. The subsequent data analysis was performed at the Medical University of South Carolina. The Institutional Review Boards (IRB) for Human Research at the University of Florida and the Medical University of South Carolina approved study protocol.

The data were limited to Veterans who had: (1) new stroke, (2) lower extremity amputation, (3) knee replacement and (4) hip replacement and who were assessed on both instruments (FIM and MDS) without any missing items. We chose distinguishable four diagnoses to minimize the possibility that the same individual would be classified into more than one functional related group in the following validation study. Also, we chose groups that were used in previous study using similar linking methodologies to allow for comparison of our results to those of the previous study. For inclusion in the study, the two assessments had to be administered within six days during the period of October 2008 to September 2010.

Statistical Analysis

SAS version 9.4 was used to merge and match data and to conduct descriptive/inferential analysis (SAS Institute; Carry, NC, USA). Mplus version 7.1 was used for factor analysis and residual correlation matrix (Muthén, & Muthén, 2014). Winsteps version 3.57.2 was used for Rasch analysis, including fits statistics, rating scale diagnoses (e.g., monotonicity), person strata, and principal component analysis (Linacre, 2014). Winsteps Rasch-Welch (logistic regression) t-test was used to identify differential item functioning (DIF) items (Linacre, 2014).

Linking Procedures

Rasch analysis common person equating method was used in this study. The co-calibration approach used in this study was based on Velozo and colleagues' (2007) first three steps of a similar study, including (a) using a pre-identified set of 26 items from the FIM and MDS measuring an equivalent construct of ADL, (b) removing invalid responses and (c) anchoring MDS and FIM person measures based on the co-calibrated FIM-MDS item difficulties and item step thresholds.

A sample of 500 Veterans were randomly stratified from a cohort of 3,000 Veterans, across four impairment groups (stroke, lower extremity amputation, knee replacement and hip replacement and) in this study. The person measures for the FIM and MDS were generated by anchoring separate analyses on item and step measures from a co-calibration of the 500 Veterans (Velozo, Byers, Wang, & Joseph, 2007). We employed Velozo and colleagues approach for removing invalid data (Velozo, Byers, Wang, & Joseph, 2007). The overall concept is to build the measurement instrument using the most valid data. For any ADL measure, a reasonable expectation is that patients should have similar scores on similar measures. For example, a patient with an overall score that represents dependence on the FIM is expected to obtain a score that represents dependence on the MDS. If this expectation is not met, the data are considered invalid and the patient's data is removed from the analysis. To accomplish this, we plotted FIM person measures against MDS person measures and excluded Veterans with person measures that fell outside of the 95% confidence interval error identity line. Through this procedure retained a sample of 371 (74.2%) Veterans for the following analyses in this study.

Item Bank Testing Based on IRT Model Requirements

The FIM-MDS item bank of 371 Veterans was examined to determine if it fulfilled the IRT model assumptions, including unidimensionality, local independence and monotonicity. We also identified items with differential item functioning (DIF) items, i.e., items showing a different probability of response from people from different age groups but the having same ADL ability. MDS data conversion procedures were based on previous Velozo and colleagues' (Velozo, Byers, Wang, & Joseph, 2007) study, from the original rating scale (i.e., 012348) to match with the rating scale of FIM (i.e., 1234567) for all the following analyses. This conversion procedure was also supported based on conceptual meanings of the items from both instruments (Jette, Haley, & Ni, 2003). Converting the rating scale enabled the scores to represent the patient's ability in the same direction from both instruments.

Confirmatory factor analysis (CFA) and Rasch fit statistics were used to determine if a scale is "essentially" unidimensional, meaning only a single construct was measured (Tennant, & Pallant, 2006). For clinical scales (Wright & Linacre, 1994), a reasonable range of mean square standardized residuals (MnSq) fit values were 0.5 to 1.7, with associated standardized fit statistics (ZSTD) of values between ±2.0 (Wright & Linacre, 1994). A CFA polychoric correlation matrix was used with a weighted least squares estimator of four model fit indices, including the comparative fit index (CFI, > 0.95), Tucker-Lewis Index (TLI, > 0.95), Root Mean Square Error of Approximation (RMSEA, < 0.06) and standardized root mean residuals (SRMR, < 0.08) (Hu, & Bentler,1996; Reeve, et al., 2007b). The factor loadings and average absolute residual correlations were also used to confirm the factor structure. We hypothesized that the FIM-MDS item bank is a one-factor model structure by measuring the same latent trait of ADL.

The Rasch residual principal components analysis (PCA) was used to assess if there were meaningful structures of residuals after extracting the primary Rasch dimension. First contrast in

the Rasch residual PCA represents the first PCA component in the correlation matrix of the residuals after extracting the Rasch dimension (Linacre, 2004, 2010 & 2012). Linacre suggested that unidimensionality of an instrument is supported when the Rasch dimension explains more than 40% variance of the data, and the first contrast of the Rasch residual explains less than 5% variance of the data (Linacre, 2004, 2010 & 2012). Local independence was identified by the residual correlation matrix produced by the factor analyses with Mplus. The items with residual correlations beyond ±0.2 were identified as violating local independence (PROMIS®, 2014; Reeve, et al., 2007b).

The rating scale structure was evaluated based on three criteria: 1) having at least ten responses in each rating category, 2) a monotonous pattern of category logit measure, and 3) the outfit mean square value for each rating scale was less than ±2.0 (Linacre, 2002). Monotonicity was examined by the increase of the probability of endorsing a rating scale response while the person ability increases. If the predicted order is reversed, it means that the item "violates" monotonicity. Rasch-Welch (logistic regression) t-test examined group differences across age (under 65 versus over 65 years). The items were identified as a DIF item if the DIF contrast ≥ 0.43 logits at significant level of $p>0.05$ (Zwick, Thayer, & Lewis, 1999).

All psychometric analyses were accomplished using the 371 Veterans. Items in the item bank that did not fit the unidimensional model, have residual correlation above ±0.2, have significant DIF values, were reviewed by the research team to determine if the items should be removed, the clinical relevance was also used to make final item elimination decisions. The final item bank, that meets the essential requirement of unidimensionality, was used for Rasch analysis to generate point-measure correlation, person strata and item-person map. Point-measure correlation is an index demonstrating the Pearson point-measure correlation coefficients between

the item observations and the corresponding Rasch measures (estimated including the current response) (Linacre, 1998). A value larger than the absolute value of 0.3 was considered acceptable. Person separation index was used to calculate the number of levels of person ability (person strata) distinguished by the item difficulties and calculated as (4Gp + 1)/3, where Gp is person separation (Wright & Masters, 1982, p. 106). An item-person map was used to determine ceiling/floor effects. Greater than 5% of the sample being at the ceiling or floor was considered as significant ceiling/floor effects.

Results

Participants had a mean age of 67.0 years old (*SD*=11.0), with a range from 22 to 90 years old. Six (1.6%) Veterans who were older or equal 90 years old were grouped as one group and were identified as 90 years old. The majority of the participants in this study were male (n=354, 95.4%), White (n=233, 62.8%) and married (n=161, 43.4%) (Table 1). The average number of days since onset was 173.4 ± 1331.3 days, about 6 months. The mean days between the administrations of the FIM and the MDS was 3.1 days (SD=2.1), with a range from zero to six days. There were 164 (44.2%) Veterans with stroke, 77 (20.8%) with lower extremity amputation, 74 (19.9%) with knee replacement and 56 (15.1%) with hip replacement (Table1).

The FIM-MDS item bank met three out of four model fit criteria (CFI/TLI=0.98> 0.95, RMSEA=0.14> 0.06, and SRMR=0.07< 0.08) for treating the item bank measuring one factor (Table 2). The PCA showed that Rasch dimension (person and item measures) explained 94.2% variance of the scale, far above 40%, and the first contrast of the Rasch residual explains 0.8% variance of the data, far less than 5% criteria. The person reliability (Cronbach alpha) of the 26-item FIM-MDS item bank was 0.98. All test items met three rating scale criteria such as monotonicity and showed local independence, except one item (MDS *walk in corridor*) which

had residual correlations above ±0.2 with two items: MDS *walk in room* (0.272) and MDS eating (-0.242) (Table 2). All items had point-measure correlations larger than 0.3 (range from 0.56 to 0.90). The raw scores of the FIM and the MDS correlated at -0.93. The measure scores of the FIM and the MDS correlated at 0.85. The raw scores and the anchored measure scores of the FIM and the MDS correlated at 0.93 and 0.85, respectively, after adjusting for rating scale direction. One item, MDS *bowel control*, had DIF contrast of 0.56, larger than criteria of 0.43 ($p<0.05$), indicating slight to moderate DIF (Figure 1).

A total of 15 items (57.7%) from the item bank showed fit statistics between 0.5 and 1.7. Misfitting items included five items with high infit values and six items with low infit values (Table 4). Items with high fit values did not fit well with the Rasch model; while the items with low fit values were Guttman-like items (fit the model too well). For practical reasons, we had more concerns about items with high fit values, which were MDS *bladder* and *bowel control*, *locomotion off unit*, *walk in corridor* and *walk in room.* The items with low fit values included FIM *dressing upper and lower body*, bathing, *toileting*, *toilet (transfer)* and *bed/chair/wheelchair (transfer)*. The items with high fit values were all MDS items and the items with low fit values were all FIM items. In general, the average person ability (Mean=0.49, SD=0.20) was higher than the item difficulty of the item bank (Mean= 0.0, SD=0.05). Person measures had skewed distribution towards the end of higher ability (Figure 2).

The item difficulty hierarchy showed eating was the easiest item and walking was the most difficult item (Table 4 & Figure 2). The range of item difficulty of the item bank is 2.76 (Min= -1.50, Max=1.26) logits while the range of person ability is 7.78 (Min=-3.57, Max=4.21) logits. Overall, the MDS items were slightly more difficult (0.55 ± 1.3) than the FIM items (0.36 ± 1.5). The MDS items covered a wider range of item difficulty (range=2.76 logits) and had the

easiest and the most difficult items in the item bank compared to the FIM items (range=1.98 logits). The person separation index was 4.51 and person strata was 6.3 (Table 3).

Discussion

This study was the first step to establish a psychometrically sound item bank prior to propose an alternative solution for developing the PAC continuum measurement by co-calibrating two existing ADL instruments currently used across PAC settings. The FIM-MDS item bank demonstrated overall good item-level psychometric properties, including good internal consistency, good person strata, good point-measure correlation, overall good model fit and acceptable fit statistics for 21 of 26 items, indicating that both instruments measure the same construct (ADL; self-care physical function). The compatibility of the FIM and the MDS was also supported by the high correlations of both the raw scores and the measure scores. One item, MDS *bowel control*, had slight to moderate DIF and one item, MDS *walk in corridor*, had high residual correlations. However, we kept both items in the final item bank in order to cover a full spectrum of item difficulty levels in the item bank because these two items were the easiest and the most difficult item. In addition, the CFA results supported 1-factor model of all 26 items. Last, we retained all 26 items in the final FIM-MDS item bank because our following studies could minimize the concerns of item redundancy by not choosing multiple items with high correlations or flagging only one of the highly correlated items since we would develop short forms from the item bank.

Compared to Velozo and colleague's study (2007), both studies used the same linking method (i.e., Rasch common person equating) and demonstrated similar psychometric properties of the FIM-MDS item bank for the similar population (i.e., Veterans with disabilities). This study had a larger sample size (371 versus 236) and was slightly more restrictive on the number of

days between administrations between FIM and MDS (6 versus 7 days), suggesting more reliable study results. The FIM-MDS item bank in this study demonstrated better internal consistency (0.98 versus 0.94), better point-measure correlations (0.56-0.90 versus 0.54-0.84), similar raw score and person measure correlations (-0.93, 0.85 versus -0.81. 0.72) but more misfitting items (eleven vs. five misfitting items). The higher percentage of misfit items may be due to Veterans having an overall higher ability than the mean item difficulty in this study compared to a more well-matched item difficulty/person ability distributions in the previous Velozo et al. (2007)'s study. However, both studies showed consistent results for four misfit MDS items, including MDS *bladder control*, MDS *locomotion off unit*, MDS *walk in corridor* and MDS *walk in room*. This finding was consistent with several studies that suggested incontinence and ambulation items should be considered as separate constructs other than ADL (Nilsson, Sunnerhagen, & Grimby, 2005; Velozo, Magalhaes, Pan, & Leiter, 1995; Velozo, Byers, Wang, & Joseph, 2007). Only current study utilized CFA, PCA and residual correlations to elaborate the determination of factor structure for the item bank while previous Velozo et al. (2007)'s study only utilized Rasch analysis to determine unidimensionality of the scale. In summary, both studies supported that the self-care physical function items of the FIM and MDS measured the same construct with acceptable to good item-level psychometric properties.

This study showed the FIM-MDS item bank had an ADL item difficulty hierarchy that was similar to that found in previous studies (Linacre, Heinemann, Wright, Granger, & Hamilton, 1994; Velozo, Byers, Wang, & Joseph, 2007; Wang, Byers, & Velozo, 2008a & b; Velozo, Magalhaes, Pan, & Leiter, 1995), supporting the previously-identified concept of global measurement system of the physical ADL functioning. This global ADL item difficulty

hierarchy has been demonstrated across diagnostic groups and different populations such as the Veterans.

The current study particularly focused on co-calibrating the FIM and MDS items and developing a psychometrically sound item bank, instead of developing a raw-score conversion table between instruments (Velozo et al., 2007). The optimal goal of current study was to generate a linked item bank that could be applied in efficient administration formats such as short forms (SFs) and computerized adaptive tests (CATs), to decrease the assessment and respondent burden for practitioners and patients, respectively. Establishing a well-developed item bank is the first step to further developing efficient delivery forms. Thus, the positive findings of this study are a crucial first step to developing a linked measurement system that can be applied across the PAC continuum. By using data collected for clinical and administrative reporting purposes, the results of this study have clear implications for future clinical applications. The results of our study suggest that a linked FIM-MDS item bank can be the foundation for SFs and CATs which would provide for continuous and efficient assessments that are practical for clinical practice, without the need to adopt a new single instrument across PAC continuum.

Study Limitations

The first limitation of the study is the possibility of functional changes between the administration of the two instruments. To reduce the influence of functional changes, this study only included the data of the same Veteran who had completed both the FIM and the MDS data within 6 days; however, it is still possible that the patients' function could change over this short period of time, which may potentially produce undesirable noise in the data. However, Wang and colleagues (2008a) found that using a 3-day window between administrations produced similar

results as a 7-day window. Based on that finding, the length of time between FIM and MDS administrations may not significantly affect the outcome measures of the current study. A second limitation of this study was that the data used were restricted to the Veterans population, which may have different demographic characteristics such as most individuals were male and tended to be older compared to the general population. Thus, the results might have limited generalizability. However, the psychometric results of the item bank may not differ across Veterans and civilians (i.e., eating items represent the easiest items and walking items represent the most difficult items for both Veterans and civilians). Furthermore, this study used the retrospective data that did not prospectively collected for the purposes of this study. Thus, the existing limitations such as rater bias could not be controlled in the data. Lastly, removing person measures that differed significantly between the FIM and MDS before co-calibrating the two instruments may favor more promising psychometric qualities. Note, that the logic behind this "cleaning" of the data, is to build the item bank using only valid responses (i.e., having the same individual scored high on one instrument indicating high functional ability and low on the other instrument indicating low functional ability is assumed to be due to invalid scoring). The second phase of our larger study, the validity testing, will use the data from all subjects (i.e., no elimination of invalid responses).

Conclusions

This study found that the FIM-MDS item bank had acceptable to good item-level psychometric properties, suggesting a single construct could be measure by these two instruments. We will use this item bank to develop short forms to decrease assessment burden for the clinical practitioners. In addition, we will conduct future studies to investigate the measurement precision and accuracy of the item bank and its multiple test forms, comparing the

item bank and the test forms against the original instrument scores (i.e., the original 13-item FIM).

Appendix

Table 1. Demographic Characteristics of Participants in this Study (n=371)

Variables	Community-Dwelling Veterans (n =371)	
	Number	%
Age (range: 22-90 y/o)	Mean=67.0	(SD=11.0)
Age Group		
≤ 65 y/o	Mean=58.8	(SD=0.39)
> 65 y/o	Mean=76.9	(SD=0.55)
Averaged number of days since onset	Mean= 173.4	(*SD*=1331.3)
Gender		
Male	354	95.4
Female	14	3.8
Missing	3	0.8
Ethnicity		
White	233	62.8
Black	83	22.4
Native American	4	1.1
Hispanic	19	5.1
Other	19	5.1
Missing	13	3.5
Diagnoses		
Stroke	164	44.2
Lower Extremity Amputation	77	20.8
Knee Replacement	74	19.9
Hip Replacement	56	15.1
Marital Status		
Single	37	10.3
Married	161	43.4
Widowed	26	7.0
Separated	18	4.9

Divorced	118	31.8
Missing	11	3.0
Days between Administrations of FIM and MDS (range=0-6)	Mean= 3.1	(*SD*=2.1)
FIM Raw Score	Mean= 63.5	(*SD*=22.8)
FIM Anchored Measure Score	Mean= 0.36	(*SD*=1.5)
MDS Raw Score	Mean= 30.0	(*SD*=25.8)
MDS Anchored Measure Score	Mean=0.55	(*SD*=1.3)

Table 2. Factor Analysis of the FIM-MDS Item Bank (n=371)

Dimensionality Analysis Criteria	FIM-MDS
CFI (>0.95)	0.98
TLI (>0.95)	0.98
RMSEA (<0.06)	0.14
SRMR (<0.08)	0.07
Local Independence (Residual correlation ≤ ±0.2)	96.2% (25/26) items
Monotonicity	100% (26/26) items

Table 3. Item-level Psychometric Properties of the FIM-MDS Item Bank (n=371)

	FIM-MDS Item Bank
Person Reliability (Cronbach alpha)	0.98
Person separation Index	4.51
Person Strata	6.3
Person Ability	Mean=0.49, SD=0.20 Min=-3.57, Max=4.21 (Range=7.78)
Item Difficulty	Mean=0, SD=0.05 Min=-1.50, Max=1.26 (range=2.76)
Misfitting Items (Both High and Low Fit)	42.3% (11/26) items
Floor Effect	0% (0/371) persons
Ceiling Effect	0% (0/371) persons

Table 4. Item Difficulty Hierarchy of the FIM-MDS Item Bank (n=371)

Items	Score		Model S.E.	Infit		Point Measure Correlation
	Raw	Measure		Mnsq	ZSTD	
walkcorridormds	1125	1.26	.04	1.81	8.5	.61
STAIRFIM	1188	1.16	.04	1.42	4.8	.67
bathingmds	1284	1.01	.04	1.10	1.2	.73
walkroommds	1313	.96	.04	1.75	7.8	.67
locomoffunitmds	1597	.48	.04	2.13	9.9	.64
dressingmds	1658	.37	.04	.78	-2.8	.84
TRANTUBFIM	1668	.35	.04	.78	-2.8	.83
toiletingmds	1706	.28	.04	.65	-4.8	.87
WALKFIM	1712	.27	.04	.86	-1.8	.82
BATHFIM	1734	.22	.04	.48	-7.7	.89
DRESSLOWFIM	1753	.19	.04	.49	-7.5	.88
TOILETFIM	1773	.15	.05	.32	-9.9	.90
TRANTOILETFIM	1779	.13	.05	.40	-9.0	.89
hygienemds	1782	.13	.05	.79	-2.6	.85
TRANCHAIRFIM	1833	.02	.05	.34	-9.9	.90
locomonunitmds	1950	-.25	.05	1.53	4.9	.78
DRESSUPFIM	1973	-.31	.05	.49	-6.6	.89
BLADDFIM	1976	-.32	.05	1.22	2.2	.82
BOWELFIM	1996	-.37	.05	1.14	1.4	.81
GROOMFIM	2028	-.46	.05	.58	-5.0	.87
bedmobilitymds	2052	-.53	.05	1.26	2.4	.82
transfermds	2089	-.64	.05	1.65	5.4	.73
eatmds	2140	-.80	.06	1.40	3.5	.78
EATFIM	2148	-.82	.06	.86	-1.4	.83
bowelmds	2193	-.97	.06	2.01	7.5	.76
bladdermds	2328	-1.50	.07	3.36	9.9	.56

Figure 1. Differential Item Functioning across Age (Age Group >65 or below)

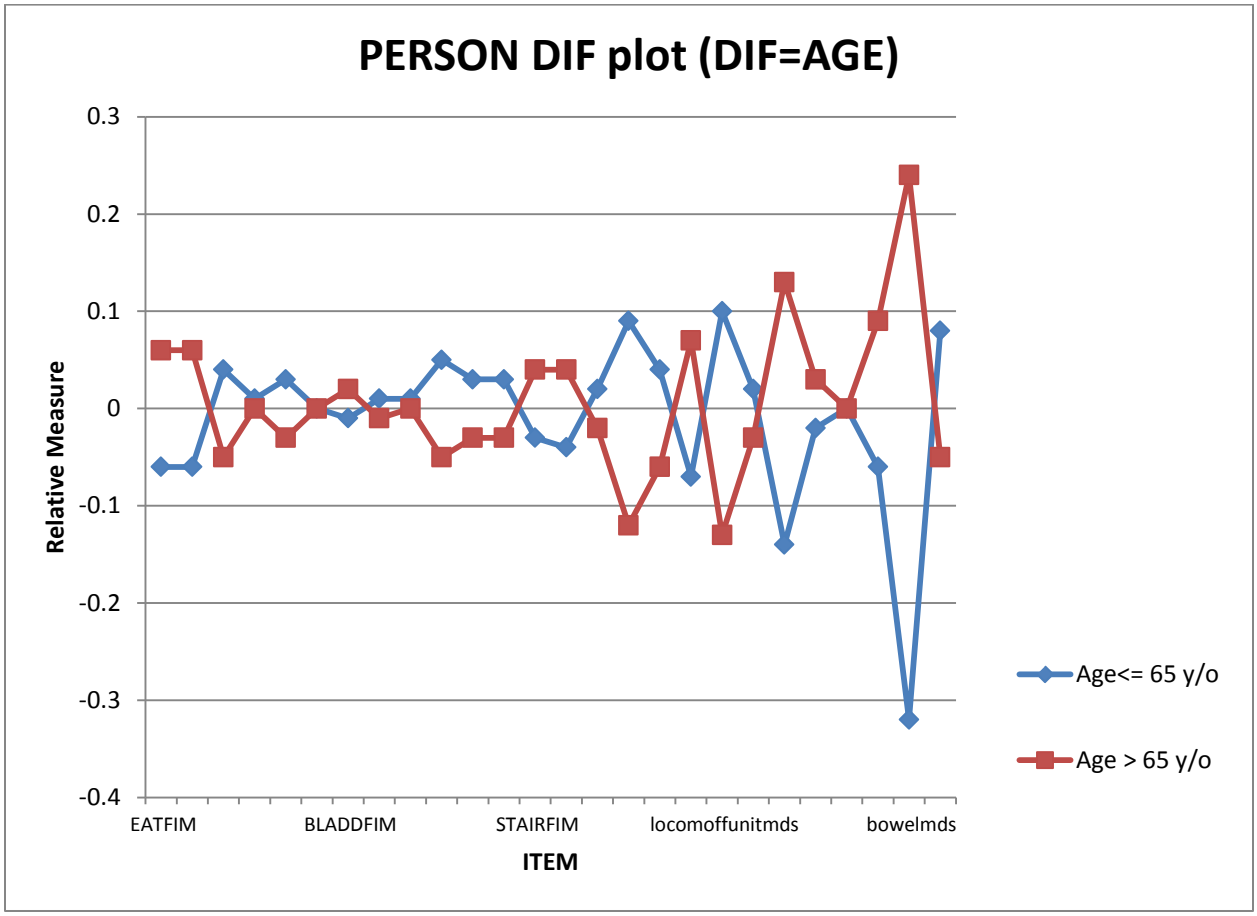


Figure 2. Item-Person Map of the FIM-MDS Item Bank

```
      PERSONS MAP OF ITEMS
              <more>|<rare>
   5                +
                    |
                    |
                    |
                    |
                .   |
   4                +
                    |
                    |
                    |
                    |
                #   |
   3                +
                . T|
                .   |
                .   |
                #   |
                #   |
   2            .##  +
               ###  |
            .##### S|
             .####  |
        .#########  |T walkcorridormds
      ############  |  STAIRFIM
   1  .###########  +  bathingmds       walkroommds
          .#######  |
           #######  |S
           ####### M|  locomoffunitmds
             .####  |  TRANTUBFIM       WALKFIM          dressingmds
                       toiletingmds
          .#######  |  BATHFIM          DRESSLOWFIM      TOILETFIM
                       TRANTOILETFIM    hygienemds
   0        .#####  +M TRANCHAIRFIM
             .####  |
               .##  |  BLADDFIM         BOWELFIM         DRESSUPFIM
                       locomonunitmds
               .##  |  GROOMFIM         bedmobilitymds
                .# S|S transfermds
                ##  |  EATFIM           eatmds
  -1             .#  +  bowelmds
                 .#  |
               .##  |T
                ##  |  bladdermds
                .#  |
               ### T|
  -2           ###  +
                .   |
                .   |
                #   |
                .   |
                    |
  -3                +
                    |
                    |
                .   |
                    |
                    |
  -4                +
              <less>|<frequ>
 EACH '#' IS 3.
```

Abbreviations: STAIRFIM=FIM_Stairs; bathingmds=MDS_Bathing; walkcorridormds=MDS_Walk_in_Corridor; locomoffunitmds=MDS_Locomotion_Off_Unit; dressingmds=MDS_Dressing; TRANTUBFIM=FIM_Tub, Shower (Transfer); walkroommds=MDS_Walk_in_Room; toiletingmds=MDS_Toilet_Use; WALKFIM=FIM_Walk/Wheelchair; BATHFIM=FIM_Bathing; DRESSLOWFIM=FIM_Dressing_Lower_Body; TOILETFIM=FIM_Toileting; TRANTOILETFIM=FIM_Toilet_(Transfer); hygienemds=MDS_Personal_Hygiene; TRANCHAIRFIM=FIM_Bed, Chiar, Wheelchair (Transfer); DRESSUPFIM= FIM_ Dressing_Upper_Body; BLADDFIM=FIM_Bladder_Management; locomonunitmds=MDS_Locomotion_on_Unit; BOWELFIM=FIM_Bowel_Management; GROOMFIM=FIM_Grooming; bedmobilitymds=MDS_Bed_Mobility; transfermds=MDS_Transfer; eatmds=MDS_Eating; EATFIM=FIM_Eating; bowelmds=MDS_Bowel_Management; bladdermds=MDS_Bladder_Management

CHAPTER FOUR (Manuscript_2)

Continuum of Care Assessment across Post-Acute Care in Veterans: Comparisons of Functional Independence Measure-Minimum Data Set Short Forms

Abstract

Objective: This study aimed to generate feasible linking assessment in efficient administration formats of short forms (SFs) to decrease assessment burden for practitioners across the post-acute care settings. We compared 4- and 8-item SFs generated from a Functional Independence Measure (FIM™) - Minimum Data Set (MDS) self-care physical function item bank.

Design: The 4- and 8-item SFs were developed based on del Toro and colleagues' (2011) procedures. This paper examined person strata, ceiling/floor effects, person fits, test standard error (SE) plot for each administration forms and 95% confidence interval (CI) of anchored person measures with the corresponding SFs.

Setting: Veterans' inpatient rehabilitation facilities and community living centers.

Participants: 2500 Veterans who completed both FIM™ and the MDS within 6 days collected by the Veterans Austin Information Technology Center during years 2008 through 2010.

Interventions: NA

Main Outcome Measure(s): FIM and the MDS

Results: The six SFs were generated with 4- and 8-items across a range of difficulty levels from the item bank, FIM and MDS. Overall, SFs with the same number of items had similar person strata and test error. The three 8-item SFs all had higher correlations with the item bank (r=0.82~0.95), higher person strata and less test error than the corresponding 4-item SFs (r=0.80~ 0.90). The three 4-item SFs did not meet the criteria of SE less than 0.3 for any theta values.

Conclusions: In general, short forms with the same numbers of items demonstrated similar precision regarding person strata and test error. The 8-item SFs appear to have the best balance between precision and efficiency.

Keywords: outcome assessment (health care), activities of daily living, Veterans

Introduction

The Improving Medicare Post-Acute Care Transformation Act of 2014 (the IMPACT Act), signed by President Obama on October 6, 2014, addressed the need to develop cross-setting quality measures, especially in the post-acute care settings of Long-Term Care Hospitals (LTCHs), Skilled Nursing Facilities (SNFs), Home Health Agencies (HHAs) and Inpatient Rehabilitation Facilities (IRFs) (Centers for Medicare & Medicaid Services (CMS), 2015). The IMPACT Act stated that "…by using common standards and definitions in order to provide access to longitudinal information for such providers to facilitate coordinated care and improved Medicare beneficiary outcomes…" (CMS, 2015).

Thus, it is crucial to establish a continuum of care measurement across post-acute care (PAC) facilities for the purposes of monitoring patients' function and ensuring fair healthcare reimbursement. While developing a single instrument across facilities to measure patients' function is a traditionally acceptable solution, this approach inevitably demands a considerable amount of money, time and resources to construct a new tool with new items, as well as extensive training that could cause a tremendous burden for the healthcare practitioners (CMS, 2011). An alternative solution to the problem is to link existing instruments to generate a continuum of care measurement, allowing different settings to keep their existing instruments, avoiding the complications of adapting a new single measure such as administration training, or the need to change the original electronic support systems. Linking existing instruments based on

item response theory (IRT) methodology has the advantage of using IRT inherent linking nature to construct an item bank, and developing efficient administration forms such as short forms or computerized adaptive tests (CAT). Based on previous findings (Buchanan, Andres, Haley, Paddock, & Zaslavsky, 2004; Wang, Byers, & Velozo, 2008a), we assumed that developing an item bank using of linked instruments would have similar error levels as using the original instruments. However, one issue arisen was that an item bank might lead to relatively large sets of items (e.g., > 25).

This concern can be resolved through the creation of efficient instruments. Thus, methods are needed in order to create short forms for clinicians and patients to use. Generating short forms from a linked item bank would reduce patients' and the healthcare practitioners' assessment burdens. However, it was not clear whether the shorter versions of the instrument could introduce more or similar error compared to the original instruments. Traditional ways researchers used to create short forms including analysis of variance such as stepwise regression (Bukstein, McGrath, Buchner, Landgraf, & Goss, 2000) and factor analysis (Landgraf, 2007). However, these traditional methods tended to create short forms with ceiling and floor effects. One way to avoid these limitations is to use the IRT-based methodologies. In addition, the advantage of IRT-based short forms could select items covering low, medium and high item difficulties that match with the range of person abilities. Thus, this study focused on investigating measurement precision of the SFs composed of different numbers of items from the item bank based on IRT methods.

In a previous study, our research team created an item bank combining the Functional Independence Measure (FIM™) and the Minimum Data Set (MDS) (Li, et al, 2015a). The developed FIM-MDS item bank showed acceptable unidimensional model fit based on

confirmatory factor analysis (CFI/TLI=0.98, RMSEA=0.14, and SRMR=0.07) and good internal consistency (Cronbach alpha= 0.98), indicating a single dominating latent trait measured in the FIM-MDS item bank. The present study examined difference and similarities of measurement precision of varied short forms generated separately from different instruments (i.e., FIM and MDS) in the same item bank. We assumed that the generated short forms with the same item numbers would have comparable measurement precisions and produce similar person measures for each patient. While item banking allows for the linking of assessments across the continuum of care, short forms are needed to facilitate the feasibility of linked instruments and reduce assessment administration burdens for the clinicians and the patients. In summary, the main purpose of this study was to develop and compare the short forms generated from the item bank. Specifically, this study aimed to: (a) generating 4- and 8-item short forms from the previously validated self-care physical function item bank composed of FIM and MDS, and to (b) comparing measurement precision of the generated short forms.

Methods

Participants

A sample of 3000 Veterans was obtained from the Veterans Austin Information Technology Center (AITC). We conducted stratified randomization of this sample as 500 Veterans for item-bank development (phase I) (Li, et al, 2015a) and 2500 Veterans for using the developed item bank to generate the short forms and validate the precisions of the short forms (phase II). We only analyzed the second sample of 2500 Veterans in this study.

The participants included were the Veterans who: (a) had a stroke, lower extremity amputee, knee replacement or hip replacement; we chose these four with the intent to compare our findings to Wang et al. (2008a)’s study and also because these were the most distinguishable

four diagnoses that could classify the same individual into only one functional-related group, (b) completed both the Functional Independence Measure (FIM) and the Minimum Data Set (MDS) within 6 days through October 2008 to September 2010, and (c) did not miss any item in both instruments.

Statistical Analysis

SAS 9.4 was used to manage data and conduct descriptive data analysis (SAS Institute; Carry, NC, USA). Winsteps version 3.57.2 was used to generate person reliability index, person separation index, person measure, person misfit, person mean/standard deviation (SD), item mean/SD, and total test standard error plots (Linacre, 2014). Microsoft Excel version 2010 was used to compare person measures of two administration forms with 95% confidence interval plots of standard errors.

Developing Six Short Forms

We developed six short forms from the FIM-MDS item bank, FIM and MDS; including: (a) full bank_8 items, (b) full bank_4 items, (c) FIM_8 items, (d) FIM_4 items, (e) MDS_8 items, and (f) MDS_4 item short forms (SFs). We referred to the 13 item instruments as the FIM_13 and MDS_13 throughout the manuscript. The six short forms were compared to the FIM_13 and MDS_13 and the full item bank.

The short forms were generated based on del Taro and colleagues' (2011) Rasch short form development procedures, including (a) excluding items with high residual correlations > ±0.2 to minimize item redundancy, (b) creating intervals with 2 standard errors apart starting at the item with mean item difficulty level to cover a full spectrum of item difficulty, and (c) choosing the items with item discrimination closest to 1 to best fit the Rasch model. We anchored the FIM and the MDS items to the item bank using the co-calibrated item difficulties

and item step thresholds prior to developing the short forms. For example, FIM_8-item SF was developed from the item bank after anchoring the FIM_8-item to the item bank.

Comparison Measurement Precision between Short Forms

We used three approaches to compare measurement precision between the item bank and the short forms. The first approach was to compare person strata calculated from the person separation index of Rasch analysis. The second approach was to generate the standard error of measurement (SEM) plot for each test form based on Rasch model. Gibbons and colleagues (2014) suggested using a cut-off value of SEM as 0.3 to represent a reliability level of 0.90 for a scale with 12 items. The SEM values were presented graphically over the challenge level of test items in order to investigate how much the scale attains measurement precision across the challenge level of the scale. The third approach was to calculate 95% confidence interval (CI) of the person measure standard error (SE) between the full-length administration form (i.e., item bank, FIM_13 and MDS_13) and the corresponding 4- and 8-item SFs.

Person- and item-level psychometrics were also reported, including: person ability (Mean ± SD), minimum and maximum of person measure, item difficulty (Mean ± SD), minimum and maximum of item difficulty, percentage of persons with maximum person measure, and percentage of persons with minimum person measure. Significant ceiling/floor effects were identified when more than 5% of the sample had the maximum/minimum person measures. We also calculated the correlations between the full-length test forms (i.e., item bank, FIM_13 and MDS_13) and the corresponding 4- and 8-item SFs (i.e., item bank_8 items, item bank_4 items, FIM_8 items, FIM_4 items, MDS_8 items, and MDS_4 items).

Results

Participants had a mean age of 67.1 years old (*SD*=11.3), with a range from 19 to 90 years old. Sixty-three patients with age older than 89 were classified as aged of 90 years old and identified in the same age group. The majority of the participants in this study were male (n=2377, 96.2%), White (n=1576, 65.6%), married (n=1064, 42.5%), admitted for initial rehabilitation (n=2362, 94.5%), and pre-living setting was at an acute medical/surgical care unit in the same rehabilitation facility (n=1113, 44.5%) (Table1). The average length of days between the administrations of the FIM and the MDS is 3.2 days, with a range from 0 to 6 days. There were 1066 (42.6%) participants with stroke, 472 (18.9%) with lower extremity amputee, 568 (22.7%) with knee replacement and 394 (15.8%) with hip replacement (Table 1).

The FIM_13 had slightly higher person ability estimated means as the MDS_13 (0.77±0.29 versus 0.57±0.28) (Table 2). We investigated the relationship between the FIM_13 and MDS_13 in the same item bank to ensure both instruments measure the individuals in the same direction. A moderate correlation was found between person measures of the FIM_13 and MDS_13 (r=0.63). The MDS_13 had a wider spectrum of item difficulties and a slightly lower measurement precision compared to the FIM_13 (person strata= 4.17 and 3.84 for FIM_13 and MDS_13, respectively) (Table 2). The correlations of the person measures between of the full bank, FIM_13, the MDS_13 and the corresponding SFs were moderate to very high (r= 0.95 and 0.91 for full bank_8-item and full bank_4-item; r=0.99 and 0.96 for FIM_8-item and FIM_4-item; r=0.89 and 0.87 for the MDS_8-item and MDS_4-item. Overall, the full-length tests (i.e., item bank, FIM_13 and MDS_13) had higher correlations with all the 8-item SFs than all the 4-item SFs (Table 3).

The full item bank had the highest person strata of 5.4 and the MDS_4 item SF had the lowest person strata of 2.2 (Table 2). The full item bank had an overall better person strata and the least test total error compared to all the other test forms, covering the widest range of theta, which was a comparison standard in this study (Table 2, Figures 3). Item bank, item bank_8 item SF and item bank_4 item SF did not show any ceiling or floor effects. However, FIM_13, FIM_8 item SF and FIM_4 item SF all had floor effects and MDS_13, MDS_8 item SF and MDS_4 item all had ceiling effects. MDS_4 item had the largest ceiling effects (18.9%) while FIM_4 item had the largest floor effects (6.72%) (Table 2).

Figures 1-3 showed SE plots for the various combinations of 13 item instruments and SF instruments relative to the full item bank. Figure 1 shows the SE plots for all test forms. FIM_13 and MDS_13 had similar standard error (SE) patterns and were the closest to the SE pattern of the item bank (Figure 1). When comparing FIM_13, MDS_13 and all three 8-item SFs, the FIM_13 had a slightly better measurement precision compared to the MDS_13 between -5 logits and .3 logits. However, the MDS_13 showed better precision at the extremes. Especially at the lower end, the MDS_13 showed the same SE as the full item bank between -3 to-2 logits (Figure 2). The FIM_13 had similar test error compared to the all three 8-item SFs (Figure 2).

For all three full-length test forms (i.e., item bank, FIM_13 and MDS_13) and all six SFs (two from each), when the number of total test items decreased, the number of person strata decreased and the total test error increased (Table 2 & Figure 1). When the number of items was the same, it showed similar person strata among different administration forms (Table 2), but the measurement precision varied across the range of person ability (Figures 2 & 3). For example, the person strata were 3.47, 3.37 and 3.16 for the item bank_8 item SF, FIM_8 item SF and MDS_8 item SF; 2.35, 2.45 and 2.2 for the item bank_4 item SF, FIM_4 item SF and MDS_4-

item SF (Table 2). Figure 2 presents the 8_item SFs relative to the item bank. For the three 8-item SFs, the MDS_8 item had the least test error at the lower theta levels (-3.8 to -2.5 logits) but the highest test error at the higher theta compared to the other two (2.5-3.8 logits) (Figure 2). However, for test error below 0.3, three 8-item SFs covered similar ranges of theta (Figure 2). Figure 3 presents the 4-item SFs relative to the full item bank. All three 4-item SFs showed similar SE patterns. The full-bank SF had two “bumps” (higher test error) at about -1 theta. All three 4-item SFs showed the test error higher than the criteria of 0.3 (Figure 3).

