

University of Northern Colorado

Scholarship & Creative Works @ Digital UNC

Dissertations

Student Research

5-2021

Development of a New Foot and Ankle Activity Level Instrument Using The Rasch Model for Orthopaedic Clinical Application

Lauren M. Matheny

Follow this and additional works at: <https://digscholarship.unco.edu/dissertations>

Recommended Citation

Matheny, Lauren M., "Development of a New Foot and Ankle Activity Level Instrument Using The Rasch Model for Orthopaedic Clinical Application" (2021). *Dissertations*. 755.
<https://digscholarship.unco.edu/dissertations/755>

This Text is brought to you for free and open access by the Student Research at Scholarship & Creative Works @ Digital UNC. It has been accepted for inclusion in Dissertations by an authorized administrator of Scholarship & Creative Works @ Digital UNC. For more information, please contact Jane.Monson@unco.edu.

© 2021

LAUREN M. MATHENY
ALL RIGHTS RESERVED

UNIVERSITY OF NORTHERN COLORADO

Greeley, Colorado

The Graduate School

DEVELOPMENT OF A NEW FOOT AND ANKLE
ACTIVITY LEVEL INSTRUMENT USING THE
RASCH MODEL FOR ORTHOPAEDIC
CLINICAL APPLICATION

A Dissertation Submitted in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy

Lauren M. Matheny

College of Education and Behavioral Sciences
Department of Applied Statistics and Research Methods

May 2021

This Dissertation by: Lauren M. Matheny

Entitled: *Development of a New Foot and Ankle Activity Level Instrument Using the Rasch Model for Orthopaedic Clinical Application*

has been approved as meeting the requirement for the Degree of Doctor of Philosophy in the College of Education and Behavioral Sciences in the Department of Applied Statistics and Research Methods

Accepted by the Doctoral Committee

Susan R. Hutchinson, Ph.D., Research Advisor

William R. Merchant, Ph.D., Committee Member

Chia-Lin Tsai, Ph.D., Committee Member

Thomas O. Clanton, M.D., Committee Member

Joyce Weil, Ph.D., Faculty Representative

Date of Dissertation Defense March 31, 2021

Accepted by the Graduate School

Jeri-Anne Lyons, Ph.D.
Dean of the Graduate School
Associate Vice President for Research

ACKNOWLEDGEMENTS

I must first express my deepest gratitude to my research advisor, Dr. Susan Hutchinson. Your immeasurable support, constant encouragement, and limitless generosity with your time and mentorship have made my doctoral experience both enjoyable and successful. Thinking back to where I was four years ago and asking, “What is Rasch?” and seeing where I currently stand is because of you. Your ability to inspire confidence is simply amazing, and any student who has been lucky enough to have you as a professor knows this deeply. A true mentor understands the mentee’s abilities and perceived boundaries. You have consistently inspired me to push myself far beyond these boundaries, which has resulted in considerable educational and personal growth, and for this I am so grateful. Thank you for believing in me, investing in me, and inspiring me, always.

I would also like to extend my sincere thanks and appreciation to Dr. Thomas Clanton, renowned orthopaedic surgeon, rigorous investigator, and innovative leader. Over the past decade I have had the utmost pleasure leading your research team, working with you side by side, and learning so many invaluable lessons along the way. Thank you for always encouraging my professional and academic growth. You have always believed in me, completely trusting in me to guide your research and represent you across the world, and for that I want to say thank you. We have had incredible success as a research team and there is no one in this world I would have rather worked with all these years.

To Dr. Chia-Lin Tsai, Dr. Will Merchant, and Dr. Joyce Weil, I am deeply grateful for the time and effort you have invested in me by being on my dissertation committee. I have

enjoyed each of your unique perspectives which has only helped me to improve upon my study. It has been a pleasure getting to know you and working with you.

I would also like to offer a special thanks to the expert panel including Karen Briggs, Marilee Horan, Dr. Norm Waldrop, Dr. Jon Backus, Dr. Thomas Clanton, Ana Robinson, and Becky DeOliveira. Without your expertise, input and time, this study would not have been possible. Thank you so much for your participation.

To Keyleigh Gurney, administrative assistant extraordinaire, thank you for being such an amazing support. Anybody who knows you, knows you are the heart and soul of the department. You are a treasure.

To my loving parents, David and Denise Matheny, your unwavering support, love and encouragement throughout my entire life is the primary reason for any success I have experienced. Thank you for always being there for me, any time, day or night. I love you more than words can say, and I only hope that I can be the type of parent for Quinn that you have been for me. He would only be so lucky.

And to my amazing partner and best friend, Kevin Gittner, I cannot believe this is our journey. From colleagues to great friends and to partners in life, grad school has been full of surprises, but you are certainly the best one.

To my son Quinn, you are only six weeks old right now, but you have already had an incredible impact on me. You inspire me every day to be my best self. I love you more than anything.

TABLE OF CONTENTS

CHAPTER

I.INTRODUCTION TO THE STUDY	1
Evolution of Patient Reported Outcomes	1
Orthopaedic Patient Reported Outcome Measures	5
Normative Values	9
Psychometric Assessment of Instruments' Scores.....	12
The Rasch Model	14
Study Purpose and Research Questions.....	14
Study Significance	16
Chapter I Summary	17
II. BACKGROUND AND LITERATURE REVIEW	19
Item Response Theory	19
Rasch Model Overview.....	34
Chapter II Summary.....	61
III. METHODOLOGY	63
Phase 1	63
Phase 2	69
Phase 3	78
Phase 4	82
Chapter III Summary	84
IV. RESULTS	86
Phase 1	86
Phase 2	87
Phase 3	99
Phase 4	123
Chapter IV Summary	126

V. DISCUSSION AND CONCLUSIONS	129
Study Summary.....	130
Comparative Literature Discussion.....	130
Score Interpretations	133
Study Issues	134
Limitations	135
Future Research	137
Study Implications	138
Conclusions.....	139
REFERENCES	141

APPENDIX

A. Institutional Review Board Approval	172
B. Interview Questions and Prompts	174
C. Panel of Orthopaedic Experts' Questionnaire.....	175
D. Reading Experts' Questionnaire	176
E. Ankle Activity Level Questionnaire.....	177
F. Remaining 77 Items with Labels	206
G. Multiple Linear Regression Assumptions.....	209

LIST OF TABLES

TABLE

1. Sociodemographic Characteristics of Phase 3 Data.....	70
2. Descriptive Statistics for FAALS Percentage.....	80
3. Eigenvalues and Total Variance Explained by Factor.....	87
4. Component Matrix for Phase 2 Data (Pilot) 77 Items	89
5. Mean-square Infit and Outfit: Phase 2 Data, 77 Items.....	92
6. Eigenvalues and Total Variance: Phase 3 Data	100
7. Component Matrix: Phase 3 Data, 66 Items	102
8. Mean-square Infit and Outfit: Phase 3 Data, 66 Items.....	105
9. Mean-square Infit and Outfit: Phase 3 Data, 62 Item	107
10. Mean-square Infit and Outfit: Phase 3 Data, 35 items	111
11. Final 22 Items for Phase 3: Highest to Lowest Difficulty	114
12. Mean-square Infit and Outfit Values: Phase 3 Data	115
13. Pearson Correlation of FAALS with Common Foot and Ankle Measures .	121
14. Multiple Linear Regression with FAALS as Dependent Variable	122
15. Normative Values Based for FAALS Aggregate Percentage	123
16. Final 22 Items: Highest to Lowest Difficulty	124
17. Raw Points and Percentages to Assign Activity Levels	125
18. Activity Level Frequencies for FAALS Scores	126

LIST OF FIGURES

FIGURE

1. Item Characteristic Curve	23
2. Category Response Curve.....	25
3. Rasch Measurement Model.....	36
4. Wright Item Person Map	50
5. Example of Item Information for Four Items	52
6. Test Information Function	53
7. Category Probability Curves.....	55
8. Scree Plot: Phase 2 Data (Pilot), 77 items	88
9. Wright Item Person Map: Phase 2 Data (Pilot)	95
10. Test Information Function for Phase 2 Data (Pilot)	96
11. Category Probability Curves for Item 1 of Phase 2 Data (Pilot)	97
12. Item Characteristic Curve for Item 1 of Phase 2 Data (Pilot).....	98
13. Wright Item Person Histogram: Phase 2 Data (Pilot).....	99
14. Scree Plot: Phase 3 Data, 66 items.....	100
15. Wright Item Person Map: Phase 3 Data, 62 Items.....	109
16. Wright Item Person Map: Phase 3 Data, 33 Items.....	112
17. Wright Item Person Map: Phase 3 Data, 22 Items.....	117
18. Test Information Function for Phase 3 Data.....	118
19. Category Probability Curves for Item 1 of Phase 3 Data.....	119
20. Item Characteristic Curve for Item 1 of Phase 3 Data	119
21. Wright Item Person Histogram: Phase 3 Data	120

CHAPTER I

INTRODUCTION TO THE STUDY

Patient-reported outcomes, commonly referred to as PROs, are one of the most valuable tools in assessing general health status of patients, health-related quality of life, functional levels, and symptoms (Abrams, 2017; Acquadro et al., 2003; Bingham et al., 2017). Understanding the patient's perspective of their own functional and symptomatic levels is important for clinicians for several reasons (Acquadro et al., 2003). PROs have been considered invaluable for understanding the impact of a disease on patients, guiding treatment decisions, evaluating treatment efficacy, and interpreting clinical outcomes (Acquadro et al., 2003; Bingham et al., 2017; Wu et al., 2013). More and more electronic patient record systems are beginning to incorporate the collection of patient-reported outcomes (PROs) for the purpose of clinical care and health-related research (Wu et al., 2013).

Evolution of Patient Reported Outcomes

There are multiple ways to measure health status, which can include perspectives from the physician or the patient. The field of health status assessment originated approximately 30 years ago (McHorney, 1999). Historically, health status has been assessed through the clinician's vantage point; however, the vantage point began to shift, with the patient's perspective gaining traction in the field of health care research (McHorney, 1999; Wu et al., 2013). Patient-reported outcomes refer to the patient's perspective of their outcome and their own assessment of themselves, rather than the assessment of the patient by the physician. Terms such as health status

and health-related quality of life (HRQOL) were used to refer to patient outcomes, and in the 1990s, the measurement of these concepts began to rapidly evolve (Wu et al., 2013). It was not until 2001 that the term patient-reported outcomes was first proposed by the Food and Drug Administration (FDA). The FDA and the PRO Harmonization Group worked collaboratively to define PROs to include any outcome based on data provided by patients or patient proxies, as opposed to data provided from other sources (Acquadro et al., 2003). In fact, research that seeks to compare effectiveness of different treatments has commonly utilized data sources such as administrative and billing data like current procedural terminology codes, clinical physician-reported data, and medical records (Wu et al., 2013). Although these data sources have been commonly used throughout clinical research, the patient's perspective is often missing from the assessment. This can be especially misleading for physicians since they need to understand the patient's assessment of their own limitations in order to determine the most appropriate intervention.

Throughout the latter half of the 20th century, PROs evolved at a rapidly increasing rate. Prior to the 1960s, health status was often measured by single-item measures, that were reported from the physician's perspective, such as Karnofsky's Performance Status Scale (Karnofsky et al., 1948), which is a single item measure of functional status on an 11-point scale, with 11 representing no evidence of disease and zero representing death (Karnofsky et al., 1948; McHorney, 1999). Single item measures can be especially problematic when attempting to estimate reliability since internal consistency cannot be measured when there is only one item (Cronbach, 1951; Lucas & Donnellan, 2012).

In the 1970s, more generic outcome measures with multi-item subscales began to emerge. In fact, from 1970 to 1981, 12 different generic outcome measures were developed (McHorney,

1999). It was not until the 1980s that the focus shifted to condensed general health measures in order to increase clinical efficiency, while also remaining cost-effective for large-scale clinical trials and population-based outcomes research (McHorney, 1999). This is the era in which the term “short-form” (SF) first appeared, and from which the commonly used SF-12 Health Survey (Ware et al., 1996) was derived in an attempt to fully measure two domains of general health, including physical and mental, while maintaining brevity (McHorney, 1999; Ware et al., 1996). In the very early days, disease-specific measures were reserved for diseases with very severe effects, such as cancer and osteoarthritis (Karnofsky et al., 1948; Steinbrocker et al., 1949). However, during the 1980s and 1990s there was another shift from focusing on general health measures to focusing on disease-specific and body region-specific measures (McHorney, 1999). Disease- and region-specific measures are necessary because general health measures are less sensitive to change following disease- and region-specific interventions (Bergner et al., 1981). These types of measures contribute a unique component to the complex composition of the patient continuum of care and recovery.

Disease- and body region-specific instruments are often used to help patients understand their functional limitations, guide decision-making for treatment, and provide documentation of changes in functional improvements or declines, which can help increase overall patient outcomes (van Cranenburgh et al., 2012). The World Health Organization (WHO) has sought to define disability, impairment, and functional limitations, with functioning referring to all body functions, activities, and participation, while disability seems to encompass impairments, activity limitations and participation restriction. The WHO has shifted away from this idea and has moved toward the thinking that individuals can have a reduction in health and thereby experience some increment of disability (World Health Organization, 2002). However, these terms are still

used interchangeably, which may be due to the fact that the word disability was previously used to describe and measure health status. The Center for Disease Control (CDC) has defined disability as “any condition of the body or mind (impairment) that makes it more difficult for the person with the condition to do certain activities (activity limitation) and interact with the world around them (participation restrictions)” (Center for Disease Control and Prevention, 2020, pp1). According to the WHO and CDC, impairment refers to one’s body structure or function (physical or mental), and disability represents how functional limitations affect an individual’s life (Center for Disease Control and Prevention, 2020; World Health Organization, 2002). Previous research that has developed a sociomedical model of disability also supports this idea that functional limitations present in daily life as disability (Verbrugge & Jette, 1994). Therefore, foot and ankle activity level is considered a scale to measure activity or functional limitations due to an impairment, which is one piece of a person’s perceived disability.

PROs that are region-specific are essential to measure the value and efficacy of conservative and surgical interventions (Abrams, 2017). In order to properly assess whether medical treatment was beneficial to the patient, it is crucial to measure changes in function, pain, and activity level (Abrams, 2017). By recording these measurements prior to treatment and throughout the process of recovery, deficits in function and symptoms of pain can be identified, enabling the physician to tailor medical care on an individual basis. When deciding which treatment option may be the most appropriate, there are a variety of factors that influence patient outcome, including a patient’s own impression of their functional disability, burden of the disease, symptoms and treatment response (Bingham et al., 2017). PROs also facilitate better communication between physician and patient, allowing physicians to manage patient

expectations by discussing the patient's priorities and realistic outcomes prior to and following intervention (Abrams, 2017; Pogorzelski & Millett, 2017; Stüber et al., 2010).

Orthopaedic Patient Reported Outcome Measures

For each area of specialization within the medical field, various patient reported outcome measures have been developed. In the field of orthopaedics, various instruments have been utilized, including some general and some anatomically specific. These evaluative instruments are used to measure an individual's change in function and activity level over time following treatment for foot and ankle injuries (Budiman-Mak et al., 2006; Button & Pinney, 2004; Carcia et al., 2008; Chen et al., 2012; Farrugia et al., 2010; Goldstein et al., 2010; Golightly et al., 2014; Herron, 2006; C. E. Hiller et al., 2006; Hung et al., 2013; Hunt et al., 2014; Hunt & Hurwit, 2013; Martin et al., 2005; McKeown et al., 2019; Muller & Roddy 2009; Niki et al., 2005; Ortega-Avila et al., 2019; Ponkilainen et al., 2020; Richter et al., 2018; Richter et al., 2006; Riskowski et al., 2011; Sadjadi et al., 2014; Sierevelt et al., 2018; Veltman et al., 2017; Yusuf et al., 2019; Zwiers et al., 2018). Aspects of lower extremity musculoskeletal injury that have been identified as most important to patients, and therefore most influential following intervention, include the patient's perception of their own functional limitations and disability, which is why quantifying deficits at the patient level is imperative. A very common evaluative instrument includes the region-specific, non-disease-specific outcome measure of foot and ankle function, the Foot and Ankle Ability Measure (Martin et al., 2005). The FAAM is comprised of an Activities of Daily Living (ADL) subscale and a Sports subscale, for which scores are reported as a percentage, with 100% representing no disability and 0% representing complete disability. Scores based on these instruments have shown evidence of excellent reliability and validity when measuring ankle function; however, this instrument does not measure activity level (Matheny et al., 2020). Additionally, the other two most commonly reported region-specific outcome

measures include the Foot Function Index (FFI) and the Foot and Ankle Outcome Score (Sierevelt et al., 2018). The FFI consists of three subscales including pain, disability, and activity limitations in the past two weeks (Budiman-Mak et al., 2006), while the FAOS consists of 42 items that comprise five subscales including pain, symptoms, activities of daily living, sport and recreation, and foot- and ankle-related quality of life (Roos et al., 2001). There are many more outcome measures that have been documented in the foot and ankle literature; however, there is no region-specific instrument that measures activity level in the foot and ankle.

Many different individuals with a large range of ages, with a large variety of orthopaedic forefoot, hindfoot, and ankle injuries and/or disorders, including arthritis, trauma, fusion, sprain, arthroplasty, arthroscopy, tendon, cartilage, infection, dislocation, paralysis, tumors, arthropathies, circulation disorders, and many more, are seen clinically by orthopaedic surgeons (Garratt et al., 2018; Golightly et al., 2014; Hung et al., 2019; Hunt et al., 2014; Niki et al., 2005). It is important to note that while many athletes are examined by orthopaedic surgeons, not all patients who are seen by orthopaedic surgeons are elite level athletes, or even athletes for that matter, and many injuries are not due to trauma (Garratt et al., 2018; Hung et al., 2019; Hunt et al., 2014; Niki et al., 2005). However, it is still very important for physicians to understand a patient's activity level, whether that be a lower level or a higher level. Understanding a patient's activity level is a clear indication of how the patient is progressing through the continuum of care. Activity level is a benchmark used by physicians to gauge where the patient is in the recovery process (Brophy et al., 2014). Patient activity level is imperative to inform physician assessment and to ascertain whether the patient has returned to their expected and/or desired activity level. This information can then be communicated to the patient, which has the potential to empower and encourage the patient or identify lags early in the recovery process (McHorney,

1999). Understanding how the patient compares to others with similar demographics, who have undergone a similar intervention, yields information that can also be used to tailor individual follow-up care and recommendations for supplementary treatments, such as physical therapy, prosthetics, pain-relieving or biologic injections, or additional surgical interventions.

An evaluative instrument that serves as a measure of activity level in the knee is the Tegner activity scale which is a numerical scale with each level (range 0 – 10) representing specific activities (Tegner & Lysholm, 1985; Tegner et al., 1986, 1988). An elite athlete is considered to have an activity level of a 10, a collegiate athlete is expected to have a value of 9, a recreational athlete may have an activity level within the 4 to 6 range, and sick leave from work or disability pension due to knee problems is considered to have a value of zero (Briggs, Steadman, et al., 2009b). Each level, 0 through 10, has a list of activities or sports that represent each activity level. The Tegner activity scale was originally developed as a lower extremity scale to measure activity level in patients with anterior cruciate ligament (ACL) injuries of the knee (Tegner & Lysholm, 1985; Tegner et al., 1986, 1988). Reliability and validity of Tegner scores were assessed in several different disease-specific samples with various knee pathologies (Briggs et al., 2006; Briggs, Lysholm, et al., 2009a). Results have shown good evidence of reliability and validity in patients with knee pathologies that were surgically treated (Briggs et al., 2006).

Although this scale has performed well in the knee, there are several limitations in its use. The Tegner activity scale is commonly used in the knee, not the foot and ankle. Although originally developed as a lower extremity instrument to measure activity level in patients with ACL injuries of the knee (Tegner et al., 1986) it is unclear whether the instrument is effective in measuring ankle-based activity level. The Tegner scale has been used for patients with foot and ankle injuries due to the lack of other instruments available for outcome measurement. The Tegner

activity scale is a one response item aimed to measure the latent construct of foot and ankle activity. When measuring a construct, it is necessary to ensure that the construct is being measured in its entirety, which is very difficult to accomplish with a single item (McHorney, 1999; Rousson et al., 2002). Since the Tegner activity scale is a single item, assessment of internal consistency cannot be performed. Therefore, reliability would need to be measured using test-retest, which is not as accurate in terms of measurement due to learner effect, nor is it cost or time effective (McHorney, 1999; Rousson et al., 2002). Foot and ankle activity is a complex construct that is comprised of various components. An activity scale that consists of multiple items, rather than a single item that lacks sufficient breadth of content to describe activity level effectively, could decrease measurement error and thereby increase reliability. When utilizing any instrument, it is crucial to have good evidence of reliability and validity, in order to ensure that the construct is consistently being measured, and that the instrument is measuring the correct construct (Messick, 1989). Additionally, the Tegner activity scale has items in the upper portion of the scale that would never be used unless studying an elite population, which is rare in the majority of foot and ankle practices. Therefore, the upper 25% of response options on the Tegner scale may never be selected in most clinical practices. This is very problematic when using the Tegner scale as a criterion variable, with the goal of making inferences based on those responses. Therefore, it is necessary to develop an instrument that can consistently and effectively measure activity level in the foot and ankle within a more typically functioning adult population rather than a combined population including elite athletes.

Therefore, the primary reasons a foot and ankle activity level assessment is needed is:

- Activity level is a benchmark to document patient recovery and what patient can and cannot do in terms of activities of daily living and recreation. Currently, there is no instrument that measures this construct.
- The Tegner activity scale was designed for use in individuals with anterior cruciate ligament injuries of the knee, not injuries of the foot and ankle. The Tegner activity scale is a one-item measure meaning that scores cannot be rigorously tested for psychometric properties, and the scores most likely have higher measurement error since the instrument is only one item. The scale is European, so it would require multiple modifications for clarity and comprehension and would require additional testing which does not make logical sense to do since the instrument would remain a one-item measure with potentially high measurement error. With all of these issues it makes logical sense to develop an instrument to specifically measure foot and ankle activity level.

Normative Values

When developing and utilizing an instrument to assess activity level in patients with foot and ankle pathologies, it is important to first determine normative values in the population of interest (Dingemans et al., 2017; Matheny et al., 2020; Pinsker & Daniels, 2011). Normative values are defined as the random sample of individuals from the general population of the United States from whom scale scores are derived, which was previously defined by the American Academy of Orthopaedic Surgeons (Hunsaker et al., 2002). By first determining normative values and assessing reliability and validity of scores in the normal, healthy population, these scores can serve as a reference group for physicians, consequently improving the interpretation

of scores (Dingemans et al., 2017; Giesinger et al., 2019; Hunsaker et al., 2002; Matheny et al., 2020; Pinsker & Daniels, 2011). While comparing baseline and post-treatment scores is imperative to evaluate treatment efficacy, scores from a normal, healthy population must also be determined in order to effectively define abnormal scores and to document full recovery (Sallay & Reed, 2003; Schneider & Jurenitsch, 2016a, 2016b). Additionally, to determine a realistic outcome for an individual patient, it is important to first establish normative values for commonly reported outcomes measures so that physicians may be better prepared to counsel patients about appropriate and realistic expectations, based on population characteristics, such as age, body mass index (BMI), and gender, as well as other factors including previous surgery (Matheny et al., 2020). With normative data serving as a reference, medical practitioners will be able to assess whether baseline outcome measurements, as well as follow-up measurements, differ from population norms, and whether the patient has returned to a “normal” level after treatment based on their specific characteristics (Hunsaker et al., 2002).