We only represented the plots of 95% confidence interval (CI) of error bands between (a) the item bank versus item bank_8 item SF, and (b) the item bank versus item bank_4 item SF in this paper (Figures 5 & 6). However, we put all the other plots of 95% CI of error bands as the supplementary materials and could be obtained by request. Table 4 presents the number and percentage of person measures outside the 95% error bands. The MDS_8 showed the highest percent of person measures outside the 95% confidence bands (8%) (Table 4). All other SFs showed less than 5% of person measures outside the error bands with the FIM SFs overall showing the lowest percentage (Table 4).

Discussion

This study generated varied 4- and 8-item SFs from the FIM-MDS item bank and compared their measurement precisions across Veterans PAC settings. The overall finding was that when the numbers of item increased, the error of the test decreased and person strata increased (e.g., 8-item SFs showed more strata and lower overall SE than 4-item SFs) regardless of which instruments were used. Similarly, correlations of the SFs with the item bank increased with the number of items increased.

The MDS_13 had a slightly lower person strata value (i.e., worse measurement precision) compared to the FIM_13, but showed lower test error in the both extreme ends of person ability levels that especially approached the item-bank error curve at the lower end; this may be due to its wider spectrum of item difficulties that was a similar characteristic as the item bank. The FIM_13 had slightly higher person strata compared to the MDS_13 and had the least error within the middle range of person ability, also for the corresponding 4- or 8-item SFs. When the number of items was the same, the test forms had similar pattern of total test error and person strata. Three 8-item SFs demonstrated comparable person strata and total test error with the item bank, FIM_13 and MDS_13. This finding supported the idea of using IRT methods to develop "equiprecise" measurements, indicating "equal" measurement precision across instruments. Thus, this finding suggested that healthcare practitioners could choose any SFs (with the same number of items) they are comfortable to use to obtain similarly precise results.

While there was an overall pattern showing more items corresponding with less error, there were some pattern differences within SFs. For instance, MDS_8-item SF had least error for the lower theta but higher error for the higher theta compared to other 8-item SFs. Overall, all 8-item SFs had person strata of 3 and all 4-item SFs had person strata of 2, indicating 8-item SFs distinguished physical self-care function better in Veterans. In addition, all three 4-item SFs showed the test error higher than the criteria of 0.3, indicating less reliability as the 8-item SFs. These findings indicate that a match between difficulty levels of the short form and ability levels of the persons determined the most precise short form. As the result, the FIM and MDS appear to match the severity levels of the patients for which they are typically used. Higher ability level persons who are typically in inpatient rehabilitation facilities are assessed with the FIM and lower ability level persons who are typically in skilled nursing facilities are assessed by the MDS.

This is further evidence that it may not be ideal to use a single instrument across all PAC settings. Rose et al. (2008) also found the precision of different tests differed at varied ranges of person ability; for instance, the Health Assessment Questionnaire (HAQ)-9 item showed highest precision with lower ability persons and the 36-Item Short Form Health Survey (SF-36) showed highest precision with higher ability persons for persons with varied disability conditions (Rose, Bjorner, Becker, Fries, & Ware, 2008).

Regarding the short form and the computerized adaptive tests (CAT), there is the possibility that CAT may have some advantages over SFs. Fries et al. (2009) and Hol et al. (2007) found CAT-based assessment offered superior performance over fixed short forms with the same numbers of item or even greater length. However, Reise and Henson (2000) found that if the SFs are designed to consist of most-administered CAT items, then the SFs showed comparable precision to the CAT. Thus, well-designed SFs may achieve the precision of CATs. Using IRT to develop SFs chooses items based on the item-level psychometrics (i.e., item difficulty), thus providing some advantages over classical test theory (CTT) methods that treat the test as a whole. The advantage of IRT-methodology used in the present study is that one can assure that items were selected across the range of person abilities.

Within the IRT-based methods, different IRT-model had different item selection criteria when developing a short form. For instance, Rose and colleagues' (2008) chose the items representing the highest discriminative values to create the short form; while Ornstein et al. (2015) developed two short forms, 5- and 10-items, from the original 20-item Family Satisfaction with End-of-Life Care scale using a selection of most informative items based on graded response model (a 2-parameter model). It is noted that the results of both studies were consistent with our results in that the longer SFs had higher precision.

There were two persons with unexpected increases in error for the Item Bank_4-item SF at the theta level approximately of -1, which did not happen in any other test forms. However, these two persons were within the fit statistics criteria of the Rasch model, indicating their responses were not erratic, which was unexpected. We also noticed that FIM_13 and relevant SFs (derived from FIM) had very high correlations, while the MDS_13 had lower correlations with its relevant SFs. However, this was as expected and we wanted to emphasize that for the FIM_13 and relevant SFs, the same individuals responded to the same instruments at the same time with the same rater; while the MDS_13 and relevant SFs, the same individuals responded to different instruments at different time and with different raters. In addition, we assumed that the modification of the MDS rating scale structure (from a four to a seven point) to match rating scales of the FIM could also contribute to more error in the MDS_13 and its relevant short forms. Also, the conversion process could also produce unexpected variance.

In summary, using existing instruments to create an item bank allows the generation of short forms with acceptable precision that would have sufficient sensitivity in detecting treatment effects (i.e., minimal clinical differences) with fewer numbers of items. The finding supported comparable measurement precision of the varied short forms with the same item numbers. Since the 4-item short forms did not meet the 0.3 or less SE criterion, in order to maximize precision and minimize assessment burden, the 8-item short forms appears to have the best balance between precision and efficiency and could be considered as a preferred instrument.

Short forms not only minimize assessment burden for the practitioners and the patients but also provides the practitioners flexibility to choose the instruments practitioners are presently using efficiently. For instance, the practitioners could choose associated short forms that may be most appropriate for the patients they evaluate, i.e, FIM for higher ability patients typically

treated in inpatient rehabilitation and MDS for low ability patients typically treated in skilled nursing facilities. The finding supported developing a continuum of measurement using existing instruments by generating an item bank and further supported developing relevant short forms to improve feasibility of the existing instruments for the practitioners and the patients.

Study Limitations

The limitations of this study included: (a) we did not compare the similarities or inconsistencies of SF development methods based on different IRT models; (b) this study was not generalizable to populations beyond the Veterans population.

Conclusions

We have demonstrated the possibility to use different existing instruments to construct an item bank and further developed varied short forms. There were three main findings in this study, including: a) test forms with the same number of items generated from different instruments showed similar precision, thus suggesting that clinicians can use the instruments they are most familiar with (i.e., FIM for inpatient rehabilitation facilities and MDS for skilled nursing facilities), supporting using existing instruments at different settings; b) the main factor in determining measurement precision appears to be the number of items (SFs with 4 items had inadequate precision); c) finally, a good balance between precision and efficiency appears to be an 8 item short form.

Appendix

Table 1. Demographic Characteristics of Participants in this Study (n=2500)

Variables	Community-Dwelling Veterans (n =2500)	
	Number	%
Age (range: 0.7-90 y/o)	Mean=67.1	(SD=11.3)
Averaged number of days since onset	Mean= 155.0	(SD= 1083.8)
Gender		
Male	2377	95.1
Female	93	3.7
Missing	30	1.2
Ethnicity		
White	1576	63.0
Black	582	23.3
Asian	8	0.3
Native American	11	0.4
Hispanic	129	5.2
Other	98	3.9
Missing	96	3.8
Diagnoses		
Stroke	1066	42.6
Lower Extremity Amputee	472	18.9
Knee Replacement	568	22.7
Hip Replacement	394	15.8
Marital Status		
Single	306	12.2
Married	1064	42.5
Widowed	160	6.4
Separated	89	3.6
Divorced	779	31.2
Missing	102	4.1
Admission Condition		
Initial Rehabilitation	2362	94.5
Short Stay Evaluation	61	2.4
Readmission	15	0.6
Unplanned Discharge Without Assessment	2	0.08

Continuing Rehabilitation	56	2.2
Missing	4	0.2
Pre-living Setting		
Home	503	20.1
Board and Care	9	0.4
Transitional Living	8	0.3
Intermediate Care	12	0.5
Skilled Nursing Facility	143	5.7
Acute Unit of Own Facility	1113	44.5
Acute Unit of Another Facility	313	12.5
Chronic Hospital	1	0.04
Rehabilitation Facility	41	1.6
Other	11	0.4
Alternate Level of Care Unit	1	0.04
Subacute Unit	3	0.1
Assisted Living Residence	5	0.2
Missing	337	13.5
Days between Administrations of FIM and MDS (range=0-6)	Mean= 3.2	(*SD*=2.1)

Table 2. Within-Subject Precision Comparisons

	Full Bank	**FIM[a] (N=13)**	**MDS[a] (N=13)**	**Full Bank_8SF**	**FIM_8SF**	**MDS_8SF**	**Full Bank_4SF**	**FIM_4SF**	**MDS_4SF**
Reliability	0.97	0.98	0.94	0.92	0.96	0.94	0.79	0.89	0.85
Person Separation Index	3.82	2.88	2.63	2.35	2.28	2.12	1.51	1.59	1.40
Person Strata	5.4	4.17	3.84	3.47	3.37	3.16	2.35	2.45	2.2
Person Ability (Mean ± SD)	0.55 ± 0.20	0.77±0.29	0.57±0.28	0.73 ± 0.34	0.77±0.35	0.50±0.35	0.78±0.49	0.87±0.52	0.46±0.44
Range of Person Measure (Min ~ Max)	7.22 (-3.34~ 3.88)	6.23 (-2.77~3.46)	6.06 (-3.09~2.97)	5.56 (-2.34~3.22)	5.49 (-2.39~3.10)	4.92 (-2.59~2.33)	4.60 (-1.98~2.62)	4.59 (-2.02~2.57)	3.87 (-1.99~1.88)
Item Difficulty (Mean ± SD)	0 ± 0.02	0.02±0.02	-0.02±0.02	0.16 ± 0.02	0.06±0.02	-0.10±0.02	0.11±0.02	0.05±0.02	0.06±0.02
Range of Item Difficulty (Min ~ Max)	2.03 (-1.13~0.90)	1.98 (-0.82~1.16)	2.76 (-1.50~1.26)	1.96 (-0.80~1.16)	1.98 (-0.82~1.16)	1.98 (-0.97~1.01)	1.96 (-0.80~1.16)	1.98 (-0.82~1.16)	1.81 (-0.80~1.01)
Percent of Persons with Maximum Scores *Ceiling Effect	0.48% (12/2500)	1.12% (28/2500)	8.96%* (224/2500)	1.08% (27/2500)	1.36% (34/2500)	17.44%* (436/2500)	1.28% (32/2500)	1.68% (42/2500)	18.88%* (472/2500)
Percent of Persons with Minimum Scores *Floor Effect	0% (0/2500)	5.76%* (144/2500)	0% (0)	3.08% (77/2500)	6.12%* (153/2500)	2.84% (71/2500)	3.72% (93/2500)	6.72%* (168/2500)	3.92% (98/2500)

FIM[a]: FIM-Anchored/MDS[a]: MDS_Anchored/SF: Short Form

* indicates significant ceiling/floor effects (greater than 5% of the total sample); Yes^: NOTE: rating scales of 3 and 6 had no values because of converted rating scale mechanism

Table 3. Correlations between Item Bank, FIM, MDS, All Three 8-item Short Forms and All Three 4-item Short Forms

	Full Bank	FIM_13	MDS_13	Full Bank_8SF	FIM_8SF	MDS_8SF	Full Bank_4SF	FIM_4SF	MDS_4SF
Full Bank	1								
FIM_13	0.889	1							
MDS_13	0.865	0.631	1						
Full Bank_8SF	0.951	0.917	0.773	1					
FIM_8SF	0.884	0.988	0.629	0.922	1				
MDS_8SF	0.824	0.635	0.892	0.742	0.623	1			
Full Bank_4SF	0.905	0.864	0.744	0.956	0.876	0.746	1		
FIM_4SF	0.865	0.956	0.621	0.904	0.974	0.611	0.753	1	
MDS_4SF	0.809	0.624	0.874	0.739	0.612	0.977	0.753	0.602	1

Table 4. Person Measure Outside of 95% Error Bands between Two Test Forms

Instrument	Short Form	Number of persons outside of 95% error bands	Percentage of persons outside of 95% error bands
Full Bank	FIM_8	33	1.3%
	FIM_4	43	1.7%
FIM	FIM_8	1	0.04%
	FIM_4	9	0.4%
MDS	MDS_8	200	8.0%
	MDS_4	80	3.2%

Figure 1. Test Error Plot of All Test Forms (n=2500)

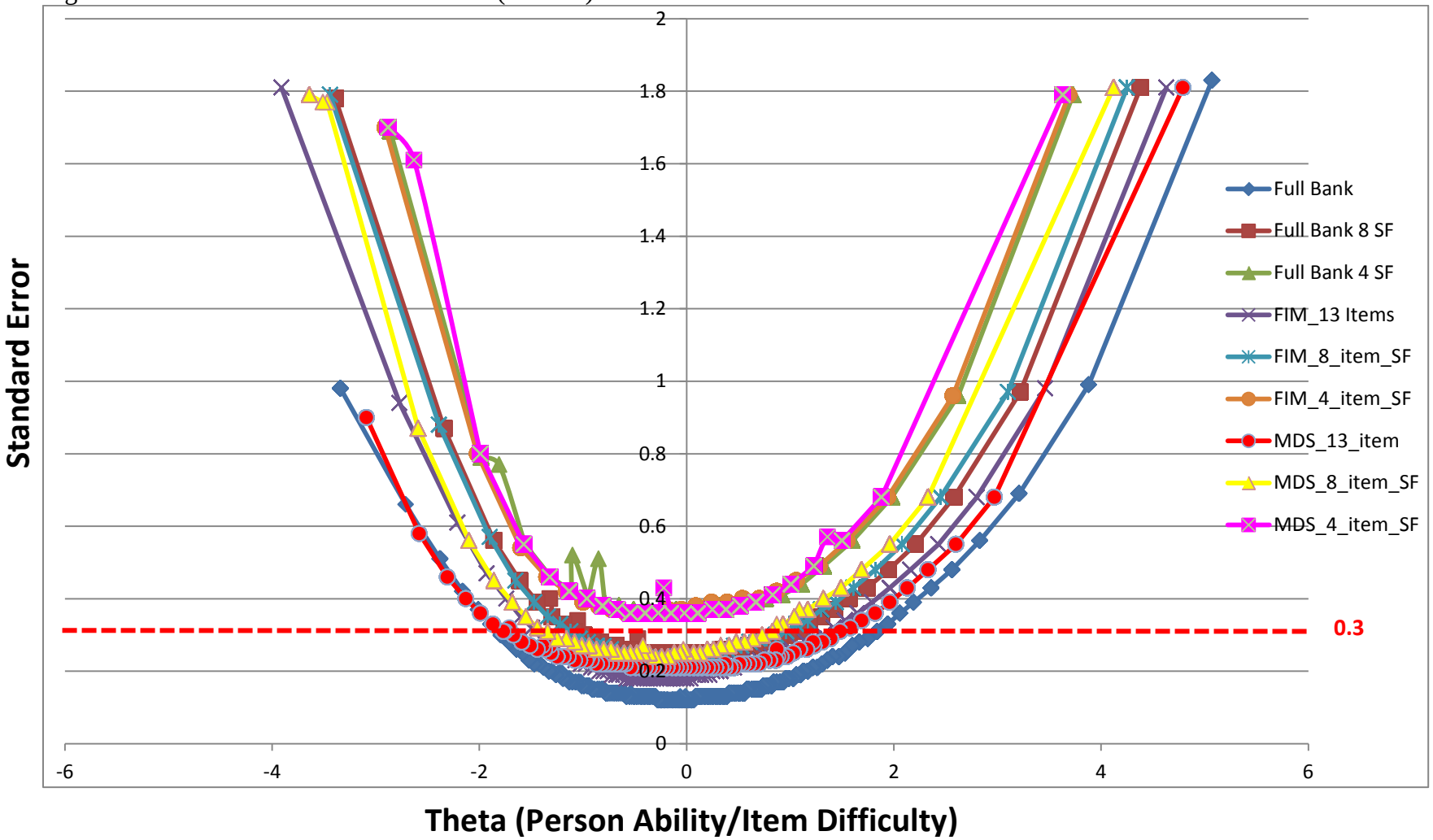


Figure 2. Test Error Plot between Full Bank and All 8-item Short Forms (n=2500)

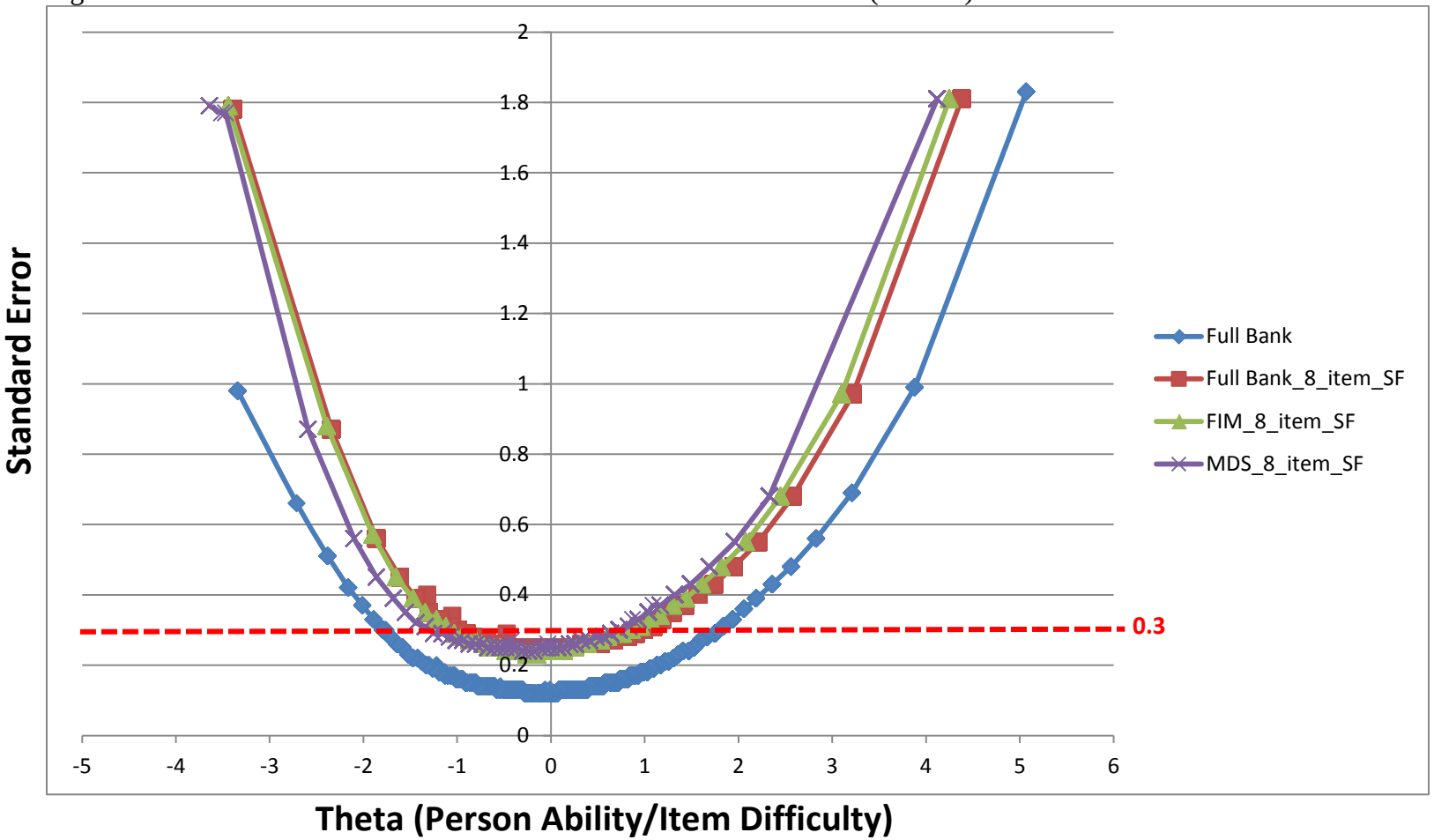


Figure 3. Test Error Plot between Full Bank an d All 4-item Short Forms (n=2500)

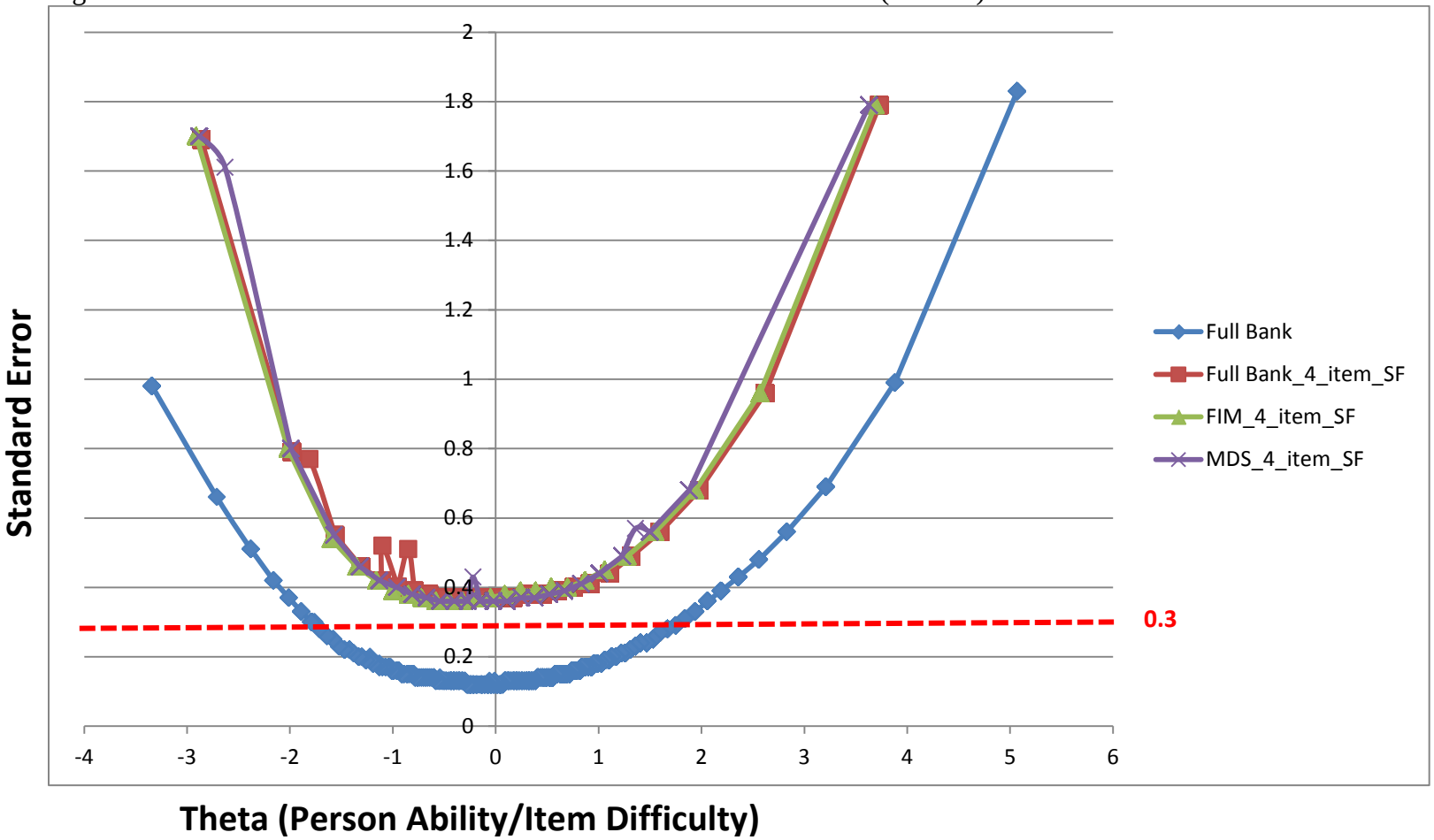


Figure 4. The 95% Confidence Interval Plot between the Item Bank and 8-item Item Bank Short Form (r= 0.95)

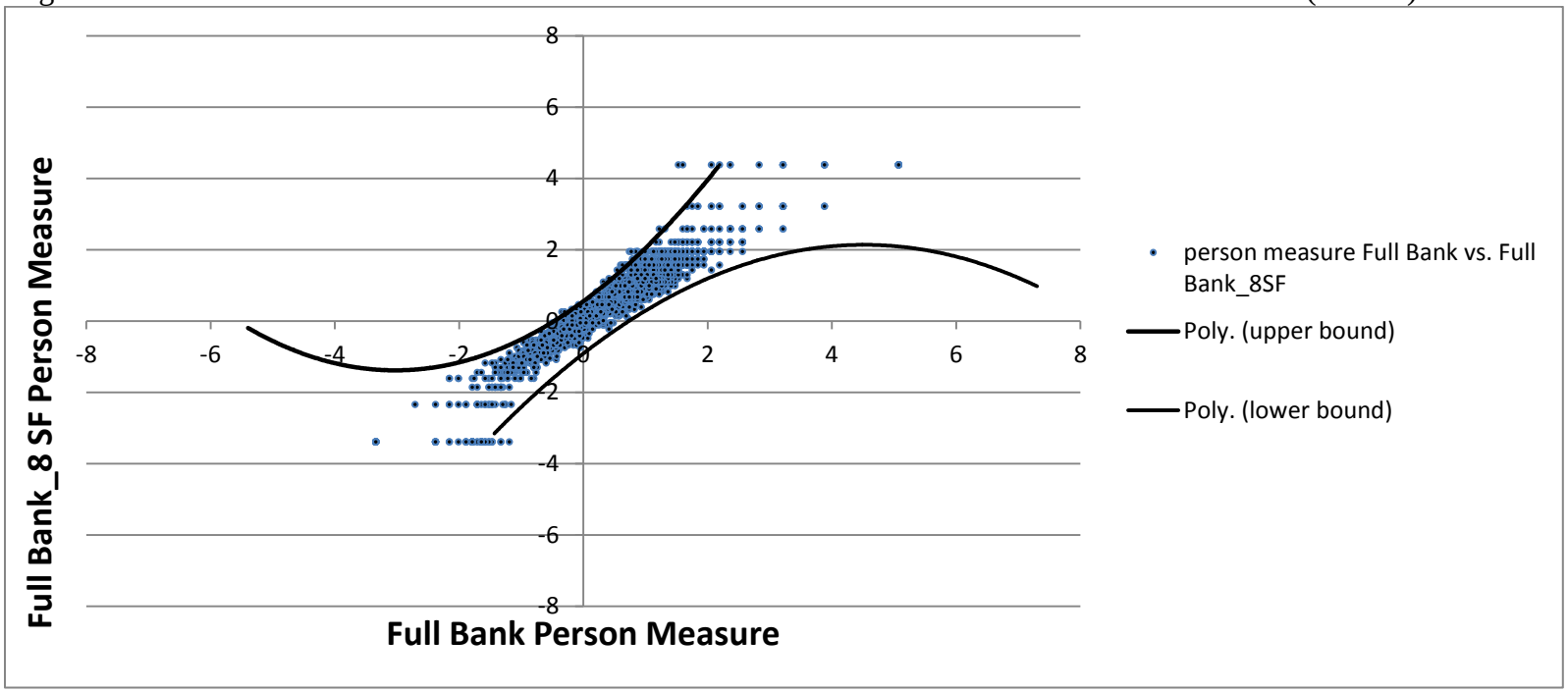
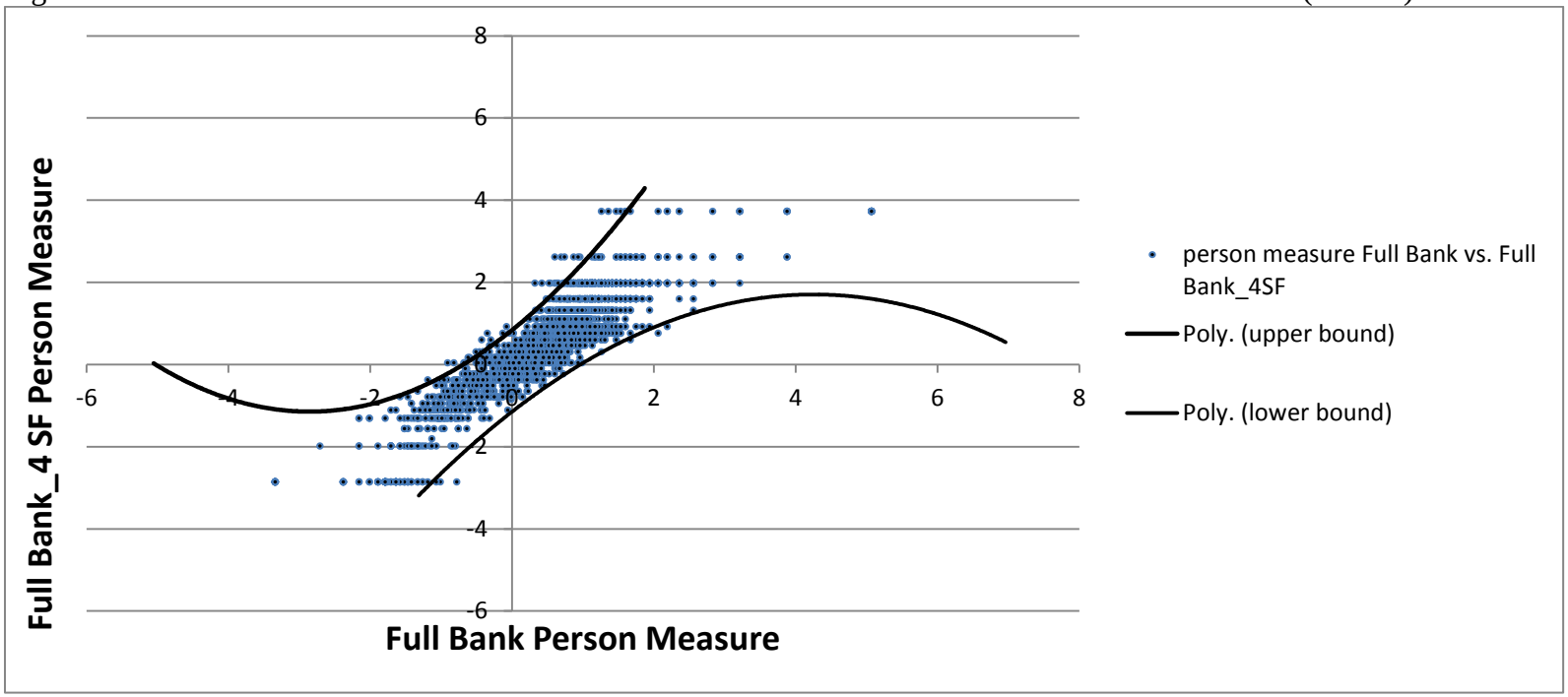


Figure 5. The 95% Confidence Interval Plot between the Item Bank and 4-item Item Bank Short Form (r= 0.90)



Supplementary Materials

Figure 6. The 95% Confidence Interval Plot between the FIM and 8-item FIM Short Form (r= 0.99)

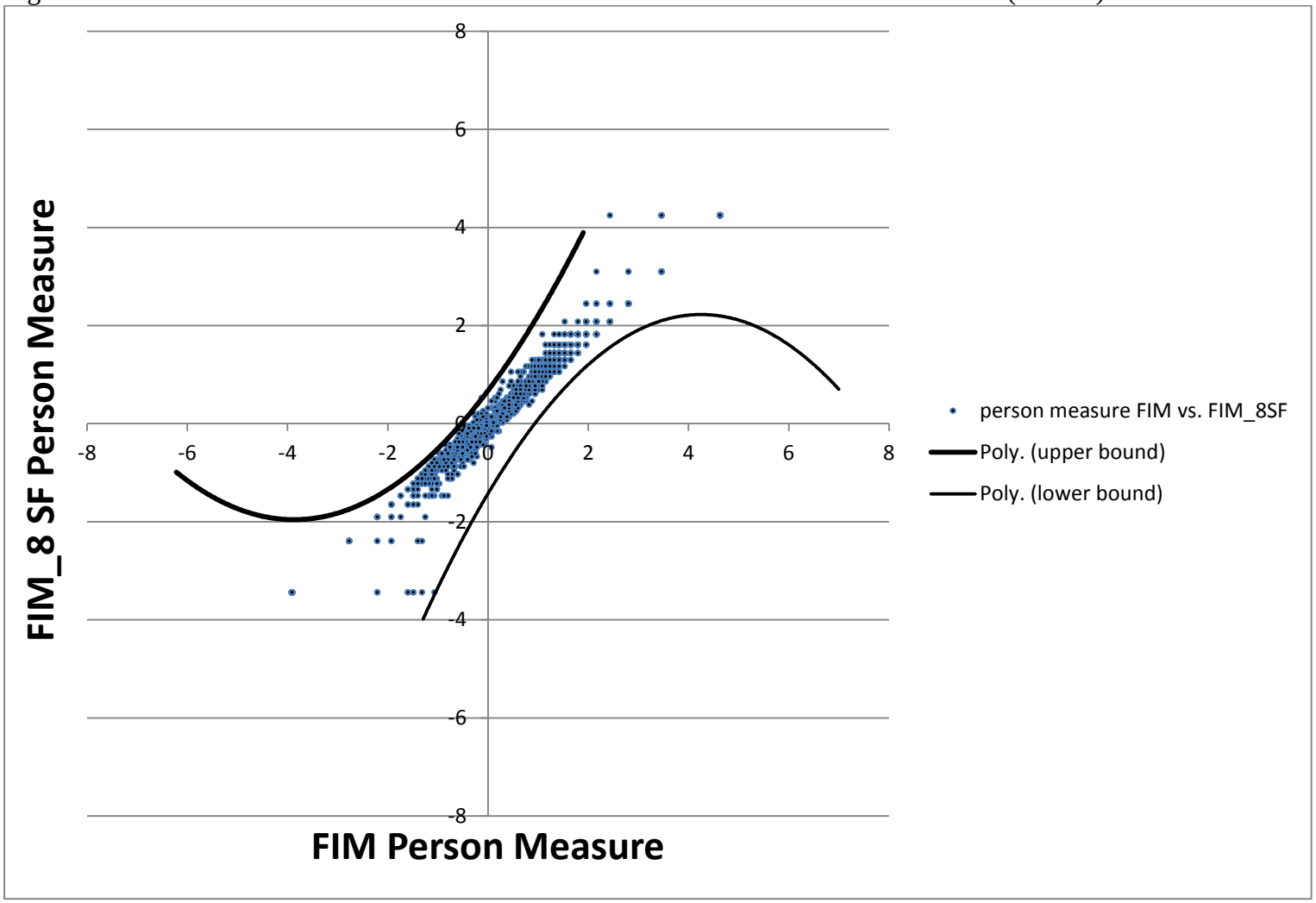


Figure 7. The 95% Confidence Interval Plot between the FIM and 4-item FIM Short Form (r= 0.96)

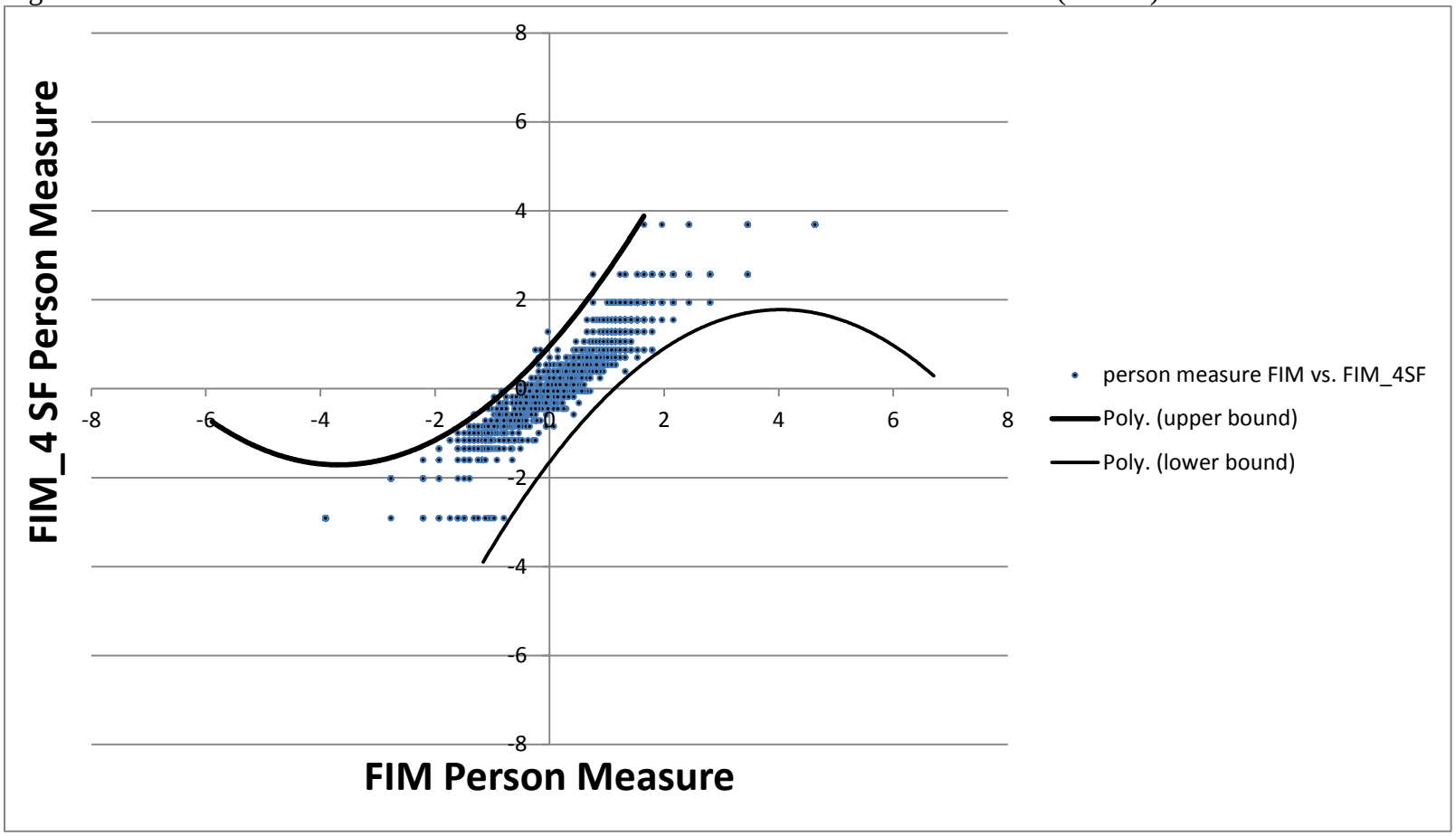


Figure 8. The 95% Confidence Interval Plot between the MDS and 8-item MDS Short Form (r= 0.89)