There has been increased attention regarding normative values for patient-reported outcomes, which may be due to the benefits that are realized when physicians can relay normative data values to their patients (Mancuso et al., 2001, 2003). A previous study reported the importance of shared decision-making in the medical care and treatment process (Henn et al., 2011). Younger patients who had shoulder injuries have been shown to have increased expectations of surgical intervention when compared to older patients, specifically in the number of expectations, as well as in the areas of nighttime pain relief, improved ability to exercise or participate in sports, and improved ability to interact with others (Henn et al., 2011). Other studies that have focused on hip arthroplasty have shown that patients with worse preoperative function, males, and older individuals tend to have higher expectations of following treatment

(Mancuso et al., 2003). Yet a study that determined expectations of patients with knee injuries revealed that patient expectations differed by diagnosis, patient characteristics, and demographics, as well as function (Mancuso et al., 2001). These studies underscore the need to determine baseline values in ankle outcomes instruments.

Normative values have been established for a variety of other outcomes scores that have previously been documented in the musculoskeletal literature (Hunsaker et al., 2002). Over the past decade, normative data have been recognized as a fundamental element in the interpretation of patient reported outcomes. In 2016, normative values for the American Orthopedic Foot and Ankle Society (AOFAS) ankle-hindfoot, midfoot and hallux, and lesser toes clinical rating system were determined. Results revealed that normative values differed based on population differences, such as age and gender (Schneider & Jurenitsch, 2016b). However, the AOFAS scores have previously demonstrated poor psychometric properties in patients who have been treated for foot and ankle injuries, especially in terms of a high ceiling effect, making interpretation of the scores quite ambiguous (Donnenwerth & Roukis, 2012; Pinsker & Daniels, 2011). In 2011, normative values for the Visual Analogue Scale Foot and Ankle were also established (Stuber et al., 2011). The American Academy of Orthopaedic Surgeons (AAOS) Outcomes Studies Committee, in collaboration with the Council of Musculoskeletal Specialty Societies and the Council of Spine Societies, conducted a study to determine normative values in 11 specific outcomes instruments that were developed by AAOS, which demonstrates the emphasis on establishing normative values by the orthopaedic community (Hunsaker et al., 2002). In fact, AAOS acknowledges the extreme importance of normative values in order to determine whether patients treated for specific conditions have returned to, or at least come closer to, normative ranges of functioning (Hunsaker et al., 2002).

Psychometric Assessment of Instruments' Scores

Assessing evidence of reliability and validity is a crucial step in the development of a new instrument used to assess PROs (Abrams, 2017; Boone, 2016; Burton & Mazerolle, 2011; Cappelleri et al., 2014; Cronbach, 1951; DeVellis, 2006; Embretson & Reise, 2013; Embretson, 1985; Hays et al., 2000; Messick, 1989; Smith, 2001; Tennant et al., 2004; Wolfe & Smith, 2007; Zanon et al., 2016). In order to conduct these assessments, classical test theory (CTT) has commonly been used. CTT is based on the principle that the observed score is partitioned into the true score plus error, such that $X = T + E$, meaning [observed score = true + error].

CTT encompasses a set of concepts and methods that have been used to develop and assess many health-related instruments. Procedures that are associated with CTT include calculating Cronbach's alpha coefficient for an estimate of internal consistency reliability, test re-test reliability or sensitivity, dimensionality assessment, correlation analyses with other measures and more (DeVellis, 2006). Although newer measurement approaches have been developed, CTT remains popular for several reasons. Advantages of CTT include familiarity of basic concepts and accessibility of software that is capable to perform procedures, relatively straightforward mathematical computations, use with relatively smaller samples, and the underlying model fits certain types of instruments fairly well, such as Likert-type items that are aggregated (Cappelleri et al., 2014). However, there are also limitations to using CTT, such that CTT has difficulty in differentiating between common themes across items that are of importance to the researcher and other themes that are not of importance, such as similar wording for multiple items (DeVellis, 2006). Methods and procedures that are CTT-based do not generally involve rigorous scrutiny of item characteristics, which can make interpretations difficult, especially if there is differential sensitivity at the center of the scale (DeVellis, 2006). In addition, parameter estimates that are produced using CTT methods are dependent on

characteristics of the sample being used for the analysis. In other words, different samples with different examinee characteristics and variability will most likely yield different item characteristics, limiting generalizability, which may be the greatest disadvantage (DeVellis, 2006).

A more modern alternative to CTT that has been found useful in the development of orthopaedic outcome instruments is the Rasch rating scale model, which is a specialized form of item response theory (Andrich, 1978b; Rasch, 1980; Tennant et al., 2004; Wright & Masters, 1982). Rasch analysis is a probabilistic mathematical technique that is used to assess psychometric properties of outcome measures. The Rasch model has two main principles that are 1) the easier the item on the scale, the more likely that the item will be chosen or “passed,” and 2) the higher the ability or trait level a person possesses, the more probable the person will “pass” or endorse an item compared to a person with a lower ability or trait level (Tennant et al., 2004). The scale of interest is measured in terms of item difficulty and generates estimates of locations of individual items (item difficulty) and ability level along a common interval-level scale (log-odds). Rasch assumes unidimensionality of either a single, primary domain or of each subdomain if the construct is multidimensional, where unidimensionality is the extent to which items measure a single construct and is a fundamental requirement of construct validity; item difficulty, which is the relative difficulty of the items when compared to one another; and person separation, which is the extent to which items distinguish between distinct levels of functioning (Tennant et al., 2004). Properties of the Rasch model can be extended to fit graded responses such as Likert-type items. The model accounts for response options not equally spaced in terms of ability and is able to identify poorly functioning items. Additionally, the calibration of the items within a scale is independent of the persons used to calibrate the instrument and vice versa.

In addition, patients and items are calibrated on the same underlying metric trait, which increases generalizability across samples. The model is also capable of examining differential item functioning (DIF), which can tell researchers whether a scale works in the same way, regardless of the sample or group being assessed (Tennant et al., 2004).

The Rasch Model

The Rasch model was originally developed for use in the education field, and was first introduced into the field of health and medicine in the late 1980s (Rasch, 1980). Specifically, physical therapists adopted the approach since the field of physical therapy is very well suited for the application of Rasch modeling in terms of language compatibility when discussing ability and difficulty. Since that time, the Rasch model has become the gold standard in assessing psychometric properties of quality of life instruments (Anselmi et al., 2015; Tennant et al., 2004). Previous studies have also utilized the Rasch model to assess psychometric properties of scores from various orthopaedic instruments in the ankle, knee, and shoulder joints (Budiman-Mak et al., 2006; Budiman-Mak et al., 1991; Conaghan et al., 2007; Franchignoni et al., 2010; Hamilton et al., 2015; C. E. Hiller et al., 2006; A. M. Keenan et al., 2007; Ko et al., 2009; Lin et al., 2009; Matheny et al., 2020; McPhail et al., 2014; Muller & Roddy, 2009; Ponkilainen et al., 2020; Sadjadi et al., 2014; Woodburn et al., 2012).

Study Purpose and Research Questions

The purpose of this proposed study was to develop a Foot and Ankle Activity Level Scale (FAALS) that will serve as a clinical tool for foot and ankle practitioners, including surgeons, physicians, nurses, and physical therapists. To develop this instrument and assess psychometric properties including reliability and validity, the Rasch model was utilized in a normal population. The instrument development process for this study was comprised of various phases. First, items for the proposed foot and ankle activity scale were developed by drawing upon knowledge from

content experts in the field of foot and ankle orthopaedics. The assumption of unidimensionality, which is required to conduct a Rasch analysis, was tested using a principal component analysis (PCA) to assess factor structure. Initial psychometric assessment, including reliability and validity, of the proposed foot and ankle score was conducted using the Rasch model. Evidence of reliability of responses to the scale was assessed using person and item reliability. Evidence of validity was assessed using mean square infit and outfit values that are indicative of acceptable or poorly fitting items. Wright item person maps and category response curves were assessed to identify poorly fitting items and assess person ability and item discrimination. Discrimination and sensitivity of the proposed scale was assessed by examining person separation. Discrimination was also assessed by examining mean differences in activity scale scores between sub-populations known to have differences in terms of previous ankle surgery status and body mass index (BMI) by utilizing a multiple linear regression analysis to account for extraneous variables. Evidence of validity was also supported by examining convergent and divergent validity by conducting a Pearson correlation to determine whether scores from the SF-12 PCS and MCS are correlated with scores from the proposed foot and ankle activity scale. Final item reduction and refinement of items was then conducted using the Rasch model.

The overarching research question was: Do scores from the proposed ankle activity scale demonstrate acceptable psychometric properties?

Research Questions:

- Q1 Does the proposed ankle activity level scale meet the assumptions of the Rasch model and demonstrate acceptable psychometric properties?
 - a. Does the proposed ankle activity level scale meet the assumption of unidimensionality?
 - b. Does the proposed ankle activity level scale demonstrate good evidence of person and item reliability with acceptable values?
 - c. Does the proposed ankle activity scale demonstrate acceptable evidence of validity?

- i. Mean square infit and outfit values will lie within acceptable limits.
 - ii. Wright-item person maps will reflect an acceptable range of ability and difficulty.
- Q2 Based on correlations between scores from the proposed ankle activity level scale and SF-12, FAAM and Tegner scores, is there evidence of convergent and divergent validity?
 - a. Are the scores from the proposed ankle activity level scale positively correlated with scores from the Short-Form 12 (SF-12) physical component summary, FAAM ADL and Sport subscales, and Tegner scores, which will provide evidence of convergent validity?
 - b. Are the scores from the proposed ankle activity level scale minimally correlated with scores from the Short-Form 12 (SF-12) mental component summary, which will provide evidence of divergent validity?
- Q3 Does the proposed ankle activity level scale demonstrate acceptable evidence of discrimination?
 - a. Will the person separation index be above the minimal threshold of 2.0?
 - b. Will there be a statistically significant difference between:
 - iii. Individuals who have undergone previous ankle surgery compared to those who have not undergone previous ankle surgery?
 - iv. Individuals who have a normal BMI compared to individuals who have a BMI indicating obese or morbidly obese?
- Q4 What are the defined thresholds of ankle activity level using a standard setting procedure?

Study Significance

Development of an ankle activity scale that can fully capture the construct of ankle activity, while remaining concise, for the purposes of clinical applications, is imperative to allow physicians to understand the functional limitations their patients possess. Activity level is a benchmark used by physicians to gauge where the patient is in the recovery process (Brophy et al., 2014). By understanding how patients perceive their ankle activity level, physicians will be better equipped to tailor post-treatment care and rehabilitation programs. PROs that are region-specific are essential to measure the value and efficacy of conservative and surgical interventions (Abrams, 2017; Acquadro et al., 2003; Bingham et al., 2017; Brophy et al., 2014; McHorney, 1999; Wu et al., 2013). Since there is no region-specific measure for ankle activity level, it was important to develop the instrument, and fulfill this need for medical practitioners and patients alike. This tool will help physicians and other practitioners identify patient progress, as well as

any physical functioning delays, throughout the continuum of care, not only on the patient's individual timeline, but also compared to other individuals who have similar characteristics and who have undergone similar treatment. Essentially, the activity scale developed for this dissertation will serve as a gauge by which to evaluate patients as they heal and attempt to return to desired activity levels. Determining normative values and assessing reliability and validity of scores in the normal, healthy population, will allow these scores to serve as a reference group for physicians, consequently improving the interpretation of scores (Dingemans et al., 2017; Giesinger et al., 2019; Hunsaker et al., 2002; Matheny et al., 2020; Pinsker & Daniels, 2011). The Rasch model is appropriate for development of an ankle activity level instrument because foundationally the Rasch model supports and requires a progression of simple to difficult activities that are then able to be ordered (Andrich, 1978a, 1978b; Bond & Fox, 2015). This progression facilitated an activity scale with specific levels that were then defined. By utilizing the Rasch model to develop this proposed ankle activity level instrument and perform psychometric assessments, this need was fulfilled in an effective and rigorous manner.

Chapter I Summary

Patient-reported outcomes are a cornerstone of the complex recovery process following treatment of ankle injuries. Although PROs have evolved over the decades, there is still a lack of an instrument that allows physicians and other medical practitioners to quickly and effectively assess activity level. Activity level is a unique outcome measure that is essential in understanding where exactly an individual lies on the spectrum of recovery. Developing an ankle activity level scale that incorporates rigor and brevity and is beneficial in clinical application will help to fill that gap in preoperative and postoperative assessment. To develop an ankle activity level instrument, normative values first needed to be established. By first assessing reliability and validity of activity scores in the normal, healthy population, these scores can serve as a

reference group for physicians (Dingemans et al., 2017; Giesinger et al., 2019; Hunsaker et al., 2002; Matheny et al., 2020; Pinsker & Daniels, 2011). To accomplish this task, various approaches exist, including classical test theory and item response theory, which is a more modern approach. Rasch, which some refer to as a special case of IRT, has several advantages over classical test theory, including the ability of the Rasch model to account for person ability and item difficulty. Person and item parameters are considered fully separable in Rasch's family of models, which is very advantageous for generalizability. Previous studies have successfully developed instruments utilizing the Rasch model. In fact, Rasch analysis is increasingly used for examination of psychometric properties of health outcome measures and the development of health-related instruments (Boone, 2016; Boone et al., 2014). Instrument development is a complex process that requires many steps; however, there is extraordinary value in the clinical application of this proposed ankle activity level instrument.

In Chapter 2, more in-depth information about IRT and the Rasch model, including foundational aspects, common models, and their corresponding parameters, is discussed. Application of the Rasch model as it applies to survey development and reliability and validity assessment is also addressed in Chapter 2, as well as how to interpret Rasch model results and commonly used graphics. Chapter 3 includes a detailed plan of how a new foot and ankle activity scale was developed in terms of methodology and materials. Chapter 4 includes the results for all phases of the study, while Chapter 5 includes a detailed discussion of the results, as well as conclusions, limitations of the study, and future implications.

CHAPTER II

BACKGROUND AND LITERATURE REVIEW

In this chapter, the fundamentals of item response theory (IRT), and how those ideas relate to the Rasch model are discussed. This discussion included common IRT models and the corresponding parameters. In addition, advantages of IRT over conventional methods by which to measure reliability and validity, such as classical test theory, are also discussed. A more in-depth discussion of the Rasch model is also discussed, as well as common Rasch model applications, Rasch model ideology, the model itself along with its parameters, and the various methods and tools with which to assess data fit to the Rasch model. Commonly used graphics to assess data fit to the Rasch model are also exemplified. In addition, previous applications of the Rasch model used in the field of orthopaedics are discussed.

Item Response Theory

While the Rasch model is fundamentally different than item response theory in various ways that are discussed later in this chapter, understanding IRT is still very beneficial. To better understand the basis of the Rasch model, additional information regarding IRT is helpful. Item response theory is a more modern approach to psychometric assessment than classical test theory and is a set of mathematical models that describe the relationship between an individual's ability (or trait) and how that individual responds to items on the scale. In fact, invariance of item and ability parameters is the foundation of IRT, and has been considered the primary advantage over CTT (Hambleton et al., 1991). This relationship can be described by a probability curve that is a monotonically increasing function called the item characteristic curve (ICC) which is unique to

IRT (Hambleton et al., 1991; Nguyen et al., 2014). In this context, monotonicity refers to the idea that as an individual's trait level increases, so does the probability of endorsing an item (Nguyen et al., 2014). As an individual's ability increases, so does the probability that the individual endorses or passes the item.

In IRT, the parameter denoted as theta (θ), which is measured on the x-axis, represents an individual's ability or trait level. Higher values of θ indicate that an individual has greater levels of the underlying trait, or greater ability. The y-axis represents the probability that an individual will endorse an item, or pass an item, which is scaled from 0.0 to 1.0 since probability is bounded between 0.0 and 1.0.

Parameters

There are three primary IRT parameters that may be included in different combination for various models. These parameters include item difficulty (b), item discrimination (a), and guessing (c), and serve as the foundation of IRT. Each parameter is described below.

Item Difficulty

Item location is synonymous with item difficulty. This parameter is symbolized as b , and in a dichotomous model, is the point on the ability scale where the probability of a correct response is equal to .5 (Hambleton et al., 1991). As the value of b increases, the examinee must have a greater ability on the underlying latent trait in order to have a 50% (or greater) chance of choosing the correct response. This implies that examinees must have greater trait ability in order to choose the correct response. In other words, item difficulty (b) denotes the “proficiency at which about 50% of the examinees are expected to answer the item correctly” (DeMars, 2010, p. 5). The general range of b lies between -2.0 and +2.0, on a standardized scale with a mean = 0 and standard deviation = to 1.0, with -2.0 representing very easy items and +2.0 representing

very difficult items (Hambleton et al., 1991). In one-parameter models, this is the only parameter, and it is assumed that all items are equally discriminating (Hambleton et al., 1991).

Discrimination

The discrimination parameter is referred to as the slope parameter, which is denoted by a , and is used to determine how well items discriminate against different levels of the latent trait. This parameter “indicates how steeply the probability of a correct response changes as the proficiency or trait increases” (DeMars, 2010, p. 5). Parameter a is proportional to the slope at the point $\theta = b$. Steeper slopes indicate higher discrimination. Since the probability of an examinee’s choosing the correct response should increase as ability increases, the general range of a lies between 0.0 and 2.0 on a standardized scale with a mean = 0 and standard deviation = to 1.0 (Hambleton et al., 1991). In a two-parameter model, the parameters include both b and a , indicating that all items are not equally discriminating, while still taking into account item difficulty. Higher discrimination indicates that a particular item is able to differentiate between individuals with a high trait ability versus those with a lower trait ability (DeMars, 2010). Therefore, the higher a discrimination value for an item, the more effective the item is at determining which individuals have higher or lower levels of the trait, which is useful in understanding, for example, who may have higher versus lower ankle ability.

Pseudo-Chance-Level

The pseudo-chance-level is the third parameter which is often referred to as the guessing parameter or the lower asymptote parameter, and is denoted by c (DeMars, 2010; Hambleton et al., 1991). Parameter c accounts for the probability of individuals who have a low ability on the latent trait (θ) answering the item correctly (Hambleton et al., 1991). When an examinee has a low ability, but still answers the item correctly, some may refer to this as guessing. This

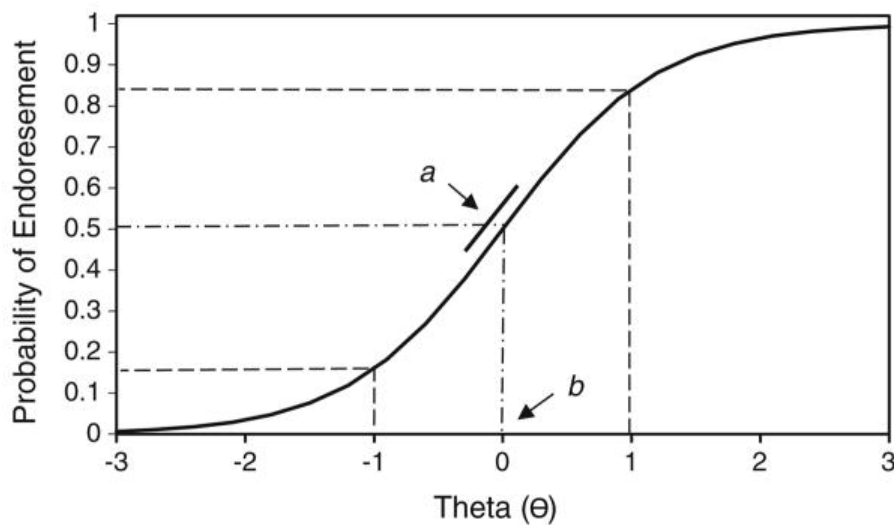
parameter is generally used in testing situations where the test-taker may guess the correct answer, rather than for instruments that measure affective traits since there is no correct or incorrect answer. The guessing parameter is also applicable for aptitude or achievement tests. Often in instruments that contain multiple choice responses, an item may contain response options that are meant to distract and are easily eliminated as a potentially correct response, or an item may contain well-designed distractors that logically seem like the correct choice, and will consequently be selected by the examinee. This may actually allow an examinee with low ability to eliminate possibly incorrect responses (easily eliminated) or choose the distractor since it may seem like a logical choice, which could consequently increase or decrease this parameter to a value greater or lower than what it would be with random chance, respectively (DeMars, 2010). It is for this reason that some authors contest the naming of the c parameter as the guessing parameter (Hambleton et al., 1991). In a three-parameter model, parameters b , a , and c are included, which account for item difficulty, person ability or trait level, and guessing.

The first two parameters (a and b) are commonly used in healthcare research. Parameter c is much less commonly used since there is no correct answer when measuring an affective trait. Depending on the data, one-, two-, and three-parameter IRT models can be used to determine reliability and validity of scores from a particular instrument. Each of these models' appropriateness depends directly on the type of instrument being evaluated, and context of instrument application (Hambleton et al., 1991). A fourth parameter has been mentioned in the literature; however, its applicability has been limited (Loken & Rulison, 2010). The fourth parameter has been suggested for use when accounting for the possibility that a respondent who has a high ability may occasionally incorrectly answer an item with low difficulty (Liao et al., 2012; Loken & Rulison, 2010). The first two parameters are meant to capture responses that

reflect the examinee's true ability, while the third parameter is meant to capture responses to items that are considered guesses, and the fourth parameter may account for incorrect responses that may be due to anxiety, carelessness, or distraction (Liao et al., 2012; Loken & Rulison, 2010). The fourth parameter is most commonly used in the context of computer-adaptive testing and requires a very large sample size which may limit its application (Culpepper & Culpepper, 2016; Yen et al., 2012). The relationship between parameters can be seen in a simple example of an ICC in Figure 1.

Figure 1

Item Characteristic Curve

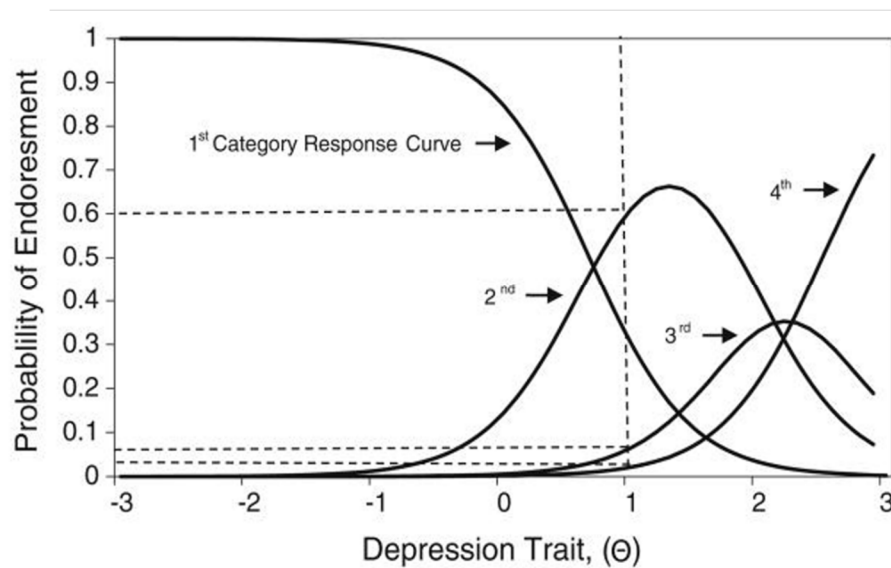


Note: Item Characteristic Curve Example for a Dichotomous Item (Nguyen et al., 2014)

The discrimination parameter, also referred to as the slope parameter, is denoted by a and, as described above, is used to determine how well items discriminate against different levels of the latent trait. The steepness in the slope or a , indicates how the probability of selecting a correct response changes as the individual's ability of the underlying trait increases (DeMars, 2010).

Also, as previously discussed, parameter b represents item difficulty and is often referred to as item location. The “location on the latent trait where the probability of endorsing an item is 50% is referred to as parameter b ” (Nguyen et al., 2014, p. 25). To find the position of b , we find the predicted probability of 50% since this is a dichotomous model and draw a straight line over to the ICC, and then draw a line downwards to the x-axis. This position of b represents the difficulty level of the item. We can see that as ability (θ) increases, so does the probability of endorsing the correct response (y-axis).

ICCs are helpful when examining data with dichotomous response options; however, polytomous data can also be examined using IRT. Polytomous models may also be referred to as graded response models (GRMs), rating scale models, or partial credit models (PCMs). Instead of ICCs, category response curves (CRCs) can be generated, with a unique curve for each response option (Hambleton et al., 1991; Nguyen et al., 2014). An example of CRCs can be seen in Figure 2.

Figure 2*Category Response Curve*

Note. Example of category response curve for PHQ-9, Item #2. Based on this figure, at a trait level of 1.0, which is seen as θ on the x-axis, the probability of endorsing the second response is approximately 60%, while endorsement of response 3 and 4 has probability of 6% and 2%, respectively.

Initial Development

IRT was initially developed by Allan Birnbaum, Fred Lord, and Melvin Novick, and documented in Lord and Novick's classic textbook (Birnbaum, 1957, 1958a, 1958b; Embretson & Reise, 2013; Lord, 1953, 1969; Lord et al., 2008). In the 1950s, Lord (1953) laid the groundwork for the work by Allan Birnbaum in the 1950s. Fred Lord was actually employed by Educational Testing Services (ETS), which explains why IRT is very well connected to testing and is considered the gold standard in test development (Embretson & Reise, 2013).

Advantages of Item Response Theory Compared to Classical Test Theory

Since the 1960s, IRT has emerged as the standard in testing psychometric properties (Hambleton et al., 1991). As indicated in Chapter 1, there are various disadvantages of using CTT compared to IRT. When using CTT, an individual's characteristics, such as person ability on the latent trait, cannot be separated from the instrument's characteristics. This is especially problematic and means that the item parameters, including difficulty and discrimination, are dependent on the characteristics of the sample, which actually limits generalizability (DeVellis, 2006; Hambleton et al., 1991). Conversely, examinees' ability estimates depend on the particular sample of items used in the instrument or test. Another shortcoming of CTT is that the standard error of measurement is assumed to be the same for all individuals. This is an unrealistic assumption, since individuals differ in ability, and their scores will therefore most likely differ in measurement error as well. With IRT, the standard error of measurement differs across scores, but is generalizable across populations (Embretson & Reise, 2013). In addition, CTT is test oriented rather than item oriented, making it difficult to predict how individuals will perform on a given item (DeVellis, 2006). In instrument development, this concept is important because IRT allows for a much richer description of the performance of each item, which allows us to see which items are performing well, and which items need improvement. This concept facilitates overall improvement of the instrument as a whole (DeVellis, 2006; Embretson & Reise, 2013; Hambleton et al., 1991; Nguyen et al., 2014). Other advantages of IRT include the fact that scores from shorter instruments or tests can be more reliable than from longer tests, and test scores can be compared across multiple forms even when forms are not parallel, or in other words, when test difficulty levels vary between persons (Gulliksen, 2013). In IRT, optimal test scores are still able to be produced from mixed item formats and change scores are able to be

compared even when baseline scores differ (Embretson & Reise, 2013). Overall, IRT has distinct advantages over classical psychometric assessment methods, that allow researchers increased flexibility due to more realistic assumptions.

Models and Assumptions

There are several IRT models that are commonly used. Models differ based on type and number of response choices, as well as number of parameters. Models may be dichotomous or polytomous in nature, and most commonly have one, two, or three parameters. There are a variety of models that are subsumed under IRT, with each model predicting the probability that a specific individual will give a specific response to a specific item. All of these models are based on measuring a person's ability (θ) relative to parameters that were previously discussed.

The primary assumptions of all IRT models are that the data are unidimensional and that the instrument administration was not speeded (Hambleton et al., 1991). Unidimensionality refers to the idea that an instrument is designed to measure only one construct. If there are multiple subscales or subdomains, these must be measured individually (Hambleton et al., 1991). For example, if seeking to measure ankle activity only, one would not use wording for items that also tested vocabulary comprehension. This would most likely result in incorrect examinee responses not based on ankle ability, but rather the examinee's lack of ability with vocabulary. Additionally, speeded refers to the idea that a test is limited to a certain amount of time to complete. This could also present an issue because an examinee may choose an incorrect response due to being hurried, rather than being due to their specific latent trait ability. Local independence is also assumed, meaning that responses to one item are not related to responses to other items, after accounting for the latent trait ability. In other words, it is possible to meet the local independence assumption when examinees' responses are related across items, as long as

the related responses are due to their ability only (Hambleton et al., 1991; Nguyen et al., 2014).

Additionally, constant or variant discrimination is an assumption that is dictated by data type and appropriateness, and reflected in model selection. Different models account for either constant or variant discrimination (Hambleton et al., 1991). Below are the most commonly used IRT models.

One-Parameter Logistic (1-PL) Model

This model is the most widely used IRT model, and is also the simplest (Hambleton et al., 1991).

$$P_i(\theta) = \frac{e^{D_a(\theta-b_i)}}{1+e^{D_a(\theta-b_i)}} \quad i = 1, 2, \dots, \quad (1)$$

Where,

$P_i(\theta)$ is the probability that a randomly chosen individual with ability θ endorses item i or chooses the correct response

b_i the item i difficulty parameter

e is the exponential function whose value is 2.718

D_a is a scaling factor equal to 1.702 and is fixed to the same value for all items in the 1-PL model

This model is used when scores on items within an instrument are dichotomous. This model predicts the probability of an item being endorsed from the interaction between individual ability, $P_i(\theta)$, and the item parameter b_i , which is referred to as the difficulty parameter. The scaling factor D minimizes the difference between the normal and logistic distribution functions and therefore makes the logistic function as close to the normal ogive function as possible

(Camilli, 1994). In the 1-PL model, the a parameter from D_a which is a scaling factor equal to 1.702, is fixed to the same value for all items in the 1-PL model. Overall, IRT models the difficulty of the item with the ability of the person. In this model, it is assumed that item difficulty is the only item characteristic that influences an individual's response choice. The primary assumption of this model, which sets it apart from the other IRT models, is that all items are assumed to discriminate equally, although there is no value of discrimination specified (Hambleton et al., 1991). Although the 1-PL model is restrictive in its assumption that discrimination is equal among all items, this model may still be useful in applications such as simple tests that can be completed with ease, or tests that are constructed from a test bank of items that are all similar in nature (DeMars, 2010). Sample size requirements have been reported as low as $N = 150$ and as high as $N = 300$, depending on instrument length (DeMars, 2010; Goldman & Raju, 1986; Guyer & Thompson, 2011; Sahin & Anil, 2017).

Two-Parameter Logistic (2-PL) Model

This model is based on the cumulative normal distribution. The 2-PL model has one additional element than the 1-PL model, a_i , which is the item discrimination parameter. The discrimination parameter is what allows an IRT model to determine which items are most able to detect differences in the ability of respondents, and therefore, an instrument with high discrimination is desirable (DeMars, 2010). For example, an item with high discrimination will mostly provide information for individuals whose ability estimates are close to the item's difficulty, whereas an item with a moderate level of discrimination will tend to provide information across a wider range of ability levels, though not as much information as the highly discriminating item has around the narrow ability range. The 2-PL IRT model can be useful in a variety of situations, especially to identify items with steep slopes or high discrimination in order

to categorize individuals based on ability level. For example, a 2-PL IRT model may be appropriate to identify students who have command of course material versus those who do not; to differentiate between individuals who have low or minimal depression and those who have higher levels of depression; or to distinguish between individuals who exhibit high risk of suicide versus low risk of suicide. The a_i parameter is allowed to vary by item, whereas in the 1-PL model, parameter a is fixed to be equal for all items. By allowing discrimination to vary by item, items that have higher or lower discrimination values are able to be identified. Items with steeper slopes are often used to separate individuals based on ability level, such as high and low proficiencies (Hambleton et al., 1991). This is reflected in the 2-PL model by the subscript i in the a_i parameter (DeMars, 2010; Hambleton et al., 1991). The a_i parameter is proportional to the slope of the ICC at the point b_i on the ability scale. High values of a_i create very steep ICCs, while low values of a_i result in less steep ICCs which increase gradually as a function of ability. Steeper slopes are considered more useful when categorizing and separating individuals by ability levels (Hambleton et al., 1991). Equation 2 represents the 2-PL model. Suggested sample sizes have been estimated at approximately $N = 500$ to $N = 750$, depending on the number of items the instrument contains (Goldman & Raju, 1986; Sahin & Anil, 2017; Thissen & Wainer, 1982; Tsutakawa & Soltys, 1988)

$$P_i(\theta) = \frac{e^{Da_i(\theta-b_i)}}{1+e^{Da_i(\theta-b_i)}} \quad i = 1, 2, \dots, \quad (2)$$

Where,

$P_i(\theta)$ is the probability that a randomly chosen individual with ability θ endorses item i or chooses the correct response

a_i	the item i discrimination parameter which is not fixed and is able to vary for all items
b_i	the item i difficulty parameter
e	is the exponential function whose value is 2.718
D	is a scaling factor equal to 1.702

Three-Parameter Logistic (3-PL) Model

The 3-PL model has one additional element than the 2-PL model, c_i , which is the item pseudo-chance-level parameter sometimes referred to as the guessing parameter as discussed earlier. Some contend that the term “guessing” is incorrect since incorrect response choices are often incorrect, but attractive (Lord, 1974). The c_i is included in the model to account for potential guessing of responses, which equates to performance at the low end of the ability continuum and is most commonly used in the context of testing with correct and incorrect answers. However, the guessing parameter is not used too often with instruments that do not have a defined set of correct answers, such as measures of affective traits. The c_i assumes values that are generally smaller than if an individual had guessed to answer the item (Hambleton et al., 1991). By assuming smaller values, the probability of endorsing a correct response is restricted, specifically when the ability of the respondent approaches $-\infty$. By accounting for the pseudo-chance parameter, items that examinees have used “guessing” to answer hold less weight, and therefore the information obtained from those items have less information which is reflected in a lower information item function peak as compared to other peaks from other items in which guessing was not used. Items that were answered with guessing indicate that an individual has a lower ability on the underlying trait, and that the difficulty of the item is greater than the ability

(DeMars, 2010). Applications of a 3-PL model may include any type of test where examinees may use guessing to answer an item, such as tests with multiple-choice responses and aptitude tests. It is also important to remember that since estimating three parameters, this model requires a large sample size and is dependent on the number of items. Sample sizes of $N = 750$ to $N = 1,000$ have been suggested to reduce bias of estimation and therefore produce more accurate results (De Ayala, 2013; Sahin & Anil, 2017; Tsutakawa & Johnson, 1990). Equation 3 displays the 3-PL model.

$$P_i(\theta) = c_i + (1 - c_i) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \quad i = 1, 2, \dots, \quad (3)$$

Where,

$P_i(\theta)$ is the probability that a randomly chosen individual with ability θ endorses item i or chooses the correct response

a_i the item i discrimination parameter which is not fixed and is able to vary for all items

b_i the item i difficulty parameter

c_i the item i pseudo-chance-level parameter

e is the exponential function whose value is 2.718

D is a scaling factor equal to 1.702

Graded Response Model

This model is utilized for polytomous data with response choices that can be ordered, such as a Likert-type scale with five possible response options (Hambleton et al., 1991). The graded response model (GRM) is an extension of the two-parameter logistic model. This model derives the probability of a response for a specific item within an instrument as a function of

latent trait ability (θ) and the item parameters (Le, 2013). There is an additional term, which is subscript k , in b_{ik} , which indicates the category (DeMars, 2010). In the GRM, there are multiple b -parameters because there are ordered categories which are referred to as thresholds. The threshold represents the point at which examinees have an equal probability of choosing either a lower or higher category than a specified k . In this model, thresholds are not fixed to equal interval widths between items (DeMars, 2010). In the GRM items are compared based on $k-1$ dichotomous categories, where k represents the number of ordered categories for an item. For example, if there are five category responses such as in a 5-point Likert-type scale, and responses are coded 0 to 4, comparisons would be made for the following: 0 vs. 1-4; 0-1 vs. 2-4; 0-2 vs. 3-4; 0-3 vs. 4, which in turn creates four dichotomous comparisons or thresholds. The probability of examinees choosing the lowest category or any of the higher categories is equal to 1 and the probability of scoring in or above category k of item i is also equal to 1; therefore, the threshold parameter for category 0 is not estimated (Zanon et al., 2016). The graded response model was first developed by Samejima (1969) and has one additional assumption than the one-, two-, and three-parameter models, and that is response choices can be ordered (Hambleton et al., 1991; Samejima, 1969; Samejima et al., 1997). In this model, discrimination varies across items and between response categories (Nguyen et al., 2014; Samejima, 1969; Samejima et al., 1997). This model is more flexible compared to the 1-PL model in that each item allows for separate discrimination and category response parameters to be estimated (Nguyen et al., 2014). This type of flexibility makes this model ideal in terms of model fit for patient-reported outcomes (Nguyen et al., 2014). Sample size requirements have been reported to be between $N = 250$ to $N = 500$ (Reise & Yu, 1990; Sahin & Anil, 2017); however, other factors such as number of items, group allocation for ratio such as in randomized controlled trials, group effects, and number of

categories play a major role in minimum sample size requirements (Doostfatemehe et al., 2016).

Equation 4 displays the graded response model.

$$P_{ik}^*(\theta) = \frac{e^{Da_i(\theta - b_{ik})}}{1 + e^{Da_i(\theta - b_{ik})}} \quad (4)$$

Where,

$P_{ik}^*(\theta)$ is the probability of scoring in or above (*) category k of item i

a_i the item i discrimination parameter which is not fixed and is able to vary for all items

b_{ik} the category boundary or threshold for category k of item i

e is the exponential function whose value is 2.718

D is a scaling factor equal to 1.702

Rasch Model Overview

Background and Guiding Principles

In 1960, a new family of models was developed by a Danish man named Georg Rasch. In his pioneering work, Rasch (1960) developed a logistic model that was initially used to develop new measures of reading and tests for use in the Danish military. The guiding principle of the Rasch model is that:

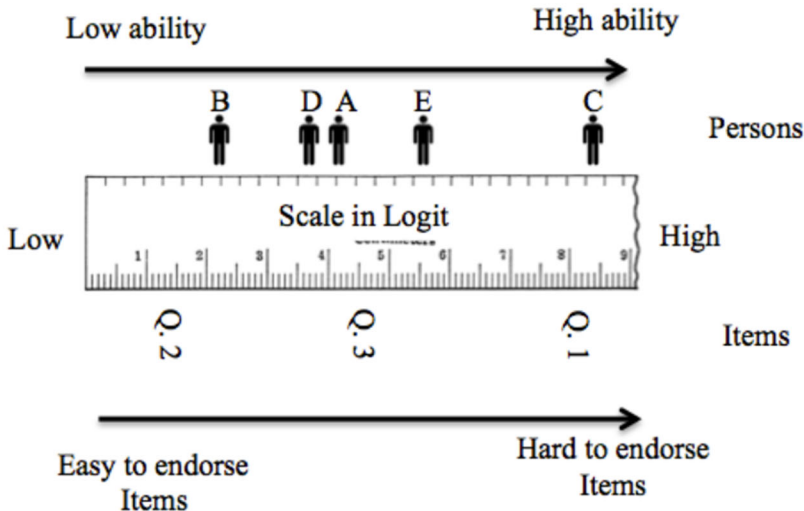
A person having a greater ability than another person should have the greater probability of solving any item of the type in question, and similarly, one item being more difficult than another means that for any person the probability of solving the second item is the greater one. (p. 117)

Rasch measurement theory was developed to address an important question that classical test theory was unable to answer, which was: What is the likelihood that an individual with a specific ability level of a particular underlying trait can get a specific instrument item with a defined difficulty level, correct? The Rasch model is capable of answering this question in terms of probability, since it is “a mathematical formula for the probability of success (P) based on the difference between a person’s ability (B) and an item’s difficulty (D)” (Bond & Fox, 2015, p. 370). The Rasch model models item responses, and not the sum total responses as in classical test theory (Green & Frantom, 2002). The underlying trait or construct is thought of as a linear continuum, which allows for person ability level and item difficulty to be assessed and mapped using the same metric units, which is a log-scale (Bond & Fox, 2015). A unique and defining characteristic of the Rasch model is that person and item parameters are considered fully separable in Rasch’s family of models, which have also been extended to fit polytomous data over the years. This idea of separating person and item parameters is referred to as specific objectivity (Bond & Fox, 2015; Rasch, 1980). The Rasch model allows persons to be ordered or ranked along a continuum, according to their ability, while items are also able to be ranked according to their difficulty. Difficult items can be defined as items that are not answered correctly or endorsed as often as other items, which is often referred to as endorsability when describing attitudinal or subjective constructs that utilize ordinal or interval-level responses such as a Likert-type scale. Since there is no correct answer per se, we consider an individual who selects a particular response as “endorsing” that response. Ability can be defined by the number of items that an individual has endorsed with more difficult items requiring greater ability on the latent construct (Bond & Fox, 2015). For example, individuals who have greater ankle ability will most likely be able to endorse more strenuous measures of activity, such as running for one

mile, as compared to individuals with lower ankle ability who may only be able to endorse walking for one mile. This concept is illustrated below (Figure 3).

Figure 3

Rasch Measurement Model



Note. Visualization of Rasch measurement model on a continuum (logit scale) for person ability and item difficulty (Ruiz-Menjivar, 2016).

For the purpose of the example in Figure 3, this is a three-item test with five participants. Both person ability and item difficulty are mapped on the same continuous scale, which is measured on the log-scale or in logits. Persons with lower trait ability can be found on the left side of the scale or continuum, while participants with greater trait ability can be found towards the right side on the scale. This is the same concept for easier versus more difficult items. Items or questions that require lower trait ability for endorsement can be found on the left side of the scale, whereas items that require higher trait ability can be found on the right side of the scale. Participant B has a very low trait ability, and is therefore more likely to endorse question 2, but not question 3 or 1, since question 2 is further to the left on the continuum compared to the other

items/questions. Participant C on the other hand, has a very high trait ability, and has a high likelihood of endorsing all three questions, which is reflected in the participant's and item's position on the continuum (far right). In terms of ability, the idea is that any participant who has a high trait ability will not only be expected to answer the most difficult items, but will also be expected to endorse any items that are considered less difficult than the most difficult item (Green & Frantom, 2002).

Rasch versus Item Response Theory

The Rasch model possesses the same general advantages that IRT models possess over classical test theory approaches, including the most important advantage which is the fact that both models allow for modeling of person ability and item difficulty along a continuum (DeMars, 2010). The Rasch model is considered a more modern approach to testing psychometric properties, as is item response theory, when compared to classical test theory (DeMars, 2010; Nguyen et al., 2014); however, there are both similarities and great differences between these families of models (Andrich, 2004). While Rasch (1960) was developing a new test theory in 1960, Lord and Novick (1969) and Birnbaum (1958b) were also developing a new test theory. There are two longstanding perspectives regarding IRT and Rasch models (Andrich, 2004). One perspective is that the Rasch model is a special case of item response theory, while the other perspective is that the Rasch model is foundationally different than item response theory models (Andrich, 2004). To this day, controversy remains regarding these vantage points (Wright, 1992).

Wright (1992) debated the inherent differences between the Rasch model proposed by Georg Rasch (1960) and the various IRT models proposed by Alan Birnbaum (1957; Lord et al., 2008). Wright contended that the Rasch model defines measures, while the Birnbaum IRT

models imitate data (Wright, 1992). Wright was referring to the idea that the Rasch model is based on a theory about the instrument and the items that comprise the instrument. Whereas IRT focuses on fitting data to a model and improving fit by adding parameters, the idea behind the Rasch model focuses on utilizing data from an instrument that will fit the theory or structure of the model, and is therefore viewed as a definition of measurement (Boone et al., 2014). While this difference may seem quite nuanced, many Rasch proponents contend that the Rasch model provides a deeper application of theory, thereby providing a more stable foundation and better justification for using the model (Boone et al., 2014).

While the Rasch model and the 1-PL model are quite similar mathematically, there is one distinct difference. In the 1-PL IRT model, discrimination values are the same across items, which supports additivity and construct stability, but those values are not specified. In the Rasch model, discrimination values are also the same across items; however, discrimination values are all equal to a value of 1.0. In IRT models that incorporate a discrimination parameter, such as the 2-PL and 3-PL, items that have the highest discrimination parameters are considered the most useful and valuable items (Masters, 1988). This is distinctly different from the Rasch model, in that items with unusually high discrimination values are thought to be problematic due to potential bias that may be present in unusually highly discriminating items (Masters, 1988). An example of how this bias may present itself is if there were two sets of students, group A and group B. Group B has been taught the material to answer item 3, but group A has not learned the material. Since these two groups are being tested as the same group, this item may appear highly discriminating; however, it would not necessarily be due to the ability of Group B, but rather the instruction or lack thereof for group A. So truly, the opportunity to learn is what would be reflected, and not the ability level of the latent construct (Masters, 1988). For discrimination

values in the Rasch model context, the amount of departure from 1.0 (above or below) is a measure of misfit of the data to the Rasch model. Discrimination values that are greater than 1.0 may indicate that the item actually discriminates between high and low performers more than expected for an item of that specific difficulty, while values less than 1.0 indicate that the item discriminates between high and low performers less than expected for an item of that specific difficulty (Linacre & Wright, 2000). However, in IRT, departures above 1.0 indicate an exceptional instrument item, while lower discrimination values are interpreted as items that are less desirable. This is the distinct difference between Rasch and IRT. In fact, this concept relates back to the idea that data or items of an instrument with a specific structure are chosen to fit the Rasch model, rather than fitting data to the model by adding parameters as in IRT. In Rasch, the average mean of the estimated discriminations should equal 1.0, and all items theoretically should be equally discriminating with the ideal value of 1.0 (Linacre & Wright, 2000).

Advantages of Rasch in the Context of Instrument Development

There are many advantages of the Rasch model over other classical methods. The Rasch model is not dependent upon sample characteristics, like classical methods, meaning the results are more generalizable (Boone, 2016; Boone et al., 2014). In addition, item difficulty and person ability can be accounted for as separate parameters, allowing for separation between person characteristics and item characteristics. This is very important not only for interpretation purposes, but specifically in instrument development. This key feature allows instrument developers to identify poorly functioning items and improve items within the instrument, while avoiding confounding information such as person characteristics. Rasch analysis also allows for instrument items to be assessed by difficulty, meaning that items can be assessed to determine

whether the instrument contains enough items that require high and low trait abilities to capture all abilities and attain acceptable separation.

Compared to IRT specifically, the Rasch model is a simple model that is based on theory. This foundation in theory allows researchers to not only rely on statistical fit, but also rely on sound theory for justification of its use and application (Boone et al., 2014; Wright, 1992). Additionally, the Rasch model requires a minimum sample size of 100 participants, while most other IRT models require more than 100 participants (DeMars, 2010; Goldman & Raju, 1986; Green & Frantom, 2002; Guyer & Thompson, 2011; Sahin & Anil, 2017)

Suitable Data for Rasch

Data that are suitable for the Rasch model include dichotomous response choices or polytomous response choices, such as a Likert-type scale. Modified versions of the Rasch dichotomous model have been developed to accommodate polytomous data, such as Likert-type scales with ordered response choices as discussed earlier. Examples of these scales include the rating scale model, the partial credit model, and the many facets model. These models are based on using ordinal or interval-level data. Data that are dichotomous, ordinal, or interval-level in nature may present problems when conducting a reliability or validity analysis that uses classical test theory techniques due to a lack of normality. However, the Rasch model avoids this potential problem by modeling probabilities that are on a log-odds scale and using scalar constants that minimize the difference between the normal and logistic distribution functions, therefore making the logistic function as close to the normal ogive function as possible. Since both person ability and item difficulty are measured in log-odds on a continuum, the measurement of ordinal or interval-level data can be derived from the relationship between the probability of a correct response reflected in the data and other attributes including person ability and item difficulty, as

long as the attributes increase in the same fashion that the probability of a correct response increases (Bond & Fox, 2015). The idea is that the best predictor of an underlying trait is the relationship between person ability and item difficulty. The flexibility of the Rasch model, coupled with its parsimony, make this model an attractive option when assessing psychometric properties of scores from an instrument that utilizes dichotomous, ordinal, or interval-level data.