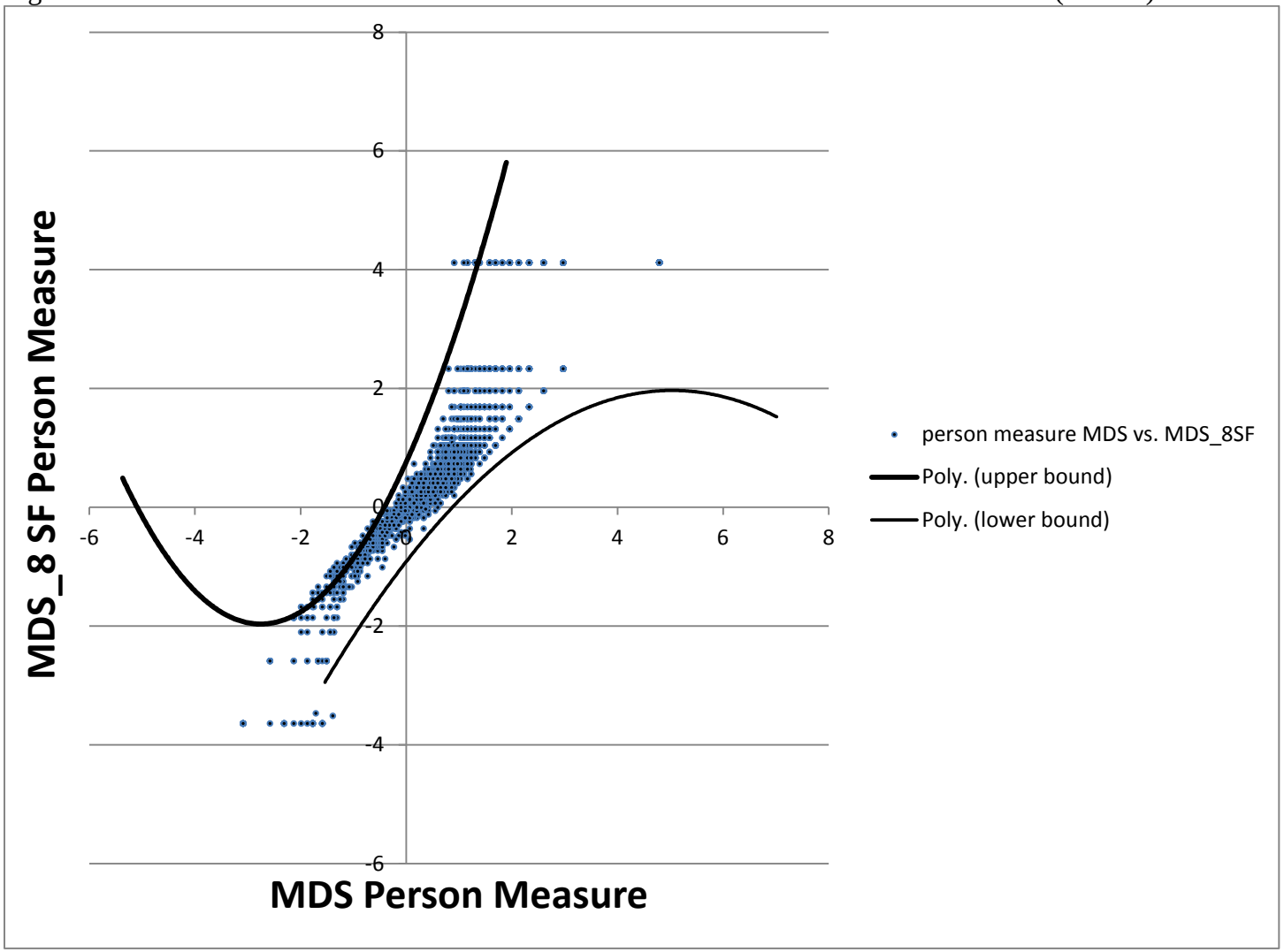
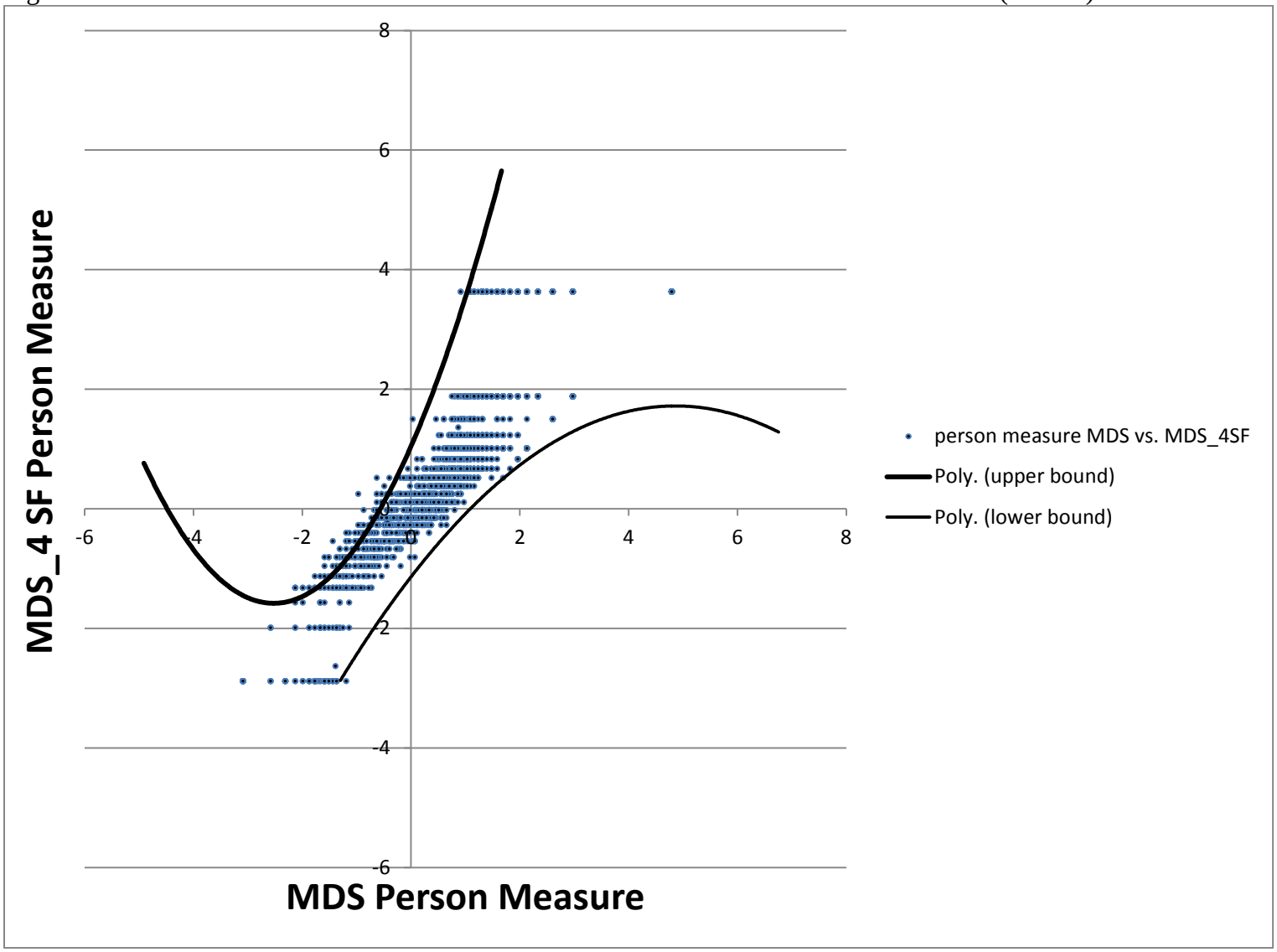


Figure 9. The 95% Confidence Interval Plot between the MDS and 4-item MDS Short Form (r= 0.87)



CHAPTER FOUR (Manuscript_3)

Continuum of Care Assessment across Post-Acute Care in Veterans: Measurement Accuracy Comparison of Short Forms Generated from Functional Independence Measure and Minimum Data Set Item Bank

Abstract

Objective: To compare measurement accuracy of varied short forms (SFs) generated from the self-care physical function item bank composed of Functional Independence Measure (FIM™) and the Minimum Data Set (MDS).

Study Design and Setting: This study used retrospective data of 2499 Veterans who completed both FIM and MDS within 6 days. We compared measurement accuracy between the converted FIM score (FIMc) generated from 4- and 8- item SFs and the original actual FIM-13-item (FIMa) motoric score at: (a) individual level using point differences, and (b) group level using functional related group (FRG) classification system.

Results: The result showed mixed findings. The differences of mean FIMa and FIMc scores generated from FIM SFs, MDS SFs and MDS_13-item were within 1.07-0.05 points. At least 55% FIMc generated from all forms were within 10 points of the FIMa. Eighty-one to ninety percent of FRGs generated by two FIM SFs were the same as those generated by the FIMa for stroke, lower extremity amputation, knee and hip replacement; 59.9-90.5% by all MDS test forms. When considering the impact of error (within one FRG difference), above 74% agreement was found by all MDS test forms across all four diagnoses. Kappa statistics demonstrated strong agreement (0.70–0.95) for all diagnoses when the data had sufficient variability.

Conclusion: Using existing instruments to generate a continuum of care measurement depends on the comparison level (i.e. individual or group level), the length of the SF and which FRG is used.

Keywords: self care, physical activity, patient outcome assessment, Veterans, classification, care continuity

1. BACKGROUND/SIGNIFICANCE

The need for developing a cross-setting measure has resulted in efforts to develop a single instrument. The Centers for Medicare & Medicaid Services (CMS) funded the development of the Continuity Assessment and Record Evaluation (CARE) Item Set, a uniform patient assessment instrument designed to provide continuum care documentation across acute to post-acute facilities, including acute hospitals, Inpatient Rehabilitation Facility (IRF), Skilled Nursing Facility (SNFs)/Community Living Center (CLC), Home Health Agency (HHA) and Long-Term Care Hospital (LTCH) (CMS, 2012 & 2015).

The CARE Item Set uses the same scoring system across the post-acute care (PAC) continuum, with the hope to generate comparable scores and standardize patient assessment data (CMS, 2012). This instrument includes a comprehensive item set and core item set as functional status quality metrics, including motor functional status (self-care and mobility) and cognitive functional status (memory, problem solving and communication), additional clinical information (e.g., skin integrity and allergies/adverse drug reactions) and demographics data (CMS, 2012). However, practical challenges regarding developing and implementing a new universal assessment tool are often underestimated. Such concerns included requiring widespread resources (e.g., money and time) for instrumental development, instrumental validation, new instrument administration training and new reimbursement software development. A new universal tool not only needs considerable researche to support its reliability and validity, but

also inevitably requires administration training, new report generation and extensive modifications of existing electronic medical records. In addition, a universal tool requires a large item set that may have inappropriate items for particular settings. As a result, some items from the universal tool will not applicable to assess some patients' functional levels. For example, easy item such as "rolling left and right on the bed" may be important to measure patients residing in the community living center but may be inappropriate to measure patients at the outpatient rehabilitation unit (Wang, Byers, & Velozo, 2008a).

We proposed an alternative cost-efficient solution of linking existing instruments into an item bank to allow for developing a measurement across the continuum of care. Using the item response theory (IRT)-based linking method allows test items from different assessments to be placed on a common scale, thus, scores of different assessments can be comparable. Linking existing instruments allows practitioners to continue using the instruments that they have been accustomed. Developing short forms from the item bank composed of existing instruments could further facilitate assessment efficiency and reduce assessment burden for the practitioners and patients.

To demonstrate feasibility of linking existing instruments to create a continuum of care measurement, we created an item bank composed of the Functional Independence Measure (FIM™) used in IRFs and the Minimum Data Set (MDS) used in CLCs in the Veterans healthcare system. This self-care physical function item bank had a total of 26 items composed of FIM and MDS motor items which have been examined for its item-level psychometric properties (Li, et al., 2015a). We developed six short forms from this FIM-MDS item bank, including item bank_4- and 8-item, FIM_4- and 8-item, MDS_4- and 8-item short forms. We have previously evaluated the measurement precision of these short forms (Li, et al., 2015b).

This study is an extension of previous linking research/ It is aimed to evaluate measurement accuracy of the developed short forms, to address the concerns about measurement accuracy of a linked item bank. Accuracy was evaluated based on whether the converted scores from different instruments could classify patients into the same disability level as the original scores. If using converted scores from the existing instrument could generate similar measurement accuracy as using the original scores, then the concept of developing a continuum of measurement using existing instruments sould be supported.

The CMS uses Case Mix Groups (CMGs), a form of Function Related Groups (FRGs), as a basis for the inpatient rehabilitation facility (IRF) prospective payment system (PPS) (Stineman, 1995). Stineman and colleagues (1994, 1995 & 1997) conducted a series of studies to develop the FRG algorithms to predict the cost of treating Medicare patients. The FRG algorithms used the FIM physical functioning (13 items) and the FIM cognitive (5 items) scores, along with patients' age at admission to the IRFs. Based on the rehabilitation impairment classification, patients are classified into one of 20 diverse impairment diagnoses (e.g., stroke) (Stineman, 1997). Each impairment diagnosis has a specific FRG algorithm resulting in different numbers of FRG categories. Patients assigned to different FRG groups are expected to have different rehabilitation outcomes and total costs of healthcare.

This study used the FRGs classification system as a pragmatic method to examine measurement accuracy at group level for the "converted" FIM score (i.e., FIM scores generated by different sets of items from the item bank). We compared the scores derived from the original FIM and different test forms, to investigate whether the converted FIM scores could classify the same patient into the same or a similar classification levels. We used the 4- and 8-item short forms from the item bank to generate FIM converted scores, and used the converted scores to

assign FRGs. We hypothesized that short forms generated from either FIM items or the MDS items will generate similar FRGs categories for Veterans compared to those generated from the original FIM.

2. METHODS

2.1. Participants

This study used a retrospective data of 2500 Veterans with diagnoses of stroke, amputation, hip replacement and knee replacement from the Veterans Austin Information Technology Center (AITC) databases. Each participant completed both full instruments of FIM and MDS within 6 days through October 2008 to September 2010. We only analyzed motor items of both FIM (n=13) and MDS (n=13) in this study. To generate FRGs, we also used FIM cognitive scores and age of each Veteran. The ability estimate based on the original FIM was considered the "gold standard" which was referred to as the FIM actual score (FIMa). In this study, we generated four FRG diagnoses: stroke, lower extremity amputation, knee replacement and hip replacement.

2.2 Instruments

We used the short forms generated from FIM-MDS self-care physical function item bank that was developed using an independent random set of Veterans (n=500). FIM_8-item, FIM_4-item, MDS_13-item, MDS_8-item, and MDS_4-item scores were converted to the FIM scores (FIM converted, FIMc). We developed the 4- and 8-item SFs based on del Toro and colleagues' (2011) short form development procedures and examined person strata, ceiling/floor effects, person fits, test standard error (SE) plot and 95% confidence interval of anchored person measures for each short form in the previous study (Li, et al., 2015b). The results showed that short forms with the same numbers of items demonstrated similar precision regarding person strata and test error.

Also, all 4-item SFs did not meet the criteria of SE less than 0.3 for any theta values (Li, et al., 2015b).

2.3 Analysis Procedures

Regarding of examining measurement accuracy of short forms, at the individual level, we used Kolmogorov-Smirnovwill statistics to test normality of the distribution. Based on the normality test results, we will use paired sample t-test for parametric data and Wilcoxon signed rank sum test for nonparametric data to compare distribution differences between FIMa and FIMc scores. Point difference was the absolute value calculated between the actual FIM (FIMa) and the converted FIM (FIMc) (|FIMa-FIMc|). We calculated the percentage of converted scores that were within 5- and 10-point differences. We also demonstrated point difference distributions of each test form. Pearson correlation coefficient was calculated between the FIMa and FIMc for all test forms. A value of 0.05 was used as the indication of significance. Intraclass correlations coefficients (ICC) were calculated between FIM_13 and all other test forms. We used two-way mixed method to calculate absolute agreement for ICC. ICC values less than .40 were classified as poor, between .40 and .59 was fair, between .60 and .74 was good, and between .75 and 1.0 was excellent (Hallgren, 2012).

At the group level, we compared FRG classifications generated from each short forms (FIM converted: FIMc) to the “actual” FRG classification by the FIM (FIM actual: FIMa). This study used three FRG classification algorithms in total because the FRG algorithm for knee replacement and hip replacement was the same. The elements of stroke, knee replacement and hip replacement FRG algorithms included FIM-motor scores, FIM-cognition scores and age. Only one element, the FIM-motor scores, was replaced and generated from the varied forms to classify FRGc. We used the original FIM cognitive scores in all FRG algorithms. After

calculating the FRGs from the FIMa and FIMc, we determined the percentage of FRGs falling into the same FRG category (perfect agreement), one category apart (± 1 level), two categories apart (± 2 levels), and also categories greater than two categories apart (± 3 ~ ±7 levels).

In addition, we quantified the strength of association of the FRG classification results from FIMa and FIMc to account for the distance between each categorical difference. We used weighted kappa to examine agreement strength for the stroke, knee replacement, and hip replacement FRG calculations. We used kappa and McNemar's test to provide a 2x2 table for the lower extremity amputation FRG calculation due to its dichotomous FRG classification algorithm. A weighted kappa statistic for categorical data ranging from 0.21 to 0.40 demonstrates a fair strength of observer agreement, from 0.41 to 0.60 represents a moderate strength of agreement, and from 0.61 to 0.80 indicates a substantial strength of agreement (Landis & Koch, 1977). Because the variability of the data could significantly bias the kappa classification results, we also examined the percentage of agreement in each diagnostic group. Finally, a two-way mixed method ICC was calculated between FRGa and FRGa for all test forms across the four diagnostic groups. It should be noted that ICC also have similar limitation as the kappa.

3. RESULTS

3.1 Participants

After removing a person with a miscoded age and thus not qualified to be classified into the FRG, a final total number of 2499 Veterans who had diagnoses of stroke (n=1065, 42.6%), lower extremity amputee (n=472, 18.9%), knee replacement (n=568, 22.7%) and hip replacement (n=394, 15.8%) was included in the study. Mean age in this sample was 67.1 (SD=11.2) years old (range=19 to 90 years old). Sixty-three (2.5%) patients were identified into the same group with the age older than 89 years old (Table 1). The majority of the sample was male (96.2%),

white (65.5%), married (42.5%) and lived at acute unit at the same rehabilitation facility (44.5%) or at home (20.1%) prior to their transition to another facility. This is representative of the Veteran population. The average length of days between the administrations of the FIM and the MDS was 3.2 (SD=2.1) days (Table 1).

3.2 Accuracy Comparisons at Individual Level- Point Difference

The FIM original and converted scores all had negatively skewed distributions for each test form (i.e., FIM_13, FIM_8, FIM_4, MDS_13, MDS_8, and MDS_4) indicating the individuals tended to have higher FIM scores (better self-care physical function). Score distributions of all test forms violated the normality assumption (all p-value <0.05). Thus, we used Wilcoxon signed rank sum test to compare score distribution difference between FIMa and FIMc. Wilcoxon signed rank sum test showed significant difference of median score distribution between the FIMa and FIMc, regardless of which test form was compared (all p-value <0.0001) (Table 2).

The distributions of absolute point difference of each test form were positively- skewed, indicating the majority of point difference was low (Figure1, (a) - (e)). Fifty-six to ninety-nine percent of the FIMc scores were within 10 points of the FIMa, while FIM short forms showed the least point differences with 95-99 percent of the scores within 10 points of the FIMa, the MDS test forms showed 57-65 percent of the scores within 10 points of the FIMa (Table 2). Thirty-one to ninety-two percent of the FIMc scores were within 5 points of the FIMa, while FIM short forms showed the least point differences with 78-92 percent of the scores within 5 points of the FIMa, the MDS test forms showed 31-39 percent of the scores within 5 points of the FIMa (Table 2).

Correlations between Original Scores and Converted Scores from Varied Test Forms

Correlations for all short forms between the FIMa and FIMc were significant (range= 0.75 to 0.99). The correlations for FIM_8-item and FIM_4-item were 0.99 and 0.97, and the correlations for MDS_13-item, MDS_8-item and MDS_4-item were 0.81, 0.78 and 0.75, respectively (Table 2). The converted scores generated from all test forms had excellent ICCs with the FIM_13 scores (Table 2).

Accuracy Comparisons at Group Level- FRG Classification

At the group level, we calculated the percentage of agreement using the FIMa (actual score) and FIMc (converted score) to classify each individual into one of the FRGs. We used FRGa to represent the FRG generated by FIMa and FRGc to represent the one generated using FIMc. Table 3 presented the percent of FRGc that were within 1 or more classifications of the FRGa. We identified agreements as exactly the same (perfect agreement), ±1 category apart, ±2 categories apart and more for each diagnosis. Overall, the FRG agreement of the FIM SF generated FRGs was higher than MDS generated FRGs. For all four diagnoses, the FIM_8-item SFs had the highest perfect agreement (85.16-97.97%) and MDS_4-item had the lowest perfect agreement (59.91-80.93%). The range of perfect agreement of stroke FRGc for all test forms was between 59.91 to 85.16 percent, agreement apart by ±1 category was 74.46 to 95.67 percent, and agreement apart by ±2 categories was 80.75 to 97.74 percent (Table 3). Ninety-five percent or greater of classifications were within 2 categories for the FIM_8-item SFs and 3 categories for the FIM_4-item SF. Above 74% of classifications were within 1 categories for the MDS_13-item, MDS_8-item SF and MDS_4-item SF. Above 81% of stroke FRGc classifications were within 2 categories for all the MDS test forms (Table 3).

The diagnosis of amputation only had two FRG groups. Thus, the range of perfect agreement of amputation FRGc for all test forms was between 80.93 to 95.34 percent. Both 4-and 8- item

FIM SFs had above 92 percent perfect agreement. MDS_13, MDS_4 and MDS_8 SFs had above 82 percent perfect agreement across diagnoses of knee/hip replacement and lower extremity amputation (Table 3). The range of perfect agreement of knee replacement FRGc for all test forms was between 78.35 to 97.71 percent, agreement apart by ±1 category was 92.26 to 98.60 percent, and agreement apart by ±2 categories was 94.9 to 99.83 percent for every test form; FIM_8, FIM_4 and MDS_13 all had above 90 percent perfect agreement (Table 3). The range of perfect agreement of hip replacement FRGc for all test forms was between 69.80 to 97.97 percent, agreement apart by ±1 category was 84.52 to 98.98 percent, and agreement apart by ±2 categories was 92.89 to 100 percent, even though there are seven FRG groups; both 4- and 8-item FIM SFs had above 94 percent perfect agreement. All MDS test forms had above 92.89 percent agreement within 2 categories (Table 3). Overall, the knee and hip replacement FRGs had the highest percent of perfect agreement for the two FIM SFs, while the stroke FRG had the lowest percent of perfect agreement. MDS_13-item had the highest perfect agreement for knee replacement FRG and lowest perfect agreement for stroke FRG. The two MDS SFs had the highest perfect agreement for amputation FRG and lowest perfect agreement for stroke FRG (Table 3).

Agreement strength was presented in Table 4. Overall, within each test forms, strength of agreement decreased with a decrease in the number of items, especially for the MDS forms. For stroke, knee replacement and hip replacement, all weighted kappa/kappa results were significant with the FIM SFs showing strong to very strong agreement and the MDS SFs showed weak to strong agreement (Table 4). Kappa statistics only provide accurate test values for the diagnoses with adequate variability. Thus, the Kappa statistics generated from the MDS test forms for the knee replacement FRGs may not be reliable. For stroke, agreement strength ranged from 0.69 -

0.93, with FIM short forms showing very strong agreement and MDS SFs showing strong agreement. The ICCs showed good to excellent for all the test forms of the stroke, amputation, hip replacement FRGs. However, for knee replacement, the MDS forms had poor-fair ICCs (Table 4).

4. DISCUSSION

The findings from the above study need to be discussed as two separate studies due to differences in data sources of the FIM and MDS scores. FIM SFs in the present study were from the same individuals, at the same time and assessed by the same raters. In contrast, the MDS SF's were the same individuals that were measured by the FIM but were assessed at different times and assessed by different raters.

Overall FIM SF's performed well at estimating the original FIM (13 items) both at the individual level (i.e., comparing point difference) and group level (i.e., comparing FRG levels). At the individual level, 78-92% of FIM_4 and FIM_8 converted scores were within 5 points from the original FIM (13 items). At the group level, across all diagnoses, 92-100% of FIM_4 and FIM_8 generated FRGs were within ± 1 of the original FIM. These findings strongly suggest that FIM SF could be effective in both measuring and classifying individuals in IRF and SNF/CLCs.

The MDS_13, MDS_8 and MDS_4 did not perform as well as the FIM SFs in generating converted scores. At least some of this decrement in performance is a function of the MDS being assessed at different times and by different raters than the FIM. At the individual level, only 31-39% of MDS_4, MDS_8 and MDS_13 converted scores were within 5 points from the original FIM (13 items). At the group level, MDS produced conversion results that were more acceptable. Across all diagnoses, 74-94% of MDS_4, MDS_8 and MDS_13 generated FRGs were within ± 1

of the original FIM. These findings suggest that while MDS converted scores are inaccurate for measuring, they may be acceptable for classifying individuals in IRF and SNF/CLCs.

The findings from the present study are similar to those of Wang and colleagues (2008a). These investigators found 33.7% of MDS_13 within 5 points of the original FIM (we found 39%). Regarding the accuracy in using converted MDS scores for generating FRG's, Wang and colleagues found 67% of stroke FRGs were within ±1 of the original FIM (we found 79%) and 83% of amputation FRGs within ±1 of the original FIM (we found 82%). Slight differences in the findings may have been due to minor differences in score conversion process and differences in the samples. In addition, our study showed slightly better agreement (60-64%) between FIM and MDS converted scores than what Buchanan and colleagues (2004) found (56% agreement) of PPS classifications between FIM and MDS-PAC-to-FIM™ scores.

Measurement accuracy at FRG group level decreased when the number of items in both the FIM and the MDS SFs decreased. For example, FIM short form accuracy for ±1 decreased from 96% to 92% for FIM_8 and FIM_4, respectively while MDS accuracy decreased from 79% to 74% for MDS_13, MDS_8 and MDS_4, respectively. Our previous precision comparison study (Li, et al., 2015b), demonstrated the decrease in FIM and MDS precision was primarily a function of the decrease in the number of items.

Across both instruments and all short forms, the stroke FRG demonstrated the lowest overall percentage agreement, the knee replacement FRG demonstrated the best agreement. This could be due to the greater variability of functional levels in stroke compared to knee replacement. For example, a patient with stroke could have a wider range of functional ability levels, e.g., being bedridden to being able to commute in the community. While a patient with

knee replacement may have less variability of functional status due to immobility. This could contribute to higher agreement of FRG results for individuals with knee replacement.

It was important to note that traditional agreement testing method of using kappa or weighted kappa statistics may provide inaccurate results when less variability was shown in the data. In this study, the higher percentage agreement contradictorily resulted in less variability in the data, leading to lower weighted kappa results especially for the knee replacement FRG. This bias may lead to the misinterpretation of the weighted kappa results. We recommended using the percent of perfect agreement analysis result to cross-validate and supplement the weighted kappa results of knee replacement to avoid potential bias.

To compare with previous crosswalk validation studies, we found those studies supported score translatability between instruments with acceptable group agreement using intraclass correlation coefficients (ICC) or Cohen's effect size at group-level comparison (Askew, at al., 2013; Bjorner, Kosinski, & Ware, 2003; Holzner, et al., 2006; Orlando, et al., 2000; Qude, et al., 2014; Ten Klooster, et al., 2013; Wang , Byers, & Velozo, 2008a). Ten Klooster and colleagues (2013) found different IRT models generated reliable crosswalks between observed and translated scores with similar agreement of ICC ranging from 0.72 to 0.82. Our study showed ICCs ranging from 0.86 to 0.99, which was slightly better. While most studies showed successful linking results at the group-level, it is noticeable that the score conversion may not work as reliable as expected at the individual-level (Askew, at al., 2013; Fischer, Tritt, Klapp, & Fliege, 2011; Holzner, et al., 2006; Ten Klooster, et al., 2013; Wang , Byers, & Velozo, 2008a). For instance, Holzner and colleagues (2006) found that the confidence intervals of translated scores for individual subjects were very large, thus the limited precision of individual scores are likely to lead to unreliable measures of individual differences. Fischer and colleagues (2011) found that

individual scores comparison was imprecise due to substantial statistical spread. Askew and colleagues (2013) recommended that individual scores derived from crosswalks should be used for the group-level analysis, not for clinical care analysis given the additional source of inherent error. In addition, Ten Koolster and colleagues (2013) found substantial discrepancies in agreement between the observed and converted scores for individual patients.

While there was considerable evidence to support translating scores between instruments, the findings have been limited to translating scores between instruments without addressing the accuracy issue. Our studies evaluated the practical concern of measurement accuracy when using the converted scores and suggested that using converted scores may be feasible to identify patients into group classification system when using the FIM SFs or MDS SFs. Since all measures have error, some acceptable range of errors should be anticipated when using converted scores. That is, while a converted scores results in one FRG level different that that generated with the original FIM, this may be largely the result of measurement error. Future studies are needed to distinguish the error associated with conversion versus the error associated with measurement.

4.1 Limitations

Since stability of patient's response is crucial to obtain reliable measurement accuracy, one of the main limitations in this study was that we assumed patients' ability did not change within 6 days. Of course, this assumption is not substantiated and the 6-day difference likely contributed to error in this study. Second, this study design was based on secondary data analysis with the data that did not intended to answer the research questions proposed in this study. Thus, the data may be subject to inherent errors from all possible uncontrollable sources in the data collection process. Finally, some current available statistics used in this study may not be truly

meaningful such as Wilcoxon Signed Rank due to the impact of sample size, or due to the lack of variability of the data that biased the kappa agreement results for the knee replacement FRG.

5. CONCLUSION

Combining existing instruments instead of generating new items to construct a universal continuum of care measure has the advantage for the healthcare policy makers, researchers the clinicians and the patients. This study found the FIM short forms showed good accuracy at both the individual measurement and group classification levels. Our finding indicate that the FIM_8-item SF provide the most accurate FRG results across the four diagnoses of stroke, lower extremity amputation, knee replacement and hip replacement and at the same time maximizes efficiency. The MDS_13-item converted scores had acceptable FRG agreement as the original FIM_13-item scores for group-level comparison. However, the two MDS SFs had the least measurement accuracy. While the MDS_13-item lacked accuracy for individual measurement, it appeared to have adequate accuracy for generating FRG classifications, especially for the FRG groups of amputation, knee replacement and hip replacement.

Appendix

Table 1. Demographic Characteristics of Participants in this Study (n=2500)

Variables	Community-Dwelling Veterans (n =2500)	
	Number	%
Age (range: 0.7-90 y/o)	Mean=67.1	(SD=11.3)
Averaged number of days since onset	Mean= 155.1	(SD= 1083.9)
Gender		
Male	2376	95.1
Female	93	3.7
Missing	30	1.2
Ethnicity		
White	1575	63.0
Black	582	23.3
Asian	8	0.3
Native American	11	0.4
Hispanic	129	5.2
Other	98	3.9
Missing	96	3.8
Diagnoses		
Stroke	1065	42.6
Lower Extremity Amputee	472	18.9
Knee Replacement	568	22.7
Hip Replacement	394	15.8
Marital Status		
Single	306	12.2
Married	1063	42.5
Widowed	160	6.4
Separated	89	3.6
Divorced	779	31.2
Missing	102	4.1
Admission Condition		
Initial Rehabilitation	2362	94.5
Short Stay Evaluation	60	2.4
Readmission	15	0.6
Unplanned Discharge Without Assessment	2	0.08

Continuing Rehabilitation	56	2.2
Missing	4	0.2
Pre-living Setting		
Home	502	20.1
Board and Care	9	0.4
Transitional Living	8	0.3
Intermediate Care	12	0.5
Skilled Nursing Facility	143	5.7
Acute Unit of Own Facility	1113	44.5
Acute Unit of Another Facility	313	12.5
Chronic Hospital	1	0.04
Rehabilitation Facility	41	1.6
Other	11	0.4
Alternate Level of Care Unit	1	0.04
Subacute Unit	3	0.1
Assisted Living Residence	5	0.2
Missing	337	13.5
Days between Administrations of FIM and MDS (range=0-6)	Mean= 3.2	(*SD*=2.1)

Table 2. Summary of FIM_13 Raw Scores and Converted FIM Raw Score Generated from Varied Test Forms (FIM_8, FIM_4, MDS_13, MDS_8, MDS_4) (n=2500)

	Median	Variance	Wilcoxon Signed Rank (Compared with FIM_13)	Correlation (Compared with FIM_13)	Intraclass Correlation Coefficients (ICC)	Point Difference (FIM_13-converted FIM)	
						≤5 points (%)	≤10 points (%)
FIM_13	77.00	523.71					
FIM_8	76.00	522.07	p<0.0001*	0.99*	0.99 (Excellent)	92.20	99.04
FIM_4	75.00	563.97	p<0.0001*	0.97*	0.98 (Excellent)	78.27	94.56
MDS_13	73.00	462.42	p<0.0001*	0.81*	0.89 (Excellent)	38.70	64.71
MDS_8	71.00	501.29	p<0.0001*	0.78*	0.88 (Excellent)	31.05	56.86
MDS_4	71.00	482.16	p<0.0001*	0.75*	0.86 (Excellent)	31.21	56.14

* significant difference < 0.05

Table 3. FRG Classification Difference between FIM_13 and Other Test Forms across Four Diagnoses (Stroke, Amputation, Knee Replacement, and Hip Replacement)

Stroke (n=1065)										
	FIM_8SF		**FIM_4SF**		**MDS_13**		**MDS_8SF**		**MDS_4SF**	
FRG Difference (FIM_13 - Δ[2])	**Percent (#No.)**	**cumulative %**	**Percent (#No.)**	**cumulative %**	**Percent (#No.)**	**cumulative %**	**Percent (#No.)**	**cumulative %**	**Percent (#No.)**	**cumulative %**
0	85.16 (907)	85.16 (907)	80.66 (859)	80.66 (859)	64.13 (683)	64.13 (683)	62.82 (669)	62.82 (669)	59.91 (638)	59.91 (638)
±1	10.51 (112)	95.67 (1019)	11.46 (122)	92.12 (981)	14.74 (157)	78.87 (840)	15.03 (160)	77.85 (829)	14.55 (155)	74.46 (793)
±2	2.07 (22)	97.74 (1041)	3.85 (41)	95.97 (1022)	6.38 (68)	85.25 (908)	7.14 (76)	84.99 (905)	6.29 (67)	80.75 (860)
±3	2.25 (24)	100 (1065)	3.66 (39)	99.63 (1061)	7.79 (83)	93.04 (991)	8.17 (87)	93.16 (992)	11.08 (118)	91.83 (978)
±4			0.28 (3)	99.91 (1064)	3.01 (32)	96.05 (1023)	3.85 (41)	97.01 (1033)	5.17 (55)	97 (1033)
±5					2.63 (28)	98.68 (1051)	1.88 (20)	98.89 (1053)	1.79 (19)	98.79 (1052)
±6					0.84 (9)	99.52 (1060)	0.75 (8)	99.64 (1061)	0.85 (9)	99.64 (1061)
±7			0.09 (1)	100 (1065)	0.47 (5)	100 (1065)	0.38 (4)	100 (1065)	0.38 (4)	100 (1065)
Amputation (n=472)										
	FIM_8SF		**FIM_4SF**		**MDS_13**		**MDS_8SF**		**MDS_4SF**	
FRG Difference (FIM_13 - Δ)	**Percent (#No.)**	**cumulative %**	**Percent (#No.)**	**cumulative %**	**Percent (#No.)**	**cumulative %**	**Percent (#No.)**	**cumulative %**	**Percent (#No.)**	**cumulative %**
0	95.34 (450)	95.34 (450)	92.37 (436)	92.37 (436)	82.42 (389)	82.42 (389)	82.84 (391)	82.84 (391)	80.93 (382)	80.93 (382)
±1	4.66 (22)	100 (472)	7.63 (36)	100 (472)	17.59 (83)	100 (472)	17.16 (81)	100 (472)	19.07 (90)	100 (472)
Knee Replacement (n=568)										
	FIM_8SF		**FIM_4SF**		**MDS_13**		**MDS_8SF**		**MDS_4SF**	
FRG Difference (FIM_13 - Δ)	**Percent (#No.)**	**cumulative %**	**Percent (#No.)**	**cumulative %**	**Percent (#No.)**	**cumulative %**	**Percent (#No.)**	**cumulative %**	**Percent (#No.)**	**cumulative %**
0	97.54 (554)	97.54 (554)	97.71 (555)	97.71 (555)	90.49 (514)	90.49 (514)	78.35 (445)	78.35 (445)	78.35 (445)	78.35 (445)
±1	1.06 (6)	98.60 (560)	0.53 (3)	98.24 (558)	3.88 (22)	94.37 (536)	15.14 (86)	93.49 (531)	13.91 (79)	92.26 (524)
±2	1.23 (7)	99.83 (567)	1.05 (6)	99.29 (564)	2.46 (14)	96.83 (550)	2.99 (17)	96.48 (548)	2.64 (15)	94.9 (539)
±3	0.18 (1)	100 (568)	0.18 (1)	99.47 (565)	0.53 (3)	97.36 (553)	0.53 (3)	97.01 (551)	0.53 (3)	95.43 (542)
±4			0.36 (1)	99.83 (567)	2.29 (13)	99.65 (566)	2.64 (15)	99.65 (566)	3.7 (21)	99.13 (563)
±5			0.18 (1)	100 (568)	0.35 (2)	100 (568)	0.35 (2)	100 (568)	0.88 (5)	100 (568)

[2] **Δ**: represents each short form in this table (i.e. FIM_8SF, FIM_4SF, MDS_13, MDS_8SF, MDS_4SF)

Hip Replacement (n=394)										
	FIM_8SF		**FIM_4SF**		**MDS_13**		**MDS_8SF**		**MDS_4SF**	
FRG Difference (FIM_13 - Δ)	**Percent (#No.)**	**cumulative %**	**Percent (#No.)**	**cumulative %**	**Percent (#No.)**	**cumulative %**	**Percent (#No.)**	**cumulative %**	**Percent (#No.)**	**cumulative %**
0	97.97 (386)	97.97 (386)	94.67 (373)	94.67 (373)	85.28 (336)	85.28 (336)	71.57 (282)	71.57 (282)	69.80 (275)	69.80 (275)
±1	1.01 (4)	98.98 (390)	1.78 (7)	96.45 (380)	6.09 (24)	91.37 (360)	15.73 (62)	87.3 (344)	14.72 (58)	84.52 (333)
±2	1.01 (4)	100 (394)	3.04 (12)	99.49 (392)	4.06 (16)	95.43 (376)	7.1 (28)	94.4 (372)	8.37 (33)	92.89 (366)
±3			0.5 (2)	100 (394)	2.03 (8)	97.46 (384)	3.04 (12)	97.44 (384)	3.3 (13)	96.19 (379)
±4					1.78 (7)	99.24 (391)	1.77 (7)	99.21 (391)	2.54 (10)	98.73 (389)
±5					0.76 (3)	100 (394)	0.76 (3)	100 (394)	1.27 (5)	100 (394)

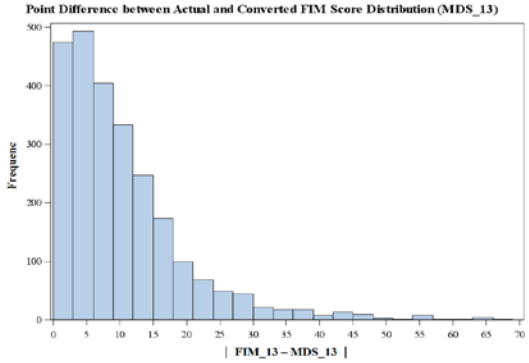
Table 4. Weighted Kappa, Kappa, McNemar's test and ICC between FIM_13 and the Varied Test Forms (FIM_8, FIM_4, MDS_13, MDS_8, MDS_4)

Stroke (n=1065)					
Test Form	**p-value**	**Weighted Kappa Statistics**	**Agreement Strength**	**ICC**	**ICC Strength**
FIM_8	<0.0001*[3]	0.93	Very Strong	0.99	Excellent
FIM_4	<0.0001*	0.90	Very Strong	0.98	Excellent
MDS_13	<0.0001*	0.73	Strong	0.90	Excellent
MDS_8	<0.0001*	0.73	Strong	0.91	Excellent
MDS_4	<0.0001*	0.69	Strong	0.88	Excellent
Amputation (n=472)					
Test Form	**p-value^[4]**	**Kappa Statistics**	**Agreement Strength**	**ICC**	**ICC Strength**
FIM_8	0.09	0.88	Very Strong	0.94	Excellent
FIM_4	0.74	0.81	Very Strong	0.89	Excellent
MDS_13	0.04**[5]	0.53	Moderate	0.70	Good
MDS_8	0.01**	0.54	Moderate	0.70	Good
MDS_4	0.01**	0.48	Moderate	0.65	Good
Knee Replacement (n=568)					
Test Form	**p-value**	**Weighted Kappa Statistics**	**Agreement Strength**	**ICC**	**ICC Strength**
FIM_8	0.0001*	0.78	Strong	0.94	Excellent
FIM_4	<0.0001*	0.70	Strong	0.87	Excellent
MDS_13	0.0001*	0.17	Weak	0.40	Fair
MDS_8	<0.0001*	0.14	Weak	0.35	Poor
MDS_4	0.0016*	0.09	Weak	0.22	Poor
Hip Replacement (n=394)					
Test Form	**p-value**	**Weighted Kappa Statistics**	**Agreement Strength**	**ICC**	**ICC Strength**
FIM_8	<0.0001*	0.95	Very Strong	0.99	Excellent
FIM_4	<0.0001*	0.85	Very Strong	0.96	Excellent
MDS_13	<0.0001*	0.55	Moderate	0.80	Excellent
MDS_8	<0.0001*	0.44	Moderate	0.76	Excellent
MDS_4	<0.0001*	0.34	Fair	0.67	Good

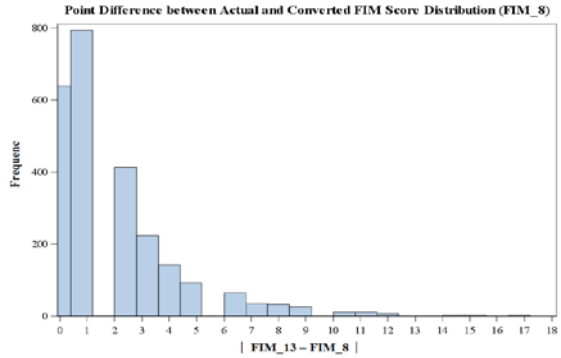
[3] ***:** Kappa agreement was significant at the level < 0.05
[4] **p-value^:** p-value from McNemar's Test for amputation FRG due to 2*2 table computation
[5] ****:** Significant difference between FRGa and FRGc

Figure 1. Point Difference between Actual and Converted FIM Score Distribution of Five Test Forms (MDS_13, FIM_8, MDS_8, FIM_4, MDS_4)

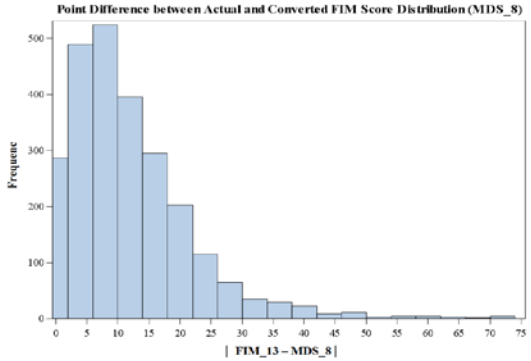
(a)



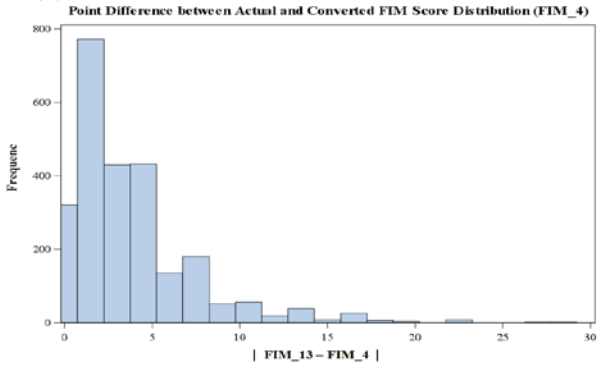
(b)



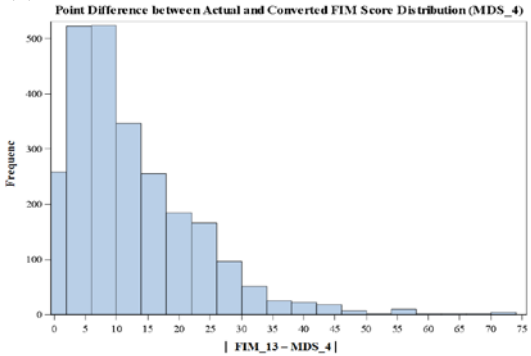
(c)



(d)



(e)



CHAPTER FIVE

CONCLUSION

Integrating the Findings

The overall goal of this dissertation was to challenge a widely accepted belief that developing a new single instrument was the only solution to assess patients' function across the continuum of post-acute care. This dissertation proposed an alternative solution by creating an item bank by linking existing instruments, Functional Independence Measure (FIM™) in the inpatient rehabilitation facilities and the Minimum Data Set (MDS) in the Community Living Centers, currently used across Veterans post-acute healthcare system.

Linking existing instruments to generate an item bank could further develop efficient administration such as short forms. To evaluate the feasibility of the 4- and 8-item short forms generated from the FIM-MDS item bank, we examined their measurement precision and accuracy compared with the original FIM_13-item motor score. To the author's knowledge, this dissertation was the first study that combined existing instruments into a single item bank and further validated precision and accuracy of the generated short forms. The importance of this study was to determine whether linking existing instruments could generate a continuity of care measurement system with precision and accuracy comparable to that of a single instrument.

Our study had five major findings:

(a) Linked instruments measuring the same latent trait can form an item bank with acceptable to good item-level psychometric properties.

(b) When the number of items of the test forms generated from the item bank decreased, measurement precision and accuracy decreased. This finding is consistent with Wright and Stone

(1979)'s formula of $SE(b_\nu) = X[L/r_\nu(L - r_\nu)]^{1/2} \cong 2.5/L^{1/2}$, indicating that when L (test length) increased then standard error (SE) of the test will decrease.

(c) MDS_13-item test form had measurement precision and measurement accuracy at the group-level that was comparable to the FIM_13-item test form.

(d) FIM_8-item had measurement precision and accuracy comparable to the FIM_13-item test form.

(e) The overall converted scores from the MDS and relevant short forms provided better group-level accuracy than the individual-level accuracy when compared to the original FIM_13-item scores.

In summary, our study results suggested the MDS_13-item could be used to obtain comparable precision and acceptable accuracy but not the MDS_4-item and 8-item short forms. In addition, the FIM_8-item instrument could potentially replace the FIM_13-item for clinical measurement, since it shows the best compromise between efficiency and precision/accuracy.

While our study results partially supported application of the MDS converted scores compared to the original FIM_13-item motor score, we raised a critical question that whether the linked instruments could produce comparable precision and accuracy to a universal instrument. In other words, if the converted scores of existing instruments measured a similar construct and showed valid results in terms of precision and accuracy as using a single instrument, then linking existing instruments using converted scores would be a cost-efficient solution to measuring patients across the continuum of care. This proposed solution could benefit healthcare policy makers and clinical practitioners regarding of maintaining fair reimbursement system across rehabilitation settings. In addition, linked measures would reduce the burden associated with adopting a new universal instrument (e.g., costs of electronic medical record software

modifications and burden of training on administering the new universal instrument) for the patients, healthcare policy makers and clinical practitioners.

Researchers have varying opinions about using converted scores to replace the scores obtained from the original instrument across the continuum of post-acute care. Buchanan and colleagues (2004) found a 56% agreement of classifications between FIM™ and converted FIM scores, and around 20% of the facilities had revenue shifts larger than 10% of the original cost with large standardized deviation (SD), thus concluded the converted scores should not be used. However, this study underestimated the impact of error variance and secondary variance on the results of their study. Wang and colleagues (2008a) found mixed results of their converted score in their validation study at individual and group levels, suggesting that error in the linked instruments could cause variance of the converted scores. In the area of rheumatoid arthritis, Ten Klooster and colleagues (2013) found that the agreements between predicted and observed scores from the Rasch-based crosswalk in the cross-validation sample had high intra-class correlation coefficients (ICCs). Oude Voshaar and colleagues (2014) replicated Ten Klooster et al.'s (2013) study and showed similar results of high ICCs, indicating the crosswalk was sufficiently reliable for group-level, even across diagnostic subgroups.

Thus, by controlling possible error sources, the results of linking instruments and using converted scores could be improved. Figure 5.1 was a visual demonstration of primary, secondary and error variance associated with using MDS_13-item converted scores as a continuity of measurement in our study. The primary variances are the consistent changes in the outcomes that we expected. Thus, the greater of the primary variance indicated a better quality of the performance of the instrument. On the other hand, secondary variance represented consistent changes in the outcomes due to the factors other than we expected but could be identified, and

the error variance were the inconsistent changes in the outcomes that could not be identified. Thus, a good instrument is expected to have greater primary variance and less secondary and error variance.

When using MDS_13-item converted scores, besides error variance such as instrumental intrinsic error that we could not control, sources of secondary variance that could impact on the outcomes may be controlled. Secondary variance may include different instrument used (i.e., MDS versus FIM), different time at administering the MDS, different raters, different rater's expectation or bias of the patients' function and patients' potential functional changes within 6 days.

Figure 5.2 demonstrated that when using MDS shorter versions, the element of "decreased number of item" could further contribute to decreasing the primary variance. When comparing short forms generated from the FIM and MDS, the main element to decrease explained primary variance of the FIM_8-item and FIM_4-item short forms was simply "decreased number of item" (Figure 5.3) compared to the MDS two short forms (Figure 5.2). This difference of involved secondary variance between the FIM and MDS short forms resulted in FIM short forms had better accuracy compared to the MDS short forms (Figures 5.2 & 5.3). Figures 5.1-5.3 also reflect the precision and accuracy comparison results between the FIM short forms and the MDS short forms presented in the previous chapter 4 of this dissertation.

Figure 5.4 visually demonstrated the assumed primary, secondary and error variance when using a single tool across the continuum of post-acute care rehabilitation settings. It is crucial to recognize that when using a single instrument across the continuum of post-acute care, this solution could simply remove one factor of "different instrument" contributing to secondary variance while other factors (e.g., different data collectors, and different time to administer the

instruments) contributing to the secondary variance still exist (Figure 5.4). Even though this single-tool-study-design may have less variance compared to our current study as shown in Figure 5.1, the main concern is the proportion of each element contributing to the secondary variance in the outcome variables. There are no studies to identify each factor contributing to the secondary variance (e.g., using different instruments would cause large or little impact on the outcomes). However, we could control certain factors with proper study design, so the impact of each factor could be minimized or identified.

Figure 5.5 demonstrated a study we proposed to identify the variance caused by using different instruments (thus also including removing the impact of different raters and rater bias) by testing the same instrument, for example, FIM_13-item, twice. In contrast to the present study, this design would eliminate the variance of having different instruments, but the design would retain, error variance and other contributors to secondary variance such as “patients’ functional change” and “different administration time.” Comparing the results of the present study (Figure 5.1), the proposed study shown in Figure 5.5 may clarify the differences between using a single instrument or a linked, item bank in measuring patients across the continuum of care.

There were several limitations of this dissertation. One was that we used retrospective data that was not designed for our study purpose. For instance, there may be potential functional change of the same patient even within 6 days between two instrumental administrations. In addition, there were inherent errors in the dataset that could not be controlled such as the level of strictness of the raters or rater bias (i.e., inpatient rehabilitation facility clinicians may be less severe raters than Community Living Center clinicians). Furthermore, the results of this study may be specific to the Veterans population due to its specific demographics, therefore limiting its generalizability.

Thus, to investigate the impact of each potential source of error upon the above mentioned limitations, we suggested future studies being designed as follows: (a) We could conduct the same study but instead of using a different instrument, testing the patient with the same instrument twice (e.g., FIM) within 6 days because FIM changes would be a function of: 1) error of the instrument and 2) impact of factors extrinsic to the instrument (e.g., changes in the patient over time). Since these parameters are similar to the conditions in which the MDS was collected, comparisons of precision and accuracy of converted scores of this proposed study would reflect the effect of using different instruments. (b) In addition, we would suggest conducting a prospective study with the same data collector to administer different instruments on the same day, which could reduce error resulting from different raters and different times for data collection. (c) Once we identified the impact of the error (i.e., error intrinsic to the instrument versus error extrinsic to the instrument), we may be able to control the impact of extrinsic error with a covariate analysis (i.e., remove the impact of the extrinsic error). Other methods of reducing error are to use computerized adaptive testing (CAT) to generate converted measures. CAT may improve the extent of error for the extreme ends of theta (i.e., person has extreme low or extreme high ability), which could potentially decrease the errors in the study. However, we hypothesized that CAT would not have a large impact in improving converted measures as compared to Item Response Theory (IRT)-based short forms since its effect is limited to the extreme scores.

In spite of advances in healthcare measurement, we are still at the beginning stages in understanding the impact of error on functional outcomes. Understanding, identifying and controlling the impact of intrinsic or extrinsic error variance and secondary variance on the healthcare instruments could improve precision and accuracy of measured outcomes and

facilitate practitioners in providing evidence-based treatment for the patients. In addition, when developing efficient tests to minimize clinician and patient burden, it is crucial to achieve a balance between test length, precision/accuracy. The ultimate goal of future studies is to establish precise and accurate functional outcome measures to monitor patients and ensure fair reimbursement across the continuum of post-acute care.

REFERENCES

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR20 index, and the Guttman scale response pattern. *Educational Psychology Research*, 9: 95–104.

Andrich, D. (2004). Controversy and the Rasch model: a characteristic of incompatible paradigms? *Medical Care*, 42:I7–I16.

Askew, R. L., Kim, J., Chung, H., Cook, K. F., Johnson, K. L., & Amtmann, D. (2013). Development of a crosswalk for pain interference measured by the BPI and PROMIS pain interference short form. *Quality of Life Research*, 22, 2769-2776.

Baker, F. B. (1993). Equating tests under the graded response model. *Applied Psychological Measurement*, 16, 87–96.

Baker, F. B., & Al-karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement*, 28, 147–162.

Bjorner, J. B., Kosinski, M., Ware, J. E. Jr, (2003). Using item response theory to calibrate the Headache Impact Test (HIT) to the metric of traditional headache scales. *Quality of Life Research*, 12, 981-1002

Bjorner, J. B., Chang, C. H., Thissen, D., & Reeve, B. B. (2007). Developing tailored instruments: item banking and computerized adaptive assessment. *Quality of Life Research*, 16 (Suppl 1), 95–108.

Bond, T. G., & Fox, C. M. (2007). Applying the Rasch model: Fundamental measurement in the human sciences. Mahwah, NJ: Lawrence Erlbaum.

Bjorner, J. B., Rose, M., Gandek, B., Stone, A. A., Junghaenel, D. U., & Ware, J. E. Jr. (2014). Difference in method of administration did not significantly impact item response: an IRT-based analysis from the Patient-Reported Outcomes Measurement Information System (PROMIS) initiative. *Quality of Life Research*, 23(1), 217-27. doi: 10.1007/s11136-013-0451-4. Epub 2013 Jul 23.

Buchanan, J. L., Andres, P. L., Haley, S. M., Paddock, S. M., & Zaslavsky, A. M. (2003). An assessment tool translation study. *Health Care Financing Review*, 24, 45-60.

Buchanan, J. L., Andres, P. L., Haley, S. M., Paddock, S. M., & Zaslavsky, A. M. (2004). Evaluating the planned substitution of the minimum data set-post acute care for use in the rehabilitation hospital prospective payment system. *Medical Care*, 42(2), 155-63.

Bukstein DA1, McGrath MM, Buchner DA, Landgraf J, Goss TF. (2000). Evaluation of a short form for measuring health-related quality of life among pediatric asthma patients. *Journal of Allergy and Clinical Immunology*, 105 (2 Pt 1), 245-51.

Calhoun, C., Haley, S., Riley, A., Vogel, L., McDonald, C., & Mulcahey, M. (2009). Development of items designed to evaluate activity performance and participation in children and adolescents with spinal cord injury. *International Journal of Pediatrics*, 854904.

Carle, A. C., Cella, D., Cai, L., Choi, S. W., Crane, P. K., Curtis, S. M., Gruhl, J., Lai, J. S., Mukherjee, S., Reise, S. P., Teresi, J. A., Thissen, D., Wu, E. J., & Hays, R. D. (2011). Advancing PROMIS's methodology: results of the Third Patient-Reported Outcomes Measurement Information System (PROMIS ®) Psychometric Summit. *Expert Review of Pharmacoeconomics & Outcomes Research*, 11(6), 677-84. doi: 10.1586/erp.11.74.

Carmody, T. J., Rush, A. J., Bernstein, I., Warden, D., Brannan, S., Burnham, D., Woo, A., & Trivedi, M. H. (2006). The Montgomery Asberg and the Hamilton ratings of depression: a comparison of measures. *European Neuropsychopharmacology*, 16, 601–611.

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., Ader, D., Fries, J. F., Bruce, B., & Rose, M. (2007). The patient-reported outcome measurement information system (PROMIS): Progress of an NIH roadmap cooperative group during its first two years. *Medical Care*, 45(5 Suppl 1), S3–S11. [PubMed:17443116]

Centers for Medicare & Medicaid Services (2011). Active Projects Report- Research and Demonstrations in Health Care Financing. 2011 Edition. Retrieved on May 22th, 2013 from https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/ActiveProjectReports/Downloads/2011_Active_Projects_Report.pdf

Centers for Medicare & Medicaid Services (2012). U.S. Department of Health and Human Services. Report to Congress: Post Acute Care Payment Reform Demonstration (PAC-PRD). January, 2012. Retrieved on May 22th, 2013 from http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Reports/downloads/Flood_PACPRD_RTC_CMS_ Report_Jan_2012.pdf

Centers for Medicare and Medicaid Services (2015). IMPACT Act of 2014 & Cross Setting Measures. Retrieved on 6/17/2015 from http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Post-Acute-Care-Quality-Initiatives/IMPACT-Act-of-2014-and-Cross-Setting-Measures.html

Chang, C. H., & Cella, D. (1997). Equating health-related quality of life instruments in applied oncology settings. *Archives of Physical Medicine and Rehabilitation*, 11(2), 397-406.

Chen, W., Revicki, D., Lai, J., Cook, K., & Amtmann, D. (2009). Linking pain items from two studies onto a common scale using item response theory. *Journal of Pain and Symptom Management*, 38, 615-28.

Choi, S. W., Podrabsky, T., Mckinney, N., Schalet, B. D., Cook, K. F., & Cella, D. (2012). Prosetta Stone® Analysis Report: A Rosetta Stone For Patient Reported Outcomes. Retrieved on 11/24/2014 from http://www.prosettastone.org/AnalysisReport/Documents/PROsettaStoneAnalysisReportVol1.pdf

Choi, S. W., Reise, S. P., Pilkonis, P. A., Hays, R. D., & Cella, D. (2010). Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. *Quality of Life Research*, 19(1): 125-36. doi: 10.1007/s11136-009-9560-5.

Cook, K. F., Taylor, P. W., Dodd, B. G., Teal, C. R., & McHorney, C. A. (2007). Evidence-based practice for equating health status items: sample size and IRT model. *Journal of Applied Measurement*, 8,175–189.

Chang, C. H., & Cella, D. (1997). Equating health-related quality of life instruments in applied oncology settings. *Archives of Physical Medicine and Rehabilitation*, 11(2), 397–406.

Davidoff, G., Roth, E. J., Haughton, J. S., & Ardner, M. S. (1990). Cognitive dysfunction in spinal cord injury patients: sensitivity of the Functional Independence Measure subscales vs neuropsychologic assessment. *Archives of physical medicine and rehabilitation*, 71(5), 326-9.

del Toro, C. M., Bislick, L. P., Comer, M., Velozo, C., Romero, S., Gonzalez Rothi, L. J., & Kendall, D. L. (2011). Development of a short form of the Boston naming test for individuals with aphasia. *Journal of Speech Language Hearing Research*, 54 (4), 1089-10100.

Dorans, N. J. (1999). Correspondences between ACT and SAT I scores. College Entrance Examination Board, New York, NY. College Board Report No. 99-1. ETS RR No.99-2.

Dorans, N. J. (2007). Linking scores from multiple health outcome instruments. *Quality of Life Research*, 16 (Suppl 1), 85-94.

Dorans, N.J., Pommerich, M., & Holland, P.W. (2007). Linking and aligning scores and scales. NY: Springer.

Dorans, N. J., Pommerich, M., & Holland, P. W. (2010). Statistics for Social and Behavioral Sciences: Linking and Aligning Scores and Scales. Springer Science + Business Media, LLC, New York, NY. Chapter 11 Concordance: The good, the bad, and the ugly, pp. 202-216.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment*, 8, 341-349.

Fiedler, R., & Granger, C. (1997). Uniform data system for medical rehabilitation SM: Report of first admissions for 1995. *American Journal of Physical Medicine & Rehabilitation*, 76(1), 76-81.

Fischer, H. F., Tritt, K., Klapp, B. F., & Fliege, H. (2011). How to compare scores from different depression scales: equating the patient health questionnaire (PHQ) and the ICD-10-

symptom rating (ISR) using item response theory. *International Journal of Methods in Psychiatric Research*, 20, 203–214.

Fischer, H. F., Wahl, I., Fliege, H., Klapp, B. F., & Rose, M. (2012). Impact of cross-calibration methods on the interpretation of a treatment comparison study using 2 depression scales. *Medical Care*, 50, 320–326.

Fisher, W. P. Jr. (1997). Physical disability construct convergence across instruments: towards a universal metric. *Journal of Outcome Measurement*, 1(2), 87-113.

Fisher, W. P. Jr., Harvey, R. F., Taylor, P., Kilgore, K. M., & Kelly, C. K. (1995). Rehabits: a common language of functional assessment. *Archives of Physical Medicine and Rehabilitation*, 76(2), 113-22.

Fisher, W. P. Jr., Eubanks, R. L., & Marier, R. L. (1997). Equating the MOS SF36 and the LSU HSI Physical Functioning Scales. *Journal of Outcome Measurement*, 1(4), 329-62.

Fong, T. G., Fearing, M. A., Jones, R. N., Shi, P., Marcantonio, E. R., Rudolph, J. L., Yang, F. M., Kiely, D. K., & Inouye, S. K. (2009). Telephone interview for cognitive status: Creating a crosswalk with the Mini-Mental State Examination. *Alzheimer's & Dementia*, 5(6), 492-7. doi: 10.1016/j.jalz.2009.02.007. Epub 2009 Jul 31

Fries, J. F., Cella, D., Rose, M., Krishnan, E., & Bruce, B. (2009). Progress in assessing physical function in arthritis: PROMIS short forms and computerized adaptive testing. *Journal of Rheumatology*, 36(9), 2061-6. doi: 10.3899/jrheum.090358.

Gibbons, R. D., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T., Kim, J. B., & Kupfer DJ. (2014). Development of the CAT-ANX: a computerized adaptive test for anxiety. *The American Journal of Psychiatry*, 171(2), 187-94. doi: 10.1176/appi.ajp.2013.13020178.

Gonin, R., Lloyd, S., Cella, D., & Gray, G. (1996). Establishing equivalence between scaled measures of quality of life. *Quality of Life Research*, 5(1), 20–6. Erratum in: *Quality of Life Research* (2001), 10(1), 104.

Granger, C., & Hamilton, B. (1993). The Uniform Data System for Medical Rehabilitation report of first admissions for 1991. *American Journal of Physical Medicine & Rehabilitation*, 72(1), 33.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. Educational measurement. New York, Macmillan. 147-200.

Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, D. R., & Gifford, J. A. (1978). Developments in Latent Trait Theory: A review of models, technical issues, and applications. *Review of Educational Research*, 48 (4), 467-510.

Haley, S. M., Ni, P., Lai, J. S., Tian, F., Coster, W. J., Jette, A. M., Straub, D., & Cella, D. (2011) Linking the activity measure for post acute care and the quality of life outcomes in neurological disorders. *Archives of Physical Medicine and Rehabilitation*, 92(10 Suppl), S37-43. doi: 10.1016/j.apmr.2011.01.026.

Hallgren, K. A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials Quantitative Methods Psychology*, 8(1), 23-34.

Hawes, C., Morris, J. N., Phillips, C. D., Mor, V., Fries, B. E., & Nonemaker, S. (1995). Reliability estimates for the Minimum Data Set for nursing home resident assessment and care screening (MDS). *The Gerontologist*, 35(2), 172-8.

Hol, A. M., Vorst, H. C. M., & Mellenbergh, G. J. (2007). Computerized Adaptive Testing for Polytomous Motivation Items: Administration Mode Effects and a Comparison With Short Forms. *Applied Psychological Measurement*, 31, 412-429.

Holzner, B., Bode, R. K., Hahn, E. A., Cella, D., Kopp, M., Sperner-Unterweger, B., & Kemmler, G. (2006). Equating EORTC QLQ-C30 and FACT-G scores and its use in oncological research. *European Journal of Cancer*, 42(18), 3169-77. Epub 2006 Oct 11.

Hu, L., & Bentler, P. M. (1996). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1-55. doi:10.1080/10705519909540118.

Jette, A. M., Haley, S. M., & Ni, P. (2003). Comparison of functional status tools used in post-acute care. *Health Care Financing Review*, 24(3),13-24.

Jones, A. L., Dwyer., L. L., Bercovitz, A. R., & Strahan, G. W. (2009). The National Nursing Home Survey: 2004 overview. National Center for Health Statistics. *Vital Health Stattistics*, 167, 1-155.

Kolen, M. J., & Brennan, R. L. (2004). Test Equating, Scaling, and Linking: Methods and Practice. 2nd edit. Springer Science + Business Media, LLC, New York, NY. Chapter 6 Item Response Theory Methods, pp. 176-205.

Lai, J. S., Cella, D., Choi, S., Junghaenel, D. U., Christodoulou, C., Gershon, R., & Stone, A. (2011). How item banks and their application can influence measurement practice in

rehabilitation medicine: a PROMIS fatigue item bank example. *Archives of Physical Medicine and Rehabilitation*, 92 (10 Suppl), S20-7. doi: 10.1016/j.apmr.2010.08.033.

Lai, J. S., Cella, D., Yanez, B., & Stone, A. (2014). Linking fatigue measures on a common reporting metric. *Journal of Pain and Symptom Management*, 48(4), 639-48. doi: 10.1016/j.jpainsymman.

Landgraf, J. M. (2007). Precision and sensitivity of the short-form pediatric enuresis module to assess quality of life (PEMQOL). *Journal of Pediatric Urology*, 3(2), 109-17. doi: 10.1016/j.jpurol.2006.04.004. Epub 2006 Jun 9.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–74.

Latimer, S., Covic, T., & Tennant, A. (2012). Co-calibration of deliberate self-harm (DSH) behaviours: towards a common measurement metric. *Psychiatry Research*, 200(1), 26-34. doi: 10.1016/j.psychres.2012.05.019. Epub 2012 Jun 22.

Lawton, M., Casten, R., Parmelee, P. A., Van Haitsma, K., Corn, J., & Kleban, M. H. (1998). Psychometric characteristics of the minimum data set II: validity. *Journal of the American Geriatrics Society*, 46(6), 736-44.

Leucht, S., Kane, J. M., Etschel, E., Kissling, W., Hamann, J., & Engel, R. R. (2006). Linking the PANSS, BPRS, and CGI: clinical implication. *Neuropsychopharmacology*, 31(10), 2318-25. Epub 2006 Jul 5.

Li, C. Y., Romero, R., Simpson, K., Simpson, A., Bonilha, H., Hong, I., & Velozo, C. (2015a). Continuum of care assessment across post-acute care in Veterans: Linking existing

instruments to develop an activity of daily living item bank. Archives of Physical Medicine and Rehabilitation (in preparation).

Li, C. Y., Romero, S., Simpson, A., Simpson, K., Bonilha, H., Hong, I., & Velozo, C. (2015b). Comparisons of Functional Independence Measure-Minimum Data Set Short Forms. Archives of Physical Medicine and Rehabilitation (in preparation).

Linacre, J., Heinemann, A. W., Wright, B. D., Granger, C. V., & Hamilton, B. B. (1994). The structure and stability of the Functional Independence Measure. *Archives of Physical Medicine and Rehabilitation*, 75(2), 127-132.

Linacre, J. M. (1998). Table 13.1 Item statistics in measure order. Rasch Measurement Forum. Retrieved on 12/12/2013 from http://www.winsteps.com/winman/index.htm?correlations.htm

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, 3:1, 85-106. Retreived from http://www.winsteps.com/a/linacre-optimizing-category.pdf

Linacre, J. M. (2004). Rasch model estimation: Further topics. *Journal of Applied Measurement*, 5, 95–110.

Linacre, J. M. (2010). Predicting responses from Rasch measures. *Journal of Applied Measurement*, 11, 1–10.

Linacre, J. M. (2012). A user's guide to Winsteps ministep 3.70.0: Rasch model computer programs. Chicago, IL: Winsteps.

Linacre, J.M. (2014). Winsteps® (Version 3.81.0) [Computer Software]. Beaverton, Oregon: Winsteps.com. Retrieved January 1, 2014. Available from http://www.winsteps.com/

Masse, L. C., Allen, D., Wilson, M., & Williams, G. (2006). Introducing equating methodologies to compare test scores from two different self-regulation scales. *Health Education Research*, 21(Suppl 1), i110–i120.

McHorney, C. A. (2002). Use of item response theory to link three modules of functional status items from the Asset and Health Dynamics among the Oldest Old Study. *Archives of Physical Medicine and Rehabilitation*, 83, 383-94.

McHorney, C. A., & Cohen, A. S. (2000). Equating heath status measures with item response theory: Illustrations with functional status items. *Medical Care*, 38 (9 Suppl), 43–59.

Muthén, L. K., & Muthén, B. O. (2014). Mplus User's Guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén.

Morris, J., Hawes, C., Fries, B. E., Phillips, C. D., Mor, V., Katz, S., Murphy, K., Drugovich, M. L., & Friedlob, A. S. (1990). Designing the national resident assessment instrument for nursing homes. *The Gerontologist*, 30(3), 293.

Morris, J., Fries, B. E., Mehr, D. R., Hawes, C., Phillips, C., Mor, V., & Lipsitz, L. A. (1994). MDS cognitive performance scale©. *Journal of Gerontology*, 49(4), M174.

NHANES (2014). National Health and Nutrition Examination Survey (NHANES) 1992-2014 Survey Content Brochure. Retrieved on 06/02/2014 from http://www.cdc.gov/nchs/data/nhanes/survey_content_99_14.pdf

Nilsson, A. L., Sunnerhagen, K. S., & Grimby, G. (2005). Scoring alternatives for FIM in neurological disorders applying Rasch analysis. *Acta Neurologica Scandinavica*, 111(4), 264-73.

Noonan, V. K., Cook, K. F., Bamer, A. M., Choi, S. W., Kim, J., & Amtmann, D. (2012). Measuring fatigue in persons with multiple sclerosis: creating a crosswalk between the Modified

Fatigue Impact Scale and the PROMIS Fatigue Short Form. *Quality of Life Research*, 21(7), 1123-33. doi: 10.1007/s11136-011-0040-3.

Orlando, M., Sherboune, C. D., & Thissen, D. (2000). Summed-score linking using item response theory: Application to depression measurement. *Psychological Assessment*, 12, 354–359.

Ornstein, K. A., Teresi, J. A., Ocepek-Welikson, K., Ramirez, M., Meier, D. E., Morrison, R. S., & Siu, A. L. (2015). Use of an item bank to develop two short-form famcare scales to measure family satisfaction with care in the setting of serious illness. *Journal of Pain and Symptom Management*, 49(5), 894-903.e4. doi: 10.1016/j.jpainsymman.2014.10.017.

Ottenbacher, K., Hsu, Y., Granger, C. V., & Fiedler, R. C. (1996). The reliability of the functional independence measure: a quantitative review. *Archives of Physical Medicine and Rehabilitation*, 77(12), 1226-1232.

Oude Voshaar, M. A., Ten Klooster, P. M., Taal, E., Wolfe, F., Vonkeman, H., Glas, C. A., & Van De Laar, M. A. (2014). Linking physical function outcomes in rheumatology: performance of a crosswalk for converting Health Assessment Questionnaire scores to Short Form 36 physical functioning scale scores. *Arthritis Care & Research*, 66(11), 1754-8.

PROMIS® (2014). Instrument Development and Psychometric Evaluation Scientific Standards* retrieved on 06/06/2014 from http://www.nihpromis.org/Documents/PROMISStandards_Vers2.0_Final.pdf

Rantz, M. (1999). The minimum data set: No longer just for clinical assessment. *Annals of Long- Term Care*, 7(9), 354-360.

Reeve, B. B., Burke, L. B., Chiang, Y. P., Clauser, S. B., Colpe, L. J., Elias, J. W., Fleishman, J., Hohmann, A. A., Johnson-Taylor, W. L., Lawrence, W., Moy, C. S., Quatrano, L. A., Riley, W. T., Smothers, B. A., & Werner, E. M. (2007a). Enhancing measurement in health outcomes research supported by Agencies within the US Department of Health and Human Services. *Quality of Life Research*, 16 Suppl 1, 175-86.

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., Thissen, D., Revicki, D. A., Weiss, D. J., Hambleton, R. K., Liu, H., Gershon, R., Reise, S. P., Lai, J. S., Cella, D. & PROMIS Cooperative Group. (2007b). Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care*, 45(5 Suppl 1), S22-31.

Reise, S. P., & Henson, J. M. (2000). Computerization and adaptive administration of the NEO PI-R. *Assessment*, 7, 347–364.

Research Triangle Institute (RTI) International (2009). Examining Post Acute Care Relationships in an Integrated Hospital System: Final Report. Retrieved on July 17th, 2013 from http://aspe.hhs.gov/health/reports/09/pacihs/report.shtml

Rose, M., Bjorner, J. B., Becker, J., Fries, J. F., & Ware, J. E. (2008). Evaluation of a preliminary physical function item bank supported the expected advantages of the Patient-Reported Outcomes Measurement Information System (PROMIS*). Journal of Clinical Epidemiology*, 61(1), 17-33.

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina.

Slavin, M., Kisala, P., Jette, A., & Tulsky, D. S. (2010). Developing a contemporary outcome measure for spinal cord injury research. *Spinal Cord*, 48(3), 262-7.

Smith, E. V., Jr. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*, 3(2), 205-231.

Smith, R. M., & Taylor, P. A. (2004). Equating rehabilitation outcome scales: Developing Common Metrics. *Journal of Applied Measurement*, 5(3), 229-242.

Stineman, M. G. (1995). Case-mix measurement in medical rehabilitation. *Archives of Physical Medicine and Rehabilitation*, 76(12), 1163-70.

Stineman, M. G., Escarce, J. J., Goin, J. E., Hamilton, B. B., Granger, C. V., & Williams, S. V. (1994). A case mix classification system for medical rehabilitation. *Medical Care*, 32, 366–379.

Stineman, M., Shea, J. A., Jette, A., Tassoni, C. J., Ottenbacher, K. J., Fiedler, R., & Granger, C. V. (1996). The Functional Independence Measure: tests of scaling assumptions, structure, and reliability across 20 diverse impairment categories. *Archives of Physical Medicine and Rehabilitation*, 77(11), 1101-1108.

Stineman, M. G., Tassoni, C. J., Escarce, J. J., Goin, J. E., Granger, C. V., Fiedler, R. C., & Williams, S. V. (1997). Development of function related groups version 2. 0: a classification system for medical rehabilitation. *Health Serveries Research*, 32, 529-548.

Tate, R. L. (1999). A cautionary note on IRT-based linking of tests with polytomous items. *Journal of Educational Measurement*, 36(4), 336–346.

Ten Klooster, P. M., Oude Voshaar, M. A., Gandek, B., Rose, M., Bjorner, J. B., Taal, E., Glas, C. A., van Riel, P. L., & van de Laar, M. A. (2013). Development and evaluation of a crosswalk between the SF-36 physical functioning scale and Health Assessment Questionnaire

disability index in rheumatoid arthritis. *Health Quality of Life Outcomes*, 11, 199. doi: 10.1186/1477-7525-11-199.

Tennant, A., & Pallant, J. F. (2006). Unidimensionality Matters! (A Tale of Two Smiths?). *Rasch Measurement Transactions*, 20(1), 1048-1051.

Thissen, D., Varni, J. W., Stucky, B. D., Liu, Y., Irwin, D. E., & DeWalt, D. A. (2011). Using the PedsQL (TM) 3.0 asthma module to obtain scores comparable with those of the PROMIS pediatric asthma impact scale (PAIS). *Quality of Life Research*, 20(9), 1497–1505.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: Test is not reliable. *Educational and Psychological Measurement*, 60(2), 174–95.

Tsan, L., Langberg, R., Davis, C., Phillips, Y., Pierce, J., Hojlo, C., Gibert, C., Gaynes, R., Montgomery, O., Bradley, S., Danko, L., & Roselle, G. (2008). Nursing home-associated infections in Department of Veterans Affairs community living centers. *American Journal of Infection Control*, 38(6), 461-6.

Tulsky, D., Kisala, P., Victorson, D., Tate, D., Heinemann, A. W., & Cella, D. (2011). Developing a Contemporary Patient Reported Outcomes Measure for Spinal Cord Injury. *Archives of Physical Medicine and Rehabilitation*, 92 suppl 1(10), S44–S51. [PubMed: 21958922]

Velozo, C. A., Byers, K. L., Wang, Y. C., & Joseph, B. R. (2007). Translating measures across the continuum of care: using Rasch analysis to create a crosswalk between the Functional Independence Measure and the Minimum Data Set. *Journal of Rehabilitation Research & Development*, 44(3), 467-78.

Velozo, C. A., Magalhaes, L. C., Pan, A. W., & Leiter, P. (1995). Functional scale discrimination at admission and discharge: Rasch analysis of the Level of Rehabilitation Scale-III. *Archives of Physical Medicine and Rehabilitation*, 76(8),705-12.