The Rasch Family of Models

The Rasch Dichotomous Model

This is the simplest of the Rasch models and was developed by Rasch (1980). This model very closely resembles the 1-PL IRT model; however, there are some stark differences (Linacre, 2005). Although both the Rasch model and the 1-PL IRT model include only the difficulty parameter and assume all items discriminate equally, as discussed above, the 1-PL model does not specify a discrimination value (Hambleton et al., 1991). In the Rasch model, a discrimination value of 1.0 is set for all items (Bond & Fox, 2015). Additionally, the Rasch model requires that the data fit the model in order to generate invariant, interval-level measures of items and persons, making it prescriptive. The 1-PL model however, dictates that a model be chosen based on the fit of the data, making the model inherently descriptive (Bond & Fox, 2015). Sample size requirements have been estimated at around 100 participants (Green & Frantom, 2002).

The Rasch dichotomous model is denoted below (Equation 5) using notation consistent with Bond and Fox (2015):

$$P_{ni}(x_{ni} = \frac{1}{B_n}, D_i) = \frac{e}{1+e^{(B_n-D_i)}} \quad (5)$$

Where,

$P_{ni}(x_{ni} = \frac{1}{B_n}, D_i)$	is the probability of individual n on item i scoring a correct ($x = 1$) response rather than incorrect ($x = 0$) response, given the individual's ability B_n and item difficulty D_i
B_n	person ability on the latent trait
D_i	the item i difficulty parameter
e	the exponential function whose value is 2.718

Rating Scale Model

A popular Rasch model, that is an extension of the Rasch dichotomous model, is the Rasch rating scale model (RSM) which is designed for use with polytomous data, such as Likert-type scales. The RSM expresses “the probability of any person choosing any given category on any item as a function of the agreeability of the person n (B_n) and the endorsability of the entire item i (D_i) at the given threshold k (F_k)” (Bond & Fox, 2015, p. 350; Wright & Masters, 1982). This model was originally developed by Andrich (1978a, 1978b) and Andersen (1972), and is considered constrained compared with the partial credit model due to the requirement that all items possess the same number of response options, and all items of the instrument have the same thresholds, leading to the set of items being structured in the same manner (Bond & Fox, 2015; Linacre, 2000). This means that there is equal discrimination across all items and that a single set of categorical location parameters is estimated for all items (Nguyen et al., 2014). The Rasch RSM has one extra feature over the Rasch dichotomous model which is that it provides one set of rating scale thresholds which is common to all items of the scale (Bond & Fox, 2015). As previously discussed, the Rasch dichotomous model produces person estimates and a difficulty threshold estimate for each item, as does the Rasch RSM. The term threshold

represents “the level at which the likelihood of failure to agree or to endorse a given response category (below the threshold) turns to the likelihood of agreeing with or endorsing the category (above the threshold)” (Bond & Fox, 2015, p. 373). The term rating scale refers to a particular format in which item response choices increase in the level of the variable (Bond & Fox, 2015). For example, an item may ask how difficult it is to walk for 10 minutes, with response choices including (in reverse order): very difficult, somewhat difficult, neutral, somewhat easy, very easy. Other response choices may include strongly disagree, agree, neutral, agree, strongly agree. Sample size requirements for polytomous data reported from various simulation studies have been suggested a range from $N = 100$ to $N = 500$ depending on the effect size and the number of test items (Alexandrowicz & Draxler, 2015; Hagell & Westergren, 2016; Smith et al., 2008a).

The Rasch rating scale model is denoted below in Equation 6 using notation consistent with Bond and Fox (2015):

$$P_{nik} = \frac{e^{(B_n - D_i - F_k)}}{1 + e^{(B_n - D_i - F_k)}} \quad (6)$$

Where,

- P_{nik} is the probability of individual n endorsing category response k of item i , given the individual's ability B_n and item difficulty D_i for each item threshold k
- B_n person ability on the latent trait
- D_i the item i difficulty parameter for the entire item
- F_k the difficulty of the threshold for category k
- e the exponential function whose value is 2.718

Partial Credit Model

This model was first developed by Masters (1982) and is considered part of the Rasch family of models. This model has been described as a version of the RSM model where threshold estimates and the number of thresholds are not constrained, and are free to vary from item to item (Bond & Fox, 2015). In this model, discrimination is equal across all items and each item has a separate category location parameter that is estimated (Nguyen et al., 2014). Therefore, the partial credit model allows each item to possess its own rating scale structure in terms of number of response categories (Bond & Fox, 2015). This model is ideal when categories have a different number of response categories or where some responses warrant partial credit for certain response choices, such as multiple choice tests where although a response may be considered incorrect for that particular test, the response may still indicate some knowledge; hence “partial credit” (Linacre, 2000); or essay responses that may warrant partial credit, respectively (Bond & Fox, 2015). The amount of partial correctness varies across responses, depending on which response is chosen (Linacre, 2000). Sample size requirements for polytomous data suitable for the partial credit model have been reported from various simulation studies and range from $N = 250$ to $N = 500$ depending on the number of the test items (Alexandrowicz & Draxler, 2015; Hagell & Westergren, 2016; Smith et al., 2008a)

The Rasch partial credit model for category probabilities is denoted below in Equation 7 using notation consistent with Bond and Fox (2015):

$$P_{nik} = \frac{e^{(B_n - D_{ik})}}{1 + e^{(B_n - D_{ik})}} \quad (7)$$

Where,

P_{nik}	is the probability that a randomly chosen individual n with ability B endorses response category k of item i
B_n	the ability of individual n along the latent trait
e	the exponential function whose value is 2.718
D_{ik}	the item i difficulty parameter for threshold k

Many Facets Model

This model allows for the most flexibility compared to the other Rasch models. The primary tenet of Rasch is that person ability and item discrimination are the primary aspects that systematically influence persons' scores; however, the many facets Rasch model also allows researchers to account for other factors that may influence scores. Some appropriate instances for use of this model may include if facets are defined as the method by which the instrument was administered, time of day the instrument was administered, individual raters, or even tasks within the item itself. If there are inherent differences in facets relating to methodological issues with instrument administration or there is a lot of variability in the conditions in which the test was taken, researchers may find the many facets model useful (Bond & Fox, 2015). To do so, a "difficulty facet" which is the additional influential factor, is calibrated based on severity of the facet, which is represented with parameter C_j .

The Rasch many-facets model is denoted below in Equation 8 using notation consistent with Bond and Fox (2015):

$$P_{nik} = \frac{e^{(B_n - D_i - F_k - C_j)}}{1 + e^{(B_n - D_i - F_k - C_j)}} \quad (8)$$

Where,

P_{nik}	is the probability of individual n on item i scoring a correct ($x = 1$) response rather than incorrect ($x = 0$) response, given the individual's ability B_n and item difficulty D_i for each item threshold k
B_n	person ability on the latent trait
D_i	the item i difficulty parameter
F_k	the difficulty of the threshold for category k
C_j	the severity of the rater j
e	is the exponential function whose value is 2.718

The Rasch model is commonly used for various applications. In the field of orthopaedics and beyond, Rasch measurement theory has been used for psychometric assessment, instrument development, and differential item functioning. The Rasch model has become an accepted, if not preferred, way to develop and assess health-related instruments (Anselmi et al., 2015; Boone, 2016; Christensen et al., 2013).

Psychometric Assessment and Model Fit

Evidence of reliability and validity of scores from the new ankle activity scale will be determined utilizing the Rasch measurement model, specifically using the rating scale model (Andrich, 1978b), which can be used for polytomous data, such as Likert scale items, when all response categories remain consistent across items. Fit of the data to the Rasch model is assessed through various fit statistics unique to the Rasch model. To identify misfitting items, outfit mean-square (MNSQ) and infit MNSQ statistics are assessed. Infit and outfit mean-square statistics are

χ^2 statistics divided by their degrees of freedom, with an expected value of 1.0 (Linacre, 2002).

Formulas for infit and outfit statistics can be seen below.

We quantify item fit by item infit and outfit. Both are aggregates of the model residuals, with departures from 1.0 representing lesser fitting items. The observed response x_{ni} of person n on item i can be 0 or 1 when referring to a dichotomous scale or measure.

The standardized residual z_{ni} shown in Equation 9 is the difference between the observed response x_{ni} and the probability of the expected response P_{ni} , which is then divided by the expected binomial standard deviation,

$$z_{ni} = \frac{x_{ni} - P_{ni}}{\sqrt{W_{ni}}} \quad (9)$$

where the expected response variance W_{ni} is calculated as

$$W_{ni} = P_{ni}(1 - P_{ni}) \quad (10)$$

Infit Formula:

$$\text{Infit} = \frac{\sum_n^N (x_{ni} - P_{ni})^2}{\sum_n^N W_{ni}} \quad (11)$$

Outfit Formula:

$$\text{Outfit} = \frac{\sum_n^N z_{ni}^2}{N} \quad (12)$$

The expected value of both infit and outfit is equal to 1.0. The value of 1.0 is considered ideal and can be interpreted as a perfect fit of the data to the model. When mean square values are greater than the expected value of 1.0, underfit or misfit is present, while values less than 1.0 represent overfit. Outfit MNSQ statistics describe the fit of items on the extremes (high, low) of the scale. MNSQ outfit is considered an outlier-sensitive fit statistic that is based on the chi-square statistic. Outfit MNSQ is more sensitive to unexpected observations by persons on items that are relatively very easy or very hard for the person. Infit MNSQ statistics describe the fit of items in the middle of the scale and MNSQ infit is considered to be an inlier-pattern-sensitive fit statistic that is based on the chi-square statistic, with each observation weighted by its statistical information (model variance). Outfit MNSQ is more sensitive to unexpected patterns of observations by persons on items. High mean-square infit/outfit are generally considered a greater threat to measurement validity than low mean-squares (Wright et al., 1994).

Mean-squares are used to demonstrate the amount of randomness or distortion of the measurement system and range from zero to positive infinity. An ideal value of 1.0 can be interpreted as observed variance equals expected variance with acceptable values ranging from .5 to 1.5 or 1.7, although various authors have proposed slightly different ranges of acceptable values (Wright et al., 1994). When infit and outfit MNSQ values are low (i.e., less than .5), redundancy in the responses may be present, indicating model overfit; however, no harm is done (Gustafsson, 1980; Martin-Löf, 1974). Values greater than 1.0 indicate unpredictability or noise in the data, which may be indicative of model underfit, which is considered more harmful. When values are greater than 2.0 for a patient, the patient may actually belong to a different population than the other patients within the sample, or perhaps the patient does not have the appropriate reading level or ability to complete the instrument accurately or honestly. When values are

greater than 2.0 for an item, that specific item may be confusing to the patients or have poor wording (Linacre, 2002; Messick, 1989; Smith, 2001).

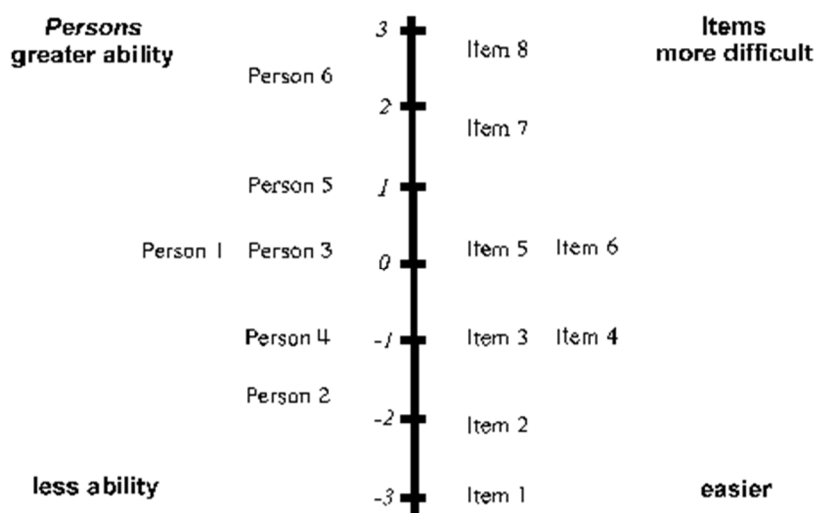
Person reliability is an estimate of reliability unique to the Rasch model. Some contend that person reliability for the Rasch model is analogous to the Cronbach's α (Fisher, 1997). The person reliability index represents the replicability of the ordering of persons if a similar, parallel set of items were administered to the same sample (Wright & Masters, 1982). To have acceptable person reliability it is necessary to have well-functioning items that target the appropriate ability estimates and items that capture a large enough range of abilities in order to ensure that there is the classic Rasch hierarchy of items, or activity levels in this case (Bond & Fox, 2015). High person reliability indicates that some individuals will consistently score high and some individuals will consistently score low (Bond & Fox, 2015). Person reliability uses estimates on the logit scale rather than raw scores for each person to calculate reliability (Fisher, 1997). Person separation, which is the ability of the scale to discriminate among different groups of individuals, is determined by the person separation index (PSI), with higher values indicating greater discrimination. It is necessary to have enough space between person abilities, which would demonstrate that there are multiple distinct person abilities. Person separation is an indication of the reproducibility of person ability ordering and essentially how efficiently the items of an instrument can separate out person abilities. Values above 2.0 have been deemed acceptable and can be interpreted as the measure being capable of separating respondents into two distinct groups (Wright & Masters, 1982). Similar to person reliability, item reliability represents the replicability of the ordering of items if a similar sample were administered the same items (Bond & Fox, 2015; Wright & Masters, 1982). High item reliability means that some items are more difficult than others, indicating that there is a hierarchy of item activity levels

with a wide difficulty range, which is desired. Item separation is used to verify the item hierarchy (Linacre, 2020). Specifically, item separation values of 3.0 that the instrument has enough separation to identify two or more distinct difficulty levels. Higher values indicate a greater range of items that capture a larger variety of abilities.

Various figures including Wright item-person maps, test information functions, category probability curves, and item characteristic curves are also produced in order to assess evidence of validity. Wright maps allow researchers to quickly identify strengths and weaknesses of scores from an instrument, based on logits of the item-person distribution (Boone, 2016). Wright maps plot persons on the left (or top) of the map and item difficulty on the right side (bottom) of the map. An example of a Wright item person map can be seen in Figure 4 (Sick, 2008).

Figure 4

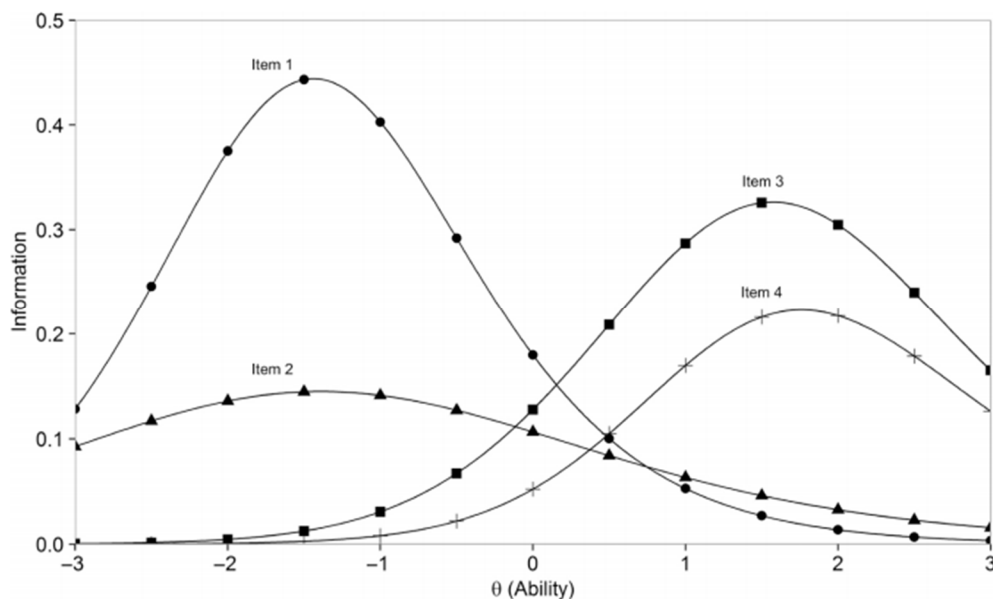
Wright Item Person Map



Note. Example of Wright item person map displaying organization of item difficulty and person ability on same scale (Sick, 2008).

The greater the person measure, the greater that person's ability. Items are plotted by difficulty, with the most difficult items located near the top of the map. The higher the item is placed on the map, the greater the difficulty level of the item. Since persons and items are plotted on the same linear scale, instrument item distribution, while accounting for person ability, can be assessed by comparing the mean person measure and the mean item measure (Boone, 2016). Essentially, an ideal item-person map should reflect the theorized order of difficulty, with little redundancy, and capture persons with various levels of ability (Cappelleri et al., 2014). The shape may resemble a normally distributed histogram for both items and persons. The item-person map can then be used as evidence that is likened to content validity (Cappelleri et al., 2014).

Item information function is a function of ability, and refers to the amount of information yielded by the test itself, given a certain ability level (Frey, 2018). For every item within an instrument, a value for item information is given. Item information identifies the ability at which the item functions most effectively, and therefore provides the greatest amount of information (Frey, 2018). An example of item information for four items can be seen in Figure 5. To find the ability level at which item 1 functions best, we find the peak and can see that it equates to approximately -1.5, which is a lower ability level. Items 1 and 2 have a lower difficulty level than items 3 and 4, based on where the item information peak lies. To determine which item provides the most information, we can look at the peak of the item information function and see that item 1 has a value of approximately .45, which is higher than the other items.

Figure 5*Example of Item Information for Four Items*

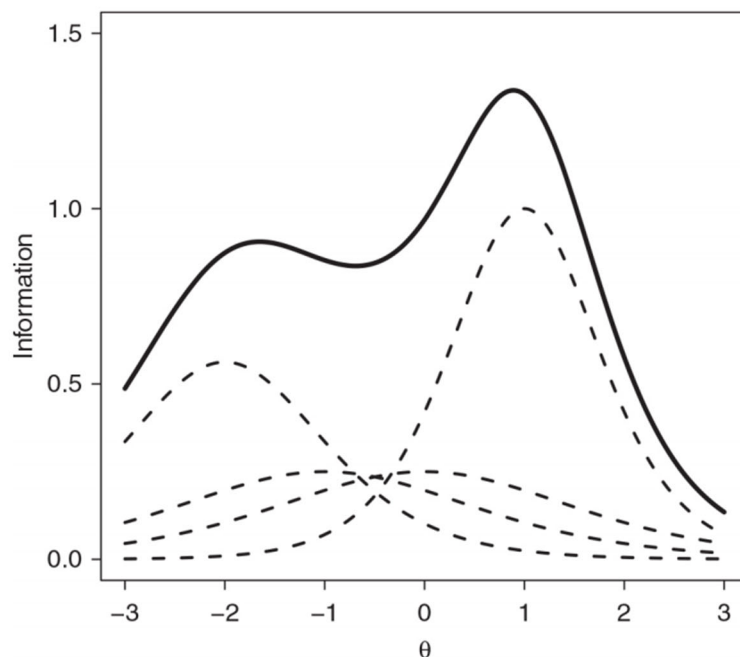
Note. The amount of information can be seen on the y-axis, which indicates how informative an item is, while ability of the trait can be seen on the x-axis (Frey, 2018).

Test information function (TIF) is similar to item information function, except TIF demonstrates how much information the entire instrument or test yields as a whole. TIF is also measured relative to ability, and serves as an estimate of instrument or test precision of the estimator of ability (Frey, 2018). Test information function is the sum of all of the item information functions and can be used to predict the accuracy to which we can measure any value of the latent ability, which is ankle activity level in this case. Information and precision of the score can be determined using test information functions. In practice, the test information function should peak at the most clinically important cut-point, which may be considered the sample mode, while the width may be considered the range of scores. If an item has a large discrimination value, the curve may be tall and narrow, representing a narrow range with high precision. If an item has a

low discrimination value, the curve may be short and wide, representing a broader range with lower precision. Some authors recommend information values of 10 or greater with standard errors (SEs) less than .5. The thought behind this general rule of thumb is that a test information function value of 10 has an approximate SE of .31, which corresponds to a reliability coefficient of .90 (Embretson & Reise, 2013). Generally, a horizontal line may be considered the ideal shape; however, this is not always the case or even possible, depending on the purpose of the instrument (Cappelleri et al., 2014). An example of test information function can be seen in Figure 6. The test information function indicates that the instrument would most reliably measure the latent trait at an ability (θ) of 1.0, which is visualized at the highest peak, which is above the average ability indicated by zero (Frey, 2018).

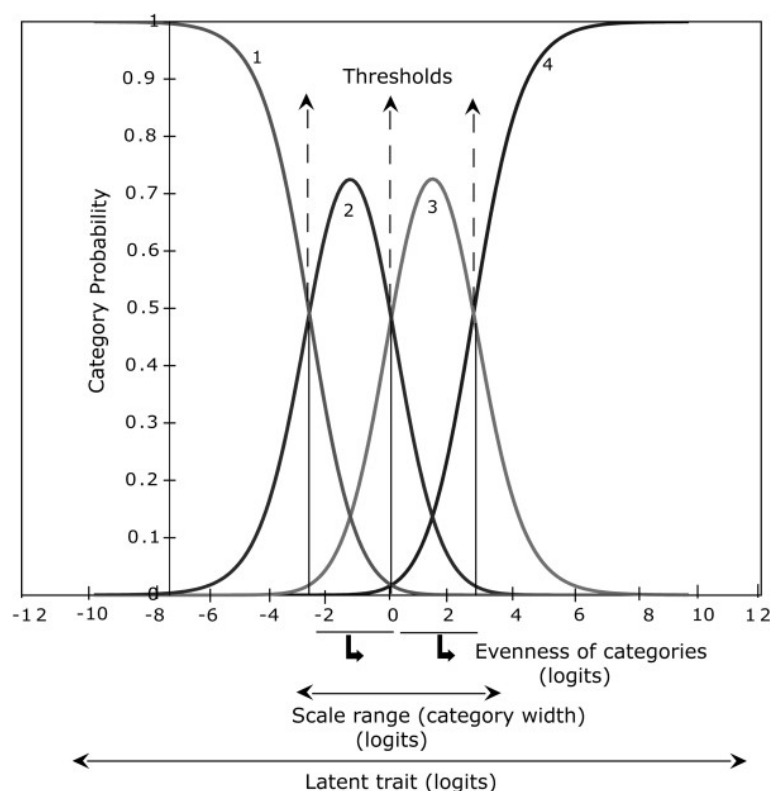
Figure 6

Test Information Function



Note. Test information function is denoted in the solid line while items 1 – 4 are denoted by the dashed lines.

Category probability curves, also referred to as category response curves for polytomous data, visualize the probability of observing each ordered category according to the Rasch model. Ideally, we would expect to observe the same probabilities for each of the five possible responses per item. Ideal category response curves for items on a measure of ankle functioning, for example, will reflect response options for each item that are monotonically related to the ankle ability, and each response option is likely to be selected at some differing range of the underlying construct (ankle activity level). Category probability curves are used to indicate ordered thresholds, which is a key characteristic of rating scales, since response choices should follow a hierarchical order (Khadka et al., 2012). In addition, these curves also provide insight as to which responses are most often selected by respondents, with smaller standard deviations indicating more even widths of curves, meaning that respondents are choosing categories equally (Khadka et al., 2012). Ideally, we want to see category probability curves that have appropriately ordered responses, with even widths, with a large width or measurement, indicating greater measurement coverage of the latent construct (Khadka et al., 2012). An example of category probability curves for a four point Likert scale can be seen in Figure 7 (Khadka et al., 2012).