VHA Directive 2000-016. (2002). Medical Rehabilitation Outcomes for Stroke, Traumatic Brian Injury, and Lower Extremity Amputation Patients.

Von Davier, A.A., Holland, P.W., & Thayer, D. T. (2004). The kernel method of test equating. NY: Springer

Wang, Y. C., Byers, K. L., & Velozo, C. A. (2008a). Validation of FIM™-MDS crosswalk conversion algorithm. *Journal of Rehabilitation Research & Development*, 45(7), 1065-76.

Ware, J. r., John, E., Gandek, Barbara; Sinclair, Samuel J.; Bjorner, Jakob B. (2005).Item response theory and computerized adaptive testing: Implications for outcomes measurement in rehabilitation. *Rehabilitation Psychology*, 50(1), 71-78. doi: 10.1037/0090-5550.50.1.71

Williams, B. C., Li, Y., Fries, B. E., & Warren, R. L. (1997). Predicting Patient Scores Between the Functional Independence Measure and the Minimum Data Set: Development and Performance of a FIM™-MDS "Crosswalk." *Archives of Physical Medicine and Rehabilitation*, 78(1), 48-54.

Wright, B. D., & Bell, S. R. (1984). Item banks: What, why, how. *Journal of Educational Measurement*, 21(4), 331–345.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8(3), 370.

Wright, B. D., & Masters, G. N. (1982). Rating scale analysis. Chicago, IL: Mesa Press.

Wright, B. D., & Stone, M. H. (1979). Best Test Design. ISBN 0-941938-00-X. LC# 79-88489.

Wu, M., & Adams, R. J. (2013). Properties of Rasch residual fit statistics. *Journal of Applied Measurement*, 14(4), 339-55.

Yu, L., Buysse, D. J., Germain, A., Moul, D. E., Stover, A., Dodds, N. E., Johnston, K. L., & Pilkonis, P. A. (2011). Development of short forms from the PROMIS™ sleep disturbance and Sleep-Related Impairment item banks. *Behavioral Sleep Medicine*, 10(1), 6-24. doi: 10.1080/15402002.2012.636266.

Zwick, R., Thayer, D.T., Lewis, C. (1999) An Empirical Bayes Approach to Mantel-Haenszel DIF Analysis. *Journal of Educational Measurement*, 36, 1, 1-28

APPENDIX- TABLES

Table 1.1. Measurement System across Post-Acute Care (PAC) Facilities

Post-Acute Care (PAC) Facilities	**Inpatient Rehabilitation Facility (IRF)**	**Skilled Nursing Facility (SNF)**	**Home Health Services (HHA)**
Measurement System	Inpatient Rehabilitation Facility Patient Assessment Instrument (IRF-PAI)*	Minimum Data Set (MDS)	Outcome and Assessment Information Set (OASIS)

* IRF-PAI includes Functional Independence Measure (FIM) and additional demographic data (ie. age, gender)

Table 1.2. Parameters Measured in the CARE Item Set, FIM and MDS

Instrument	**Continuity Assessment and Record Evaluation (CARE) item set**	**Functional Independence Measure (FIM)**	**Minimum Data Set (MDS)**
Parameter I: ADL/Motor Skill	Eating	Eating	Eating
	Oral Hygiene	Grooming	Personal Hygiene
	Wash Upper Body	-----	-----
	Shower/ Bathe Self	Bathing	Bathing
	Dressing- Upper Body	Dressing- Upper Body	Dressing
	Dressing- Lower Body	Dressing- Lower Body	-----
	Toileting Hygiene	Toileting	Toilet Use
	-----	Bladder Management	Bladder Continence
	-----	Bowel Management	Bowel Continence
	Put On/ Take Off Footwear	-----	-----
	Bed to Chair/Wheelchair Transfer Sit to Lying Sit to Stand	Bed, Chair, Wheelchair (Transfer)	Transfer
	Toilet Transfer	Toilet (Transfer)	-----
	-----	Tub, Shower (Transfer)	-----
	-----	Stairs	-----
	Roll Left to Right Lying to Sitting On Side of Bed	-----	Bed Mobility
	Walking or Wheeling (in room, 50 feet, 100 feet, 150 feet) * One Step Curb * Four Steps *	Walk/Wheelchair	Walk in Room
	Twelve Steps	-----	Walk in Corridor
	Walk 50 feet With 2 Turns *	-----	Locomotion on Unit
	Walk 10 feet On Uneven Surfaces *	-----	Locomotion off Unit
	Pick Up Object	-----	-----

	Car Transfers	-----	-----
Rating Scale	6= Complete Independence	7= Complete Independence	0= Independent
	5= Setup or Cleanup Assistance	6= Modified Independence	-----
	4= Supervision or Touching Assistance	5= Supervision	1= Supervision
	-----	4= Minimal Assistance (>75% independence)	2= Limited Assistance
	3= Partial or Moderate Assistance	3= Moderate Assistance (>50% independence)	-----
	2= Substantial or Maximal Assistance	2= Maximal Assistance (>25% independence)	3= Extensive Assistance
	1= Complete Dependence	1= Total Assistance	4= Total Dependence
	M= Unable to Perform the Activity due to Medical Issues S= Unable to Perform the Activity due to Safety Issues N= Non-Applicable P= Patient Refuses A= The Activity was Attempted but Not Completed **	-----	8= Activity Did Not Occur During Entire 7-Day Period

“ * ” means this activity may be considered as either “Locomotion on Unit” or “Locomotion off Unit”.

“ ** ”: All letter codes are recoded to 1 (totally dependent).

Table 2.1. Literature Reviews of Linking Methods Used in Healthcare Professions (Classical Testing Theory) (ordered by year) (n=6)

Author	Title	Aims	Methods	Instruments/ Population	Results	Conclusions
Williams, B. C., Li, Y., Fries, B. E., & Warren, R. L. (1997)	Predicting patient scores between the Functional Independence Measure and the Minimum Data Set: Development and performance of a FIM™-MDS "crosswalk"	Establish and validate a crosswalk between FIM™ and MDS across acute rehab settings and nursing homes	• Prospective study • An expert panel of 7 rehab experts chose and rescaled MDS items to create "Pseudo-FIM™" • The relationships between Pseudo-FIM™ and FIM™ were compared using Wilcoxon Rank Sum tests • Rescaled the MDS based on two methods: the expert panel (FIM™(E)) determinations and observed relationships in development data set (FIM™(O))	• Functional Independence Measure (FIM™) • Minimum Data Set (MDS) • 173 Rehab patients admitted to six nursing homes (same population of patients)	1. Items of walking/ locomotion and social interaction were excluded because the authors considered no corresponding MDS items found in the FIM™. The final were 13 out of 18 FIM™ items having corresponding MDS items (but two dressing items were combined; so the final total number of item is 12) 2. Mean Pseudo-FIM™ (E) and FIM™ scores of five items (out of 12 items); and eight items of Pesudo-FIM™ (O) were not significantly different (p <.05). 3. Intraclass correlation coefficients between the FIM™ and Pseudo-FIM™ (E) motor and cognitive subscales were both 0.8l. 4. Crosswalk values defined as implausible by the expert panel generally occurred for middle levels of limitations. 5. FIM™ and MDS-based rescaled items were	From the Article: FIM™ and MDS can predict item and subscale scores interchangeably with reasonable accuracy, which could compare the effectiveness (degree of improvement among similar patients) and efficiency (cost of care to obtain a given degree of improvement) of rehabilitation care in different settings. Relevant to Dissertation: This study partially supports the assumption of creating a crosswalk between instruments (i.e., FIM™ and MDS) based on CTT methods by developing corresponding items between instruments and compare their differences

					more similar when using the method of FIM™(O) than using the method of FIM™(E) 6. The absolute differences in group means FIM™(E) for the two instruments were within 0.5 points for 6 items and within 0.8 points for 11 of the 12 items	
Buchanan, J. L., Andres, P. L., Haley, S. M., Paddock, S. M., & Zaslavsky, A. M. (2003)	An assessment tool translation study	Aims to examine if it is feasible to substitute the minimum data set post-acute care (MDS-PAC) into the planned prospective payment system (PPS) for inpatient rehabilitation hospitals instead of currently used tool [the functional independence measure (FIM™™)] from a large scale effort using classical testing theory methods	• Prospective study • Raters of both FIM™ and MDS-PAC completed training with development group trainers before scoring the patients • MDSPAC scores of 1 (set up help only) and 2 (supervision) were mapped to a FIM™™ score 5 (Supervision) • The linking method used included: (a) realigning the seven scoring levels; (b) incorporating ADL assist codes; (c) item-specific translation revisions • Factor analysis was used on the	• Functional Independence Measure (FIM™) • Minimum Data Set post-acute care (MDS-PAC) • Fifty FIM™™-certified rehab facilities (representing rehab hospitals across the country; 16% were rural and 28% were freestanding facilities) • Over 3,200 FIM™™ and MDS-PAC pairs. One or more of three highly trained calibration teams visited each participating hospitals and rescored both the FIM™™ and the MDS-PAC for 38	1. The comparison between the actual FIM™™ motor scale and item scores with those obtained from the MDS-PAC translations and summated scales: (a) mean FIM™™ motor scale score differed from the mean MDS-PAC motor scale translation by nearly 5 points (45.46 vs. 50.26); (b) mean FIM™™ cognitive scale score was close to the mean MDS-PAC translation (28.50 vs. 28.51) 2. The revised translation reduced the mean difference in motor scores between the FIM™™ and the MDS-PAC by 50 % from the original translation	From the Article: 1. Scoring differences varied by hospital and this variation was not explained by any other independent variables, indicating this was a substantial effect to be of concern for the comparability of scoring procedures across facilities; the authors suggested more training is needed to adequately standardize assessment process 2. Under all potential adjustments, the level of classification agreement of translated scores was low and clearly not adequate for payment purposes 3. The authors also found substantial proportion of the facilities would experience potentially important shifts in revenue. Thus, policymakers opted to retain the FIM™™ 4. The authors concluded that the need for a unified common conceptual framework and a rigorous standardized approach

			combined set of motor items from both the FIM™™ and the MDS-PAC • Scoring agreement was measured with Pearson correlation and weighted/ unweighted kappa statistics • Regression analysis to analyze scoring differences across facilities	current cases. Thus approximately 200 cases had two FIM™™ and two MDS-PAC ratings	3. Neither the raw items nor those from the original translation all loaded onto the same factors as the corresponding FIM™™ items (while items from the revised translation did) 4. The agreement between the instruments for institutionally-based scoring teams was only moderate and absolute agreement was worse compared to the calibration teams scored patients using both instruments (notably higher levels of agreement) 5. Regression analysis found that after controlling for administrative factors, patient, and hospital characteristics, that a random effect for hospitals was significant	to the content of functional assessment measures and to the assessment techniques used. 5. Translation of scores between instrument may need quality monitoring and outcomes management and we should be cautious regarding our ability to substitute one instrument to another Relevant to Dissertation: 1. This study showed the translation effort between AM-PAC and the FIM™ failed to achieve sufficient accuracy for use in the planned payment system based on classical testing theory methods
Buchanan, J. L., Andres, P. L., Haley, S. M., Paddock, S. M., & Zaslavsky, A. M. (2004)	Evaluating the planned substitution of the Minimum Data Set-Post Acute Care (MDS-PAC) for use in the rehabilitation hospital	To assess agreement of PPS case-mix groups (CMGs) classifications using FIM™ and MDS-PAC translated "FIM™-like" items using classical testing theory methods	• Prospective cross-sectional design using consecutive sampling • All participants completed both the FIM™ and the MDS-PAC • Eighteen items from the MDS-	• Functional Independence Measure (FIM™) • Minimum Data Set-Post Acute Care (MDS-PAC) • All Medicare admissions with stays of 3 days or more over a 2-	1. The mean differences between the FIM™ motor and cognitive scales and MDS-PAC translations were 2.4 (mean =45) and 0.0 (mean=28), with scale correlations of .85 and .84 respectively. 2. Weighted kappas on	From the Article: The MDS-PAC should not be substituted for the FIM™ instrument in determining the rehabilitation hospital PPS due to poor payment cell agreement and substantial revenue shifts (even though with better item-level agreement than previous observed).

	prospective payment system (PPS)		PAC were combined and translated to "FIM™-like" items. • Hierarchical regression models were used to analyze motor score differences	month period • Fifty inpatient rehabilitation hospitals in 22 states • 2959 cases with both MDS-PAC and FIM™ data were analyzed	individual items ranged from .32 to .64 of motor and cognitive scales between FIM™ and MDS-PAC. 3. Substantial hospital-specific differences in scoring were found. 4. A 56% agreement of PPS CMGs classifications between FIM™ and MDS-PAC. 5. Around 20% of the facilities had revenue shifts larger than 10% of the original cost with large SD differences ($1,960), even though the mean payment difference between these two instruments was not significantly different from zero	Relevant to Dissertation: This study does not support using the translated scores between instruments (i.e., FIM™ and MDS-PAC) to decide payment classifications based on CTT methods
Leucht, S., Kane, J. M., Etschel, E., Kissling, W., Hamann, J., & Engel, R. R. (2006)	Linking the PANSS, BPRS, and CGI: clinical implication	The authors conducted previous study to examine the associations between the percentage BPRS/PANSS change from baseline and the CGI-improvement This study is to link the absolute change of the BPRS/PANSS to the CGI-improvement and CGI-severity scores using equipercentile	• Secondary data analysis; the same databases from the original clinical trial study (PANSS and PABPRS database, composed of seven randomized, double-blind trials that compared olanzapine or amisulpride with other antipsychotics or placebo) was used • The method used	• Brief Psychiatric Rating Scale (BPRS) • Positive and Negative Syndrome Scale (PANSS) • Clinical Global Impressions Ratings (CGI) • Patients data used in this study included who had a PANSS and a CGI rating at baseline so that they could be	1. Associations between various CGI and BPRS/ PANSS/PABPRS (PANSS-derived BPRS; PABPRS) scores for the whole sample at baseline and at weeks 1–6 ranged between 0.52 and 0.74, reflecting moderate to strong associations between scores 2. Replication of the linking functions 'CGI-severity score vs BPRS total score' and 'CGI-improvement score and	From the Article: 1. It is important to translate research results into practice and by translating the given scores among the Brief Psychiatric Rating Scale (BPRS) and the Positive and Negative Syndrome Scale (PANSS) to the Clinical Global Impressions Ratings (CGI) mean could facilitate clinical implications 2. Less severely ill patients required less BPRS/PANSS total score reduction to achieve the same CGI-improvement score than more severely ill patients, implying that the CGI-

		linking method (CTT-based method). The authors also replicated the previous analysis linking the BPRS with the CGI using PABPRS scores from their PANSS database Three goals of this study: (a) compared the absolute change of the BPRS/PANSS with the CGI - improvement score and the change of the CGI severity score, (b) analyzed whether the severity of illness at baseline had an impact on the latter association, and (c) attempted to replicate previous BPRS findings using a completely different data set based upon the PANSS-derived BPRS	was equipercentile linking of BPRS and CGI ratings from 14 drug trials in acutely ill patients with schizophrenia (n=5970) • SAS program, EQUIPERCENTILE (the algorithms for equipercentile linking described by Kolen and Brennan in 1995), was used to compare the BPRS/PANSS with the CGI • All patients with valid values on both measures were analyzed • Spearman correlation coefficients were used for examine correlations between tests	included in at least one linking function • The trials included patients with schizophrenia, schizoaffective disorder, or schizophreniform disorder (according to DSM-III-R or DSM-IV). All patients had sufficient symptoms, and most of them had a minimum of positive symptoms	percentage BPRS reduction' using the PABPRS showed similar results reported previously for the original BPRS (Leucht et al, 2005). There was a time effect, with more percentage PABPRS reductions needed at later weeks to link with the same CGI-C score (expectation effects are a likely reason for these time effects) 3. Linking of the CGI-improvement score to the absolute change of the BPRS/PANSS/ PABPRS from baseline: An absolute reduction of the BPRS/PANSS by approximately 10/15 points corresponded to a CGI change of 'minimally improved' and to a change of the CGI severity score by one severity step 4. A percentage reduction of the BPRS/PABPRS by approximately 28 percentage points (range BPRS 26–30, PABPRS 27–30) reflects a reduction of the CGI-severity score by one severity step. The same number for	improvement score associates with the severity of symptoms at baseline 3. This effect of initial severity was attenuated using percentage rather than absolute BPRS/PANSS reduction scores. The linking analysis between the absolute BPRS/PANSS reduction and the CGI may have an implication for the interpretation of efficacy differences found in clinical trials, and for sample size estimations. Clinicians seem to base CGI ratings on relative change rather than on absolute change of symptoms Relevant to Dissertation: The authors used absolute BPRS/PANSS reduction scores instead of the percentage (despite they conducted previous researches using the percentages) for better comparing the scores from the BPRS/PANSS to the CGI-improvement and severity scores based on the equipercentile linking, a CTT-based method, for the purpose to translate scores between instruments for the use of clinical trial study. The major difference of this study from my dissertation is that this study focuses more on how to connect the changed score from one instrument to represent improvement change score of the other instrument, instead of simply testing if the comparable score translated between instruments are validate or not.

					the PANSS was 25 percentage points (range 24–28) 5. Linking analyses depending on the initial severity of illness: For less severely ill patients (≤ median of the BPRS/PANSS at baseline), a smaller change of the absolute BPRS/PANSS was associated with a certain degree of CGI-improvement than in the more severely ill patients	
Fong, T. G., Fearing, M. A., Jones, R. N., Shi, P., Marcantonio, E. R., Rudolph, J. L., Yang, F. M., Kiely, D. K., & Inouye, S. K. (2009)	Telephone interview for cognitive status: Creating a crosswalk with the Mini-Mental State Examination	To link comparable cut-point scores from a standard global cognitive function test to another additional tests using percentile equivalents equating (traditional CTT) methods	• A cross-sectional analysis of baseline data from a longitudinal study • Secondary data analysis • Direct comparisons of scores were performed using an equipercentile equating method • Equipercentile equating method scores from two different measures (i.e., ADAMS TICS-30 and MMSE,) may be considered equivalent to one another if their corresponding percentile ranks	• Mini-Mental State Examination (MMSE) • Telephone Interview for Cognitive Status (TICS): 30 items • Telephone Interview for Cognitive Status-Modified (TICS-M): 40 items • 746 community-dwelling elders who were participants in the Aging, Demographics, and Memory Study (ADAMS) (a random subsample age ≥ 70 years old)	1. The MMSE and TICS (also TICS-M) are highly correlated 2. The majority of the sample in this study was diagnosed as normal/nondemented (306; 41%), and 81 (11%) and 77 (10%) participants were diagnosed as having possible and probable Alzheimer's Disease (AD), respectively 3. The mean score on TICS-30 was 17 (SD=6; median= 18; range= 0–29), and the mean score on TICS-40 was 21 (SD=9; median= 22; range=0–39); while the mean	<u>From the Article:</u> 1. This study used equipercentile equating to develop a crosswalk between scores on MMSE and those on the ADAMS TICS-30 and TICS-40 successfully 2. This study provides cut points for the TICS that mirror these commonly accepted cut points of the MMSE, with which clinicians and researchers alike are familiar and comfortable <u>Relevant to Dissertation:</u> This study provides equivalent scores of cut points for cognitive impairment between a widely accepted standard tool (MMSE) and two different versions of another phone interview screening tool (TICS) based on equipercentile equating, a classical testing theory-based method, on a large, nationally representative sample. The purpose

			in any given group are equal • Because equipercentile equating leads to irregular score distributions when actual values are graphed; thus, a log-linear method was used to smooth the raw scores of MMSE and TICS, and create a regular distribution • In order to facilitate accurate calculation and interpretation of statistical estimates, the respondent-level sampling weights derived from the national population sample used in ADAMS were used • All analyses were conducted using SAS		score on MMSE was 23 (SD=6; median=24; range= 3–30) 4. The intraclass correlation coefficient for MMSE versus TICS-30 was 0.80 (95% confidence limits of 0.78 to 0.83); for TICS-40 was also 0.80 (95% confidence limits of 0.78 to 0.83) 5. For each cut-point category in MMSE, a correlation was calculated with the corresponding cut points for TICS-30 and TICS-40. This yielded weighted k-values of 0.69 for both, indicating substantial agreement exceeding chance. 6. The calculated correct classification for TICS-30 was 87.6%, and for TICS-40, 88.1%	of this study is to promote the utilization of TICS instead of MMSE due to its several limitations (i.e., rely heavily on verbal response, require reading and writing ability, remarkable ceiling effects in highly educated older adults, poor sensitivity to detect mild cognitive impairment when using MMSE). I would suggest the authors to conduct a validation study to use the cut scores in order to further support the possibility to use the TICS instead of MMSE. The limitation of using Equipercentile equating method may include the difficulty to translate the cut-off point to different populations besides the older adults with cognitive impairments.
Noonan, V. K., Cook, K. F., Bamer, A. M., Choi, S. W., Kim, J., & Amtmann, D. (2012)	Measuring fatigue in persons with multiple sclerosis: creating a crosswalk between the Modified Fatigue Impact	To (a) identify an appropriate linking method; (b) create cross-walk tables to associate scores for the Modified Fatigue Impact Scale (MFIS) with scores for the Patient Reported Outcome	• Prospective study (part of a longitudinal study) by sending letters • Single-group linking design (the same person completed both the MFIS and PROMIS Fatigue	• Modified Fatigue Impact Scale (MFIS) • Patient Reported Outcome Measurement Information System (PROMIS) Fatigue Short Form (SF)	1. Correlations between deviations and fatigue level for the PROMIS Fatigue SF and MFIS were -0.31 and -0.30, respectively, indicating moderately greater deviations with lower fatigue scores. That is, the cross-walks are	<u>From the Article:</u> 1. The cross-walk tables developed in this study enable to link and compare scores between the MFIS and PROMIS Fatigue SF 2. When sample sizes are 150 or greater, scores of the MFIS and PROMIS Fatigue SF can be cross-walked with relatively small estimation error in sample

	Scale and the PROMIS Fatigue Short Form	Measurement Information System (PROMIS) Fatigue Short Form (SF); and (c) validate the linking results at a follow-up time point in persons with Multiple Sclerosis (MS)	SF at both time points) • Cross-walk tables were created using equipercentile linking (a method to identify pairs of raw scores that corresponded to the same percentile rank) and allows data from studies using different measures of fatigue to be combined to achieve larger sample sizes and compare their results [this is a traditional linking method] • Deviations between estimates and actual scores were compared across levels of fatigue • The impact of sample size on the precision of sample mean estimates was evaluated using bootstrapping (a method of random sampling with replacement) • Five participants with missing item responses were removed from the sample (list-wise	• Survey invitation mails were sent to 7,806 persons from the NMSS National Multiple Sclerosis Society (NMSS) mailing list (eligibility criteria: over age of 18 and self-reported having been diagnosed with MS by a physician) • 1597 (of the 1629) respondents were eligible and were mailed a paper survey • 1,271 subjects in the first survey (Time 1) and a random subset of 562 subjects was invited to participate in the longitudinal study that required completing five additional surveys, with approximately 4 months between the repeated administrations. For the current study, data from the fifth and sixth time points were used • Data collected at first time point (5th;	more accurate at higher than at lower levels of fatigue 2. Estimated sample means were impacted by sample size; with larger sample sizes, the impact of deviations in individual scores may average out, but with smaller sample sizes, the cross-walking tables are less likely to closely approximate sample mean scores 3. Scores for the MFIS and PROMIS Fatigue SF in the cross-validation data were very similar to those in the linking data 4. For group-level analyses, with larger sample sizes, estimates of sample means were much less variable, especially with sample sizes of 150 or greater	mean estimates; on the other words, the cross-walk tables are not suitable for use at the individual level or with small samples 3. Cross-walking will allow data from studies to be combined to examine effectiveness of MS intervention studies and will support meta-analytic studies 4. Though the linking function successfully associated scores from the two instruments, cross-walked scores are not equivalent and should not be considered interchangeable Relevant to Dissertation: 1. The authors used the same sample to validate the linking and suggested a stronger design of cross-validate is to use an independent sample, however, my dissertation study design will also use the same sample for developing linking and validating the linking, so is not a stronger design 2. The results of this study positively supported the linking results between two instruments under some certain linking conditions: (a) determine the most appropriate linking strategy based on data characteristics (i.e., similarity of constructs measured, strength of the empirical relationship between the scores, and invariance of scores across sub-populations); and (b) sample size is larger than

			deletion) during the creation of the cross-walk tables and for the cross-validation • Quantile–Quantile plots to show score distribution	linking data) in a longitudinal study of persons with MS (N = 458). Validation of the tables was conducted using data collected at a subsequent time point (N = 444) (6th ; cross-validation data)		150 3. The authors used the traditional procedures (e.g., equipercentile)

Table 2.2. Literature Reviews of Linking Methods Used in Healthcare Professions (Item Response Theory) (ordered by year) (n=25)

Author	Title	Aims	Methods	Instruments/ Population	Results	Conclusions
Fisher, Harvey, Taylor, Kilgore, & Kelly, (1995)	Rehabits: A common language of functional assessment	To develop a single rehabilitation-measuring unit, *rehabit*, by co-calibrating motor scales from 2 instruments	• Prospective study • Two steps of cocalibration: (a) analyzed the motor skills items from the two instruments together; (b) compared the theoretically common-unit measures from the two instruments • Rasch partial credit model	• Functional Independence Measure (FIM™) • Patient Evaluation and Conference System (PECS) • 54 participants with 5 physical disability diagnoses (brain injuries, neuromuscular, musculoskeletal, spinal cord and stroke), to increase variations of the sample	1. The authors found common 9 motor skills items measured by both the FIM™ and PECS with similar item calibration order supported by Silverstein and colleagues (1989). These nine items included feeding/eating, upper extremity (UE) bathing, UE dress, lower extremity (LE) bathing, LE dress, toilet, transfer, walk and stairs. 2. The easiest item is "feeding/ eating" and the most difficult item is "stairs/environment barriers." 3. In general, upper extremity functions are easier than lower extremity functions 4. the persons measured are spread along 5. The measurement continuum with a reliability of 0.95, meaning that the 35 FIM™/PECS items have distinguished six statistically distinct levels of functional independence (strata) in the persons' abilities	From the Article: The results demonstrate that item difficulty estimates of the FIM™ and PECS are stable sufficiently to support the use of the common "function metric" unit: rehabit. Relevant to Dissertation: This study supports the concept of developing a common metric measuring physical self-care activities between instruments (i.e., FIM™ and PECS) based on IRT methods

					6. The FIM™ items range across 27.3 rehabits, from 32.6 to 59.9 (error = 1.2), and the PECS ranges across 34.2 rehabits, from 34.2 to 68.4 (error = 1.3), a difference of 6.9 rehabits, or more than 5 times the average item error 7. The two calibrations correlate 0.89, with an R^2 of 0.79, which supports the contention that the same construct is being measured in both samples	
Gonin, R., Lloyd, S., Cella, D., & Gray, G. (1996)	Establishing equivalence between scaled measures of quality of life	This is a very early study aims to demonstrate (a) how equivalence of QOL across different measures can be established (b) how to link two QOL instruments based on equivalent linear relationship	• Secondary data analysis • Used IRT to generate logit scores and then used CTT to compare the equivalence • Patients completed both the FACT and FLIC in the same sitting • Raw scores from instruments will be transformed into linear measures using the Andrich rating scale model • All the logit calculations were done using the BIGSTEPS scaling	• Functional Assessment of Cancer Therapy (FACT; n=7); general version • Functional Living Index Cancer (FLIC; n=27) • 447 patients (both inpatient and outpatients) with cancer (heterogeneous with respect to type and stage of cancer)	1. The Pearson correlation coefficients between the raw and corresponding logit measures were 0.91 and 0.86 for the FACT and FLIC score respectively. The correlation coefficient between the two logit measures was 0.74 2. Only 15 data points out of 447 (3%) fall outside of the control lines. The correlation between the differences and means is r = 0.086 (p = 0.071) which is essentially zero, indicating no association between the differences and the size of the measurements 3. Estimates and standard	<u>From the Article:</u> 1. The authors demonstrated systematic methodology to provide comparability and compatibility of two commonly-used QOL instruments using standard QOL scores as a way to translate raw scores between instruments <u>Relevant to Dissertation:</u> 1. This is the first study that used IRT method, Andrich rating scale model, to transform the raw scores across instruments to the linear scores in oncological area measuring quality of life 2. This study used a linear conversion method to translate the FLIC logit measures to equivalent FACT logit measures

			program • The remarkable property of this model is that the patient QOL and item position on the QOL dimension can be estimated independently by means of conditional maximum likelihood estimation • Denote the ith measurements on the two scales by Xi1 and Xi2 respectively. The quantities (Xi1 - Xi2) are plotted against (Xi1 + Xi2)/2. This plot is then examined for any tendency for the amount of variation (Xi1 – Xi2) to change with the magnitude of the measurements (Xi1 + Xi2)/2 • In the event of no association (zero trend implying zero bias) a paired t-test or non-parametric Wilcoxon signed		errors for the slope and intercept were 0.26193 (SD= 0.0437) and 0.92431 (SD=0.0525) respectively 4. Only two subjects out of 447 had FACT or FLIC scores high enough to be truncated at a QOL of 100 and no subjects had scores even close to the low ends of the scales	

			rank test can be performed on these differences • Equating logit measures using orthogonal linear least squares regression to convert the FLIC logit scale values into equivalent FACT logit scale values			
Chang, C. H., & Cella, D. (1997)	Equating health-related quality of life instruments in applied oncology settings	To demonstrate how equivalence of quality of life (QL) across different measures could be established and to develop a standard metric (called Q-score) for five commonly used quality of life measures	• Prospective study • Rasch rating scale analysis • BIGSTEPS computer program • Five separate Rasch analyses were conducted to obtain Rasch statistics results • Patients' QL measures for each instrument were then estimated using anchored item difficulties and step difficulties obtained from the simultaneous calibration • Cronbach's alpha reliability coefficients for subscales of the five instruments were examined	• Cancer Rehabilitation Evaluation System-Short Forms (CARES-SF) • European Organization for Research and Treatment of Cancer Quality of Life Questionnaire-Core (EORTC QLQ-C30) • Functional Assessment of Cancer Therapy (FACT) Scales • Spitzer's Quality of Life Index (QLI) • RAND 36-Item Health Survey 1.0 (known as SF-36)	1. Total item number = 140 2. The ranges of the internal consistency coefficients of the subscales are 0.69-0.87, 0.64-0.90, 0.73-0.88, 0.68, 0.77-0.93 for the CARES-SF, EORTC QLQ C-30, FACT, QLI, and RAND-36, respectively 3. Person separation statistics indicate that these five QL instruments are moderately comparable, with one exception: QLI has the lowest person separation statistic (0.48) 4. Item reliabilities are quite similar, except for QLI RAND-36 has the highest item separation statistic (15.86), followed by the EORTC (12.78) and FACT-G	From the Article: 1. The five test ogives demonstrate that each instrument retains different degrees of precision in relation to corresponding test-centered logits 2. This study demonstrates the compatibility of five commonly used QOL measures Relevant to Dissertation: 1. This is an extended study of previous Gonin and colleague's (1996) study by increasing the linked QOL instruments from two to five based on similar linking methodologies; both results supported the development and feasibility of linking tools using IRT-based methods 2. Test precision could be an important criterion for selecting appropriate QL instruments

			(because acceptable internal consistency is an important prerequisite for establishing equivalence)	• Data were collected from five different performance sites located in hospital settings in different parts of the country • Eligibility criteria include a diagnosis of cancer of all types or HIV, a period of at least 2 months after diagnosis of any particular cancer or HIV infection, a life expectancy of at least 3 months, and sufficient fluency in English to complete forms	(12.26) 5. The slopes of the CARES and FACT are deeper than those of EORTC, RAND-36, and QLI, particularly in the regions between -1and1.5 logits, meaning that these two instruments have better precision in measuring the QL continuum in that range	
Fisher, W. P. Jr. (1997)	Physical disability construct convergence across instruments: Towards a universal metric	The purpose is to indicate whether formal equating of instruments calibrations would be likely to succeed	• Retrospective study • Rasch measurement model • Pseudo-common item equating methods to calibrated items • Four instruments provided data from ten reviewed articles presenting Rasch analyses of	• Functional Independence Measure (FIM™) • Katz AOL Index (Katz) • Levels of Rehabilitation Scale - III (LORS) • Patient Evaluation and Conference	1. The 21 original correlations among the LORS, two PECS, FIM™WPECS, FIM™RST, FIM™LIN, and the FIM™LRI with seven pseudo-common items was .92 on average (an average p= .02). 2. Measures based on these calibrations should be linearly transformed on the same metric with the	From the Article: 1. The results supported that physical functioning construct is stable and can be treated as one construct across the instruments and samples. 2. Measures from different instruments could be linearly transformed based on the calibrations. 3. The results supported the concept that quality and stability of psychosocial measures are not noticeably less consistent than results from the physical sciences

			physical functioning scales • The item orders were examined by correlation coefficients and scatter plots of the 7 pseudo-common item values • For each pair of calibrations, items lying outside bounds of the 95% confidence intervals were omitted • To be conservative, the authors used under estimate error and over-estimate reliability to avoid inflate correlations (because of removing more error from them than actually exists)	System (PECS) • Sample sizes range from 53 to almost 30,000 subjects across studies (Note: different instruments on different samples)	final overall average correlation for error is .93 (with an average of 7 pseudo-common items), and p-value on average is .01. 3. After removing values outside 95% confidence intervals, 53 (96%) of the 55 correlations over .80, and 43 (78%) over .87. The average correlation for all 55 pairs increases to .91, with an average of seven pseudo-common items, and an average p-value of .01.	Relevant to Dissertation: This study used existing literatures to validate and supports the concept of a universal metric measuring physical functioning among instruments (i.e., FIM™, Katz, LORS and PECS) based on IRT methods
Fisher, Eubanks, & Marier (1997)	Equating the MOS SF36 and the LSU HSI physical functioning scales	The purpose is to equate the physical functioning subscales of two instruments (SF36 and LSU HIS)	• Prospective study • Rasch rating scale model was used to create a common metric • Graphical display and correlation calculation were used to evaluate the relationship of two instruments • BIGSTEPS	• Medical Outcomes Study Short Form 36 (SF36)- the physical functioning scale (PF10) • Louisiana State University Health Status Instruments	1. The PF10 had 86% greater calibration error, and 175% greater measurement error. 2. Eight-two cases with the highest outfit were removed from analysis, reducing the sample size to 153. 3. Data from the SF36's 10-item physical functioning scale, the	From the Article: 1. This study highlights the demand and importance of sample-free and scale-free measurement to fulfill the needs for accountability, outcome comparability, and a consumer-oriented focus increasing in health care. 2. The PF10 best person separation reliability of .90 is identical with that obtained for the same set of items in the reference data set published in the

			program (a Rasch calibration program for two-facet data) • SF36 and the LSU HIS were first analyzed separately and then co-calibrated into the same item pool by Rasch analyses • The cases with highest outfit statistics (least consistent) were removed over the course of several subsequent analyses, until there was no further improvement in person separation reliability	(LSU HSI) Physical Functioning Scale (PFS) • The PF10 has only 3 rating categories, where the PFS has 6 categories • 285 convenient sample (patients waiting for appointments in a public hospital general medicine clinic)	PF10, and the LSU HSI's 29-PFS-item, were fit to separate and co-calibrated Rasch rating scale models. 4. The paired-sample t-test between the PFS and the PF10 is .95 (p=.34) with the PFS mean and standard deviation (SD) at .27 and 2.2, and the PF10 mean and SD at .14 and 2.5. 5. The PFS had lower error, better model fit, and higher reliability coefficients than the LSU HSI. 6. Eight of the PF 10 items have corresponding items in the PFS addressing similar areas of physical functioning 7. The difficulty estimates for the items from both the separate and combined analyses of the different instruments correlate at .95, indicating that the items from the two scales measure the same variable. 8. The person separation reliability of initial PF10 is 0.80 (after removing errors becoming 0.90) and for the PFS is 0.95 (after removing errors becoming 0.97); for	HSQ 2.0 manual (Fisher, et al., 1995). 3. Since the items do not represent identical areas of physical functioning, this result does not deny the possibility of equating the two instruments, but does present an opportunity for understanding more about the effects of the instruments' differing numbers of rating categories and items. 4. Both instruments measure physical functioning; implying that common unit of measurement is feasible. <u>Relevant to Dissertation:</u> 1. This study used similar co-calibration methods (Rating scale methods) to validate and supports the concept of developing a universal metric measuring physical functioning using the same quantitative unit between two instruments (i.e., MOS SF36 and LSU HSI). 2. This study compared differences between with and excluding errors in the model fit

					combined two instruments, the person separation reliability became 0.98 9. The items' difficulty estimates of the items from the separate and combined calibrations are not statistically identical.	
McHorney, C. A., & Cohen, A. S. (2000)	Equating heath status measures with item response theory: Illustrations with functional status items	To develop an item bank of physical functioning items and equated them using item response theory	• Prospective study • Common-item equating design (anchor test) • A self-administered survey of functional status • Two mailing survey (first one has 61% response and second one has 58% response rate), with n=3358 total mailing surveys • The graded response model (a 2-parameter IRT model, assuming that (a) item discrimination is not equal across all items, (b) differences between each of the response categories are not the same across all items, and (c) that all categories in an item are ordered)	• A total of 162 published articles, books, and book chapters that focused on the measurement of physical functioning and functional status were obtained as the items in the item bank • Individuals>65 years of age who had >1 ambulatory visit across a 3-month sampling frame to a Veterans Administration Medical Center or its affiliated university medical center • Patients were sampled from the outpatient ambulatory	1. The average age was 75.5 years, and consistent with a 75% of the sample was male 2. Principal components analyses conducted separately for the 71 common items on each of the 3 forms and on the 3 forms combined revealed a first factor that accounted for .40% of the variance; and the magnitude of the first eigenvalue to the second was large (>7.0) 3. The 5 most discriminating items were to put underclothes on, manage clothes after toileting, move between rooms, take pants/ slacks off, and get into bed. Most of the items were located on the easier end of the ability continuum. Six items were classified as being very difficult 4. A total of 28 items were detected as DIF	From the Article: 1. Item response theory could equate and calibrate a large number of activities of daily living items on the same scale; which could be further expanded to generic, disease specific or mixed item banks; as well as linking different age-specific functional measures 2. Co-calibrating items can better understand the structure and order of domain-specific items across scales, and also the interrelations among items across the ability continuum Relevant to Dissertation: 1. This article mentioned potential/additional important concerns in terms of measuring patients' function such as the comprehensibility of the item for the elder population. In fact, the strategies and resources used to perform activity may play more important roles compared to simply measuring the "difficulty" level. 2. Combined with Fisher and colleagues' (1995) study and McHorney and Cohen's (2000) study, feeding and eating is the easiest item and