Figure 7*Category Probability Curves*

Note. Category probability curves with four response categories, with 1 representing “not at all” and 4 representing “a lot.” The y-axis represents the probability of a response category being selected, while the x-axis represents ability of the latent construct. Since there are four response categories, there are three thresholds. These thresholds are spaced evenly, have a small standard deviation, and have a good range of ability (Khadka et al., 2012).

The Rasch Model Applications

Rasch analysis is increasingly used for examination of psychometric properties of scores obtained from health outcome measures (Boone, 2016; Boone et al., 2014). Instrument development is a complex process and utilizes a multi-faceted approach. After developing the instrument items, a pilot study should be conducted to collect initial data for the first assessment.

Data should be checked for the assumption of unidimensionality, which can be tested using a principal component analysis (PCA) as previously suggested (Andrich, 1978b). Next, a Rasch analysis can be performed, and model fit can be assessed. By looking at the Wright item-person map, items can be ordered or ranked in terms of difficulty and person ability. This will help to ensure that the instrument is capturing a variety or range of items that represent high and low traits of the latent construct. By assessing MNSQ infit and outfit statistics, poorly functioning items can be identified and flagged for removal, which will further refine the new instrument. Following refinement, a second wave of data should be collected, and a Rasch analysis can be conducted to determine whether the new instrument has been refined enough to produce acceptable psychometric properties.

Secondary Assessment: Differential Item Functioning

Rasch analysis with a focus on Differential Item Functioning (DIF) is increasingly used for examination of psychometric properties of health outcome measures among many other applications of Rasch analysis (Andrich, 1978a, 1978b; Hagquist et al., 2017). DIF is often used to determine whether differences in measurement exist between two different groups, such as patient groups who represent preoperative and postoperative patients, or males and females, etc. This concept is also referred to as measurement bias. Measurement bias has been commonly defined as differences in the way a test item functions across demographic groups that are matched on the attribute measured by the test or the test item (Hagquist et al., 2017). Modeling DIF is often thought of as a “group x item location” interaction, meaning that item location is dependent upon group membership. If the interaction is significant, the latent variable may be defined differently based on group membership, which is problematic for reasons previously discussed (Cauffman & MacIntosh, 2006).

Rasch Model Applications in Orthopaedic Instrument Development

The Rasch model has been recognized for its great potential and application in the field of health care and medicine (Aryadoust et al., 2019). In fact, the Rasch model has been considered the gold standard in health-related instrument development (Anselmi et al., 2015; Tennant et al., 2004). Application of the Rasch model in the health care field for the development of mobility instruments has increased from only one in 1987 to 48 in 2007 (Belvedere & de Morton, 2010). A key advantage in utilizing Rasch modeling to develop instruments that assess activity, function, or mobility is that there is a clear hierarchical progression of abilities for these constructs, so items can be structured as such. Since the structure of the scale and the data for these types of constructs are considered additive, the Rasch model is very appropriate (Belvedere & de Morton, 2010). This also applies to orthopaedic research, which is a subset of health care research.

There is a wide variety of health-related and orthopaedic instruments that have utilized the Rasch model for psychometric assessment (Aryadoust et al., 2019; Balalla et al., 2019; Balsamo et al., 2014; Belvedere & de Morton, 2010; Bravini et al., 2017; Budiman-Mak et al., 2006; Christensen et al., 2019; Comins et al., 2008, 2018; Conaghan et al., 2007; Franchignoni et al., 2010, 2012; Hamilton et al., 2015; C. Hiller et al., 2006; Hung et al., 2016; Hunnicutt et al., 2019; Kallinger et al., 2019; A.-M. Keenan et al., 2007; Kersten et al., 2014; Ko et al., 2009; Korenevskiy et al., 2016; Lin et al., 2009; Matheny & Clanton, 2020; Moreton et al., 2012, 2015; Muller & Roddy, 2009; Niama Natta et al., 2019; Nishigami et al., 2017; Oude Voshaar et al., 2017; Perera et al., 2020; Perruccio et al., 2008; Ponkilainen et al., 2020; Repo et al., 2017; Sadjadi et al., 2014; Smith et al., 2010; Smith et al., 2008b; Tennant et al., 2004; Turner et al.,

2017; Vrotsou et al., 2016; Wang et al., 2017, 2018, 2020; Warholak et al., 2011; Woodburn et al., 2012; Yorke et al., 2012; Yuksel et al., 2018). The Rasch model has been applied to foot and ankle instruments, as well as other joint-specific instruments, in different ways, including developing instruments, revising instruments, and assessing psychometric properties of instruments. As time progresses, the Rasch model has become more and more commonly used due to its advantages over classical methods (Belvedere & de Morton, 2010; Budiman-Mak et al., 2006; Comins et al., 2008).

The Cumberland Ankle Instability Tool (CAIT) was developed, using the Rasch model, to measure functional ankle instability in bilateral injuries in 236 individuals from the normal population (C. Hiller et al., 2006). Two items were shown to have unacceptable fit statistics, indicating the potential need for removal. The Wright item-person map revealed a nice range of items and person ability resembling a bell curve on either side, demonstrating good evidence of validity. Hiller et al.'s (2006) study demonstrated the strengths of the Rasch model in that poorly functioning items are able to be identified and removed, if necessary, to improve internal validity.

Budiman-Mak et al. (1991, 2006) first developed a Foot Function Index (FFI) using classical test theory methods in 1991 in order to measure the impact of foot pathology on foot function. In 2006, the FFI was revised using the Rasch model. The developers of the FFI cited shortcomings of classical test theory as their reasoning, further justifying improvements in the instrument utilizing more modern techniques that allowed assessment of individual items (Budiman-Mak et al., 2006). The end result was a revised instrument that was much more efficient at measuring the construct of foot function. In fact, some original items were removed that were found to be poorly fitting by the Rasch analysis and were then able to be replaced by

better fitting items. Although the Rasch model was not used in the original development of the FFI, it was used in the extensively revised version, which provided an incredible amount of improvement in the composition and structure of the instrument (Budiman-Mak et al., 2006). Another instrument that used the Rasch measurement model for revisions was the Knee injury and Osteoarthritis Outcome Score (KOOS), which is comprised of five subscales (Comins et al., 2008). This scale is another scale that was originally developed using CTT; however, this instrument was also re-assessed using the Rasch model (Comins et al., 2008). Although the KOOS was not re-assessed by its original developers, as the FFI was, another group of researchers assessed the KOOS, citing the same shortcomings of CTT (Comins et al., 2008). All five subscales were analyzed as separate constructs in order to adhere to the assumption of unidimensionality. Upon analysis, the Rasch model revealed that only two of the five subscales supported use in the intended target population of patients who underwent anterior cruciate ligament (ACL) reconstruction at 20 weeks status-post (Comins et al., 2008). These findings support the use of the Rasch measurement model as a more rigorous assessment of psychometric properties of orthopaedic instruments that are used to assess function and activity.

There are many ankle scores that have utilized to Rasch model to assess evidence of reliability and validity. In 2020 the Foot and Ankle Ability Measure (FAAM), which is an instrument that was designed to assess ankle function, was re-assessed using the Rasch measurement model in patients who underwent surgery for a variety of ankle pathologies (Matheny & Clanton, 2020). The FAAM was originally developed using item response theory (Martin et al., 2005); however, the FAAM had not been assessed in a surgical population. Findings revealed two poorly functioning items out of the original 21 items for the activities of daily living subscale and acceptable fit of all items for the sport subscale (Matheny & Clanton,

2020). Reliability was high for both subscales. Recommendations were made to consider removal of the two items to help improve internal consistency of responses to the FAAM (Matheny & Clanton, 2020). The Rasch model is helpful in identifying items as acceptable or poorly functioning, allowing researchers to modify scales based on data fit. Since there are multiple ways to assess fit with Rasch analysis, researchers can feel confident in decision-making and scale modification. Another instrument that was assessed using the Rasch measurement model is the Foot Posture Index which was originally intended to quantify variation in the position of the foot easily and quickly in a clinical setting (Keenan et al., 2007). A Rasch analysis was performed on a sample of 143 individuals with varying foot types. Findings demonstrated poor fit for two of the eight items, which was indicated by mean-square values and an overall significant chi-square test. Upon removing the two problematic items, the Foot Posture Index revealed acceptable internal validity of scores (Keenan et al., 2007). A similar study was conducted for the Manchester Foot Pain and Disability Index (FPDI) using a Rasch analysis to assess psychometric properties of scores in 149 adults aged 50 years or greater (Muller & Roddy, 2009). The FPDI is comprised of three subscales including function, pain, and appearance. Results revealed that the function and pain subscales had acceptable fit; however, the appearance subscale was unable to be tested since there were too few non-extreme scores reported from participants. This study highlights the importance of rigorous psychometric assessment (Muller & Roddy, 2009). Although some instruments have been used for a very long period of time since their original development, it is still important to assess psychometric properties of scores over time. Since the FPDI was tested again, problems were revealed that may not have otherwise surfaced if psychometric properties were not assessed again. In 2020 psychometric properties of scores from the modified Gait Efficacy Scale (eGES) were assessed

using the graded-response Rasch model (Perera et al., 2020). The eGES is comprised of nine items; however, after performing the Rasch analysis, researchers discovered that two items could be removed, while maintaining similar, acceptable psychometric properties as the nine-item scale. An additional Rasch analysis was performed to determine how many more items could be removed to reduce responder burden, with results showing that three items could remain with minimal effect on psychometric properties (Perera et al., 2020). Studies like this highlight the utility of the Rasch measurement model. Reducing responder burden is important, since participants are more likely to complete the entire questionnaire when less burden is placed on them.

The Rasch model has demonstrated its utility in scale development with specific attention paid to item assessment, allowing instrument developers to identify acceptable items that provide a large amount of information, as well as poorly functioning items that do not fit the model and should therefore be removed to improve internal consistency. The Rasch model has also proven useful in re-assessing instruments that were originally developed using CTT methods. Although CTT has been incredibly useful over the years, contributing to the assessment and improvement of psychometric properties, the Rasch model has demonstrated its superior capabilities in assessing individual items, as well as person ability.

Chapter II Summary

The Rasch model has been applied to assess psychometrics in numerous ways, with the most common including instrument development, instrument revision, simply assessing psychometric properties to document internal consistency and reliability, and to determine differential item functioning. The Rasch model offers advantages over CTT methods, making it a popular choice in the field of orthopaedics. The Rasch model is based on theory and therefore provides sound justification for its use when the data are deemed appropriate (Masters, 1988). In

orthopaedics, and specifically regarding foot and ankle activity instruments, there is a clear hierarchy of activity that lends its structure perfectly to the Rasch model. The Rasch model also requires a relatively small sample of approximately 100 participants. In addition, the Rasch model facilitates improvements and optimization in instruments by allowing developers to identify both items that provide the greatest amount of information, and also items that have poor fit and should be removed (Bond & Fox, 2015). This reduction in poorly fitting items also helps reduce respondent burden, which is important for our participants, and valuable in a clinical setting where time is often limited (Perera et al., 2020). Overall, the Rasch measurement model is a simple model that allows researchers to separately assess items within an instrument and person ability. This helps to improve an instrument's overall internal consistency and therefore improve accuracy in measurement of the latent construct of interest which is of course the primary goal.

Chapter 3 includes a detailed plan of how a new foot and ankle activity scale was developed using the Rasch measurement model in terms of methodology and materials. There were three phases of data collection discussed in detail.

CHAPTER III

METHODOLOGY

Through this study I sought to develop a new foot and ankle activity scale instrument that may be applied in the field of orthopaedic patient-reported outcomes and research. The primary purpose of this study was to develop an instrument that will be used to measure patient-reported activity level in the normal population, and eventually be used to measure activity level following an ankle or foot injury, as well as following surgical treatment for an ankle or foot injury. The scores from this instrument were tested for acceptable psychometric properties, including reliability and validity, utilizing the Rasch measurement model. Since instrument development entails multiple rounds of data collection, this study was also comprised of multiple phases of data collection. Specifically, three rounds of data collection from three different samples of participants are included. Details for each phase are explained below.

Phase 1

The overall goal of the first phase of data collection was to establish the breadth and depth of the construct, ankle activity level, and operationally define this construct. In order to accomplish this task, initial content domain was identified through a detailed literature review regarding foot and ankle activity level, which I have already provided in Chapter 2. I also drew from my experience in the field of ankle orthopaedic outcomes research over the past decade. The literature review is considered the foundational step in instrument development that allows the researcher to operationally define the construct of interest (Gable & Wolf, 1993). Next, an expert panel was interviewed, using knowledge regarding ankle activity level from the literature

review as prompts to start the conversation. After conducting interviews, all data were aggregated to develop an initial pool of items. Following item generation, the original expert panel was asked to assess items based on content appropriateness and difficulty, a reading expert was then asked to assess items based on appropriate reading level, and two stakeholders from the normal population were asked to conduct think-aloud protocols to identify any issues with interpretability or usability. These steps within phase 1 are discussed in greater detail below.

Sample

Sample 1 was comprised of various stakeholders in the orthopaedic ankle and foot field. I have various contacts in the field of orthopaedics whom I asked to participate in the study via phone call and email. Specifically, there were three ankle and foot orthopaedic surgeons, two orthopaedic researchers, and one ankle and foot physical therapist who comprised the expert panel. The surgeons have a combined 65 years in orthopaedic surgery and treatment. The researchers have a combined experience of 60 plus years in outcomes and survey research. The ankle and foot physical therapist has over 12 years of experience in ankle and foot physical therapy. The ankle and foot literature review was then used to inform questions that were posed to stakeholders during interviews. The goal was to include at least five stakeholders in order to effectively capture content and variability of the domain, which was accomplished (Zamanzadeh et al., 2015). This content expert panel was used to help determine the breadth and depth, and operationally define, the construct of ankle activity level. This process has also been referred to as a Delphi process (McPhail et al., 2014). A reading expert was asked to assess readability of items for a seventh-grade reading level to improve comprehension among all participants. The reading expert was asked to flag any items that may require a reading ability above the seventh grade. Two stakeholders from the normal population were asked to conduct a think-aloud

protocol via Zoom to identify and items that may be difficult for the normal population to interpret due to clarity or usability issues. These stakeholders from the normal population had no more than “some college with no degree” in order to represent the majority of the United States population who is 25 years or older ("Educational Attainment in the United States: 2019," 2020). This study was approved by the institutional review board at the University of Northern Colorado. Each individual provided consent prior to participation in the study.

Data Collection and Procedures

Step 1: Define Construct and Content Domain

An expert panel was comprised and included various experts in the orthopaedic community as detailed above. The interview process was a semi-structured interview with a more conversational and collaborative focus. These interviews more resembled colleagues discussing a familiar topic in their common field. Interviews were essential to fully and effectively understand content domain and to generate instrument items (Gable & Wolf, 1993; Tilden et al., 1990; Zamanzadeh et al., 2015). Interviews included a variety of questions dependent upon the type of expert and their clinical and professional experiences (i.e., surgeon, researcher, physical therapist). A sample of questions can be found in Appendix A. The goal of conducting the interviews was to have participants discuss ankle activity level from their perspective in order to ensure the breadth and depth of the content domain was captured, while filling in any gaps that may have been missed in the initial literature review.

Interviews. Interviews lasted approximately 45 to 60 minutes. Interviews were conducted via telephone and video conference (Zoom). It is important to note that some individuals previously expressed their desire to participate as a content expert, and some of these individuals know one another very well and have an established relationship with one another.

The community of ankle orthopaedics is quite small, and many experts either know each other or know of one another.

Information from the interviews was recorded electronically. I typed the ideas and thoughts of each interviewee in an electronic document at the time of the interview. This allowed me to construct a list of activities, and essentially items, that the interviewees felt represents the construct of ankle activity level. This type of interaction was more of a collaborative effort, which was the goal. I created an electronic summary of ankle activity level activities that were captured from each interview. Each interviewee was labeled with a number and recorded in my electronic research journal. The information from the interviews was aggregated and used to confirm or deny the use of items that I had already generated from the literature review, as well as to generate new items to address the identified gaps and incorporate content expert input. These steps helped to capture ankle activity level effectively and completely (Gable & Wolf, 1993). The goal was to generate a minimum of 30 to 50 items in order to operationally define the affective characteristic (Gable & Wolf, 1993; Green & Frantom, 2002).

Step 2: Initial Item Generation and Orthopaedic Expert Assessment

Initial items were generated by pulling from the literature review, including other foot and ankle items (Martin & Irrgang, 2007; Martin et al., 2005) and aggregated data from the expert panel collected through interviews to ensure good content coverage. I then constructed a questionnaire that was reviewed by an expert survey methodologist for item clarity, and then distributed to the expert panel via email, which included the complete list of items from the pool. I asked the experts a variety of questions that address construct representativeness and difficulty level to perform the tasks described in each item using a 5-point Likert scale. Question 1 was used to rate the items with respect to the extent that the items describe ankle activity level. This

allowed experts to quantify the intensity with which the item represents the construct (Gable & Wolf, 1993). Question 2 allowed experts to quantify the difficulty level of the tasks indicated by each item to ensure that items with a wide range of ankle activity levels were included to fully capture the construct. A Likert-type scale was used to assess difficulty. Although ranking of all items in terms of content domain representativeness and difficulty has been suggested (Gable & Wolf, 1993), the cognitive processing it would have required to rank the large number of items (101 items) would have been incredibly overwhelming for participants, and may not have actually reflected accurate information (Alwin & Krosnick, 1985; Dillman et al., 2014; Smyth et al., 2018). For question 3, I requested that experts only rank their top 10 items for high, moderate, and low activity levels to avoid overwhelming cognitive processing. A question was also included to ensure participants could leave any desired comments or suggestions. The questionnaire can be seen in Appendix B.

All data from questionnaires were evaluated. If an item was not considered to represent the construct of ankle activity level (question 1) by at least 75% of orthopaedic experts, the item was flagged as a poor item, and was further assessed in the initial factor structure assessment in Phase 2. If determined to be a misfitting item in Phase 2, the item was removed prior to Phase 3. For question 2 regarding difficulty level, the majority of responses determined the level of difficulty as decided by the panel of experts. For example, if three of five experts deemed item one as extremely difficult, that item was considered to be an item that requires the highest ankle ability, and therefore the highest ankle activity level. Although much of this initial process of item generation was based on judgement, the categorizing of items based on content appropriateness and difficulty has been shown to be an effective empirical procedure and is commonly used in item generation (Gable & Wolf, 1993). I removed items that were deemed

inappropriate in terms of content representativeness and difficulty by the majority of the expert orthopaedic panel as described above, and a revised list of items was distributed in the next step.

Step 3: Readability Assessment

The newly revised item pool was distributed to the reading expert and she was asked to assess readability for a seventh-grade reading level via a questionnaire that was distributed via email (Appendix C). The goal of this assessment was to improve comprehension of items among all participants who are from the normal population. The reading expert was asked to flag any items that may require a reading ability above the seventh grade, and these items were modified according to her recommendations. The idea was to have an easy to moderate reading comprehension to facilitate reading comprehension by all participants. Previous studies have shown that the average American has a seventh to eighth grade reading ability (Kirsch, 1993). In other marginalized or vulnerable populations, such as minorities, those living in poverty, homeless, and those older than 65 years, reading comprehension may be lower (Calderón et al., 2016).

Step 4: Cognitive Interviewing

After the readability of all items was assessed and items were modified, a revised item pool was distributed to two stakeholders who were from the normal population. The two individuals were asked to perform a think-aloud protocol via Zoom. Think-aloud protocols have been deemed useful in gaining insight into an individual's thought process, which can be helpful when trying to understand why a particular item may not be functioning the way it was intended (Prior et al., 2011; Van Someren et al., 1994). As the individuals were performing the think-aloud protocol, I took notes, and then asked them about items that seemed to present

interpretability issues or had clarity issues. I used that information to improve the item or remove the item if necessary.

Phase 2

Following Phase 1, the revised pool of items was tested in the normal population, which comprised the second sample. This second sample was used to pilot the newly generated items and had a sample size of 100 participants which is the minimum sample required for the Rasch analysis (Green & Frantom, 2002). Average age was 46.0 years ($SD = 17.8$), and average BMI was 27.3 ($SD = 6.4$). Sociodemographics can be seen in Table 1.

Table 1*Sociodemographic Characteristics of Phase 3 Data*

Characteristic	n	%
Gender		
Female	260	51.5
Male	245	48.5
Level of Education		
Less than high school degree	19	3.8
High school graduate (high school diploma or equivalent)	191	37.8
Some college but no degree	34	6.7
Associate degree in college (2-year)	104	20.6
Bachelor's degree in college (4-year)	36	7.1
Master's degree	12	2.4
Doctoral Degree or Professional degree (JD, MD)	3	0.6
Race		
American Indian or Alaska Native	24	4.8
Asian	58	11.5
Black or African American	91	18.0
Black or African American, American Indian or Alaska Native	3	0.6
Native Hawaiian or Pacific Islander	8	1.6
Other	50	9.9
White	271	53.7
Income		
Less than \$10,000	44	8.7
\$10,000 to \$19,999	60	11.9
\$20,000 to \$29,999	73	14.5
\$30,000 to \$39,999	55	10.9
\$40,000 to \$49,999	49	9.7
\$50,000 to \$59,999	34	6.7
\$60,000 to \$69,999	28	5.5
\$70,000 to \$79,999	23	4.6
\$80,000 to \$89,999	10	2.0
\$90,000 to \$99,999	6	1.2
\$100,000 to \$149,999	76	15.0
\$150,000 or more	47	9.3
Region		
Midwest	89	17.6
Northeast	90	17.8
South	228	45.1
West	98	19.4

The data collected from this sample served as the pilot data in which to identify items with the best fit to the Rasch model, and therefore, the best items that captured the construct of activity level. The data also served to identify poorly functioning items, which were subsequently flagged for removal.

Sample

The defined target population was adults, 18 years of age and older in the general public. A variety of individuals with a large range of ages and orthopaedic ailments are seen by foot and ankle orthopaedic surgeons as previously discussed; therefore, it was important to sample the normal population. Individuals who are considered part of the normal population of the United States were included as participants in this study. For the purpose of this study, normal was defined as was done previously by the American Academy of Orthopaedic Surgeons (AAOS), which is the random sample of individuals from the general population of the United States from whom scale scores are derived (Hunsaker et al., 2002). Participants who have had previous foot or ankle surgery, as well as participants who have not had previous surgery, were included. Individuals were only excluded if they were not at least 18 years old. A convenience sampling method that stratified participants by age, gender, ethnicity, education, income (a proxy for socioeconomic status) and region in the United States was used to capture a sample that was representative of the general population based on these six demographics.

Data Collection Procedures

The newly generated items that comprised the survey questionnaire were administered via email through the web-based survey platform Qualtrics (Drive Provo, Utah). The questionnaire also included demographics such as age, sex, height, weight, pregnancy status in order to account for above average body mass index (BMI) in pregnant women, previous ankle

surgery status and laterality, previous ankle injury status and laterality, previous knee surgery status and laterality, previous knee injury status and laterality, and previous and current Covid-19 status. Commonly used foot and ankle outcome scores have been reported to have different normative values based on age, sex, BMI, previous ankle injury, and previous ankle surgery status which is why these variables were included in the study (Matheny et al., 2020). No information identifying people who completed the survey was included. All participants provided consent prior to participation in the study. The average time to complete the survey was approximately 15 minutes.

Instrumentation

The questionnaire included items for the new ankle activity scale instrument and demographic items as discussed above, as well as commonly reported outcomes measures used in the foot and ankle, including a measure of general physical and mental health, the Short-Form 12 (SF-12) Physical Component Summary (PCS) and Mental Component Summary (MCS) scores (Ware et al., 1995), the FAAM ADL and Sport scales (Martin et al., 2005), and the Tegner activity scale (Tegner & Lysholm, 1985; Tegner et al., 1986). (Appendix D).

Foot and Ankle Ability Measure

A common measure of foot and ankle function is the Foot and Ankle Ability Measure (FAAM; (Martin & Irrgang, 2007; Martin et al., 2005). The FAAM is comprised of two subscales: the activities of daily living subscale (ADL) and the sport subscale. The ADL subscale consists of 21 questions that pertain to various functional activities one would encounter in normal daily activity. The sport subscale contains eight questions pertaining to different activities that are related to sport participation. Each question is based on a Likert type rating scale, with a possibility of 0 to 4 points per question, for a total of 84 possible points for the ADL

subscale and 32 possible points for the sport subscale. Anchor points include no difficulty at all (4), slight difficulty (3), moderate difficulty (2), extreme difficulty (1), and unable to do (0). N/A (not applicable) is also an option for each question on the FAAM Sport Subscale.

Participants were given the following instructions: “Please answer every question with one response that most closely describes your condition within the past week. If the activity in question is limited by something other than your foot or ankle, mark N/A.” The Foot and Ankle Ability Measure scores are recorded as a percentage of 84 points. The Foot and Ankle Ability Measure Sport scores are recorded as a percentage of 32 points. The Foot and Ankle Ability Measure Total score is recorded as a percentage of 116 points (Hale & Hertel, 2005). A previous study of patients who underwent surgical treatment for an ankle injury (n=456) reported excellent reliability and validity of FAAM ADL and Sports subscale scores utilizing the Rasch measurement model (Matheny & Clanton, 2020). A systematic review that examined patient-assessment instruments designed for patients with chronic ankle instability (39 articles, 17 instruments) also found the FAAM to be a useful instrument to quantify chronic ankle instability and reported Cronbach’s alpha coefficients for the ADL and the sport subscale scores to be .98 and .96, respectively (Eechaute et al., 2007). These values have been deemed to be within an acceptable range for the orthopaedic literature (Martin & Irrgang, 2007; Martin et al., 2005). Previous studies revealed good psychometric properties, evidence for good reliability and validity for these scores, providing evidence to support the use of this instrument in the assessment of foot and ankle injuries (Martin & Irrgang, 2007; Martin et al., 2005).

Tegner Activity Scale

Participants were given the following instructions: “Please choose ONE of the following that best describes your current activity level.” The scale ranges from 0 to 10, with 10

representing an elite athlete and 0 representing disability leave at work. The Tegner activity scale was designed as a lower extremity scale to utilize in individuals with foot, ankle, or knee problems (Tegner & Lysholm, 1985; Tegner et al., 1986, 1988). Most commonly, the score has been used in knee outcomes research; however, Tegner and Lysholm, the original developers of the scale, intended use of the scale among lower extremity problems (Briggs, Lysholm, et al., 2009; Briggs, Steadman, et al., 2009). Briggs, Lysholm, et al. (2009) reported test-retest intra-class correlation coefficients for scores on the Tegner activity scale to be .80 in patients who underwent an anterior cruciate ligament reconstruction. demonstrating reliability estimates to be above the commonly accepted value of .70 in orthopaedic literature. Previous studies have also demonstrated good psychometric properties for use of this score in the knee; however, no study has addressed psychometric properties of scores from this instrument in the ankle (Briggs, Steadman, et al., 2009).

Short-Form 12 (SF-12)

This instrument is a general measure of health comprised of two subscales: the Physical Component Summary (PCS) and the Mental Component summary (MCS; (Tucker et al., 2016; Ware et al., 1995). The two subscales indicate general physical health and general mental health, which are standardized with a mean of 50 and a standard deviation of 10 in the general population. The SF-12 has 12 questions that measure eight health domains: physical functioning, role-physical, bodily pain, general health, vitality, social functioning, role-emotional, and mental health. The scale of items was created using a Likert-type scale with different standardized scores for the PCS and MCS. The scores are computed and range from 0 to 100, where a zero score indicates the lowest level of health measured by the scales and 100 indicates the highest level of health. The SF-12 PCS and MCS scores were included in this study in order to provide

evidence of convergent and divergent validity which has been commonly done in previous studies that sought to assess other foot and ankle instruments (Martin & Irrgang, 2007; Richter et al., 2006). The SF-12 PCS and MCS, as well as the SF-36 PCS and MCS, which is an older, longer version of the SF-12, have been shown to be moderately (PCS) and weakly (MCS) correlated to other foot and ankle functional scores (Martin & Irrgang, 2007; Martin et al., 2005; Matheny et al., 2020; Richter et al., 2006). In addition, the SF-12 had often been collected to assess activity limitations in foot and ankle orthopaedic samples before other more specific and appropriate measures were developed (Herron, 2006; Richter et al., 2006). While specific foot and ankle functional instruments have since been developed, traditionally, SF-12 scores are still collected in other studies for comparison purposes (Martin & Irrgang, 2007). In the general population, the test-retest reliability coefficient of scores from the SF-12 PCS was .89 and the test-retest reliability coefficient of scores from the SF-12 MCS .76 in a general United States population (n=2,329). To assess evidence of validity, F-statistics, which represent a ratio of the amount of separation either between groups or between assessments over time relative to the within-group variance, were compared to those of the SF-36, a general health instrument that has been commonly used prior to the development of the SF-12. From the F-statistics, an RV coefficient, which is a multivariate generalization of the squared Pearson correlation coefficient, was produced. The RV coefficient for each test represents empirical validity of SF-12 scores relative to the SF-36. For all tests based on criterion variables defining differences in physical health, statistical conclusions based on the SF-12 PCS agreed with the SF-36 PCS 30 out of 36 times, indicating that the SF-12 PCS was able to measure the same construct of physical health that was originally measured using the 36 questions from the SF-36 PCS. Additionally, RV

coefficients ranged from .43 to .93 in a large sample of individuals with various severity levels of heart conditions (Ware et al., 1995, 1996, 2008).

Data Analysis

Assumptions

For sample 2, pilot data for the new instrument were analyzed using the Rasch measurement model. An exploratory factor analysis of observed responses using principal components analysis (PCA) as the extraction procedure was first conducted to test for unidimensionality, which is an assumption of the Rasch model (Bond & Fox, 2015; Lafave et al., 2016). PCA has been recommended to assess unidimensionality (Bond & Fox, 2015; Christensen et al., 2013). Unidimensionality in this study's context means that the items are only measuring ankle activity level or that they are measuring foot and ankle activity level as a unitary trait with no subdomains. A forced-factor extraction for one factor was used to identify any poorly fitting items with lower structure coefficients. Cumulative variance was assessed to determine whether there was a dominant first factor. Previous studies have recommended 75% of variance on the first factor is desired; however, reasonable values may be closer to 50% (Henson & Roberts, 2006). A scree plot was also produced to assess unidimensionality. Structure coefficients were assessed for strength. Any item that had a low coefficient (less than .30 to .40) was flagged as misfitting or poorly performing. Items that have coefficients less than .30 to .40 may indicate poorly performing items (Costello & Osborne, 2005; Henson & Roberts, 2006; Howard, 2015). SPSS version 25.0 (Chicago, Illinois) software was used for this portion of the analysis. A PCA of the Rasch standardized residuals was also conducted to further investigate unidimensionality and assess whether a single latent trait explained the majority of the variance in the data (Lo Martire et al., 2017; Muller & Roddy, 2009; Smith, 2002). Cumulative variance was assessed,

with values greater than 50% considered acceptable (Franchignoni et al., 2013; Linacre, 2000). Standardized residual correlations were also assessed to identify items with excessive co-variation, with values greater than .7 indicating potential multidimensionality (Linacre, 1998; Linacre & Wright, 2000). To perform this portion of the analysis, WINSTEPS version 4.0.1 (Beaverton, Oregon) was used. Misfitting items were assessed again in Phase 3 and removed if they were still considered to have poor fit.

Reliability

Person reliability was assessed. Possible reliability estimates range from 0 to 1.0, with higher values indicating greater internal consistency reliability. Values above .80 have been deemed acceptable (Linacre, 2020). Person reliability uses estimates on the logit scale rather than raw patient scores for each person to calculate reliability (Fisher, 1997). Person separation was also assessed, with acceptable values greater than 2.0, which was also discussed in Chapter 2. Item reliability was assessed. Values above .90 have been deemed acceptable (Linacre, 2020). Item separation was assessed, with values above the recommended threshold of 3.0 indicating acceptable item hierarchy or item difficulty range (Linacre, 2020). To perform this portion of the analysis, WINSTEPS version 4.0.1 (Beaverton, Oregon) was used.

Validity

Items were tested for fit by utilizing the Rasch measurement model, specifically using the rating scale model (RSM; Andrich, 1978a, 1978b), which can be used for polytomous data, such as Likert scale items, when all response categories remain consistent across items. To identify misfitting items, outfit MNSQ and infit MNSQ statistics were assessed, with acceptable values between 1.5 and .5, which was also described in Chapter 2. WINSTEPS version 4.0.1 (Beaverton, Oregon) was used to perform this portion of the analysis.

Various figures including Wright item-person maps, test information function, category probability plots, and item characteristic curves were produced to assess evidence of validity of scores from the new ankle activity scale. The ideal shape of a Wright map may resemble a normally distributed histogram for both items and persons. Wright item-person maps were assessed and considered as an additional piece of evidence for validity assessment, in addition to infit and outfit statistics. A test information function was generated for the new ankle activity scale. The information function is the sum of all of the item information functions and can be used to predict the accuracy to which we can measure any value of the latent ability, which is ankle activity level in this case. A cut-point of 10 with a standard deviation of .5 was used to assess test information, with recommended values of 10 or greater with standard errors (SE) less than .5 (Embretson & Reise, 2013). Category probability plots were generated and assessed. Ideal category response curves for items reflect response options for each item that are monotonically related to the ankle activity level, and each response option is likely to be selected at some differing range of the underlying construct (ankle activity). To perform this portion of the analysis, WINSTEPS version 4.0.1 (Beaverton, Oregon) was used.

Based on MNSQ infit and outfit statistics, Wright item-person maps, category response curves, and test information, individual items and overall instrument performance was assessed as a whole, for evidence of validity. Each of these pieces of information plays a role in determining how responses to the instrument are performing. Therefore, all information was carefully assessed, items were flagged for redundancy, and misfitting items were removed.

Phase 3

After identifying and/or removing poorly functioning items from the instrument derived from Phase 2, the new subset of items was tested in the normal population, which comprised the

third sample. This third sample was used to assess factor structure and psychometric properties of scores from the new instrument. Phase 3 addressed research questions one, two, and three.

Sample

The defined target population was identical to sample 2. The sample size was $N = 505$, which exceeded the recommended minimum sample size of 300 participants (Clark & Watson, 1995). Although 100 participants is the minimum sample required for Rasch analysis, it was important to capture more individuals to better represent the normal population (Green & Frantom, 2002). The same stratified convenience sampling method from sample 2 was employed using Qualtrics (Drive Provo, Utah).

Data Collection

The survey was identical to the survey administered in sample 2. The questionnaire was administered via email through the web-based survey platform Qualtrics. No information identifying people who completed the survey was included. Expected time to complete the survey for participants was approximately 10 to 15 minutes.

Data Analysis

Phase 3 sample data were analyzed by following all of the steps that were outlined and included for Phase 2. All data were assessed for unidimensionality and reliability and validity of scores from the ankle activity level instrument using the same criteria as Phase 3. The primary goal of this phase was to create a unidimensional measure of foot and ankle activity level. Unidimensionality was assessed at the beginning of the Phase 3 analysis for all 77 items, and again after final item reduction. Final item reduction was conducted based on MNSQ infit and outfit statistics, which also provided further evidence of unidimensionality, as well as Wright

item-person maps, category response curves, and test information. All figures that were generated in Phase 2 were also generated in Phase 3. Wright item-person maps were assessed for redundancy of well-fitting items. Items that were well-fitting, but redundant, were removed with the intent to retain a variety of items that required different physical movements (i.e., stairs, balance, sports, exercise, walking, etc.) rather than one type of physical movement (i.e., balance only). Ability levels that had three items grouped at the same level were assessed first as previously recommended (Green & Frantom, 2002). Ability levels that had two items grouped at the same level were assessed next. Redundant items that had the greatest deviation from 1.0 for infit values were removed first. Following final item reduction, a subsequent Rasch analysis was conducted to determine improvements in psychometric properties. WINSTEPS version 4.0.1 (Beaverton, Oregon) was used to perform this portion of the analysis.

After the final structure of the ankle activity level scale was determined, descriptive statistics including means, standard deviations, skewness, and kurtosis were examined (Table 2).

Table 2

Descriptive Statistics for FAALS Percentage

<i>N</i>	505
Mean	67
Standard Deviation	22
Minimum	5
Maximum	100
Skewness	-0.63
Kurtosis	-0.16

Additionally, correlations between scores on the new ankle activity level instrument and FAAM subscale scores, SF-12 subscale scores, and Tegner activity level scores were conducted to assess convergent and discriminant validity using the Pearson product moment correlation coefficient. Positive, linear correlations of .6 and above were considered as evidence of convergent validity, and weak correlations of .2 and below were considered as evidence of divergent validity (Matheny et al., 2020; Matheny & Clanton, 2020). A multiple linear regression analysis was conducted to assess discrimination and determine mean differences in activity level scores between individuals who have had previous ankle surgery and those who have not, as well as between those with a normal BMI and those who are considered obese or morbidly obese. Since BMI and previous surgery are factors that affect foot and ankle function, a multiple linear regression analysis was utilized to adjust for each of those factors (Matheny et al., 2020). Assumptions of multiple linear regression were tested, including linearity, the mean of residuals is equal to zero, homoscedasticity of residuals, independence of residuals, and normal distribution of residuals. Multicollinearity was also assessed. Residual plots, histograms, and a quantile-quantile (QQ) plot were produced to assess assumptions. To assess multicollinearity, variance inflation factor (VIF) and condition indices were produced. Generally accepted values of VIF are less than or equal to 10.0, and condition indices are ideally less than 15 to 30 (Chatterjee & Hadi, 2015). Significance level was set at an alpha of .05 for all statistical tests. Software used for this portion of the analysis was SPSS version 25.0 (Chicago, Illinois). Normative values, including means and standard deviations of the FAALS aggregate percentages, for BMI (normal versus overweight/obese), previous ankle surgery, age category (<25, 25 – 30 and \geq 40) and sex were documented since these variables have been shown to

affect ankle function. The age and BMI categories were used since they have been shown to be meaningful in the foot and ankle in a clinical setting (Matheny et al., 2020).

Phase 4

Standard Setting and Instrument Item Structure

To answer research question four, a criterion-referenced standard setting procedure was used to determine thresholds for assigning activity levels in a hierarchical fashion. This provided a simple clinical assessment tool for ankle and foot practitioners. Responses for all items were structured as five-point Likert-type responses, on a scale of zero to four, with zero representing no ability on the activity and four representing the highest ability on the activity.

There are various standard-setting procedures that are commonly used including holistic standard setting which entails judgement from experts who examine the instrument as a whole rather than by focusing on individual items. Judges provide a passing standard estimate and the results are then averaged to determine a final cut score for passing (Kellow & Wilson, 2008). Another commonly used standard setting procedure is the content-based basic Angoff method which entails assembling a panel of judges who are presented with instrument items. The judges are then asked to think of a reference group of 100 individuals comprised of those who are considered “minimally acceptable,” “borderline,” or “barely proficient” in the construct of interest and estimate the number of people who should answer each item correctly. Truly what the judges are doing is estimating the probability of success, which is then equated to a cut-score for passing by averaging all the judges’ reported probabilities or percentages. There are also other modified versions of this method (Kellow & Wilson, 2008). Another procedure, the bookmark method, is a performance-based standard setting procedure that relies on a combination of empirical data and expert judgement (Beretvas, 2016; Cizek & Bunch, 2007).

This criterion-referenced method is founded on item-mapping and requires output that allows difficulty to be rank-ordered from lowest to highest item difficulty, such as based on the Rasch model and other IRT models. A booklet of items is then created, starting with the easiest item on the first page and the most difficult item on the last page. Experts are then given the booklet and asked to place a bookmark between thresholds. The thresholds would then represent the transition points at which the ankle ability level changes in terms of clinical significance and would then represent activity levels for the FAALS. After the first round of bookmark placement, experts are allowed to discuss bookmark placement in a small group. The small groups of experts will then rectify any thresholds/levels that have not reached a consensus. In the final round, all experts are allowed to discuss findings, and final modifications will be made to reach a consensus (Kellow & Wilson, 2008).

It is important to note that most standard setting procedures are based in educational assessments, where there are correct and incorrect answers to items (Cella et al., 2014; Kellow & Wilson, 2008; Lewis & Cook, 2020). However, the construct for this study, ankle activity level, is not based on correct and incorrect answers, but rather varying physical abilities, which makes applying many of these standard setting procedures difficult for health-outcomes measurement. A study that measured physical function in cancer patients using the Patient-Reported Outcomes Measurement System (PROMIS) used a modified version of the bookmark method. Item responses were originally calibrated using a graded-response model and derived scores were transformed to a T-score metric from previously collected PROMIS score data from the normal population. Modifications were based on item and response structure, with varying levels of physical function that were identified based on the distribution of the scores (Cella et al., 2014; Rothrock et al., 2019).

For the current study, a modified version of the bookmark method was utilized. This method combines the Rasch output that utilizes logits to order items in terms of difficulty, as well as recommendations from experts in the field of interest (Kellow & Wilson, 2008). The finalized list of 22 items was disseminated to the orthopaedic experts, and they were asked to confirm or deny the order of difficulty of the items based on the Rasch analysis. If experts disagreed with the order of items, they were asked to re-order items. To determine a final order of items, the majority of expert input was utilized. For example, if four experts decided that Item 21 was more difficult than Item 22, then Item 21 would move to the higher difficulty level and Item 22 would move to the next highest difficulty level. If there were any ties for item placement majority, the order of that item would default to the order of items from the Rasch analysis. For example, if Item 1 had three votes to remain Item 1, and Item 1 also had three votes to be moved and re-ordered as Item 2, Item 1 would remain in its original placement as Item 1 as determined by the Rasch analysis. If item placement was tied for two different placements that did not include the original placement, experts would be asked to vote again and come to consensus. For example, if Item 6 had two votes to be re-ordered as Item 8, two votes to be re-ordered as Item 9, one vote to be re-ordered as Item 5, and one vote to reordered as Item 6, all experts would be asked to come to consensus amongst themselves. After activity levels were determined, raw point and percentage ranges were determined in order to assign activity levels for each participant. Point ranges for activity level categorization was dependent on the number of items assigned to each activity level, with each item being worth a possible four points each.

Chapter III Summary

This chapter has outlined the four phases of development for the ankle activity level instrument. The goal of Phase 1 was to establish the breadth and depth of the ankle activity level construct, operationally define this construct, generate items, and assess items for various

characteristics including content appropriateness, level of difficulty, readability and interpretability, by drawing on various expert judgements and opinions. The goal of Phase 2 was to pilot the pool of generated items in the normal population that were developed in Phase 1. The data from Phase 2 was used to identify item fit using the Rasch model. Poorly functioning items were then then flagged for removal to improve the instrument. The goal of Phase 3 was to assess psychometric properties of scores from the new ankle activity level instrument by addressing each of the study hypotheses. First the assumption of unidimensionality was tested, then reliability was assessed, including item and person reliability. Next, evidence of validity was determined by assessing mean square infit and outfit values and Wright-item person maps. In addition, convergent and divergent validity of scores were examined by conducting a Pearson correlation between the new ankle activity level scores and the SF-12 PCS and SF-12 MCS, respectively. Separation of groups, or discrimination, was also assessed by determining the person separation index, and conducting a multiple linear regression analysis to examine differences in activity level scores between individuals who have undergone previous ankle surgery compared to those who have not undergone previous ankle surgery, as well as between individuals who have a normal BMI compared to individuals who have a BMI indicating obese or morbidly obese. After assessing all psychometric properties of scores from the new ankle activity level instrument, a standard setting procedure was applied in Phase 4, in order to assign activity levels to corresponding items and scores. By completing all of these procedures and steps, the goal was to develop a simple ankle activity level instrument that can be used in a simplistic manner, in a clinical setting by various foot and ankle practitioners to quickly and accurately assess ankle activity level. Study results are presented in Chapter 4.

CHAPTER IV

RESULTS

Data were recorded and analyzed in four phases as outlined in Chapter 3. The purpose of Phase 1 was to define the construct of ankle activity level and complete item generation for the ankle activity level scale. The purpose of Phase 2 was to pilot the items generated in Phase 1 and flag poorly performing items for item reduction. The purpose of Phase 3 was to finalize item selection and test sensitivity and convergent and divergent evidence of validity of the ankle activity level scale. The purpose of Phase 4 was to finalize difficulty order of items and determine thresholds for activity levels of the ankle activity level scale.

Phase 1

Interviews with six experts in the orthopaedic community were conducted in order to develop a representative list of items that represent the construct of ankle activity level. Item generation resulted in a list of 101 items. After reviewing the orthopaedic expert panel's description of items in terms of content representativeness and difficulty, 20 items were removed, retaining 81 items total. The reading expert suggested reading comprehension levels above the seventh grade for 31 items and adjustments were made to accommodate a seventh-grade reading level for these items. The reading expert also helped determine the most appropriate wording for all 81 items. Two individuals from the normal population identified ambiguity or clarity issues with four items. These items were removed due to severe clarity issues, leaving a total of 77 items (Appendix E).

Phase 2

There were 100 participants included in the pilot analysis of the 77 items for the new ankle activity scale. The initial PCA of observed responses revealed that there was a lack of unidimensionality, which was expected with the large number of items. The cumulative variance on the first factor was 51.7%, which was slightly above the previously defined acceptable threshold of 50%, with nine factors having an Eigenvalue greater than 1.0 (Table 3).

Table 3

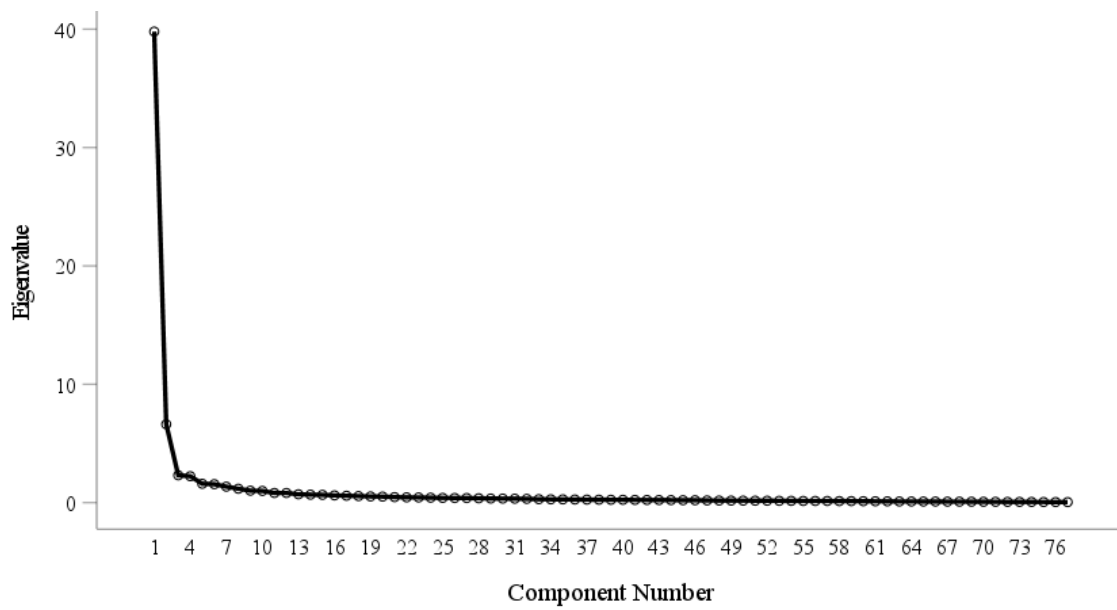
Eigenvalues and Total Variance Explained by Factor

Component	Eigenvalue	Percent Variance
1	39.8	51.7
2	6.6	8.6
3	2.3	3.0
4	2.2	2.9
5	1.6	2.1
6	1.6	2.0
7	1.4	1.8
8	1.2	1.5
9	1.0	1.3

The scree plot below also demonstrated the lack of unidimensionality, with the bend (elbow) in the scree plot at three (Figure 8).