			was use for equating • Concurrent calibration (estimating item and ability parameters in both the base group and the target group simultaneously) was used using MULTILOG • DIF detection for the graded response model with the likelihood ratio test has been found to provide control over type I errors when 1 item at a time is compared between 2 groups • The non-DIF items were used to anchor the subsequent concurrent calibration run. Then, the values of the common items were fixed and used to anchor the calibration of the parameters for the unique items on each form.	clinics of the Madison, Wisconsin, Veterans Administration Medical Center (VAMC) and the University of Wisconsin Hospitals and Clinics (UWHC) • In total, 1,588 items were banked; elimination of redundancies resulted in 206 candidate items for Health of Seniors Survey. The 206 items cover 7 domains of function (toileting, bathing, cooking/ eating, dressing, mobility, household and community activities, and recreation); 3 forms of the survey were created. Sample members were randomly assigned to 1 of	5. 60 items were constrained between forms 1 and 2 and 54 items between forms 1 and 3 after removing DIF items 6. About two thirds of the items provided maximum information at or below theta=0, meaning most of the items were located on the easier end of the ability continuum 7. Only 6 items had locations < -1.50 and thus would classify as being very difficult 8. The dressing items were the most discriminating (across domains) and toileting is the least discriminating item 9. Bathing, dressing, and mobility items provide the most information	Stairs is the most difficult item; while dressing, bathing and mobility are the most discriminating items

				the 3 forms. These forms were equated with the use of IRT • 39 unique items (across all domains) and 89 common items		
Orlando, M., Sherboune, C. D., & Thissen, D. (2000)	Summed-score linking using item response theory : Application to depression measurement	To calibrate a modified scale (added 10 new items) to the standard scale based on the item response theory (IRT) summed scores approach	• Secondary data analysis with longitudinal study design • IRT summed scores approach; the 2 scales were linked on the basis of derived summed-score-to-IRT-score translation tables • MULTILOG was used for calibrating 30 items • Samejima's (1969, 1997) graded IRT model was used (because of the ordered nature of the CES-D item responses) to calibrate the original and modified scales • A recursive algorithm that builds the joint likelihood for each score group item	• A modified 23-item version of the Center for Epidemiologic Studies Depression Scale (CES-D) • The standard 20-item CES-D • Data are from the Depression Patient Outcomes Research Team, II, which used a modified CES-D to measure risk for depression. • 1,120 participants responded to items on both the original and modified versions (total 30 items because of redundancy of two scales)	1. The first eigenvalue (13.4) was substantially greater than the next four (1.6, 1.5, 1.4, 1.1). In addition, 29 of the 30 items had standardized factor loadings greater than .35, (ranging from .28 to .81, with an average of .65), indicating that the factor structure of the 30 items is sufficiently unidimensional for application of IRT 2. The authors also examined the validity of the cut score generated from the sum-score translation method by comparing the classification rates of respondents at the 18-month wave as depressed using both the 20 CES-D items (cut score of 16) and the 23-item scale (corresponding cut score of 20); and the result	<u>From the Article:</u> 1. The IRT summed-score is a straightforward and valid linking approach that can be applied in a variety of situations, such as questionnaires of various lengths, dichotomous, Likert-type, or combinations of response formats as long as the scales measure the same construct and there is some degree of item overlapping <u>Relevant to Dissertation:</u> 1. This study used IRT summed-score linking approach to translate scores from the original scale to the new one (with 10 newly added items) and found the classification rates of identifying patients as depression had 95% agreement at the 18-month wave, indicating this linking method can be successfully used for translate comparable scores between the original and revised scales; also, this summed-score IRT linking method can be applied to different response formats such as dichotomous or the Likert-type scales

			by item was used; this algorithm has effectively collected all the pattern likelihoods corresponding to each summed score to form a joint likelihood • The average (or EAP) value as the IRT score associated with that summed score were calculated • Before linking the two scales, similarity were examine with: (a) correlation with 36-Item Short-Form Health Survey (SF-36) (b) the correlation between the 20-item scale and the 23-item scale at 18 months, as well as the correlation between the non-overlapping items on the two scales • A principal components analysis was used to establish the unidimensionality of the 30 items • The item	• Both scales use a four-category, Likert-type response scale in which participants are asked to indicate the extent to which they have experienced the feeling or condition expressed in the item stem during the past week	showed nearly 95% of the sample are classified in the same way regardless of which criterion was used 3. The established cut score of 16 on the standard CES-D corresponded most closely to a summed score of 20 on the modified version 4. The cut score of 20 demonstrated acceptable concordance rates with the Composite International Diagnostic Interview (CIDI) at two time points (baseline and 24-month) 5. The sensitivity (the probability of screening positive given that the diagnosis is present) of the 23-item scale is slightly higher and the specificity is lower than the summary measures of the CES-D reported by Mulrow et al. (1995)	

			parameters from the 20-item scale and from the 23-item scale were separately input into the program SS_IRT[2] to estimate the IRT score corresponding to each summed score for each scale			
McHorney, C. A. (2002)	Use of item response theory to link 3 modules of functional status items from the Asset and Health Dynamics among the Oldest Old Study	To link three modules of functional status items in the Asset and Health Dynamics Among the Oldest Old (AHEAD) study by using item response theory (IRT)	• Secondary data analysis • Participants completed 16 common functional status items in the AHEAD study, and were randomly assigned to complete 1 of 2 modules containing different functional status items • A 2-parameter (PM) model for dichotomous items (common-item design) was used to link the 3 modules of items between LSOA and NLTCS • The 16 common items were distributed as 6 ADLs, 4 higher order ADLs, and 6 IADLs • Used the marginal	• US baseline data from 4655 respondents (a nationally representative panel study of elderly; all participants are 70 years old or older) • The first (baseline) wave of AHEAD data was collected in 1993 (N= 8223, 80% response rate) • Two of the modules contained functional status items from the Longitudinal Study on Aging (LSOA) and National Long-Term Care Survey	1. Disability in doing basic ADLs ranged from 1.8% to 9.1% 2. The 6 common ADL items had a single dominant dimension, accounting for 48% of the variation 3. The 6 ADLs from the LSOA, the first eigenvalue was 2.30 and accounted for 38% of the variance 4. The first eigenvalue for the 9 ADL items from the NLTCS was 4.06, accounting for 45% of the variance 5. Higher-order ADLs (n=4) and the IADLs (n=6) both lacked of unidimensionality so were not used for linking in this study; additional 13 items were added to common item bank (n=6) and principal components analysis	<u>From the Article:</u> 1. Both sets of supplemental items were successfully linked to the common items, allowing the placement of all items on the same underlying measure of ability 2. IRT-based linking methods were a useful way to overcome test dependency and place items on a common metric even if different respondents answer different sets of items 3. Numerous important design features can degrade linking results and should be restricted in the future linking studies <u>Relevant to Dissertation:</u> The authors used 2-parameter model to link the instruments based on the reason that 2-P model fits the data better compared to the 1-P model, which allows item difficulty and item discrimination to be different

			maximum likelihood estimation (MMLE) algorithm with concurrent calibration by MULTILOG	(NLTCS) • The base sample was limited to respondents who had at least 1 disability on the 16 common items (5368/8223 respondents)	showed these 19 items had a single underlying dimension 6. The 2-P fit data better than the 1-P, so was chosen as the IRT model for the linking 7. Three items were identified as functioning differentially between the base and NLTCS samples 8. Most of the 19 items are at the easy end of the functioning continuum 9. The items on toileting were among the most discriminating item for groups with different abilities	
Bjorner, J. B., Kosinski, M., Ware, J. E. Jr, (2003)	Using item response theory to calibrate the Headache Impact Test (HIT) to the metric of traditional headache scales	To develop and assess the calibration of IRT-based scores on the Headache Impact Test (HIT) into the metrics of the traditional headache scales; and also to examine if the calibrated HIT scores can lead to the same conclusions in group comparisons	• Secondary data analysis • For each of the traditional scales, agreement between calibrated HIT scores and observed scores were assessed by intraclass correlation (ICC) and the agreement of mean scores and the relative validity (RV) in discriminating among groups differing in migraine diagnosis, headache severity,	• Headache Impact Test (HIT) • Migraine Specific Questionnaire (MSQ) • Headache Disability Inventory (HDI) • Headache Impact Questionnaire (HIMQ) • Migraine Disability Assessment Score (MIDAS) • Telephone interview data	1. ICC's of calibrated HIT and the observed traditional scores were between 0.80 and 0.94 2. In RV analyses, the maximum mean difference between the observed and expected scores was 1.7 points on a 0–100 scale for comparisons at one point in time 3. ICC's were between 0.56 and 0.61 and the maximum mean differences were 2.9 (on a 0–270 scale) and 3.8 (on a 0– 450 scale) in RV analyses at one point in time	<u>From the Article:</u> 1. The high agreement between the calibrated IRT scores and the traditional sum scale scores is noteworthy for group comparisons 2. Analyses of change over time and analyses calibrating scores from the fixed-form HIT-6 to the metric of other questionnaires showed satisfactory but less precise results 3. The ability of the calibrated scale scores to discriminate between groups was at least as good as the ability of the observed sum scales and often remarkably better 4. The theoretical advantage of IRT models in scale calibration is supported by the study results 5. This study supported the IRT approach to achieve comparability of

			and change in impact over time were evaluated • For test of responsiveness (ability to detect change over time) the follow-up interviews were completed after three months for initial respondents sampled randomly from mild, moderate and severe strata • A generalized partial credit model (GPCM) was used with the Parscale and Multilog software • One HIMQ item had 11 response categories and the Graded Response Model (GRM) was used • Model-based approach	(n=1016) and Internet data (n=1103) from general population surveys of recent headache sufferers • 300 (out of f 365) completed the entire interview (105 with mild headaches, 113 with moderately severe headaches and 82 with severe headaches) completed the follow-up interviews	4. The HIMQ item had the lowest threshold of 2.35 5. There is more variation in the slope parameters than in the binomial model, but still the MIDAS items are more discriminant than the HIMQ item(s) 6. The GRM gives slightly higher expected values for HIT scores in the middle range (around 50), while for high HIT scores, the binomial model has far higher expected scores than the GRM 7. The largest difference between observed and graded response calibrated scores were 15% of the difference between minor and moderate headache sufferers 8. The MIDAS and the HIMQ, the agreement between calibrated and observed scores was less good because of different item types 9. Although there are some individual differences, the MSQ and the HDI scales seem to follow the same overall pattern and show most variation in the range of HIT scores from 40 to 80, indicating	new and widely-used scales 6. This study supported the implications for the applications of IRT based scoring methods in health outcomes research, because it can make 'backwards compatibility' for the IRT scores feasible 7. Overall, the calibrated HIT-6 scores did slightly worse than the calibrated total IRT scores. IRT scoring of the HIT-6 gave better calibrations in terms of mean scores for groups, but agreement in terms of ICCs were similar for the standard HIT-6 scoring and IRT scoring Relevant to Dissertation: This study linked one sum score scale to another sum score scale using the approach of calculating the expected IRT score for a given sum score, which supported the "backward" score translations from logit scores to the raw scores.

					that the scales are sensitive to roughly the same levels of headache impact	
Smith & Taylor (2004)	Equating rehabilitation outcome scales: Developing common metrics	Replication of the Fisher et al. (1995) study by comparing interval measures from two instruments measuring the same underlying construct, but with different wording of the items and different rating categories (FIM™ and PECS)	• Prospective study • Rasch partial credit model to co-calibrate items • BIGSTEPS program (a Rasch calibration program for two-facet data) • To assess the relative equivalence of corresponding PECS and FIM items, the Expected Score Maps were compared for pairs of items	• Functional Independence Measure (FIM™); 14 Motor items with 7 point rating scale • Patient Evaluation and Conference System (PECS); 20 Motor skills items with 7 point rating scale • 500 patients on admission and at discharge to a free-standing rehab hospital in 1998 (five diagnostic: brain injuries, neuromuscular, musculoskeletal, spinal cord injuries and stroke)	1. The average measure of 44.9 suggests that the PECS items as a whole are harder than the PIM items. 2. The mean of the standardized INFIT and OUTFIT item statistics, -2.1 and -1.8 (expected value of 0.0) suggests an extreme negative skew in the distribution of item fit statistics. 3. Seventeen of the 34 items have standardized OUTFIT values less than -3.0, while nine of the items have values greater than 3.0; the final ends in 6 most misfitting items 4. Person correlation of person measures between the PECS and FIM™ is 0.92 (without counting measurement error) 5. Four category (standby assistance) on the PECS and the 5 category (supervision and set-up) on the FIM. These two categories have approximately the same definition and represent the last step before achieving some form of	From the Article: 1. The results suggested a common equal-interval translation between the PECS and FIM™ could be constructed, even when instruments had different rating scales and different number of items. 2. Measures on the common metric can be based to either scale and are independent of the number of items completed. 3. The results implied the use of anchored scales could allow institutions using either the PECS or FIM™ to make direct comparisons of clinical outcomes with other institutions. Relevant to Dissertation: This study supported the perspective that developing a common metric outcome measure could allow hospitals and consumers to compare outcomes from different locations without imposing a single measurement scale on all institutions and programs. As well as to improve the measurement quality of the data and reducing administration burden of the clinicians and researchers

					independence	
Carmody, T. J., Rush, A. J., Bernstein, I., Warden, D., Brannan, S., Burnham, D., Woo, A., & Trivedi, M. H. (2006)	The Montgomery Asberg and the Hamilton ratings of depression: A comparison of measures	To provide both CTT and IRT results on two distinctly different depressed outpatient samples, and also provide an empirical basis for converting one scale total score into another scale total score; also, the item response pattern and the psychometric features were compared for all three depressive instruments	• Secondary data analysis • Classical test theory (CTT) examined consistency: Cronbach's alpha and item-total correlations (not corrected for item/total overlap), for the HRSD17, MADRS, and HRSD6 • Effect sizes were computed for each total score and item for each measure within each study • Samejima's graded IRT model (Samejima, 1997) based on Orlando et al. (2000)'s procedures was used to equate total scores for each pair of scales; item parameters were estimated for each item of each measure • The graded IRT model was also used to compute the test information function (TIF) for each scale in each study	• Hamilton Rating Scale for Depression (HRSD17; n=17) • Hamilton Rating Scale for Depression (HRSD6; n=6) • Montgomery Asberg Depression Rating Scale (MADRS; n=10) • Two datasets were analyzed for this study (a) The first sample (n =233) generated from a 12-month uncontrolled, long-term study of adult outpatients (18-75 years old) with highly treatment-resistant, nonpsychotic major depressive episodes (MDEs) who participated in a study of adjunctive vagus nerve	1. In Study 1, the correlation between the HRSD17 and HRSD6 total scores was 0.89; between the HRSD17 and MADRS was 0.88, and between the HRSD6 and MADRS was 0.86. In Study 2, all the correlations were slightly higher: HRSD17 vs. HRSD6 was 0.94, HRSD17 vs. MADRS was 0.92, and HRSD6 vs. MADRS was 0.91. 2. Internal consistency: For the HRSD17, the Cronbach's alpha values were 0.81 (Study 1) and 0.88 (Study 2). For the MADRS, values were slightly higher: 0.90 (Study 1) and 0.92 (Study 2). Finally, for the HRSD6, the values were 0.78 (Study 1) and 0.86 (Study 2) 3. Item-total correlation: Most items on the MADRS correlated with the total score at ≥ 0.60 (both studies); median item-total correlations were 0.75 (Study 1) and 0.78 (Study 2) for the MADRS. For the HRSD17 median item total correlations were lower (0.50 for Study 1	From the Article: 1. All three measures were highly correlated with each other and Cronbach's alpha showed highly acceptable internal consistency for all measures 2. Both the MADRS and the HRSD6 were unidimensional; and the HRSD17 had two factors 3. All MADRS items had acceptable effect sizes, and were therefore sensitive to change over time 4. These results support the conclusion that the MADRS is preferred over the HRSD17 in measuring depression severity and change in depression severity over time given its unifactorial structure, the high and consistent relationship between items and the measured concept of depression (by IRT) or to total score (by CTT), and its greater precision Relevant to Dissertation: I would question about the linking between MADRS vs. HRSD17 and the HRSD6 vs. HRSD17 because HRSD17 was not unidimensional; and I think the authors should examine the validation of both cross tables.

			• The principal components factor analysis was conducted to assess the dimensionality of each measure • Parallel analysis was used to infer how many real factors/dimensions were present by comparing the eigenvalues from a principal components analysis (PCA) of the real data to eigenvalues that might be expected to arise by chance alone; the number of principal components for which the real eigenvalues exceed the simulated eigenvalues defines the dimensionality • A series of simulated datasets consisting of random numbers (where correlations between all variables are zero) using the same number of observations and variables (items) as the real data;	stimulation added onto ongoing diverse medication regimens; (b) The second sample (n =985) included only outpatients with nonpsychotic major depressive disorder (MDD) defined by DSMI-V	and 0.56 for Study 2) 4. For the HRSD17 in Study 1, two factors were identified using parallel analysis to determine the number of factors. The average of the first three eigenvalues from the simulated datasets were 1.50, 1.39, and 1.31, which were compared to the first 3 real eigenvalues of 4.33, 1.73, and 1.19 5. The HRSD17 in Study 2 also revealed two factors based on the comparison of the first 3 simulated data eigenvalues of 1.23, 1.19, and 1.15 to real data eigenvalues of 5.77, 1.30, and 1.11 6. For the MADRS, only one factor was identified for Study 1 because the first real eigenvalue of 5.41 was much larger than the first simulated eigenvalue of 1.33, while the second real eigenvalue of 1.06 was smaller than the second simulated eigenvalue of 1.23 7. The MADRS was about 2 times as precise as the HRSD17 8. The more treatment-resistant sample (Study	

			eigenvalues of the principal components for each simulated dataset are computed and averaged over replications		1) had lower overall item and total score effect sizes with each of the three measures 9. An HRDS17 total of 20 approximated a MADRS of 27, which were comparable to those reported by Hawley et al. (1998)'s recommendations based on a regression analysis	
Holzner, B., Bode, R. K., Hahn, E. A., Cella, D., Kopp, M., Sperner-Unterweger, B., & Kemmler, G. (2006)	Equating EORTC QLQ-C30 and FACT-G scores and its use in oncological research	To examine the equivalence of the European Organization for Research and Treatment of Cancer Core Questionnaire (EORTC QLQ-C30) and the Functional Assessment for Cancer Therapy – General (FACT-G) on the basis of corresponding subscales, and where appropriate to derive a scheme for converting QLQ-C30 scores into FACT-G scores and vice versa for use in oncological research in Germany	• Prospective study • Applied both classical test theory and the Rasch measurement model • Correlation analysis (Pearson r) was performed to check if corresponding subscales of the two instruments measure the same construct • The internal consistency (Cronbach's alpha) of the subscales served as an approximate upper limit for the correlation r of corresponding subscales and thus as a criterion for assessing agreement of	• European Organization for Research and Treatment of Cancer Core Questionnaire (EORTC QLQ-C30) • Functional Assessment for Cancer Therapy – General (FACT-G) • A calibration sample of 737 (89% of total recruited participants) cancer patients who filled in both quality of life (QOL) questionnaires was Used • Participants inclusion	1. For the participants, the mean age= 51.4 ± 7.6 (SD), 63% female, 25% with current chemotherapy 2. Three of the four subscales common to both QOL instruments (physical, emotional, functional) proved suitable for equating (acceptable inter-correlations of corresponding subscales physical (r = 0.77), emotional domain (r = 0.60) role/functional (r = 0.63) relative to their internal consistency, sufficient unidimensionality of pooled subscales, satisfactory fit to the Rasch model) 3. Physical domain: The internal consistency (Cronbach's alpha) of the two subscales, is 0.84	<u>From the Article:</u> 1. The physical, emotional and functional/role domains of the two instruments (FACT-G and EORTC) were found to be equitable; but for the social domain, serious discrepancies between the corresponding subscales were detected and therefore equating of these subscales had to be discarded 2. The conversion tables developed in this study (physical, emotional and functional/ role domain) appear promising for the comparison between EORTC QLQ-C30 and FACT-G scores in oncological research 3. This study accomplished the main objective which was to derive direct conversion tables for the EORTC QLQ-C30 and the FACT-G <u>Relevant to Dissertation:</u> 1. This study linked two QOL instruments in the field of oncological research for the purpose to enable the investigators of clinical trials to compare information across studies that use different instruments.

			subscales • Confirmatory factor analysis (CFA) vis Mplus was used to confirm unidimensionality before conducting Rasch analysis • Rasch analysis was conducted with Winsteps • The pooled set of items in each pair of corresponding EORTC QLQ-C30 and FACT-G subscales was fitted to the Rasch model • Patients' QOL measures for each instrument were then estimated separately using anchored item and threshold measures obtained from the joint calibrations • In order to investigate if the conversion is largely independent of the sample used, the whole equating procedure was done once for the total patient sample and once separately for two subgroups	criteria: a diagnosis of cancer, age between 18 and 85 years, German speaking, no cognitive impairments, expected survival time of at least 3 months and completed informed consent • Clinical data were extracted from the medical records	for EORTC QLQ-C30 and 0.89 for FACT-G 4. Emotional domain: The internal consistency (Cronbach's alpha) of the two subscales, is 0.80 for EORTC QLQ-C30 and 0.75 for FACT-G 5. Role/Functional domain: The internal consistency (Cronbach's alpha) of the two subscales, is 0.89 for EORTC QLQ-C30) and 0.87 for FACT-G 6. The distribution of the raw scores is skewed towards higher values (as common for most QOL questionnaires) 7. Social domain: correlation between corresponding subscales was very low, (r = 0.09); with Cronbach's alpha is 0.82 and 0.64, respectively; thus not eligible for equating 8. Based on the residual correlations in a one-factor CFA, all residual correlations between –0.25 and + 0.25 for physical and emotional domains. Only one item in the functional/role domain has value < -0.25 9. The numbers of misfitting items were 2, 1 and 2 for the physical,	2. Besides rehabilitation, the early efforts using linking as a method also found in the field of clinical trail especially in the area of quality of life (QOL) assessments. Three articles were found linking QOL instruments. Gonin and colleagues (1996) initially used Rasch rating scale model to equate scores of different QOL-questionnaires to demonstrate 'equitability' between the total scores of FACT-G and Functional Living Index for Cancer (FLIC27). Gonin and colleagues (1996) also derived 'standard QOL scores' as a link between the raw scores of the two instruments (FACT-G and FLIC27). Follow-up, Chang and Cella (1997) also used the Rasch rating scale model to investigate equitability across total scores of five different QOL-instruments questionnaires and compared the total scores of the Cancer Rehabilitation Evaluation System (CARES), the FACT-G, the EORTC QLQ-C30, the Spitzer's Quality of Life-Index and the Short Form Health Survey (SF-36). Finally, Holzner and colleagues (2006) applied both classical test theory and the Rasch measurement model to investigate the equivalence of the European Organization for Research and Treatment of Cancer Core Questionnaire (EORTC QLQ-C30) and the Functional Assessment for Cancer Therapy – General (FACT-G) References:

			(patients receiving current treatment and patients without current treatment) • 95% Confidence intervals (95% CI) for the converted QOL scores were calculated for (a) individual subjects and (b) mean scores of samples of size N		emotional and functional/role domains, respectively 10. There was only one item out of 57 with a fit statistic exceeding 1.5 (FACT item 'proud of coping', infit = 1.52) suggesting that the data acceptably fit the Rasch model, so the authors decided to keep all items for equating purpose 11. For all of the three domains the differences between the two subsamples (patients with and without current treatment) were almost negligibly small, indicating a certain amount of stability of the conversion across various groups of patients 12. Confidence intervals for individual subjects were very large, thus, score conversion appears to be of very limited use; but for samples of size 25 the intervals become substantially smaller; thus, the conversion tables are of limited use for score conversion of individual subjects and may be most appropriate for comparing QOL scores of groups of	Gonin R, Lloyd S, Cella D, Gray G. Establishing equivalence between scaled measures of quality of life. Qual Life Res 1996;5(1):20–6. Erratum in: Qual Life Res 2001;10(1):104. Chang CH, Cella D. Equating health-related quality of life instruments in applied oncology settings. Arch Phys Med Rehabil 1997;11(2):397–406. 3. This study also supported using IRT-based Rasch linking method could generate sample-free common metric based on the results showed that separate analysis of the subsample of patients (with and without current oncological treatment) led to very similar results regarding the conversion of FACT-G to QLQ-C30 scores.

					patients across different studies using either the EORTC QLQ-C30 or the FACT-G	
Masse, L. C., Allen, D., Wilson, M., & Williams, G. (2006)	Introducing equating methodologies to compare test scores from two different self-regulation scales	To demonstrate the usefulness of item response modeling linking methodology in comparing groups of participants who were administered different scales intended to measure the same underlying constructs	• Secondary data analysis • Both groups received all 15 TSRQ items • The authors simulated conditions in which different groups receive different sets of items by selecting the items for which the responses were analyzed as a set and 'eliminating' all other item	• Treatment Self-Regulation Questionnaire (TSRQ): two 8-item TRSQs with only four items in common • Data collected as part of the Behavior Change Consortium (BCC) were analyzed for this study, including	1. The principal component analysis results indicated that the eight items assigned to OHSU and UR explained 40.3 and 41.6% of the total variance, respectively 2. The DIF analysis on the 15-item scale was significant and indicated that DIF indeed was present (x2 = 56.073, df = 14, P = 0.000) 3. Scale reliability was reduced when there were fewer items in the scale: 0.81 for 15 overlapping	<u>From the Article:</u> 1. The results showed that two eight-item TSRQ scales can be linked if they have at least four items in common 2. Varying the number of linking items did not affect the reliability of the results; however, it significantly affected the relative rating with respect to the 15-item scale <u>Relevant to Dissertation:</u> 1. This study suggested that linking methodologies can be used to compare results across studies in health behavior and health education research, that use slightly different versions of a scale to measure the

			responses for a particular analysis • The partial credit model was used for all the analyses, and all IRM analyses were conducted using ConQuest software • A linear transformation then is used to link the metrics of the groups by using the common items as an anchor for the linking	firefighters aimed at improving dietary and physical activity behaviors from Oregon Health Sciences University (OHSU; n=627) and adult smokers in a tobacco dependence treatment and diet intervention study from the University of Rochester (UR; n=355)	items and 0.64 for eight overlapping items 4. The reliability of 8-item and 4-item TSRQ are both 0.64 5. The impact of varying the number of items on the reliability index was minimal; reliability was found to range from 0.62 to 0.65	same construct
Velozo, Byers, Wang & Joseph (2007)	Translating measures across the continuum of care: Using Rasch analysis to create a crosswalk between the Functional Independence Measure and the Minimum Data Set	To develop a crosswalk between two instruments (FIM™ and MDS) across inpatient rehabilitation and the skilled nursing facilities	• Secondary data analyses • Rasch partial credit model • Common person equating • Winsteps Rasch programming	• Functional Independence Measure (FIM™) • Minimum Data Set (MDS) • A sample of 236 patients (original n=254) from four Department of Veterans Affairs' facilities who completed both the FIM™ and the MDS within 7 days • The major diagnostic	1. Eighteen patients with FIM™-MDS measures that fell outside the 95 percent confidence interval around the scatter plot identity line were eliminated from all further analyses (final n=236) 2. The mean ± standard deviation (SD) days between the administration of the FIM™ and the MDS is 3.7 ± 1.9 days 3. The combined FIM™-MDS showed good internal consistency (Cronbach alpha = 0.94) 4. The average item infit of	From the Article: 1. Ambulation/locomotion items and incontinence items may represent a construct separate from other motor items. 2. This study demonstrated a practical methodology for connecting scores from similar healthcare instruments 3. This study demonstrated Rasch analysis for linking the motor components of the FIM™ and the MDS using existing VA databases and six linking steps. 4. The overall psychometrics of the cocalibrated analysis indicated that the motor activity items of the FIM™ and MDS appear to be measuring the same construct. Relevant to Dissertation:

				groups included stroke (25%), orthopedic (22.5%), medically complex (11.4%), and amputation (8.5%)	the combined instrument was 1.1 (ideal is 1.0) 5. Twenty-one of the 26 items showing acceptable fit statistics 6. FIM™ and MDS raw scores correlated at -0.81 and the measures correlated at 0.78., slightly higher than the 0.72 correlation found by Williams et al. (1997) in comparison of the FIM™ with rescaled motor activity MDS (Pseudo-FIM™(E)); and slightly lower than the 0.85 correlation found by Buchanan et al. (2004) between the FIM™ and the MDS-PAC motor scales. But in Fisher's prospective study (1995) of crosswalk between instruments FIM™ and PECS had correlations of 0.90 7. Point-measure correlations for the items ranged between 0.54 and 0.84 8. The average item difficulty (mean ± SD = 0.00 ± 0.56 logits) was well matched with the mean of person measures (mean ± SD = 0.01 ± 0.9 logits)	This study used similar data source and linking method to establish a common metric between FIM™ and MDS, which supports the feasibility of using Rach common person equating to link different instruments.
Wang,	Validation of	To achieve score	• Secondary	• Functional	1. Both the MDS and the	From the Article:

Byers, & Velozo. (2008a)	FIM™-MDS crosswalk conversion algorithm	compatibility by validating a crosswalk that converts Functional Independence Measure (FIM™) scores to Minimum Data Set (MDS) scores (and vice versa) developed by Velozo et al. (2007)	retrospective data analyses • Rasch partial credit model • Common person equating • Point differences assigned to the FRG classification groups and Kappa • Winsteps IRT programming • The conversion algorithm was tested its validity at the: (1) individual patient (2) classification, and (3) facility levels • Two data sets: phase (I) were used to generate the FIM™-MDS crosswalk motor and cognition tables; phase (II) were used to perform the validity testing • Individual level: the absolute value of point differences between the actual FIM™ (FIM™a) scores and the MDS-derived FIM™ (FIM™c) scores (|FIM™a – FIM™c|) and the percentage of	Independence Measure (FIM™) • Minimum Data Set (MDS) • 2,130 patients (out of 151,770 original available records) were obtained from the Department of Veteran Affairs' Austin Automation Center who completed both the FIM™ and the MDS administered within 5 days (between June 1st 2002 ~ December 31th, 2004) • Three major impairment groups in the database were selected for analysis: stroke, amputation, and orthopedic impairment • Individual level: paired t-tests was used to test the equivalence of the score distributions to	FIM™ motor score distributions showed slightly skewed toward higher functioning individuals 2. The FIM™a motor and cognition scales correlated with the actual MDS motor and cognition scales at–0.80 and –0.66, respectively. 3. Wilcoxon signed rank test showed significant differences between the FIM™a and FIM™c motor score distributions ($z = -4.11$, $p < 0.001$); with 0.79 Pearson correlation coefficient 4. The mean FIM™c scores were within 1.3 and 0.1 points of the mean FIM™a scores for the motor and cognition scales, respectively. 5. (a) Stroke: Chi-square showed a significant association between the classification results ($\chi 2 = 1{,}232.6$, degrees of freedom [df] = 64, $p < 0.001$); Kappa demonstrated a fair strength of agreement ($\kappa = 0.37$). 44.0% were classified into the same FRGs, 67.0% into within ±1 FRG level, and 80.5 % into FRGs within ±2 FRG levels. (b)	1. Kappa statistics demonstrated a fair to substantial (0.37–0.66) strength of agreement between functional-related group classifications generated from the MDS-derived FIM™ and actual FIM™ scores. 2. "Mixed" findings from the validity testing of the FIM™-MDS motor and cognition crosswalks=> While sample distributions were similar, individual score comparisons fell short of expectations. Also, nonparametric results did not support the hypothesis that the actual and converted scores had the same score distributions 3. In general, the crosswalk algorithm showed feasibility of score comparisons across rehab settings. 4. Several results in this study supported the feasibility of developing FIM™-MDS crosswalks. 5. The effectiveness of a single measure or crosswalk conversions may depend on the quality of the data. 6. Low "individual equivalence" (i.e., relatively low percentage of actual and converted scores being within 5 points of each other), suggests that the crosswalks do not have adequate accuracy to monitor individual patients who transfer from facilities that use the FIM™ (e.g., IRFs) or from facilities that use the MDS (e.g., SNFs) Relevant to Dissertation: Compared with Buchanan and colleagues' (2004) findings, this study used different methodologies and sampling (e.g., differences in FRG vs

			FIM™a-FIM™c scores within 5 and 10 points were calculated. • Classification level: functional-related group (FRG) classification system was used to examine whether the FIM™c would classify the same patient into the same classification level as the FIM™a.	compare whether the score distributions were similar between the actual/converted scores • Classification level: Chi-square statistics were used to test whether any association existed between classification results based on the actual and converted scores. • Kappa statistics were used to quantify the strength of association • Phase (I): 654 subjects. The mean age is 68.0 y/o (SD= 12.0); 96% was male and 74% was white. • Phase (II): 1476 subjects. The mean age is 70.2 y/o (SD= 11.7); 97% was male and 69% was white.	Amputation: Chi-square showed a significant association; Kappa showed a substantial strength of agreement (κ = 0.66). 83.1% were classified into the same FRG. (c) Orthopedic Impairment: Chi-square showed a significant association; kappa showed a fair strength of agreement (κ = 0.37). 55.0 % were classified into the same FRGs, 69.2% into FRGs within ±1 level, and 87.4 % into FRGs within ±2 levels. 6. Four of the five facilities had an average point difference of 2.4 between the mean FIM™c and FIM™a scores. 7. Individual score comparisons are worse than expected with only 37~ 67% of the translated scores were within 5 points of the FIM™ actual scores.	CMG calculations, also, secondary analysis of VA data vs prospective data collection) and the study showed mixed results of using translated scores to classify patients for reimbursement purpose

Chen, W., Revicki, D., Lai, J., Cook, K., & Amtmann, D. (2009)	Linking pain items from two studies onto a common scale using item response theory	This study examined two approaches to linking items from two pain surveys to form a single item bank with a common measurement scale	• Secondary data analysis • Two approaches : (a) common item non-equivalent group design; OR multiple groups simultaneous calibration (all items were calibrated to an item response theory (IRT) model simultaneously); and (b) items were calibrated separately and then the scales were transformed to a common metric by using "scale transformation" • Samejima's Graded Response Model as implemented in MULTILOG was used to calibrate the items (MULTILOG uses marginal maximum likelihood method to estimate the item parameters) • Four transformation methods were used to obtain the transformation constants by using the computer	• Two independent surveys: (a) from Initiative on Measurement, and Pain Assessment in Clinical Trials (IMMPACT) Survey with Main Survey (959 chronic pain patients; 42 pain items) and Pain Modules (N=148; 36 pain items); and (b) Center on Outcomes, Research and Education (CORE) Survey (400 cancer patients; 43 pain items) • The two surveys included items measuring three pain domains: pain intensity (n=29), pain quality (n=10) and pain interference (n=38); but pain quality domain was excluded from this study because no	1. Simultaneous IRT Calibration: There were 29 pooled items and 1,364 subjects for the pain interference domain with the slope parameters were all reasonable large from 1.84 to 3.74, and all the threshold parameters were monotonically increasing. The item characteristic curves suggest that 10 response categories may be too many. IMMPACT sample reported higher levels of pain interference, which is reasonable because CORE subjects were cancer patients and not all of them experienced significant pain. 2. Separated IRT Calibration: There were 7 common items between the IMMPACT Main survey (n=959) and CORE items (n=400). The IMMPACT Pain Module and CORE surveys shared 12 common items and had 148 and 400 subjects, respectively; the slope parameters ranged from 1.20 to 2.99 for the items in the IMMPACT	<u>From the Article:</u> 1. The two linking approaches produced similar linking result across the two sets of pain interference items because there was sufficient number of common items and large enough sample size 2. The results suggested that simultaneous IRT calibration method produced more stable item parameters across independent samples (which is consistent as other simulation studies) than separated calibration when the IRT model fits the data, so this method is recommended for developing comprehensive item banks 3. When the items were calibrated separately, extreme item parameter estimates (threshold parameters estimates as high as 16.6 and 37.0) were obtained and some of the threshold parameters were not monotonically ordered correctly <u>Relevant to Dissertation:</u> 1. This study demonstrated how pain items from separated surveys can be linked to the same measurement scale to form a single item bank with shared common items, even for different populations (cancer and chronic pain) 2. This study recognized the importance of the number of the sample and the numbers of the items because these two factors may affect the linking results and the authors suggested that

			program STUIRT, an equating and scale transformations computer programs (available at www.uiowa.edu/~casma) • The extreme response categories for some of the items were excluded from the IRT calibration because no patients endorsed these categories • The graded response model assumes that the response options are monotonically ordered, thus, the threshold parameters are ordered (Non-ordered threshold parameters, in graded response model, are indication of non-convergence or problematic model fitting)	common items existed • Eight common items among the three data sets (7 pain interference, 1 pain intensity) • Evaluation of the two approaches: examine whether the calibration converged (by evaluating the value of the item parameters and the order of the threshold parameters)	sample, and from 2.49 to 5.96 for the CORE sample; the threshold parameters for the IMMPACT items ranged from −5.56 to 0.66, and ranged from −0.11 to 1.92 for the CORE items. 3. The correlation between the slope parameters of two approaches was 0.923; the correlations between the threshold parameters ranged between 0.911 ~ 0.992, except the first threshold parameter 4. The two scales differed by a factor of 0.784, the ratio of the standard deviations for the IRT scores of the CORE sample (1.047/0.821). The correlations between the IRT scores of the two approaches were as high as 0.999 for the IMMPACT and CORE samples, and for overall; meaning the two calibration approaches produced very similar item characteristics 5. For pain intensity, simultaneous calibration yielded more stable results; while the separated calibration	with smaller sample sizes and fewer common items, simultaneous calibration is preferable when linking sets of item from two surveys 3. There is no fixed rule regarding the number of common items across two samples/ instruments

					yielded unsatisfactory (because of a single common item with small sample size)	
Fischer, H. F., Tritt, K., Klapp, B. F., & Fliege, H. (2011)	How to compare scores from different depression scales: equating the patient health questionnaire (PHQ) and the ICD-10-symptom rating (ISR) using item response theory	To compare the ISR depression scale to the PHQ depression scales PHQ-9 and PHQ-2; and link both questionnaires on a common scale, providing data to enable the conversion of test scores in Germany	• Secondary data analysis • A General Partial Credit Model was applied to data from two different depression scales to check for unidimensionality • R 2.8.1 software was used for all statistical procedures (the packages included (a) nFactors for parallel analysis, (b) SEM: Structural Equation Models for confirmatory factor analysis (CFA) and (c) ltm for IRT model fitting • Method of maximum likelihood was used by assuming multinomial errors • Goodness of fit, the modification indices and the matrix of standardized residuals was examined • To compare quality	• Patient Health Questionnaire (PHQ-9) • Only first two items of PHQ-9 (PHQ-2) • ICD-10-Symptom Rating (ISR) • depression-CAT (D-CAT) was used as an external validation criteria • All three instruments were used for routine psychometric diagnostics at the Clinic for Internal Medicine, Department of Psychosomatic Medicine and Psychotherapy, Charité University of Medicine, Berlin • A consecutive sample in clinical settings with 4517	1. The mean age was 44.2 (SD = 14.8) years, with ages ranging from 14 to 86 years 2. Mean ISR depression score in the sample is 1.59 (SD = 1.06, range = 0–4), mean PHQ-9 score is 10.56 (SD = 6.22, range = 0–27) and mean PHQ-2 score is 2.70 (SD = 1.84, range = 0–6) 3. Unidimensionality: The first eigenvalue of the correlation matrix is 6.99 and is substantially greater than the second eigenvalue (which is 1.00); the first factor accounts for 54% of the total variance 4. A good fit for a unidimensional model of the ISR depression scale and an acceptable fit for the PHQ-9 depression scale were found. Both combined models had strong SRMR values for absolute fit, whereas RMSEA and CFI had poor fit 5. In the two-factor model, both factors correlated very highly (0.95) and the goodness of fit	<u>From the Article:</u> 1. Both instruments were constructed to measure the same construct and their estimates of depression severity are highly correlated 2. The predicted scores provided by the conversion tables are similar to the observed scores in a validation sample 3. The PHQ-9 and ISR depression scales measure depression severity across a broad range with similar precision 4. While the PHQ-9 shows advantages in measuring low or high depression severity, the ISR is more parsimonious and also suitable for clinical purposes 5. The equation tables derived in this study enhance the comparability of studies using either one of the instruments, but due to substantial statistical spread, the comparison of individual scores is imprecise <u>Relevant to Dissertation:</u> 1. This study used two sample method with one as a construction sample and the other as a validation sample to decrease study bias 2. This study also found that individual scores comparison is imprecise due to substantial statistical spread was observed 3. This conversion table of measuring depression showed suitability for patients with a wide variety of