Figure 8

Scree Plot: Phase 2 Data (Pilot), 77 items



By assessing the component matrix, there were no items that had a coefficient with a value less than the established threshold of .40 (Table 4).

Table 4*Component Matrix for Phase 2 Data (Pilot) 77 Items*

Items	Component 1
AAS_STAIRS3	.82
AAS_DAILY11	.82
AAS_DAILY12	.82
AAS_CUTTING3	.81
AAS_SPORT3	.81
AAS_STAIRS6	.80
AAS_CUTTING2	.80
AAS_STAIRS8	.80
AAS_DAILY10	.78
AAS_STAIRS4	.79
AAS_STAIRS1	.79
AAS_EXERCISE2	.79
AAS_SPORT7	.79
AAS_SPORT6	.78
AAS_STAIRS2	.78
AAS_CUTTING5	.78
AAS_DAILY9	.78
AAS_EXERCISE1	.77
AAS_SPORT5	.77
AAS_STAIRS5	.77
AAS_DAILY16	.76
AAS_EXERCISE3	.76
AAS_WALK4	.76
AAS_WALK8	.75
AAS_EXERCISE4	.75
AAS_WALK5	.75
AAS_WALK7	.75
AAS_CUTTING4	.75
AAS_WALK6	.75
AAS_CUTTING1	.75
AAS_DAILY8	.74
AAS_SPORT2	.74
AAS_BALANCE12	.74
AAS_DAILY1	.73
AAS_SPORT11	.73
AAS_BALANCE9	.73
AAS_SPORT8	.72
AAS_DAILY2	.72
AAS_DAILY5	.71
AAS_WALK3	.71
AAS_DAILY15	.71

Table 4, *continued*

Items	Component 1
AAS_SPORT13	.71
AAS_RUN1	.71
AAS_BALANCE3	.71
AAS_DAILY3	.71
AAS_BALANCE10	.70
AAS_BALANCE2	.70
AAS_STAIRS7	.70
AAS_BALANCE8	.70
AAS_BALANCE11	.70
AAS_SPORT4	.70
AAS_DAILY13	.68
AAS_RUN4	.68
AAS_EXERCISE5	.68
AAS_SPORT9	.67
AAS_SPORT14	.67
AAS_SPORT12	.66
AAS_BALANCE1	.66
AAS_DAILY4	.66
AAS_SPORT1	.66
AAS_SPORT10	.66
AAS_BALANCE4	.66
AAS_SPORT15	.66
AAS_RUN5	.66
AAS_BALANCE7	.65
AAS_WALK2	.65
AAS_RUN2	.65
AAS_SPORT16	.64
AAS_DAILY14	.63
AAS_BALANCE6	.62
AAS_BALANCE5	.62
AAS_DAILY6	.62
AAS_WALK1	.60
AAS_RUN3	.59
AAS_SPORT17	.59
AAS_SPORT18	.55

The Rasch analysis was conducted with all 77 items. The observed cumulative variance from the PCA of the Rasch residuals was 62.5% and the expected cumulative variance was 63.6%, which are very similar values, and are above the acceptable threshold of 50%, providing evidence of unidimensionality and indicating that the data fit the Rasch model. There were seven

standardized residual inter-item correlations that were greater than the threshold of .7, indicating a lack of local independence or unidimensionality. Person reliability was .95 and person separation was 4.15, which were both above the recommended thresholds. Item reliability was .98 and item separation was 6.81, which were also both above the recommended thresholds. Three items had mean-square infit values greater than or equal to the recommended threshold of 1.50, indicating poor fit. Ten items had mean-square outfit values greater than or equal to the recommended threshold of 1.50, indicating poor fit. Overall, there were 11 items that indicated poor fit for both infit and outfit values (Table 5).

Table 5*Mean-square Infit and Outfit: Phase 2 Data, 77 Items*

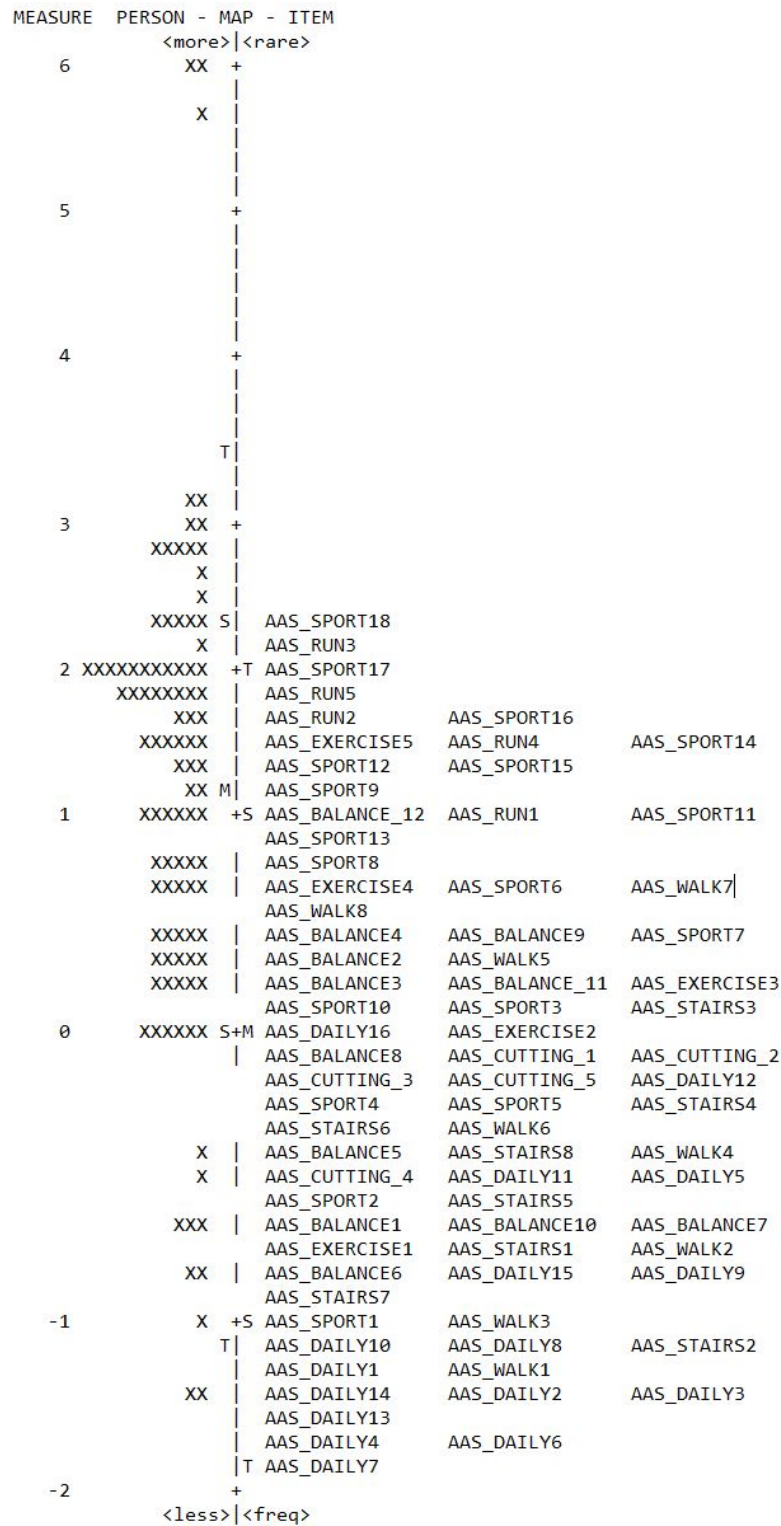
Item	Infit MNSQ	Outfit MNSQ
AAS_RUN3	1.77*	1.67*
AAS_RUN4	1.54*	1.39
AAS_BALANCE5	1.50*	1.82*
AAS_BALANCE6	1.47	1.78*
AAS_SPORT1	1.43	1.18
AAS_WALK1	1.40	2.02*
AAS_SPORT9	1.40	1.45
AAS_RUN2	1.39	1.28
AAS_SPORT18	1.34	1.54*
AAS_SPORT10	1.30	1.14
AAS_SPORT2	1.29	1.09
AAS_SPORT14	1.28	1.29
AAS_BALANCE9	1.27	1.54*
AAS_SPORT17	1.27	1.25
AAS_RUN5	1.27	1.16
AAS_SPORT4	1.23	1.55*
AAS_WALK2	1.22	1.52*
AAS_BALANCE3	1.19	1.42
AAS_BALANCE8	1.18	1.90*
AAS_BALANCE7	1.17	2.31*
AAS_BALANCE2	1.16	1.29
AAS_SPORT16	1.15	1.04
AAS_BALANCE4	1.13	1.47
AAS_SPORT5	1.13	1.12
AAS_DAILY14	1.13	1.07
AAS_CUTTING1	1.11	0.83
AAS_RUN1	1.09	1.01
AAS_STAIRS2	1.09	0.81
AAS_STAIRS1	1.08	1.05
AAS_SPORT8	1.08	1.04
AAS_SPORT15	1.07	1.04
AAS_SPORT13	1.02	0.92
AAS_WALK5	1.01	0.87
AAS_WALK4	1.00	0.77
AAS_CUTTING2	0.99	0.79
AAS_CUTTING5	0.98	0.95
AAS_DAILY6	0.97	1.04

Table 5, *continued*

Item	Infit MNSQ	Outfit MNSQ
AAS_STAIRS7	0.97	0.82
AAS_DAILY5	0.96	1.05
AAS_BALANCE1	0.96	1.02
AAS_WALK8	0.96	0.90
AAS_DAILY7	0.91	0.68
AAS_DAILY12	0.90	0.73
AAS_SPORT6	0.89	0.90
AAS_STAIRS8	0.88	0.88
AAS_DAILY9	0.88	0.79
AAS_EXERCISE5	0.87	0.94
AAS_SPORT12	0.87	0.79
AAS_CUTTING4	0.87	0.69
AAS_EXERCISE1	0.87	0.67
AAS_WALK3	0.87	0.64
AAS_EXERCISE4	0.85	0.83
AAS_DAILY3	0.84	0.81
AAS_DAILY8	0.83	0.70
AAS_SPORT7	0.82	0.90
AAS_STAIRS3	0.81	0.84
AAS_DAILY11	0.81	0.79
AAS_STAIRS5	0.79	0.95
AAS_STAIRS6	0.79	0.81
AAS_EXERCISE3	0.79	0.78
AAS_SPORT3	0.79	0.75
AAS_DAILY13	0.79	0.51
AAS_DAILY15	0.78	0.75
AAS_WALK7	0.78	0.73
AAS_SPORT11	0.78	0.71
AAS_DAILY4	0.77	0.52
AAS_BALANCE10	0.76	1.07
AAS_STAIRS4	0.75	0.79
AAS_EXERCISE2	0.75	0.74
AAS_CUTTING3	0.75	0.66
AAS_DAILY10	0.74	0.85
AAS_DAILY16	0.73	0.71
AAS_DAILY1	0.70	0.92
AAS_WALK6	0.66	0.53
AAS_DAILY2	0.63	0.68

*Values above acceptable threshold

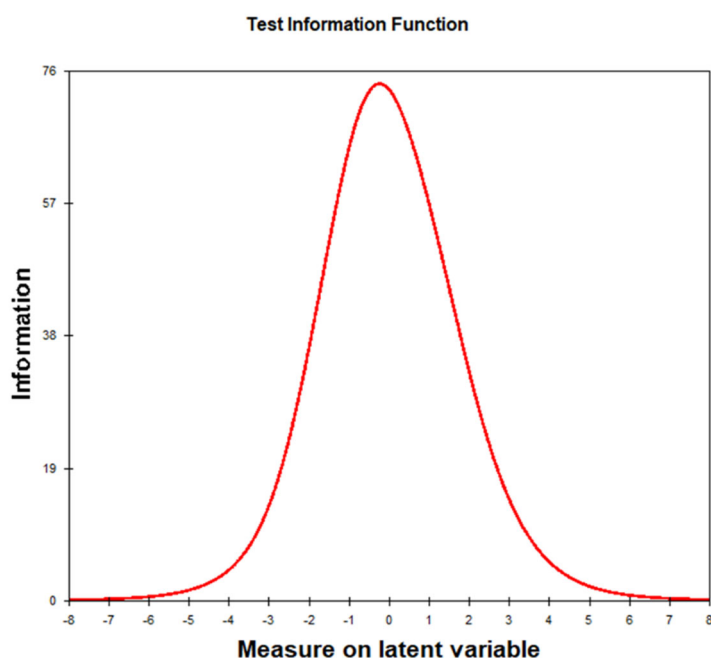
The Wright item-person map was then assessed for multiple items plotted at the same level of difficulty (Figure 9). This map demonstrates redundancy in the instrument where triplicate items are found at the same difficulty level. This map also shows a lack of extremely difficult items. The most difficult item that is included in the scale is AAS_SPORT18, which represents “Completing a half-marathon in 2 hours or less (9.2 minutes per miles).” Although the map shows a lack of difficult items, the map may be more indicative of an abundance of individuals with extremely high abilities, especially since the ability levels that seem to be lacking items are at approximately 3 to 6 standard deviations above the average ability level. This will be discussed further in Chapter 5.

Figure 9*Wright Item Person Map: Phase 2 Data (Pilot)*

Difficulty levels that had multiple items at the same difficulty level were identified for redundancy and denoted for future removal. Many difficulty levels had two and three items grouped together. These items were flagged and then reassessed for redundancy during Phase 3. Test information function was visualized to assess evidence of validity (Figure 10). Test information was approximately 75, which was greater than the acceptable threshold of 10, as previously defined, and the shape was tall and narrow which is generally indicative of high precision for the instrument in the middle ability range and provides some evidence of validity.

Figure 10

Test Information Function for Phase 2 Data (Pilot)



Category probability plots and item characteristic curves were also produced to assess evidence of validity. Each item of the ankle activity level scale has a corresponding category probability plot and an item characteristic curve, which were extremely similar for all items. An example of a category probability plot and item characteristic curve, using item 1, is visualized (Figures 11 and 12, respectively). Category probability plots were similar for all items. Based on step

calibrations of Andrich threshold, all categories were ordered and increased monotonically, as demonstrated by item 1. The category measure, which represents the average ability estimate in logits was -2.11 for persons who chose response category 0, -0.87 for category 1, -0.05 for category 2, 0.83 for category 3, and 2.26 for category 4. This will be discussed further in Chapter 5.

Figure 11

Category Probability Curves for Item 1 of Phase 2 Data (Pilot)

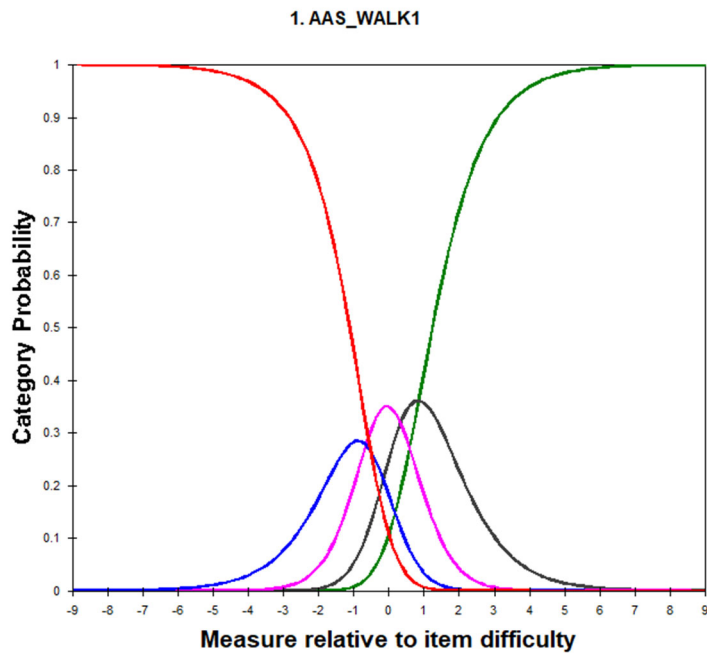
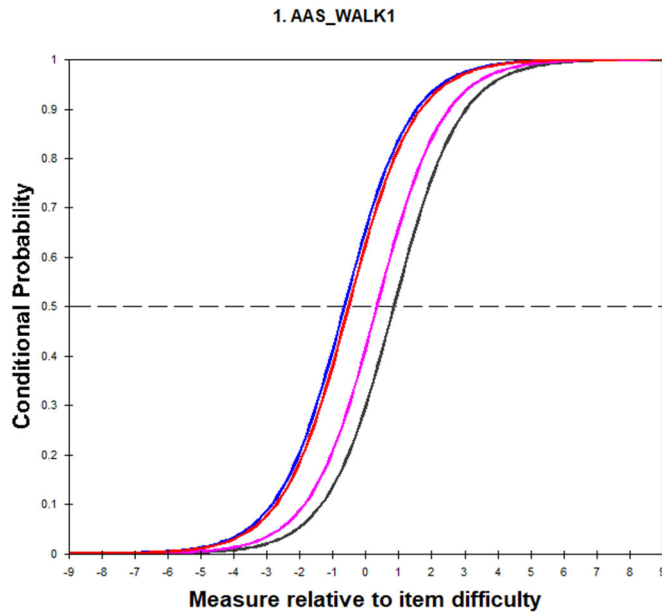


Figure 12

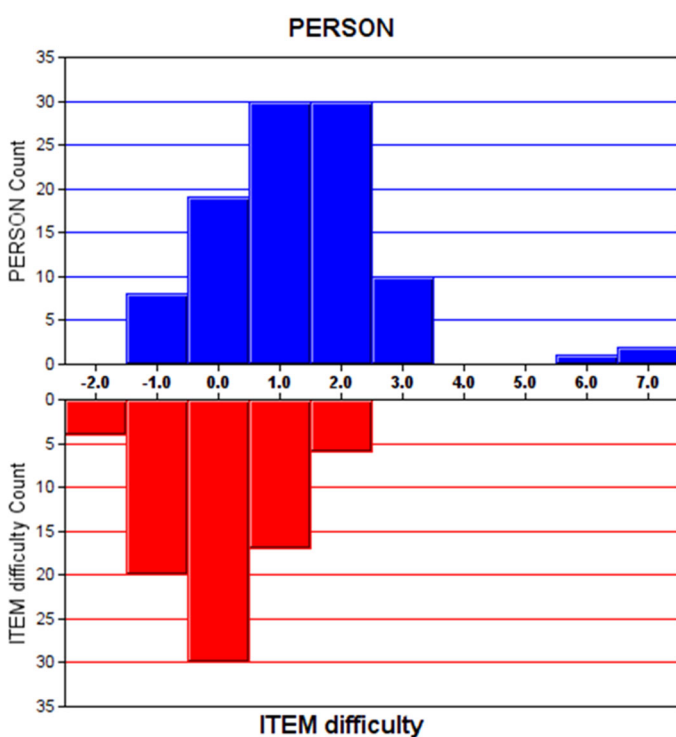
Item Characteristic Curve for Item 1 of Phase 2 Data (Pilot)



The item characteristic curve demonstrates that as person ability increases, so does the probability of endorsing an item (Figure 12). A Wright item person histogram was also produced (Figure 13). This histogram also demonstrates the lack of more difficult items at six to seven standard deviations above the mean.

Figure 13

Wright Item Person Histogram: Phase 2 Data (Pilot)



After evaluating all of the results, the 11 misfitting items were removed leaving 66 items available.

Phase 3

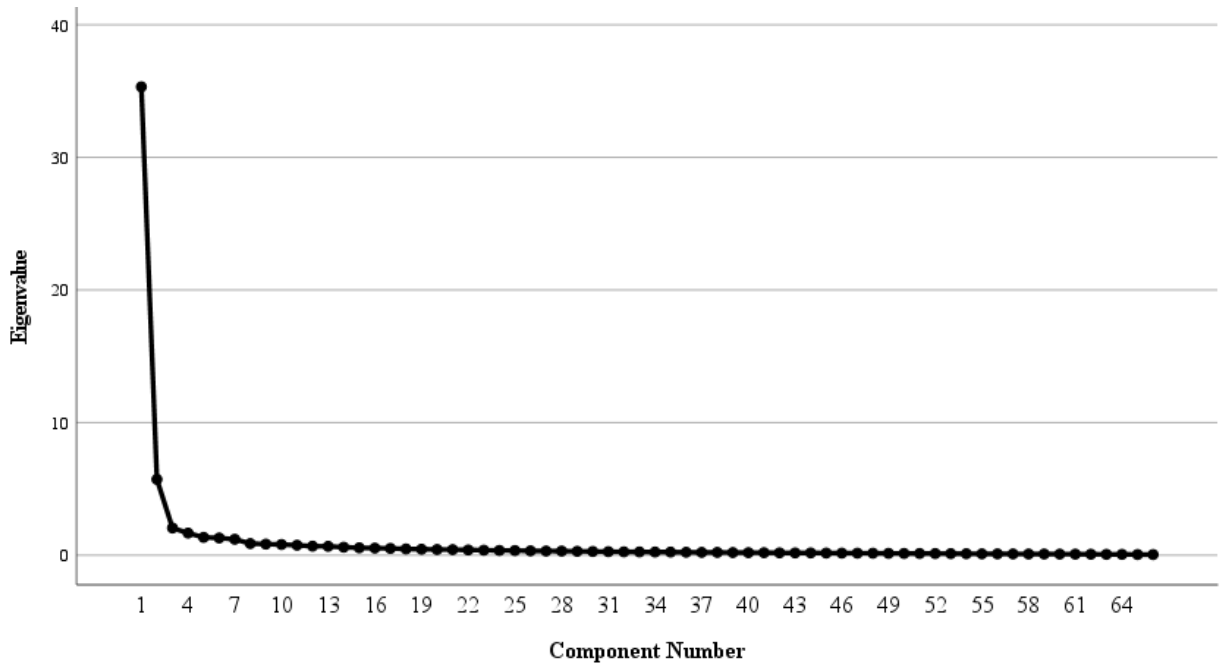
Rasch Analysis

There were 505 participants included in the phase 3 analysis of the 66 items for the new ankle activity scale. The PCA revealed that there was a lack of unidimensionality, which was still expected with a large number of items. The cumulative variance on the first factor was 53.5%, which was above the acceptable threshold of 50%, with seven factors having an Eigenvalue greater than 1.0 (Table 6).

Table 6*Eigenvalues and Total Variance: Phase 3 Data*

Component	Eigenvalue	Percent Variance
1	53.5	53.5
2	8.7	62.2
3	3.1	65.3
4	2.5	67.9
5	2.1	69.9
6	2.0	71.9
7	1.8	73.7

The scree plot below still demonstrated a lack of unidimensionality (Figure 14).

Figure 14*Scree Plot: Phase 3 Data, 66 items*

As seen in the initial PCA in the pilot data, the component matrix was produced by forcing a one-factor solution which contained no items that had a coefficient with a value less than the established threshold of .40 (Table 7).