			of measurements, respondents' scores of depression severity and measurement error were calculated in the validation sample by using (a) information provided solely from one questionnaire (ISR theta, PHQ-9 theta, PHQ-2 theta) or (b) information (Overall theta) based on an empirical Bayes method • Using corresponding theta values to create conversion table	observations from a total of 2999 inpatients and outpatients of a psychosomatic clinic • The sample was randomly divided in to a construction sample (n = 2258) and a validation sample (n = 2259) • About 5% of the patients do not complete the psychometric evaluation at each time of assessment, mainly due to reading difficulties or language issues	measures were comparable to the one-factor model 6. A correlation of 0.85 was found for estimated thetas from the four ISR items and the nine PHQ-9 items. 7. Differences between theta Estimates by ISR and PHQ-9 are distributed around zero (mean = 0.03, SD = 0.48). 8. In 77% of the 2259 cases, the absolute value of the difference is below or equal to 0.5. 9. The converted ISR scores and the means of the actual scores of the instruments, as well as intervals which contain about 66% (mean±1 SD) and 95% (mean± 2 SD) of the observed scores.	somatic and mental symptoms and diseases 4. The authors implied that equating questionnaires by calibrating the scores on a common scale could be more helpful in applied research than the use of a linear regression estimation of scores
Haley, Ni, Lai, Tian, Coster, Jette, Straub & Cella. (2011)	Linking the activity measure for post acute care and the quality of life outcomes in neurological disorders	To link physical functioning items from two instruments (AM-PAC and Neuro-QOL) using item response theory (IRT) methods	• Secondary data analysis • Nonequivalent sampling(group) design with 36 core items (Mobility (n=25) and activity of daily living (ADL) items (n=11)) common to both instruments (using linking	• Activity Measure for Post-Acute Care (AM-PAC) • Quality of Life Outcomes in Neurological Disorders (Neuro-QOL) • AM-PAC sample (n=1041) and	1. EFA: (a) 37 mobility Neuro-QOL items showed 2 factors explaining 59% of the variance for mobility; (b) 44 ADL Neuro-QOL items explained 79% of the item variance for ADL 2. Four items (3 items in mobility with moderate DIF, and 1 item as large	From the Article: 1. The AM-PAC and Neuro-QOL mobility and ADL scores could be placed on a common metric 2. The linking allowed score translations between instruments (i.e., estimation of AM-PAC mobility and ADL subscale scores could be based on Neuro-QOL mobility and ADL subscale scores and vice versa) Relevant to Dissertation:

			coefficients from common items to develop score conversions) • Neuro-QOL were linked to the AM-PAC by using the generalized partial credit model (An IRT-based linking method) • Stocking-Lord method, a test characteristic curve transformation method, to develop linking coefficients for the conversion scores • Linking was conducted with both raw and scaled AM-PAC and Neuro-QOL scores	community-dwelling adults (n=549) for the Neuro-QOL sample • AM-PAC were administered in post–acute care (PAC) settings. • Neuro-QOL items were administered to a community adults through the internet	DIF level: taking off a pullover shirt, chopping or slicing vegetables, shaving your neck and face safely and thoroughly with an electric razor, holding a screw and screwing it in tight with a manual screwdriver) in ADL had DIF 3. The final set of common items included 25 mobility and 11 ADL items 4. AM-PAC had many more items requiring less ability than Neuro-QOL 5. In both the mobility and ADL domains, common items were located in the middle of the scale	The authors suggested that future prospective study should ask participants to respond both instruments in order to replicate and validate the accuracy of the results from this study, and my dissertation will use equivalent group design (the same person answers both instruments) and partial credit model to link two instruments (FIM™, MDS) in measuring ADL
Thissen, D., Varni, J. W., Stucky, B. D., Liu, Y., Irwin, D. E., & DeWalt, D. A. (2011)	Using the PedsQL (TM) 3.0 asthma module to obtain scores comparable with those of the PROMIS pediatric asthma impact scale (PAIS)	To provide evidence of validity for one of the PROMIS measures, the Pediatric Asthma Impact Scale (PAIS), and to link the PedsQL™ Asthma Symptoms Scale with the metric of the PAIS	• Secondary data analysis • Samejima's graded IRT model, a calibrated projection, was used to link scores • Expected a posteriori (EAP) estimates for response patterns were computed for each respondent • Root mean squared deviation (RMSD)	• Pediatric Asthma Impact Scale (PAIS) • PedsQL™ Asthma Symptoms Scale • Approximately 300 children ages 8–17 • The two test forms containing PROMIS pediatric asthma	1. The estimated correlation between theta 1 (the underlying construct measured by the PAIS) with theta 2 (underlying construct measured by the PedsQL Symptoms Scale) is 0.96 2. All of the a parameter estimates exceed six times of standard errors, indicating that the corresponding relationships differ significantly from zero	<u>From the Article:</u> 1. The PAIS exhibited strong convergent validity with the PedsQL™ Asthma Symptoms Scale, and less strong relations with the other five scales (Treatment, Worry, and Communication Scales, and the DISABKIDS Asthma Impact and Worry Scales); indicating only one of the legacy scales was linked to the metric of the PAIS; the other five scales appear to measure constructs too different from that of the PAIS to link 2. The linkage system uses scores on

			statistic to check invariance of subgroup differences between scales to be linked • In calibrated projection, the multidimensional version of Samejima's graded model is fitted to the item responses from the two measures: theta 1 represents the underlying construct measured by the PAIS; while theta 2 represents the underlying construct measured by the PedsQL Symptoms Scale • IRTPRO software with two-tier methods	items were completed by a diverse sample of 622 respondents • Participants were recruited in hospital-based outpatient general pediatrics and subspecialty clinics and in public school settings between January 2007 and May 2008 in North Carolina and Texas	3. The likelihood ratio test for the difference in fit between the unidimensional model and the two-dimensional model was significant ($\chi^2(1) = 50.9$, $P < 0.0001$), meaning rejecting the unidimensional model	the PedsQL™ Asthma Symptoms Scale to produce relatively precise score estimates on the metric of the PAIS Relevant to Dissertation: 1. This study used calibrated projection to provide linkage of scores on the PedsQL Symptoms Scale with the metric of the PROMIS tool, PAIS by taking into account the slight difference between the constructs measured by the two scales 2. This study aims to integrate HRQoL measurement and suggested that calibrated projection may be useful to link other legacy scales to the PROMIS metrics as well
Fischer, H. F., Wahl, I., Fliege, H., Klapp, B. F., & Rose, M. (2012)	Impact of cross-calibration methods on the interpretation of a treatment comparison study using 2 depression scales	To evaluate the validity of an IRT-based cross-calibration approach that compares treatment outcomes from 2 clinics	• Prospective study • ISR scores and estimated latent trait values were transformed to PHQ-9 scores by using previously established conversion tables (Fischer, et al. 2011) using ISR response patterns in the Berlin	• Patient Health Questionnaire (PHQ) • ICD-10-Symptom Rating (ISR) • Data were collected within clinical practice settings at two different departments for psychosomatic	1. No difference in variance between the original PHQ-9 scores and the PHQ-9 scores transformed from ISR scores ($F = 1.0$, numerator df = 1561, denominator df = 1561, P value = 0.76). But a significant difference in means (difference = 0.19, $t = 2.03$, df = 1561, P value = 0.04, effect	From the Article: 1. There was no substantial change in the interpretation of the study results when different instruments were used. However, F- values, P-values, and effect sizes in the analysis of variance changed significantly. This might be attributed to differences in the content or measurement properties of the instruments. But no difference was observed between use of transformed sum scores and latent trait values

			sample • ISR scores and ISR latent trait estimates were compared with PHQ-9 scores and PHQ-9 latent trait estimates • Paired-t tests were used for examining mean differences and differences in variance (F-test) • Bland-Altman plots were used to examine to assess agreement between ISR and PHQ-9 scores • Differences against average scores of both measures and the limits of agreement were calculated • Pearson's correlations between sum scores and latent trait estimates from both instruments are also reported • Generalized Partial Credit Model was used to estimate individual depressive severity on latent trait level • Latent trait levels were estimated	medicine in Germany • 1066 patients were recruited during admission (within the first 3 days) and discharge (the last 3 days before discharge) with some type of mental and/or behavioral disorder and all patients received multimodal psychotherapeutic treatment	size = 0.03) was found, with original PHQ-9 scores being slightly higher than ISR scores that were transformed to PHQ-9 scores (11.09 vs. 10.90) 2. The correlation between original PHQ-9 sum scores and transformed PHQ-9 sum scores was 0.82 (P < 0.001) 3. Bland-Altman plots shows only poor concordance of observed and transformed individual PHQ-9 sum scores 4. The 95% limits of agreement were -7.05 and 7.43; differences between observed and transformed individual scores are beyond clinical importance, given the PHQ-9 scale ranges from 0 to 27 5. 95% limits of agreement latent trait estimates ranged from -0.99 to 1.03. Latent trait estimates from ISR scores differed from latent trait estimates from PHQ-9 scores at both admission (mean difference = -0.08; t= -4.39; df = 780; P-value < 0.01; effect size = 0.09) and at discharge (mean	2. Using ISR instead of PHQ-9 to estimate depressive severity also led to lower scores at admission and higher scores at discharge. Therefore, the influence of clinic on the improvement of depression severity was accentuated Relevant to Dissertation: 1. Although IRT cross-calibration methods are a convenient way to enhance the comparability of questionnaire data in applied clinical settings, the authors in this study implied that it seems IRT-method could not be able to overcome differences in measurement properties of the instruments. As these differences can lead to biased results, further study may need additional advanced techniques

			from the observed response patterns of each instrument with an expected a posteriori scoring algorithm • 2 x 2 ANOVAs were used to examine the impact on statistical results when using different cross-calibrated measures in a treatment outcome		difference = 0.04, t=2.43, df = 780, P-value = 0.01, effect size = 0.05) 6. When PHQ-9 was used in both clinics, a nonsignificant main effect of clinic, a significant main effect of assessment time, and a significant clinic-by-assessment time interaction were found	
Latimer, S., Covic, T., & Tennant, A. (2012)	Co-calibration of deliberate self-harm (DSH) behaviours: towards a common measurement metric	To explore a hierarchy of deliberate self-harm (DSH) behaviors and also produce a raw score conversion table between six DSH scales based on Rasch model	• Prospective study • Both samples contained the SHI-22 and SHIF-16 to provide a common item equating structure • Rach analysis was used to put six existing DSH scales into one single matrix and constructed an item pool by calibrating all items together • All items were examined by appropriate stochastic ordering (fit) and local independence assumptions, resulting in an 82-item set that fitted with the Rasch	• Six DSH scales containing 82 items • Self-Injury Questionnaire Treatment Related (SIQTR) • Self-Injurious Thoughts and Behaviours Interview (SITBI) • Deliberate Self-Harm Inventory(DSHI) • Inventory of Statements About Self Injury (ISAS) • Self-Harm Information Form (SHIF) • Self-Harm	1. For (a): Initially all 82 items were considered together and fit to the model was poor with significant residual correlations 2. For (b): The core linking scale, SHI-22 and SHIF-16, showed fit to the model (chi-square= 28.053, d.f.= 16, P= 0.031), using a Bonferroni adjusted p-value of 0.025 (0.05 divided by 2); and the principal component analysis (PCA) test showed strong support for unidimensionality (1.49% of the t-tests were significant). The PSI estimate was 0.666 and the Cronbach's Alpha was 0.827, with The mean logit value of	From the Article: 1. A raw score conversion table and a validated hierarchy of DSH behaviors were generated and all items from six DSH scales represented a unidimensional scale Relevant to Dissertation: 1. Latimer, Covic., & Tennant (2012)'s study used common item non-equivalent group design to co-calibrate six scales into one common metric using Rasch analysis, and the result showed that a raw score conversion table can be created when measuring patients' self-harm behaviors; however, I would suggest to have a follow-up study with an independent sample to validate this developed crosswalk 2. Latimer, Covic., & Tennant (2012)'s study used chi-square to test model fit and principal component analysis to test unidimensionality instead of fit statistics used in Velozo et al. (2007) and Wang et al. (2008a)'s

			model • Rasch analysis was used to examine unidimensionality with software, RUMM 2030 • Five Rasch analyses were conducted: (a) all items, (b) SHI-22 and SHIF-16, (c) ISAS-12, SHI-22 and SHIF-16, (d) SHI-22, SHIF-16, SITBI-11, DSHI-16 and SIQTR-5, and (e) ISAS-12, SHI-22, SHIF-16, SITBI-11, DSHI-16 and SIQTR-5 • The chi-square and residual fit statistics were used to test if the data meet with model expectations • Person Separation Index (PSI), which is analogous to Cronbach's Alpha, has the advantage of being provided when there are missing cases	Inventory (SHI) • The population was 568 Australians aged 18-30 years old (62%university students, 21% mental health patients, and 17% community volunteers) • The ISAS-12, SHI-22 and SHIF-16 were administered to 332 participants (Sample1). The SITBI-11,SIQTR-5, DSHI-16,SHI-22 and SHIF-16 were administered to 236 participants (Sample2)	the respondents was -1.881, suggesting the sample were at much lower level of DSH 3. All three co-calibrations of (c) ISAS-12, SHI-22 and SHIF-16, (d) SHI-22, SHIF-16, SITBI-11, DSHI-16 and SIQTR-5, and (e) ISAS-12, SHI-22, SHIF-16, SITBI-11, DSHI-16 and SIQTR-5 showed fit to the model (chi-square= 18.928, d.f.= 12, P= 0.090 for (c); chi-square= 16.137, d.f.= 12, P= 0.185 for (d); chi-square= 36.35, d.f.=32, P= 0.273 for (e)). 4. For (c): PSI= 0.774, Cronbach's Alpha =0.827; For (d): PSI= 0.748, Cronbach's Alpha =0.821; For (e): PSI= 0.690, Cronbach's Alpha =N/A due to missing cases 5. The resulting calibration shows that the different scales occupy different ranges on the hierarchy of DSH (prevalence estimates ranging from 47.7 to 77.1%) 6. The least frequently endorsed item is was 'dropping acid on skin', and the most frequently endorsed item is 'picking	studies; however, both studies supported the possibility to develop the linking crosswalks by Rasch analysis

					at a wound' 7. Some of the individual DSH items showed Differential Item Functioning (DIF) by age, gender, and group (clinical versus non-clinical)	
Askew, R. L., Kim, J., Chung, H., Cook, K. F., Johnson, K. L., & Amtmann, D. (2013)	Development of a crosswalk for pain interference measured by the BPI and PROMIS pain interference short form	To develop and test a crosswalk table to transform Brief Pain Inventory pain interference scale (BPI-PI) scores to PROMIS-PI short form (PROMIS-PI SF) scores for the multiple sclerosis (MS) patients	• Secondary data analysis • Unidimensionality is assessed by one-factor confirmatory factor analysis • Two-parameter logistic graded response model was used to derive item difficulty and discrimination parameters for each BPI-PI item • The calibration was anchored on the established parameters for the PROMIS-PI SF items to maintain direct comparability with the US general population • Two BPI-PI scores for each person: (a) obtained from the PROMIS metric using the IRT calibrated item parameters; and (b) obtained using	• Brief Pain Inventory pain interference scale (BPI-PI) • PROMIS-PI short form (PROMIS-PI SF) • The BPI-PI and the PROMIS-PI SF were administered in two studies that included persons with MS • Two samples: one served as a developmental calibration sample (n=369); and a separate one served as a validation sample (n=360) • Participants in this study were community dwelling individuals with MS primarily recruited	1. For BPI-PI summary scores ranging from 0 to 10, corresponding T scores ranged from 38.6 to 81.2 on the PROMIS metric 2. The mean difference between observed and crosswalked T scores was 0.51 (95 % CI = 0.11–0.91) (SD = 3.9) in the calibration sample and -1.47 (95 % CI = -1.91 to -1.04) (SD = 4.2) in the validation sample 3. Approximately 80 % of crosswalked scores in the calibration sample were within four score points of the observed PROMIS-PI SF scores, and 70 % were within four points in the validation sample 4. The largest differences were at lower levels of the pain interference continuum 5. Differences between observed and crosswalked T scores were compared in both	<u>From the Article:</u> 1. Crosswalked pain interference scores adequately approximated observed PROMIS-PI SF scores in both the calibration and validation samples 2. MS researchers and clinicians interested in adopting the PROMIS instruments can use this table to transform BPI-PI scores to enable comparisons with other studies and to maintain continuity with previous research 3. Regression-based score linking leads to larger errors in prediction and often fails to meet important criteria for score linking 4. The crosswalk was applied to a different dataset, the average difference in prediction error was greater in the validation dataset than in the calibration dataset <u>Relevant to Dissertation:</u> 1. The authors also found that individual scores derived from crosswalks are recommended for group-level analysis and are not intended for use in clinical care given the additional source of potential bias inherent to any crosswalking procedure

			traditional scoring of the BPI, averaging over individual item scores • To assess variability in the performance of the crosswalk, the standardized root mean square difference (RMSD) was compared across multiple subgroups (gender, race, age, education, type of MS, mobility) • Multiple F tests were carried out to assess variability in the performance of the crosswalk by subgroups • Bland-Altman plots were used to examine differences across all levels of trait • IRT-based analyses were carried out with IRTPRO v2.1	through the Northwest Chapter of the National Multiple Sclerosis Society • The validation sample (n=360) completed both the BPI-PI and the PROMIS-PI SF	samples 6. The estimates of internal consistency also supported scale calibration with nearly identical Cronbach's a coefficients (PROMIS SF = 0.94; BPI = 0.93) 7. In the validation dataset, 70 % of predicted scores were within four points of actual scores and 87 % were within six points 8. Subgroup comparisons indicated that RMSD estimates ranged from 0.01 to 0.06, indicating that the crosswalk table functioned well across subgroups in the validation sample	
Ten Klooster, P. M., Oude Voshaar, M. A., Gandek, B., Rose, M.,	Development and evaluation of a crosswalk between the SF-36 physical functioning scale and Health	To develop and evaluate a crosswalk between scores on the PF-10 and HAQ-DI in patients with rheumatoid arthritis (RA) (because these two are the most	• Retrospective, secondary data analysis • The same patient completed both instruments • The maximum likelihood	• SF-36 physical functioning scale (PF-10) • Health Assessment Questionnaire disability index (HAQ-DI)	1. Total scores on the PF-10 and HAQ-DI were strongly correlated ($r = -0.75$) 2. The Rasch-based co-calibration of the HAQ-DI adequately fitted the data according to the LM	<u>From the Article:</u> 1. The crosswalk developed in this study allows for converting scores from one scale to the other and can be used for group-level analyses in patients with RA 2. The HAQ-DI can measure levels of extremely poor function that are not

Bjorner, J. B., Taal, E., Glas, C. A., van Riel, P. L., & van de Laar, M. A. (2013)	Assessment Questionnaire disability index in rheumatoid arthritis	frequently used instruments for measuring self-reported physical function in RA); this study also examined the appropriateness of different IRT models by comparing the calibrations and performance of a crosswalk based on a one-parameter Rasch model with the two-parameter and multidimensional extensions	estimation procedure was utilized to estimate the structural model parameters • The latent disability levels of patients were estimated using the expected a posteriori (EAP) method throughout all IRT analyses. • Model fit of all estimated models was assessed using Lagrange Multiplier (LM) item fit statistics specifically targeted at polytomous items • Absolute differences (effect sizes; ES) between expected and observed item scores for high, average and low scoring individuals were computed • All IRT analyses were performed with the MIRT software package • Agreement between patients' observed and predicted scores on the PF-10 and	• Two independent datasets were used for this study: (a) Data from 1791 RA patients, a large and clinically diverse sample from Dutch Rheumatoid Arthritis Monitoring (DREAM) online registry from 2003-2012 was used for IRT calibrations and development and comparison of the crosswalks; (b) Patients from the DREAM remission induction cohort (n=276) were used for accuracy and validity of the final crosswalk [note: The accuracy of the final crosswalk was cross-validated using baseline (n = 532) and 6-month follow-	tests, with all accompanying ESs <0.10 3. Both PF-10 and HAQ-DI measured an approximately equally wide range of physical functioning with high precision. But overall, the PF-10 was slightly more precise at better levels of physical functioning 4. The PF-10 and HAQ-DI adequately fit a unidimensional Rasch model. Both scales measured a wide range of functioning (although the HAQ-DI tended to better target lower levels of functioning) 5. The Rasch-based crosswalk performed almost identically to crosswalks based on the two-parameter (GPCM) and multidimensional IRT models; with high correlations between predicted scores based on the different crosswalks (r's >0.988) 6. The crosswalks based on the two-parameter and multidimensional models did not perform substantially better in terms of agreement between observed and	represented in the PF-10 and, conversely, that some levels of extremely good PF can be measured with the PF-10, but not with the HAQ-DI 3. Rasch-based crosswalk was adequate for converting total scale scores because the agreement between observed and predicted scale scores did not improve much in the more general models (GPCM and multidimensional GPCM models) 4. Agreement between predicted and observed scale scores from the Rasch-based crosswalk was acceptable for group-level comparisons 5. The longitudinal validity in discriminating between disease response states was similar between observed and predicted scores 6. Results showed that it was possible to develop a straightforward Rasch-based crosswalk between both scales in patients with RA 7. The Rasch-based crosswalk performed similarly to crosswalks based on its two-parameter and multidimensional extensions. Relevant to Dissertation: The study design of Ten Klooster et al. (2013)'s study is very similar as Wang et al. (2008a)'s validation study that aimed to validate the developed crosswalk between FIM™ and MDS by investigating ICCs, using Kapps to classify patients with the translated scores into FRG groups and comparing the

			HAQ-DI was assessed by computing intraclass correlation coefficients (ICCs) with 95% confidence intervals (95% CI) using two-way mixed effects models with absolute agreement for single measurements (type A,1) with ICCs were considered adequate for group level comparisons when ≥0.70 • Bland-Altman plots of the difference against the mean of predicted and observed scores were constructed • Observed and predicted change scores and total effect sizes (ES) (Cohen's d) were calculated for patients who completed both measures at baseline and 6-month follow-up (n = 276) • The relative	up (n = 276) data from an independent cohort of early RA patients in a treatment-to-target study] • 1-, 2-, and 3-parameter models was used to develop crosswalks and the crosswalks were compared • 1-parameter model: Polytomous Rasch partial credit model (PCM) • 2-parameter model: Generalized partial credit model (GPCM) is a two-parameter IRT model for polytomous data which includes a discrimination parameter that accounts for the different reliability of individual items with respect to measuring the underlying	predicted total scores on the PF-10 and HAQ-DI 7. Agreement between predicted and observed scale scores from the Rasch-based crosswalk in the cross-validation sample had high ICCs (95% CI) for both HAQ-DI (0.72 to 0.81) and PF-10 (0.75 to 0.82) 8. Bland-Altman plots showed intra-individual differences were similarly distributed above and below the mean and not related to the magnitude of the measurement 9. However, the limits of agreement were wide for both scales and showed substantial discrepancies in agreement within individual patients 10. Regarding the observed 6-month change scores in the total cross-validation sample, standardized improvements were largest for the HAQ-ADI (ES =0.55), closely followed by the HAQ-SDI (ES = 0.49) and the PF-10 (ES = 0.40) 11. In terms of differentiating between levels of longitudinal treatment response, the	differences between the actual scores and the translated scores at group and individual levels. By using independent sample to test the validation of the developed crosswalk's, both Wang et al. (2008a)'s and Ten Klooster et al. (2013)'s studies supported using straightforward Rasch-based linking methods could create validated crosswalk between two instruments, so the estimate scores on one scale could be validly translated from scores on the other scale, even though these two studies used different software to run Rach analysis (Winsteps vs. MIRT), different patient diagnosis groups (stroke, amputation, and orthopedic impairment vs. RA), different instruments (FIM™, MDS vs. PF-10, HAQ-DI) and different group classification methods used (using FRG classification systems vs. using ICCs for between observed and predicted scores at group-level). Holzner and colleagues (2006) also had similar results by finding confidence intervals for individual subjects were very large, thus, the score conversion is of limited use for individual subjects and may be most appropriate for comparing QOL scores of groups of patients.

			efficiency of the change scores to discriminate between responder status was analyzed using one-way analysis of variance (ANOVA) tests	latent trait • 3-parameter model: Multidimensional GPCM models • The 28-joint Disease Activity Score (DAS28), a pooled index, that includes a tender joint count, a swollen joint count, the erythrocyte sedimentation rate, and the patient's global assessment of general health, was used as the external criterion for determining response to treatment • The standard disability index (SDI) adjusts category scores upwards for the use of aids or devices or help from others • The alternative disability index (ADI) does not take the use of	HAQ-ADI was slightly more efficient than the HAQ-SDI and PF-10 12. Relative validity coefficients of the predicted scores were close to, and not significantly different from, those of the actual observed scores for all three scales	

				aids and devices into account		
Lai, J. S., Cella, D., Yanez, B., & Stone, A. (2014)	Linking fatigue measures on a common reporting metric	To report the methods used to develop linking (crosswalk) tables to enable the direct comparison of fatigue scores from three instruments (most widely used measure of fatigue) and link fatigue scores to the same metric in order to facilitate interpretation of fatigue outcomes	• Retrospective study: using the sample recruited from previous study (Lia et al., 2005) • Two item response theory (IRT)-based linking methods: (a) the Stocking-Lord calibration (produces additive and multiplicative constants to transform item parameters); and (b) fixed-parameter calibration (places non- PROMIS items on the same metric as PROMIS items), were used to establish linking between measures • The IRT calibrations were derived using the graded response model (GRM) implemented in MULTILOG software • Confirmatory factor analysis was used to assess the unidimensionality of the combined scales before	• Patient-Reported Outcomes Measurement Information System (PROMIS)-Fatigue with Functional Assessment of Chronic Illness Therapy-Fatigue Scale (FACIT-F) 13 items (*note: FIB: Fatigue Item Bank, has 95 items) • The Medical Outcomes Study Short Form-36 (SF-36) four-item Vitality Scale • The Quality of Life in Neurological Disorders Fatigue Scale (Neuro-QOL Fatigue Scale) 19 items • Participants were recruited from two data sets (n=803 and n=1120)	1. Factor analysis confirmed the assumption of unidimensionality of the combined scale (SF-36 + PROMIS; Neuro-QOL + PROMIS) 2. The correlations between instruments are high (r= 0.89 for SF-36 and the PROMIS FIB; r= 0.88 for Neuro-QOL and PROMIS) 3. The correlations between the combined score and the measures were 1.0 and 0.90 (for PROMIS FIB and SF-36 Vitality Scale, respectively); and 0.98 and 0.99 (for PROMIS FIB and Neuro-QOL, respectively) 4. SF-36 + PROMIS: the correlations of the parameters (slope/ threshold parameters) from two methods (Stocking-Lord &fixed-parameter calibration) ranged from 0.94 to 0.99; and the person-scaled scores from these two methods were almost identical (r=1, p < 0.001). 5. The T-score discrepancies (Stocking-	From the Article: 1. Both the Stocking-Lord calibration and fixed-parameter calibration linking methods produced comparable results (The final crosswalk tables are reported for the fixed-parameter calibration) 2. Findings can facilitate comparison of scores across some of the most widely used fatigue measures and assist in comparing patient-reported fatigue outcomes in clinical trials, comparative effectiveness research, and clinical practice Relevant to Dissertation: 1. When considering using linking strategies, multiple linking strategies are available, including both traditional procedures (e.g., equipercentile) and IRT (e.g., fixed-parameters; Stocking-Lord linking) and I will use IRT methods to link different scales 2. It may be important to recognize whether the differences of scale content or differences in psychometric properties of the scales would affect the linking results or quality of linking

			linking (Mplus 6.0) • The Stocking-Lord as implemented in Plink (a package for R), was used to link IRT-estimated parameters from different scales using two steps • Fixed-parameter calibration: fixed the PROMIS Fatigue item parameters and calibrated only SF-36 Vitality Scale and Neuro-QOL Fatigue items using GRM model • Crosswalk tables to convert the SF-36 Vitality Scale and Neuro-QOL raw scores to the PROMIS FIB using the PROMIS scoring system as described in Lai, et al (2011) article		Lord minus fixed-parameter) ranged from -0.30 to 1.10 with a mean of 0.06 (SD =.01), and only one participant had a discrepancy greater than 1 T-score unit (0.1 SD) 6. Neuro-QOL+ PROMIS: the correlations of the parameters (slope and threshold parameters) from two methods (Stocking-Lord &fixed-parameter calibration) ranged from 0.99 to 1.00; and the person-scaled scores from these two methods were almost identical (r=1, p < 0.001). The T-score discrepancies ranged from -0.87 to 1.24 with a mean of 0.01 (SD =.30), and only one participant had a discrepancy greater than 1 T-score unit (0.1 SD)	
Oude Voshaar, M. A., Ten Klooster, P. M., Taal, E., Wolfe, F., Vonkeman, H., Glas, C. A., & Van De Laar,	Linking physical function outcomes in rheumatology: performance of a crosswalk for converting Health Assessment Questionnaire	To evaluate the reliability of a crosswalk, developed in the Netherlands, between the Health Assessment Questionnaire (HAQ) disability index (DI) and the Short Form 36 physical functioning scale	• Retrospective study • Reliability of the crosswalk was evaluated by calculating intraclass-correlation coefficients (ICCs) with 95% confidence intervals using	• Short Form 36 (SF-36) physical functioning scale (PF-10) • Health Assessment Questionnaire (HAQ) disability index (DI)	1. Patients reported mild to moderate levels of disability, on average 2. The crosswalk produced reliable conversions for both the HAQ DI (ICC range 0.70–0.77) and PF-10 (ICC range: 0.73–0.78) in all 3 disease groups. 3. The mean difference	<u>From the Article:</u> 1. The crosswalk produced reliable conversions at the diagnostic-subgroup level in a cross-cultural setting and can be used to convert HAQ DI to PF-10 scores and vice versa in the US patients with RA, FM, or SLE. 2. For all 3 disease groups, the limits of agreement were fairly wide and conversion at the level of individual

M. A. (2014)	scores to Short Form 36 physical functioning scale scores	(PF-10)	two-way mixed-effects models with absolute agreement for single measurements (type A,1) [ICCs are generally considered adequate for group-level comparisons when ≥ 0.70] • Agreement between observed and predicted scores was evaluated using the Bland-Altman approach (a plot of the difference against the mean of predicted and observed scores) *note: ICC and Agreement are the same as their previous Ten Klooster et al.'s 2013 article • SPSS version 21 was used for all analyses	• A sample of patients with various rheumatic diseases in the National Data Bank for Rheumatic Diseases data in the US • Baseline data from patients with rheumatoid arthritis (RA; n=29,020; majority is RA in this study), fibromyalgia (FM; n=3,776), and systemic lupus erythematosus (SLE; n=1,609) participating in the National Data Bank for Rheumatic Diseases were Analyzed [a large-scale open cohort; total of 34,405 patients]	between observed and expected scores was close to zero in US patients with RA. 4. ICCs between predicted and actual scores ranged from 0.70–0.78, indicating that the crosswalk was sufficiently reliable for group-level use across diagnostic subgroups in the US data 5. Visual inspection of the Bland-Altman plots revealed that individual errors appeared to be unsystematically distributed across the observed PF levels 6. Mean differences between observed and predicted scores were small in magnitude across diagnostic groups on both scales 7. Bias was marginally higher (slightly less reliable) in FM and SLE patients than it was in RA patients; but the magnitude of the mean difference between observed and predicted scores was smaller than 1 total score level for both the HAQ (i.e., 0.125 units) and the PF-10 (i.e., 5 units) in SLE and FM and thus may	patients is not recommended 3. The study results suggest that the crosswalk can be used for descriptive purposes (i.e., systematic reviews), group-level inferential purposes (i.e., calculate standardized treatment effects on PF in meta-analyses), or to evaluate trends in longitudinal studies (when different measures were used at different time points) Relevant to Dissertation: 1. Even with the assumption that a crosswalk may differ between patients with different cultural backgrounds or diseases, thus, the generalizability of a crosswalk needs to be tested before it can be used in a new setting (since patients with gout, osteoarthritis, and RA function differently on the HAQ DI), the results demonstrated that accurate group-level conversions can be obtained using the crosswalk in the setting of US patients with RA with the crosswalk developed in the Netherlands 2. Ten Klooster et al. (2013)'s study had consistent results in examining both the individual-level and group-level classifications using translated scores compared to Wang et al. (2008a)'s crosswalk validation study. Both studies consistently showed that the linking crosswalk could provide better/more identical group-level classification results using translated scores compared to the actual scores but not for the individual-level classifications. 3. The study supported that the

					not be clinically relevant 8. The limits of agreement were fairly wide for both scales and showed substantial discrepancies in agreement within individual patients across conditions 9. The crosswalk slightly underestimated mean PF levels for converted HAQ DI scores and slightly overestimated mean PF levels for converted PF-10 scores in SLE and FM 10. It should be noted that any estimate of a sample's mean using the crosswalk will be affected by measurement error associated with converting scores	observed reliability of the crosswalk reflected the reliability of the instruments used for developing crosswalk (the assumption is that the measurement error of the crosswalk is a function of the reliability of the crosswalked instruments suggested by Ten Klooster, Oude Voshaar, Gandek, Rose, Bjorner, et al., 2013) 4. Although Ten Klooster et al (2013) showed that estimated effect size statistics in a sample of 276 RA patients were quite close to the actually observed effect sizes, use of the crosswalk for inferential purposes is not recommended in small sample sizes; this was consistent with Noonan et al. (2012)'s study in terms of having appropriate sample size to conduct linking

Table 2.3. Literature Reviews of Comparing Measurement Precisions among Item Bank, Short Forms (SFs) and Computerized Adaptive Tests (CATs) Used in Healthcare (ordered by year) (n=3)

Authors	Title	Aims	Population/ Methods	Instruments	Results/Conclusions
Choi SW, Reise SP, Pilkonis PA, Hays RD, Cella D (2010)	Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms	To assess the efficiency of static short forms and computer adaptive testing (CAT) using data from the Patient-Reported Outcomes Measurement Information System (PROMIS) project	• 6,213 general population subjects • 7,844 clinical subjects • Post-hoc simulations based on the PROMIS calibration sample	The 28-item PROMIS depressive symptoms bank	1. Short-form patient-reported outcome measures can minimize test burden 2. All short forms and CAT produced highly correlated scores compared with full-bank scores, but CAT performed better than static short form in almost all criteria. 3. Short-form selection strategies performed only marginally worse than CAT. 4. A two-stage branching test format in static short form can increase measurement precision. 5. The efficiency of a two-stage semi-adaptive testing strategy was similar to CAT, therefore, the two-stage short form can have further consideration and study.
Lai, J. S., Cella, D., Choi, S., Junghaenel, D. U., Christodoulou, C., Gershon, R., & Stone, A. (2011)	How item banks and their application can influence measurement practice in rehabilitation medicine: a PROMIS fatigue item bank example	This article used fatigue item bank developed by the NIH PROMIS Cooperative Group as an example to demonstrate the item bank and its further applications, including CATS and short forms	• For “dimensionality evaluation”: 803 people • For “item calibrations”: 14,931 people • (U.S. general population representative sample collected by internet)	• 112 PROMIS fatigue items • 13-item of Functional Assessment of Chronic Illness Therapy-Fatigue • 4-item of SF-36 Vitality scale	1. The PROMIS FIB consists of 95 items demonstrated acceptable psychometric properties. 2. Computerized Adaptive Testing (CAT) showed consistently better precision than short-forms. 3. All three short-forms showed good precision for the majority of participants, in that more than 95% of sample could be precisely measured with a reliability greater than 0.9. 4. Measurement practice can be advanced by using a psychometrically sound CAT

					and short forms. 5. CAT and short-forms derived from the PROMIS FIB item bank can reliably estimate fatigue reported by the US general population. 6. Evaluation in clinical populations is warranted before the item bank can be used for clinical trials
Bjorner, J. B., Rose, M., Gandek, B., Stone, A. A., Junghaenel, D. U., & Ware, J. E. Jr. (2014)	Difference in method of administration did not significantly impact item response: an IRT-based analysis from the Patient-Reported Outcomes Measurement Information System (PROMIS) initiative.	To test the impact of method of administration (MOA) on the measurement characteristics of items developed in the PROMIS	• IRT methods were used to develop two non-overlapping parallel static 8-item forms from each of three PROMIS domains (physical function, fatigue, and depression) to ensure two short forms have similar item information function • 923 adults (age 18-89) with three diagnostic groups (chronic obstructive pulmonary disease, depression, or rheumatoid arthritis) • A randomized crossover design • Subjects answered one form by interactive voice response (IVR) technology, paper questionnaire (PQ), personal digital	• To construct parallel static forms reflecting the content of the larger PROMIS item banks, the items were selected for each domain based on the number of items per content category within each form was proportional to the number of items per category in the full item bank. • The categories included: upper, central, and lower extremity functions and instrumental activities of daily living (for physical function), experience and impact (for fatigue), and mood and cognition (for depression)	1. Multigroup confirmatory factor analysis supported equivalence of factor structure across MOA 2. No differences in item location parameters were found, which strongly supported the equivalence of scores across MOA 3. No statistically or clinically significant differences were found in score levels in IVR, PQ, or PDA administration compared to PC. 4. Potential adjustment is far below the pre-specified minimal important difference, indicating that the implied mean score levels are equivalent (no minimal important difference was specified for slope effects prior to analysis) 5. Item discrimination was significantly lower for IVR administration in the depression domain, which is one of a few significant effects of MOA on score precision

			assistant (PDA), or personal computer (PC) on the Internet, and a second form by PC, in the same administration. • Confirmatory factor analysis and item response theory methods were used to assess structural invariance, equivalence of item responses, and measurement precision		

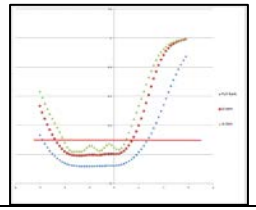
Table 3.1. Physical Items Measured in the FIM and MDS

Instrument	**Functional Independence Measure (FIM)**	**Minimum Data Set (MDS)**
Parameter: ADL/Motor Skill	Eating	Eating
	Grooming	Personal Hygiene
	-----	-----
	Bathing	Bathing *
	Dressing- Upper Body	Dressing
	Dressing- Lower Body	-----
	Toileting	Toilet Use
	Bladder Management	Bladder Continence †
	Bowel Management	Bowel Continence †
	Bed, Chair, Wheelchair (Transfer)	Transfer
	Toilet (Transfer)	-----
	Tub, Shower (Transfer)	-----
	Stairs	-----
	-----	Bed Mobility
	Walk/Wheelchair	Walk in Room
	-----	Walk in Corridor
	-----	Locomotion on Unit
	-----	Locomotion off Unit
Rating Scale	7= Complete Independence	0= Independent
	6= Modified Independence	-----
	5= Supervision	1= Supervision
	4= Minimal Assistance (>75% independence)	2= Limited Assistance
	3= Moderate Assistance (>50% independence)	-----
	2= Maximal Assistance (>25% independence)	3= Extensive Assistance
	1= Total Assistance	4= Total Dependence
	-----	8= Activity Did Not Occur During Entire 7-Day Period

Note: (from Wang, et al., 2008a)
* Separate rating scale in MDS: 0 = independent, 1 = supervision, 2 = physical help limited to transfer only, 3 = physical help in part of bathing activity, 4 = total dependence, 8 = activity did not occur during entire 7 days.
† Separate rating scale in MDS: 0 = usually continent, 2 = occasionally continent, 3 = frequently incontinent, 4 = incontinent.

Table 3.2. A Summary Table of Hypothesis, Methods and Final Products for Each Specific Aim

BUILDING INSTRUMENT	**<u>Specific Aim I</u>** **Create a FIM-MDS item bank** that meets Item Response Theory (IRT) model requirements	**Hypotheses**	**Methods**	**Products**
		N/A Prior to this SA, the hypothesis is that based on the latent trait model, we could link FIM and MDS (Velozo, Byers, Wang, & Joseph, 2007; Wang, Byers, & Velozo, 2008a)	• Rasch fit statistics • Confirmatory factor analysis • Differential Item Functioning (DIF) • 95% confidence interval plots to eliminate "invalid" data	The FIM-MDS item bank meets IRT assumptions and criteria: (a) unidimensional (b) local independence (c) monotonicity; and remove DIF items [Note: Essential DIF items may be kept]
	<u>Specific Aim II</u> **Generate IRT-based short forms and computer adaptive tests** from the item bank	**Hypotheses**	**Methods**	**Products**
		N/A IRT-based short forms and computer adaptive tests can be established	• Short Form o del Toro and colleagues' (2011) Short Form development procedures	Short Form FIM, Short Form _MDS, Short Form Full Bank
VALIDATION INSTRUMENT	**<u>Specific Aim III</u>** **Compare measurement precision** of the IRT-based short forms and MDS converted scores to the original FIM measure	**Hypotheses**	**Methods**	**Results**

		The varied short forms and MDS converted score will have similar measurement precision compared to the original FIM	• Descriptive Statistics
			• Precision
			  o Rasch analysis person strata calculation
			  o Test Error Plots
				Test Information/Error Plots
				 

	<u>Specific Aim IV</u> **Assess measurement accuracy** of the IRT-based short forms and MDS converted score in classifying Veterans into Function Related Groups (FRGs) compared to the original FIM	**Hypotheses**	**Methods**	**Results**

		The varied short forms and MDS converted score will have similar accuracy in determining FRGs categories for patients compared to the original FIM (standards)	Assess the accuracy of linking tools and original FIM in classifying Veterans into Function Related Groups (FRGs)
			• McNemar's and kappa statistics (for amputation)
			• Weighted kappa statistics (for stroke, knee/hip replacement)
			• Intraclass correlation coefficient (ICC)
			 
				• Individual-level: Significant test of median for score distribution
				• Group-level: The percentage of individuals being classified into the same FRG category
				  o One category apart (±1 level)
				  o Two categories apart (±2 levels)

Table 3.3. Comparison Table of the Proposed Study with Other Three Different Study Designs

Research Project	Research Design	Advantages	Limitations
Item Banking Across the Continuum of Care (VA FIM-MDS item banking project)	Retrospective, secondary data analysis (using longitudinal data in a format as cross-sectional data analysis)	a. <u>Sampling Frame</u> 1. Big sample size 2. Homogeneity of the sample (Veterans using post-acute care) 3. Real-life data b. <u>Required Resources</u> 1. Save time, cost, and resources in terms of collecting data compared to prospective study c. <u>Internal Validity</u> 1. Two instruments are "real" different tests developed independently and used currently 2. Subjects are blind to the study	a. <u>Characteristics of the Dataset</u> 1. Not public accessible database 2. Narrowed breadth of available data; not flexible; only approved variables could be obtained b. <u>External Validity (Generalizability)</u> 1. Restricted to the Veterans population; may not be able to generalize to the general population c. <u>Miscellaneous Factors</u> 1. Even though we limited to the same patients taking the FIM and MDS within 7 days, it is possible that the patients' functional status may change over these 7 days [secondary variance] 2. May take more time to receive the dataset after getting the approval in real-world situation
National Health and Nutrition Examination Survey (NHANES)	Retrospective, secondary data analysis (using longitudinal data in a format as cross-sectional data analysis)	a. <u>Sampling Frame</u> 1. Big sample size 2. Community dwelling sample b. <u>Characteristics of the Dataset</u> 1. Free, public accessible database 2. Wide breadth of available data 3. Have potential and flexibility to conduct longitudinal study c. <u>Required Resources</u> 1. Save time, cost, and resources in terms of collecting data compared to prospective study d. <u>Internal Validity</u> 1. Subjects are blind to the study	a. <u>Sampling Frame</u> 1. Not real-life data b. <u>Internal Validity</u> 1. Subjects whose responses are inconsistent (invalid person data) between 2 scales were not excluded in the analysis 2. The process to divide 20 items into 2 scales may not be theoretical valid based on Crimmins' categories, thus 2 scales may not be conceptually equivalent c. <u>External Validity (Generalizability)</u> 1. Restricted to the subjects who answered at least 75% of the total items (15 items); may not be able to

			generalize to the general population (because this population may have higher functioning) d. <u>Miscellaneous Factors</u> 1. Two scales were established from the same questionnaire thus have identical rating scale and contextual structures, which may produce results in favor of our hypothesis [secondary variance] 2. Data were collected not based on our research purpose; so the variables may have been defined or categorized differently than the research purpose [error] 3. The researcher/analyst does not know the exact data collection process (i.e., how the process was done and how well was done). Thud the researcher is not aware of important information such as if respondents understand specific survey questions.
Medicare Data	Retrospective, secondary data analysis (using longitudinal data in a format as cross-sectional data analysis)	a. <u>Sampling Frame</u> 1. Big sample size 2. Community dwelling sample 3. Real-life data b. <u>Characteristics of the Dataset</u> 1. Wide breadth of available data 2. Have potential and flexibility to conduct longitudinal study c. <u>Required Resources</u> 1. Save time, cost, and resources in terms of collecting data compared to prospective study d. <u>Internal Validity</u> 1. Subjects are blind to the study	a. <u>Characteristics of the Dataset</u> 1. High cost: expensive to purchase (especially for the government-monitoring database) 2. Not public accessible database b. <u>Miscellaneous Factors</u> 1. Data were collected not based on our research purpose; so the researcher/analyst does not know the exact data collection process [error] 2. Some important information may be lacking (i.e., drop-out rate) and these may lead to false results

		e. <u>External Validity (Generalizability)</u> 1. High generalizability to the general population	
Prospective Study to examine differences between the Continuity Assessment Record and Evaluation (CARE) item set and linking tool (FIM-MDS)	Prospective, cross-sectional	a. <u>Sampling Frame</u> 1. Community dwelling sample 2. Real-life, first-hand data b. <u>Internal Validity</u> 1. Data were collected based on research purpose; so the variables are defined based on the research purpose 2. Important information during data collection process can be recognized (i.e., drop-out rate) and help valid result interpretations c. <u>External Validity (Generalizability)</u> 1. High generalizability to the general population	a. <u>Sampling Frame</u> 1. May be difficult to recruit big sample size b. <u>Required Resources</u> 1. High cost 2. Require more time and resources c. <u>Internal Validity</u> 1. Difficult to “blind” subjects

APPENDIX- FIGURES

Figure 1.1. Continuum of Care in the United States HealthCare System (this picture is based on 5.0 percent national sample of 2006 Medicare claims)

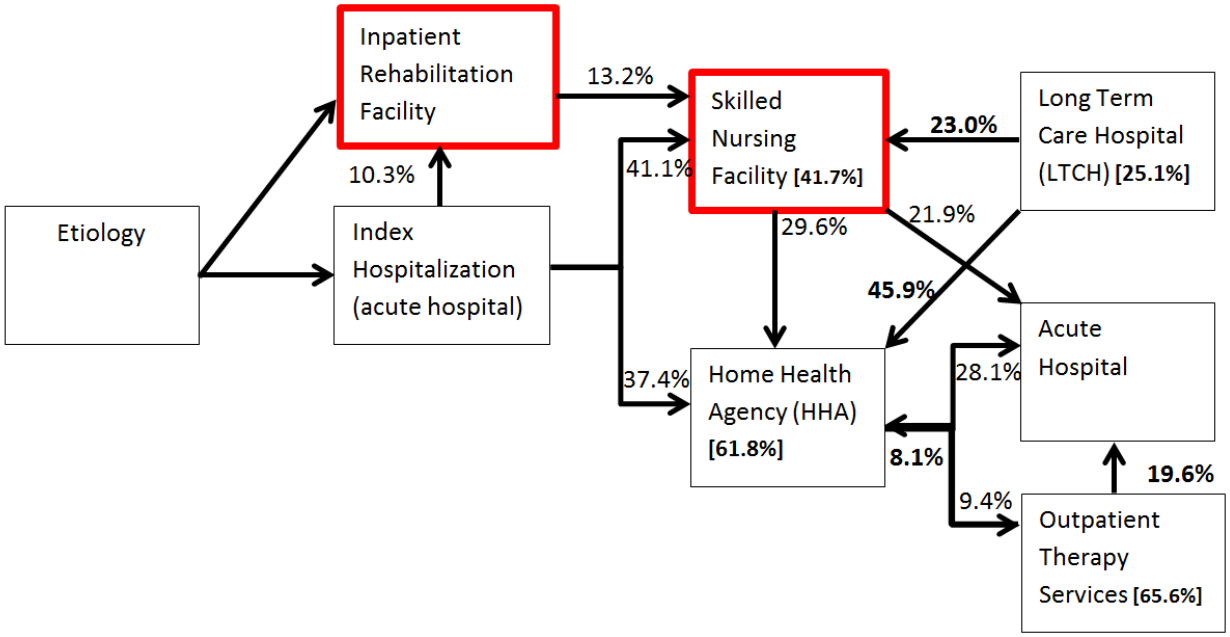
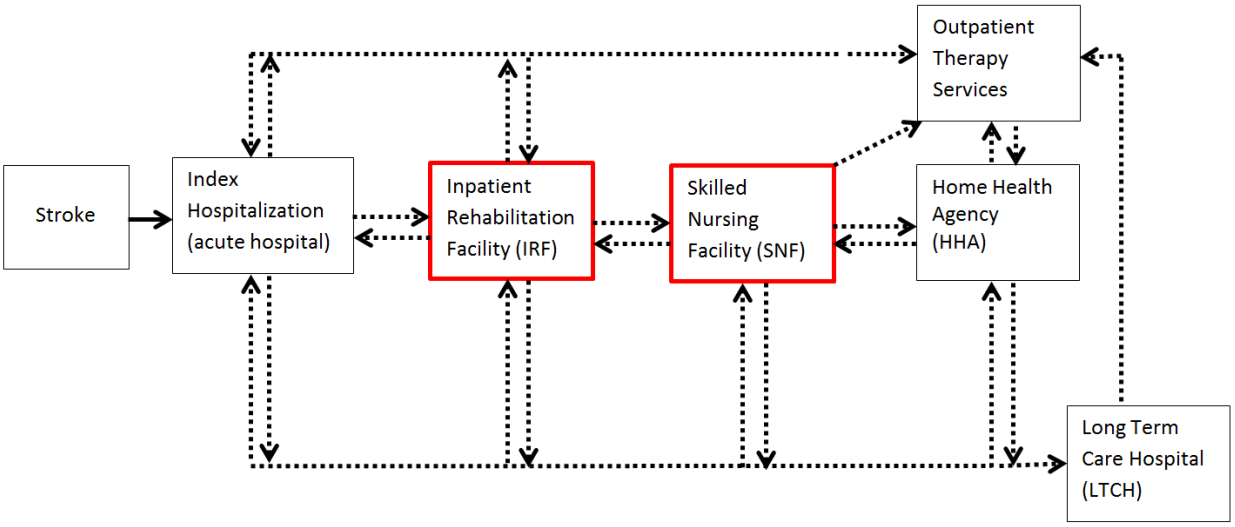


Figure 1.2. An Example: A trajectory of care for a person with stroke



NOTE: Dotted line: possible path

Figure 3.1. Study Procedure Diagram

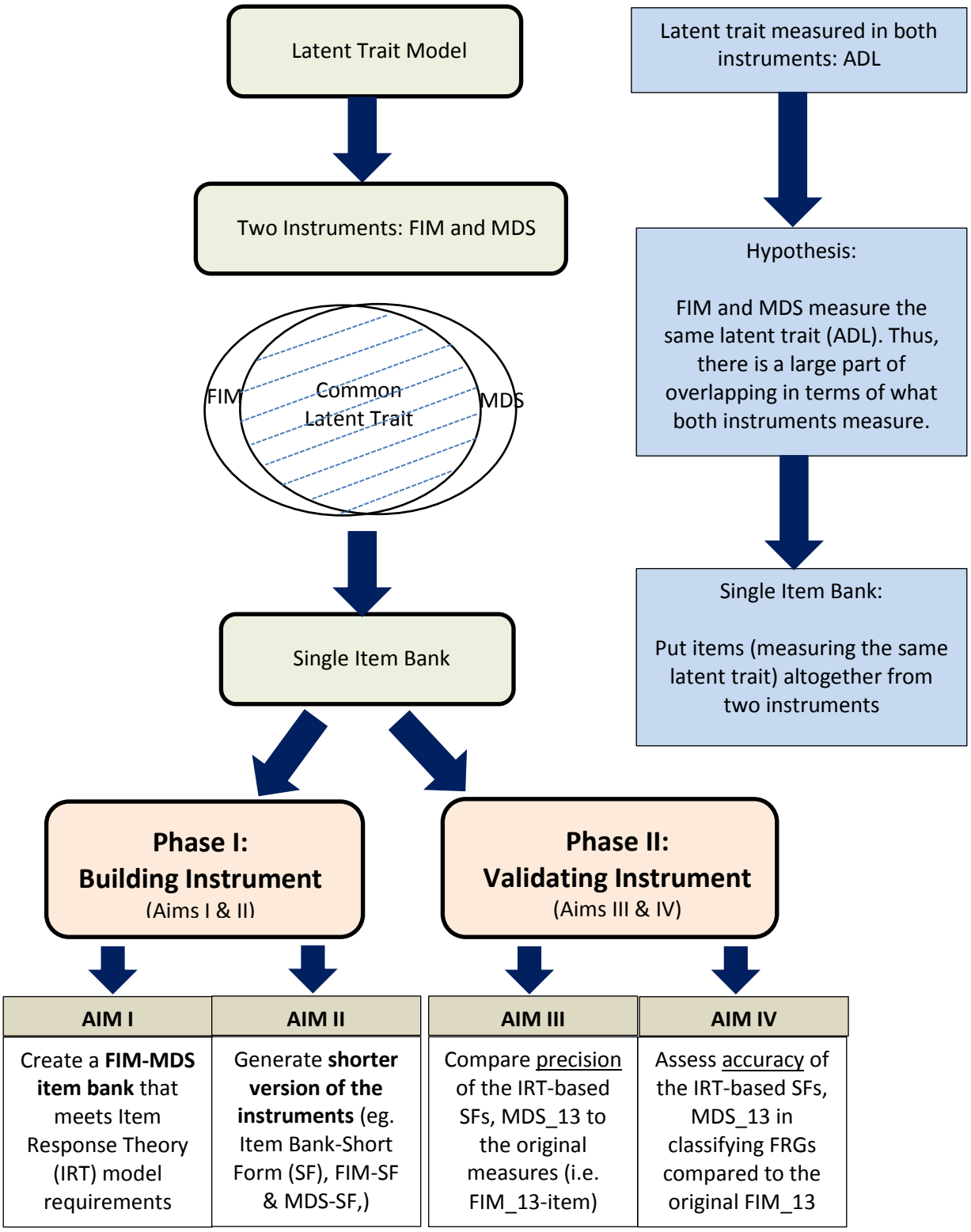


Figure 3.2. Rehabilitation Impairment Classification (RIC) for Stroke: Function Related Groups (FRGs) Algorithm (Impairment code: 1.1 to 1.9)

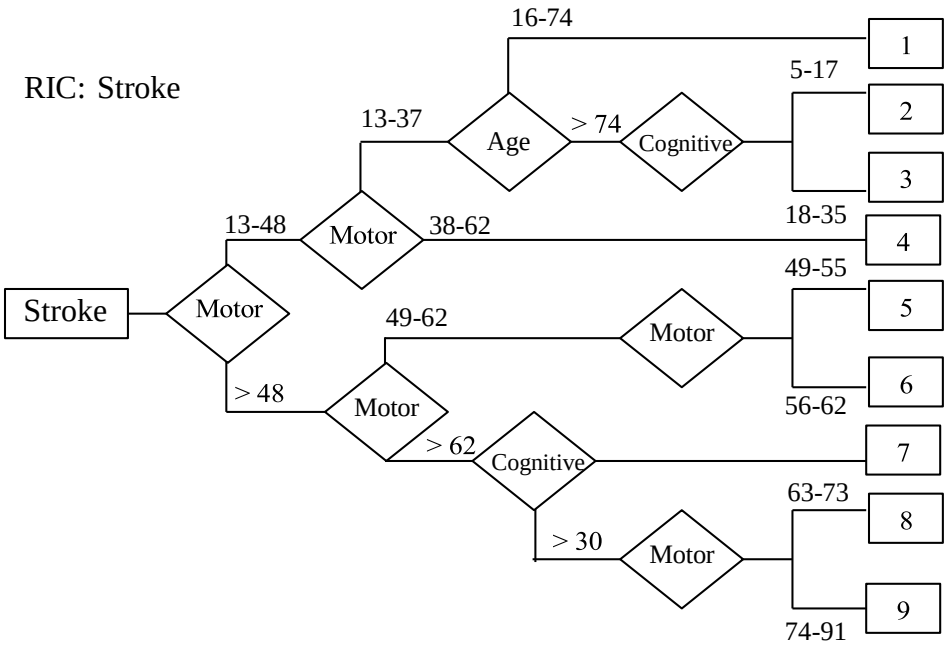


Figure 3.3. Rehabilitation Impairment Classification (RIC) for Lower Extremity Amputation: Function Related Groups (FRGs) Algorithm (Impairment code: 5.3 to 5.9)

RIC: Lower Limb Amputation (LLA)

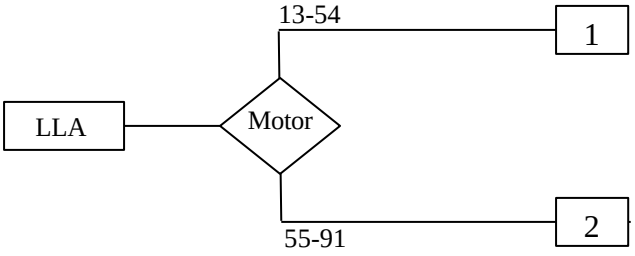


Figure 3.4. Rehabilitation Impairment Classification (RIC) for Knee Replacement: Function Related Groups (FRGs) Algorithm (Impairment code: 8.6 to 8.62)

RIC: Status Post Knee Replacement (SPKR)

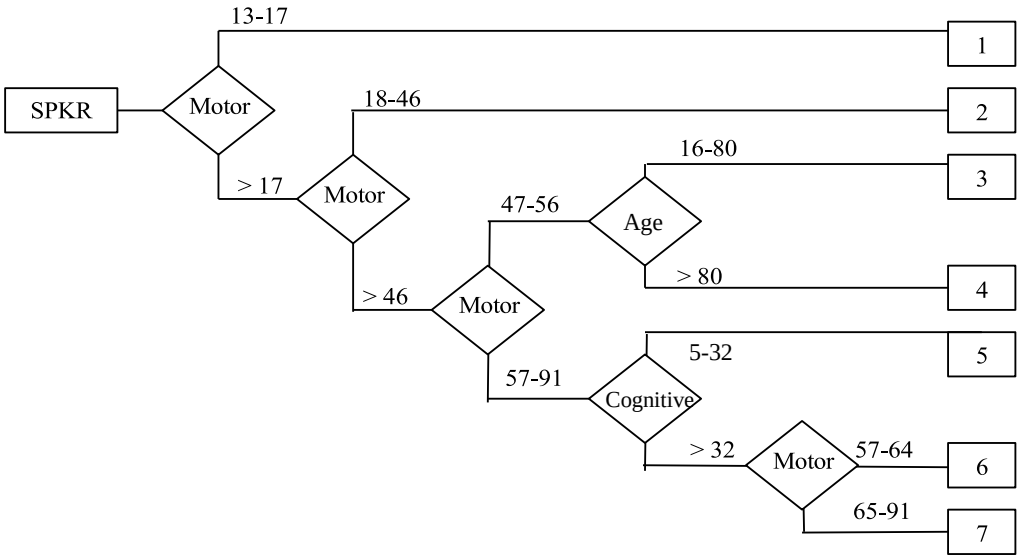


Figure 3.5. Rehabilitation Impairment Classification (RIC) for Hip Replacement: Function Related Groups (FRGs) Algorithm (Impairment code: 8.5 to 8.52)

RIC: Status Post Hip Replacement (SPHR)

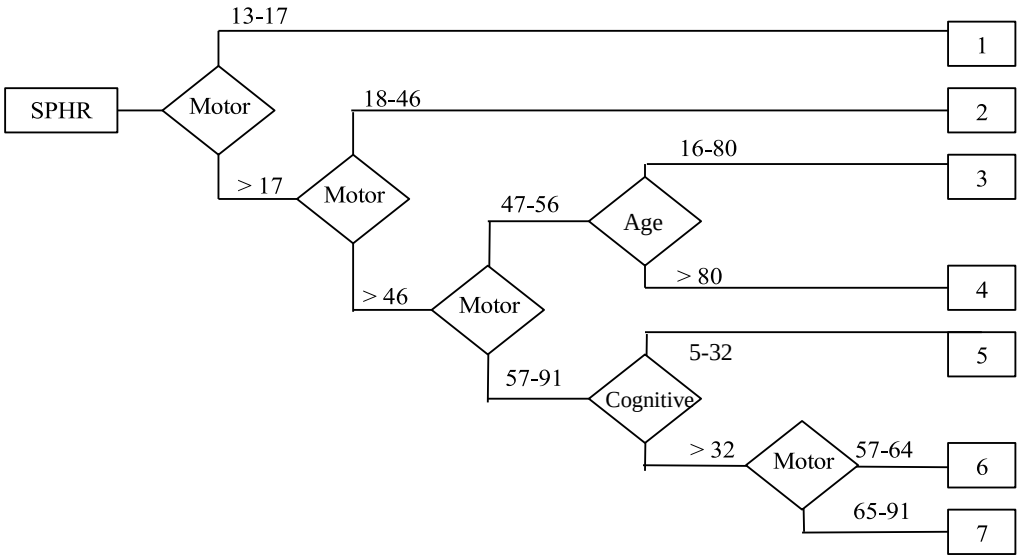


Figure 5.1. Visual Demonstration of Primary, Secondary and Error Variance in the Current Study Using MDS_13-item Converted Score

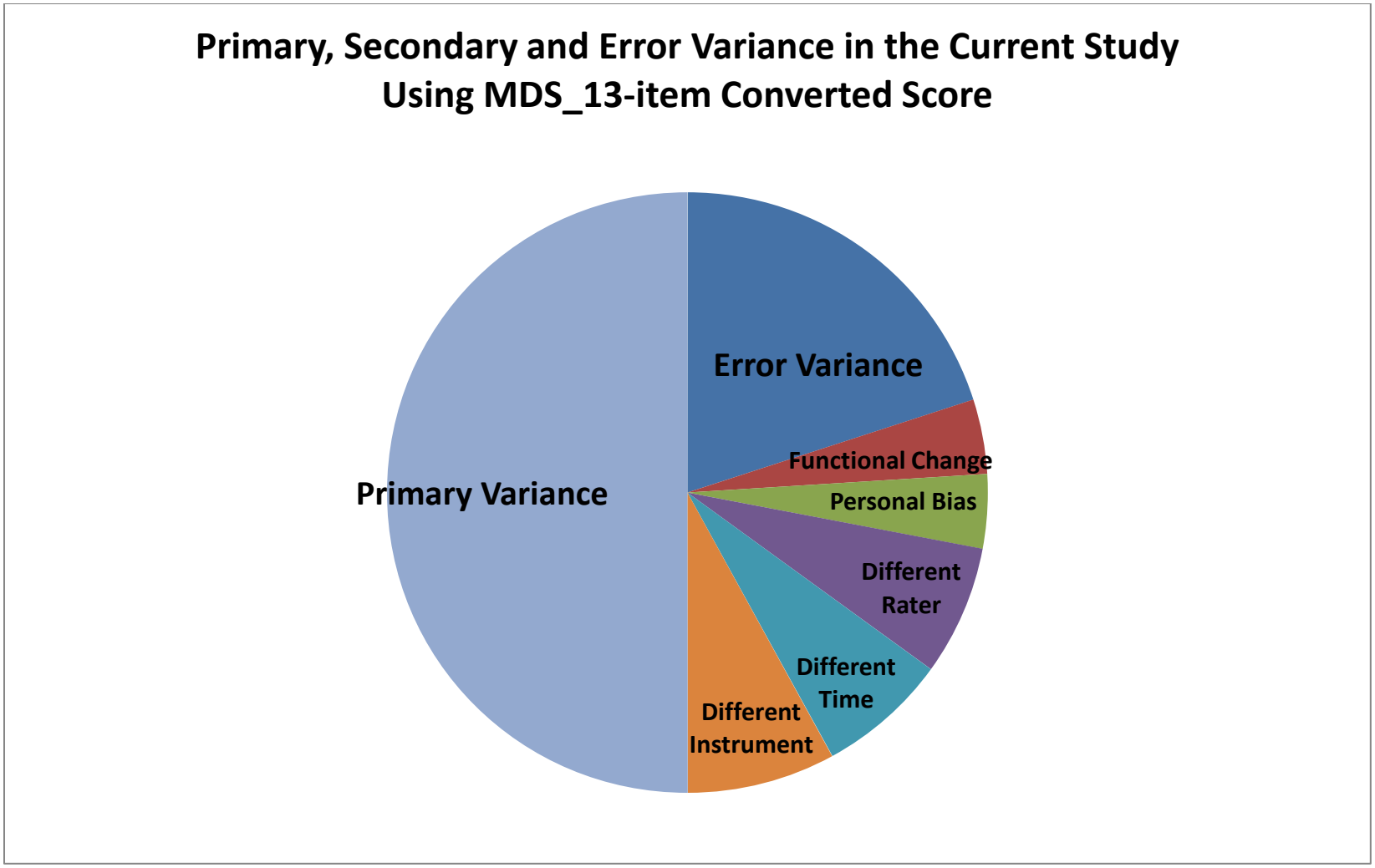


Figure 5.2. Visual Demonstration of Primary, Secondary and Error Variance in the Current Study Using MDS_4-item and 8-item Short Forms Converted Score

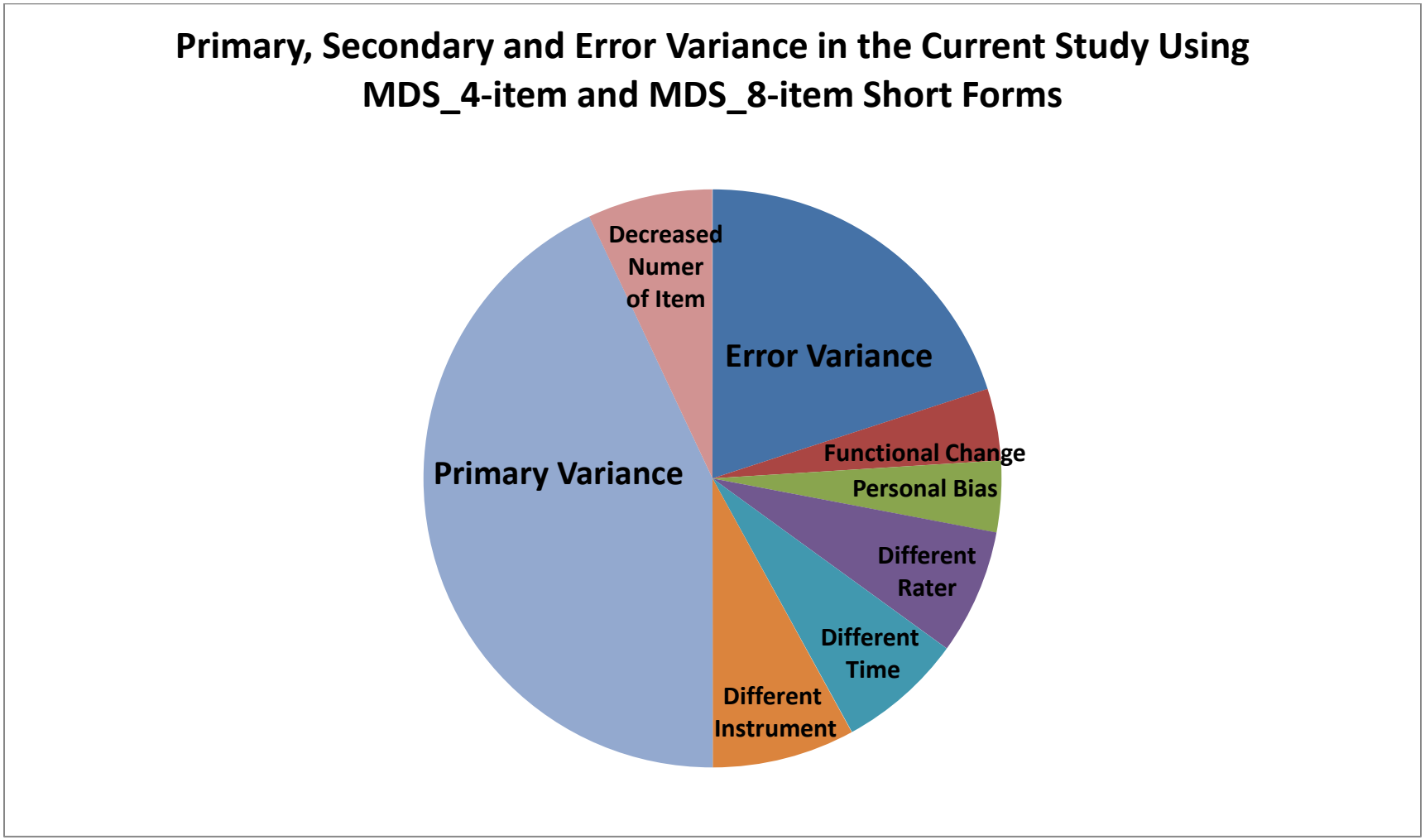


Figure 5.3. Visual Demonstration of Primary, Secondary and Error Variance in the Current Study Using FIM_4-item and FIM_8-item Short Forms Converted Score

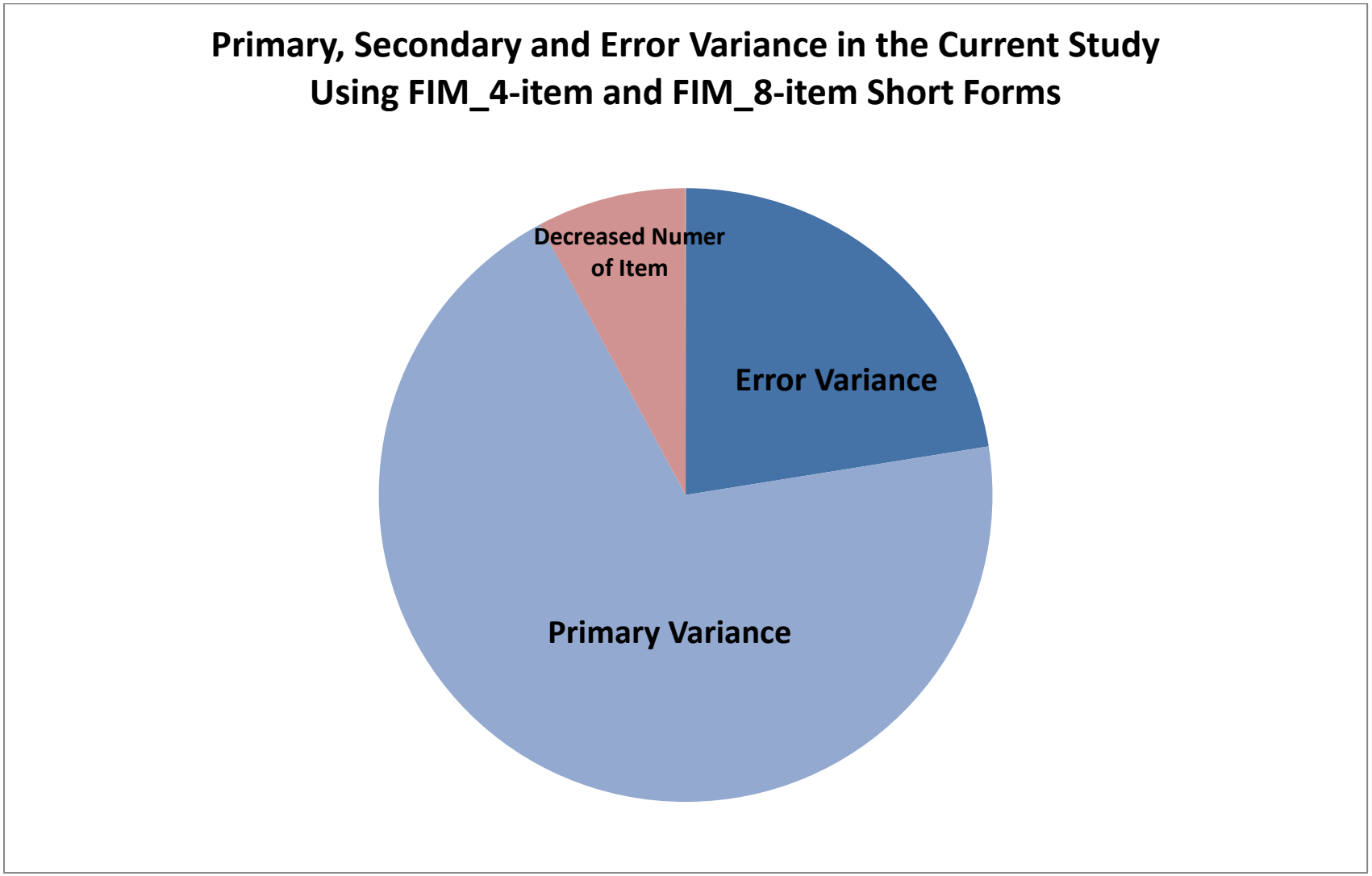


Figure 5.4. Visual Demonstration of Primary, Secondary and Error Variance in the Study Using a Single Universal Tool (e.g., CARE Item Set) across the Continuum of Post-acute Care

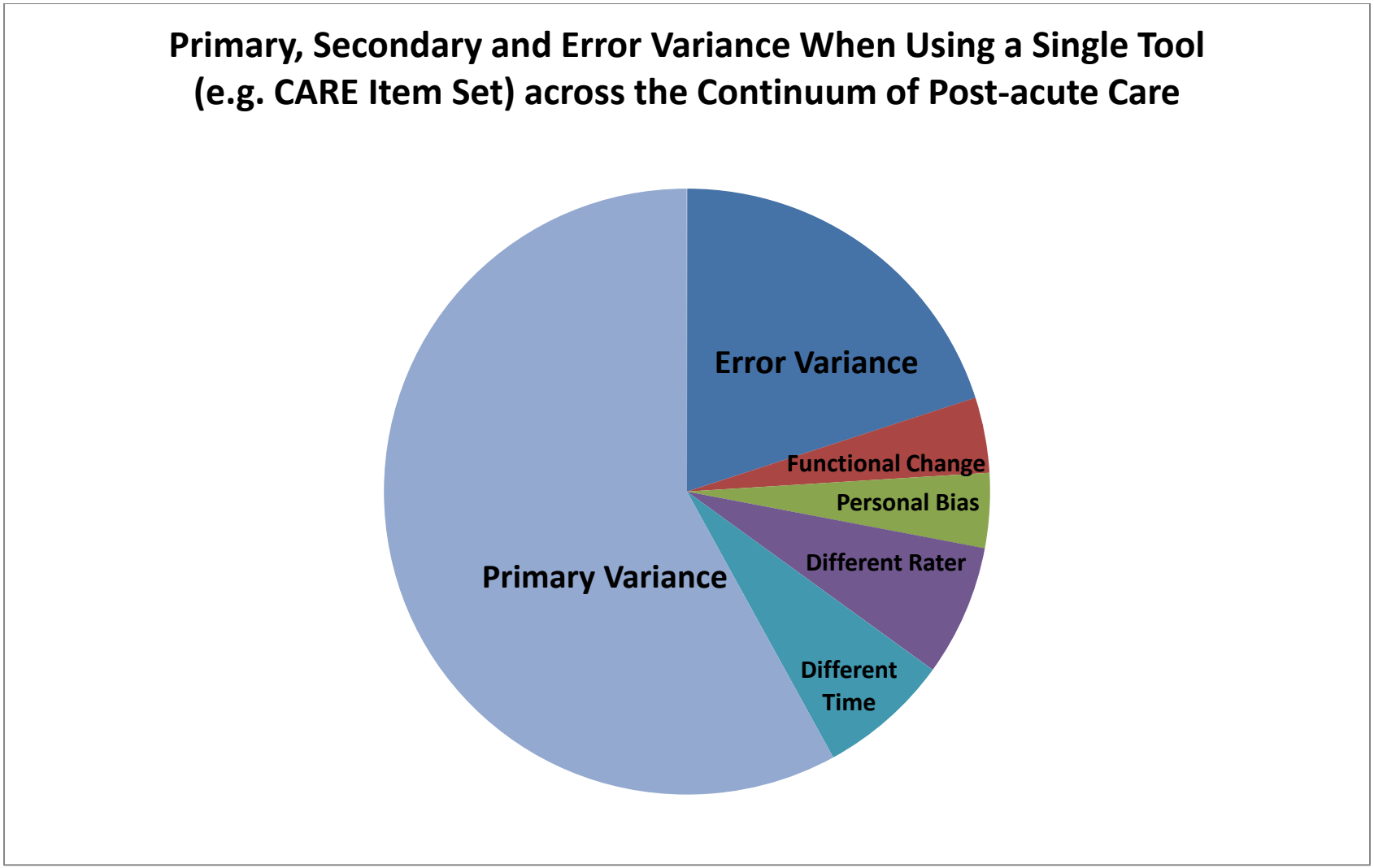


Figure 5.5. Visual Demonstration of Primary, Secondary and Error Variance in the Future Proposed Studying Using Two FIM Data for the Same Patient at the Same Facility across the Continuum of Post-acute Care