Table 7*Component Matrix: Phase 3 Data, 66 Items*

Items	Component 1
AAS_STAIRS3	.83
AAS_DAILY12	.83
AAS_DAILY11	.83
AAS_CUTTING3	.81
AAS_STAIRS6	.81
AAS_CUTTING2	.81
AAS_SPORT3	.81
AAS_DAILY10	.80
AAS_STAIRS4	.80
AAS_STAIRS8	.80
AAS_STAIRS1	.79
AAS_EXERCISE2	.79
AAS_STAIRS2	.79
AAS_SPORT7	.79
AAS_DAILY9	.78
AAS_CUTTING5	.78
AAS_SPORT6	.78
AAS_EXERCISE1	.78
AAS_STAIRS5	.77
AAS_DAILY16	.77
AAS_EXERCISE3	.77
AAS_SPORT5	.77
AAS_WALK4	.76
AAS_DAILY8	.76
AAS_EXERCISE4	.75
AAS_CUTTING4	.75
AAS_CUTTING1	.75
AAS_WALK8	.75
AAS_WALK5	.75
AAS_WALK7	.75
AAS_WALK6	.74
AAS_SPORT2	.74
AAS_DAILY1	.74
AAS_BALANCE12	.73
AAS_SPORT11	.73
AAS_DAILY2	.72
AAS_DAILY5	.72

Table 7, *continued*

Items	Component 1
AAS_SPORT8	.72
AAS_DAILY15	.72
AAS_DAILY7	.72
AAS_DAILY3	.71
AAS_WALK3	.71
AAS_SPORT13	.71
AAS_STAIRS7	.71
AAS_RUN1	.70
AAS_BALANCE10	.70
AAS_BALANCE_11	.69
AAS_BALANCE3	.69
AAS_BALANCE2	.69
AAS_DAILY13	.69
AAS_EXERCISE5	.68
AAS_SPORT14	.67
AAS_SPORT9	.67
AAS_DAILY4	.67
AAS_SPORT12	.66
AAS_SPORT1	.66
AAS_SPORT10	.66
AAS_SPORT15	.66
AAS_BALANCE1	.65
AAS_RUN5	.64
AAS_BALANCE4	.64
AAS_SPORT16	.64
AAS_RUN2	.64
AAS_DAILY14	.63
AAS_DAILY6	.63
AAS_SPORT17	.58

Therefore, the Rasch analysis was conducted with all 66 items. The observed cumulative variance from the PCA of the Rasch residuals was 66.9% and the expected cumulative variance was 68.2%, which are similar values, and provide evidence of unidimensionality and indicate that the data fit the Rasch model. There were six standardized residual inter-item correlations that were greater than the threshold of .7, indicating a lack of local independence or

unidimensionality. Person reliability was .95 and person separation was 4.27, which were both above the recommended thresholds of .80 and 2.0, respectively. Item reliability was 1.00 and item separation was 16.89, which were also both above the recommended thresholds of .90 and 3.0, respectively. One item had a mean-square infit value greater than or equal to the recommended threshold of 1.50, indicating poor fit. Three items had mean-square outfit values greater than or equal to the recommended threshold of 1.50, indicating poor fit. There were no items with infit values less than .50. Overall, there were four items that indicated poor fit for both infit and outfit values (Table 8).

Table 8*Mean-square Infit and Outfit: Phase 3 Data, 66 Items*

Item	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
AAS_SPORT10	1.52*	6.75	1.45	4.38
AAS_BALANCE4	1.47	6.26	1.63*	6.08
AAS_RUN1	1.42	5.74	1.28	3.09
AAS_SPORT1	1.41	4.02	1.15	0.99
AAS_RUN2	1.37	5.08	1.39	3.83
AAS_BALANCE3	1.35	4.49	1.66*	5.42
AAS_BALANCE1	1.29	3.23	1.28	1.83
AAS_CUTTING1	1.28	3.61	1.15	1.38
AAS_RUN5	1.25	3.48	1.21	2.08
AAS_BALANCE2	1.25	3.37	1.57*	5.19
AAS_SPORT16	1.24	3.41	1.23	2.24
AAS_SPORT2	1.22	2.66	1.08	0.66
AAS_SPORT13	1.21	3.15	1.06	0.76
AAS_DAILY6	1.20	1.74	0.94	-0.27
AAS_SPORT17	1.20	2.60	1.21	1.77
AAS_BALANCE10	1.19	2.42	1.37	2.97
AAS_SPORT9	1.19	2.85	1.09	1.01
AAS_SPORT15	1.19	2.68	1.10	1.09
AAS_CUTTING5	1.17	2.29	1.00	0.08
AAS_SPORT12	1.15	2.21	1.11	1.16
AAS_SPORT5	1.13	1.60	0.97	-0.21
AAS_SPORT14	1.12	1.76	1.01	0.12
AAS_BALANCE11	1.09	1.32	1.38	3.80
AAS_WALK5	1.07	1.03	1.10	1.16
AAS_DAILY14	1.07	0.68	1.73	3.27
AAS_SPORT8	1.06	0.89	1.00	0.09
AAS_CUTTING4	1.05	0.60	1.00	0.04
AAS_WALK4	1.04	0.59	0.94	-0.49
AAS_DAILY5	1.04	0.59	1.17	1.51
AAS_STAIRS7	1.03	0.37	1.69	3.82
AAS_DAILY4	1.03	0.34	0.92	-0.41
AAS_EXERCISE5	1.03	0.49	1.18	1.91
AAS_CUTTING2	1.00	-0.01	0.85	-1.42
AAS_DAILY8	1.00	0.01	0.73	-1.96
AAS_WALK3	0.98	-0.19	0.84	-1.03
AAS_DAILY16	0.98	-0.25	1.13	1.35
AAS_SPORT11	0.98	-0.27	0.88	-1.49
AAS_BALANCE12	0.97	-0.38	1.01	0.16
AAS_STAIRS8	0.96	-0.48	0.82	-1.65
AAS_DAILY13	0.96	-0.36	0.77	-1.25
AAS_DAILY12	0.95	-0.65	0.77	-2.41
AAS_EXERCISE1	0.95	-0.62	0.86	-1.06

Table 8, *continued*

Item	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
AAS_WALK8	0.94	-0.93	0.99	-0.10
AAS_DAILY2	0.94	-0.54	0.71	-1.83
AAS_WALK7	0.93	-1.06	1.00	0.07
AAS_DAILY15	0.93	-0.74	0.92	-0.49
AAS_EXERCISE3	0.93	-1.08	0.94	-0.63
AAS_WALK6	0.92	-1.14	1.03	0.30
AAS_DAILY3	0.92	-0.85	0.78	-1.37
AAS_SPORT7	0.89	-1.66	0.84	-2.02
AAS_DAILY7	0.88	-1.18	0.59	-2.47
AAS_EXERCISE4	0.87	-2.16	0.96	-0.40
AAS_SPORT6	0.87	-2.05	0.85	-1.75
AAS_STAIRS4	0.86	-1.98	0.88	-1.12
AAS_STAIRS5	0.86	-1.82	0.91	-0.71
AAS_EXERCISE2	0.85	-2.16	0.79	-2.18
AAS_CUTTING3	0.84	-2.35	0.82	-1.69
AAS_STAIRS1	0.81	-2.38	0.72	-2.06
AAS_DAILY1	0.79	-2.19	0.61	-2.56
AAS_DAILY11	0.79	-3.00	0.69	-2.92
AAS_SPORT3	0.78	-3.55	0.83	-1.84
AAS_DAILY10	0.77	-2.85	0.64	-2.76
AAS_STAIRS2	0.74	-3.03	0.65	-2.51
AAS_STAIRS6	0.73	-4.16	0.77	-2.41
AAS_DAILY9	0.72	-3.62	0.79	-1.47
AAS_STAIRS3	0.69	-5.11	0.72	-3.25

The four poorly fitting items, which were based on mean-square infit values greater than 1.5, were removed and the Rasch analysis was performed again on the remaining 62 items. Person reliability was .95 and person separation was 4.26, which were both above the recommended thresholds. Item reliability was 1.00 and item separation was 17.51, which were also both above the recommended thresholds. There were no misfitting items after removing the four poorly fitting items and re-running the Rasch analysis (Table 9).

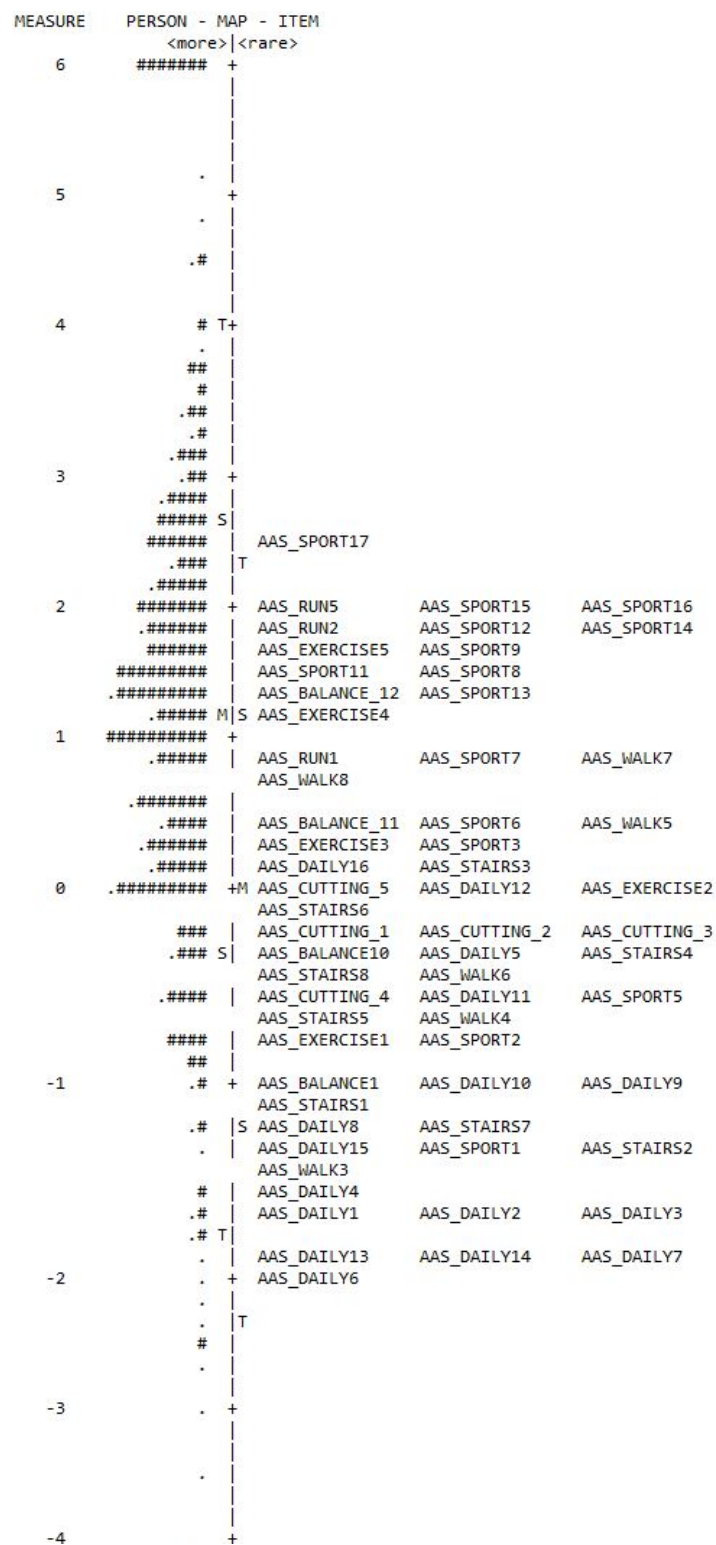
Table 9*Mean-square Infit and Outfit: Phase 3 Data, 62 Items*

Item	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
AAS_RUN1	1.45	6.12	1.31	3.40
AAS_SPORT1	1.44	4.27	1.19	1.19
AAS_RUN2	1.40	5.43	1.43	4.12
AAS_BALANCE1	1.36	3.86	1.40	2.56
AAS_CUTTING1	1.32	4.06	1.18	1.73
AAS_RUN5	1.28	3.85	1.24	2.32
AAS_BALANCE10	1.27	3.34	1.55	4.25
AAS_SPORT2	1.25	2.90	1.11	0.89
AAS_SPORT16	1.25	3.49	1.24	2.31
AAS_SPORT13	1.24	3.53	1.08	1.00
AAS_SPORT9	1.22	3.20	1.12	1.36
AAS_DAILY6	1.21	1.78	0.97	-0.06
AAS_SPORT17	1.21	2.76	1.25	2.03
AAS_CUTTING5	1.20	2.64	1.03	0.36
AAS_SPORT15	1.20	2.82	1.11	1.17
AAS_BALANCE11	1.17	2.42	1.55	5.30
AAS_SPORT12	1.17	2.42	1.13	1.32
AAS_SPORT5	1.16	1.98	0.99	-0.01
AAS_SPORT14	1.14	2.02	1.03	0.33
AAS_WALK5	1.10	1.47	1.14	1.60
AAS_DAILY14	1.09	0.82	1.79	3.39
AAS_CUTTING4	1.08	0.97	1.03	0.32
AAS_SPORT8	1.08	1.20	1.02	0.30
AAS_DAILY5	1.07	0.99	1.24	2.04
AAS_WALK4	1.06	0.84	0.95	-0.38
AAS_STAIRS7	1.06	0.67	1.74	4.09
AAS_EXERCISE5	1.06	0.97	1.22	2.34
AAS_DAILY4	1.05	0.51	0.96	-0.21
AAS_BALANCE12	1.03	0.51	1.09	1.10
AAS_CUTTING2	1.02	0.34	0.89	-1.11
AAS_DAILY8	1.00	0.05	0.73	-1.90
AAS_DAILY16	1.00	0.06	1.15	1.57
AAS_SPORT11	1.00	0.06	0.89	-1.26
AAS_WALK3	0.99	-0.03	0.85	-0.94
AAS_WALK8	0.98	-0.28	1.04	0.47
AAS_STAIRS8	0.98	-0.20	0.85	-1.42
AAS_WALK7	0.96	-0.55	1.05	0.66

Table 9, *continued*

Item	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
AAS_DAILY12	0.96	-0.54	0.77	-2.42
AAS_DAILY13	0.96	-0.30	0.79	-1.13
AAS_EXERCISE1	0.96	-0.43	0.87	-1.04
AAS_WALK6	0.95	-0.71	1.09	0.81
AAS_DAILY2	0.95	-0.46	0.72	-1.72
AAS_DAILY15	0.95	-0.58	0.95	-0.30
AAS_EXERCISE3	0.95	-0.70	0.98	-0.22
AAS_DAILY3	0.92	-0.83	0.79	-1.28
AAS_SPORT7	0.92	-1.22	0.86	-1.70
AAS_SPORT6	0.90	-1.50	0.88	-1.39
AAS_EXERCISE4	0.89	-1.77	0.99	-0.13
AAS_STAIRS4	0.88	-1.67	0.91	-0.77
AAS_STAIRS5	0.88	-1.59	0.96	-0.33
AAS_EXERCISE2	0.88	-1.76	0.82	-1.93
AAS_DAILY7	0.87	-1.24	0.60	-2.40
AAS_CUTTING3	0.86	-2.02	0.86	-1.28
AAS_STAIRS1	0.81	-2.27	0.73	-1.98
AAS_SPORT3	0.81	-2.99	0.88	-1.36
AAS_DAILY1	0.80	-2.08	0.62	-2.44
AAS_DAILY11	0.80	-2.81	0.71	-2.72
AAS_DAILY10	0.78	-2.68	0.65	-2.69
AAS_STAIRS2	0.75	-2.96	0.66	-2.38
AAS_STAIRS6	0.75	-3.79	0.80	-2.04
AAS_DAILY9	0.73	-3.43	0.84	-1.10
AAS_STAIRS3	0.71	-4.65	0.75	-2.89

Next, the Wright item-person map from the 62 item Rasch analysis was assessed for multiple items plotted at the same level of difficulty in order to identify redundancy in items (Figure 15).

Figure 15*Wright Item Person Map: Phase 3 Data, 62 Items*

Based on the Wright item person map, redundancy in items was revealed, with two and three items grouped together at various ability levels (Figure 16). There was still a lack of more difficult items as demonstrated by the map. Next, redundant items from the scale were removed based on infit values. Ability levels that had three items grouped at the same level were assessed for redundancy first, as this has been previously recommended (Green & Frantom, 2002). Redundant items that had the greatest deviation from 1.0 for infit values were removed first. This process was repeated three more times, leaving 35 items in the scale (Table 10).

Table 10*Mean-square Infit and Outfit: Phase 3 Data, 35 items*

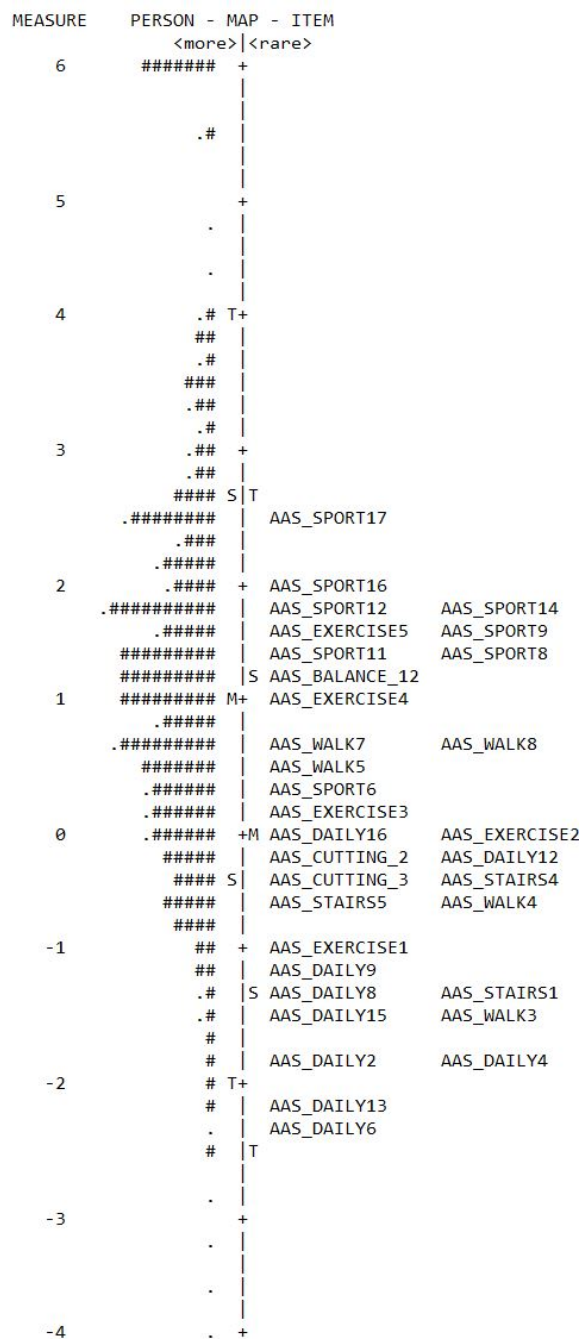
Item	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
AAS_DAILY6	1.32	2.67	1.11	0.58
AAS_SPORT17	1.23	2.98	1.19	1.50
AAS_SPORT16	1.23	3.24	1.20	1.89
AAS_SPORT13	1.20	2.97	1.05	0.59
AAS_CUTTING2	1.18	2.35	1.04	0.39
AAS_SPORT9	1.14	2.13	1.04	0.43
AAS_DAILY4	1.14	1.43	1.10	0.63
AAS_DAILY2	1.10	1.03	0.85	-0.83
AAS_BALANCE12	1.10	1.48	1.15	1.73
AAS_WALK5	1.08	1.25	1.12	1.39
AAS_SPORT14	1.07	1.11	0.95	-0.46
AAS_SPORT12	1.07	1.08	0.99	-0.13
AAS_DAILY12	1.07	0.99	0.87	-1.33
AAS_WALK4	1.04	0.52	0.94	-0.47
AAS_DAILY16	1.04	0.56	1.12	1.28
AAS_DAILY8	1.03	0.36	0.81	-1.31
AAS_SPORT8	1.02	0.33	0.93	-0.77
AAS_DAILY15	1.02	0.21	1.00	0.03
AAS_DAILY13	1.02	0.19	0.96	-0.13
AAS_WALK3	1.00	0.01	0.92	-0.46
AAS_EXERCISE5	1.00	0.06	1.10	1.14
AAS_CUTTING3	0.99	-0.16	1.17	1.56
AAS_STAIRS5	0.98	-0.27	1.08	0.73
AAS_WALK8	0.97	-0.41	1.02	0.25
AAS_SPORT6	0.97	-0.42	0.95	-0.52
AAS_EXERCISE1	0.97	-0.39	0.86	-1.13
AAS_STAIRS4	0.96	-0.58	1.03	0.29
AAS_SPORT11	0.96	-0.56	0.85	-1.82
AAS_WALK7	0.94	-0.86	1.02	0.21
AAS_WALK6	0.94	-0.83	1.10	0.94
AAS_EXERCISE3	0.93	-1.10	0.96	-0.39
AAS_STAIRS1	0.88	-1.39	0.83	-1.20
AAS_EXERCISE2	0.86	-2.13	0.79	-2.33
AAS_EXERCISE4	0.83	-2.74	0.89	-1.35
AAS_DAILY9	0.82	-2.29	0.95	-0.33

There were no items that had unacceptable values for infit or outfit statistics, indicating no items were misfitting (Table 10). Person reliability was .93 and person separation was 3.76, which were both above the recommended thresholds. Item reliability was 1.00 and item separation was

19.29, which were also both above the recommended thresholds. The Wright item-person map revealed redundancy with two items grouped together at various ability levels (Figure 16).

Figure 16

Wright Item Person Map: Phase 3 Data, 33 Items



Items from the scale were removed based on redundancy. Ability levels that had two items grouped at the same level were assessed. Redundant items that had the greatest deviation from 1.0 for infit values were removed first. If infit values from the redundant items were equal, outfit values were then assessed. After this process, there were 22 items remaining (Table 11).

Table 11
Final 22 Items for Phase 3: Highest to Lowest Difficulty

Item	Label
AAS SPORT 17	Completing a half-marathon
AAS SPORT 16	Rock climbing or bouldering outdoors
AAS SPORT 14	Competitive sports like martial arts or boxing
AAS EXERCISE 5	Moderate to heavy exercise for 1 hour with no breaks (may include running at a medium to fast pace or heavy use of a fitness machine such as a treadmill or elliptical)
AAS SPORT 11	Competitive sports like basketball, soccer, lacrosse, or tennis
AAS BALANCE 12	Standing or walking for an entire 8-hour workday
AAS EXERCISE 4	Moderate to heavy exercise for 30 minutes with no breaks (may include running at a medium to fast pace or heavy use of a fitness machine such as a treadmill or elliptical)
AAS WALK 8	Walking up a steep hill
AAS SPORT 6	Moderate cycling for 30 minutes, some hills
AAS WALK 5	Walking 5 miles at your normal pace
AAS EXERCISE 3	Moderate exercise for 15 minutes with no breaks (may include running at an easy to medium pace or moderate use of a fitness machine such as a treadmill or elliptical)
AAS DAILY 16	Heavy household work (such as mowing the grass, raking leaves, weeding, moving furniture)
AAS DAILY 12	Running a short distance (such as to catch a bus, chase after a child)
AAS CUTTING 3	Turning quickly or making quick sideways movements (cutting)
AAS STAIRS 5	Squatting to pick up something off of the floor
AAS EXERCISE 1	Light exercise for 15 minutes with no breaks (may include walking or light use of a fitness machine such as a treadmill or elliptical)
AAS DAILY 9	Putting on pants while standing
AAS DAILY 8	Getting into or out of the bathtub
AAS DAILY 15	Moderate household work (such as vacuuming flat surfaces, mopping, carrying in shopping bags)
AAS DAILY 2	Stepping down from curbs
AAS DAILY 13	Walking to the end of the driveway to get the mail
AAS DAILY 6	Getting on and off the toilet

The PCA of observed responses revealed a cumulative variance on the first factor of 55.4% which was consistent with the initial PCAs in Phase 2 and Phase 3, and the component matrix

had no coefficients less than .60. The observed cumulative variance from the PCA of the Rasch standardized residuals was 72.1% and the expected cumulative variance was 72.9%, providing evidence of unidimensionality and indicating that the data fit the Rasch model. There were no standardized residual inter-item correlations that were greater than the threshold of .7, further indicating unidimensionality. None of the 22 items had unacceptable values for infit or outfit statistics, indicating no items were misfitting (Table 12).

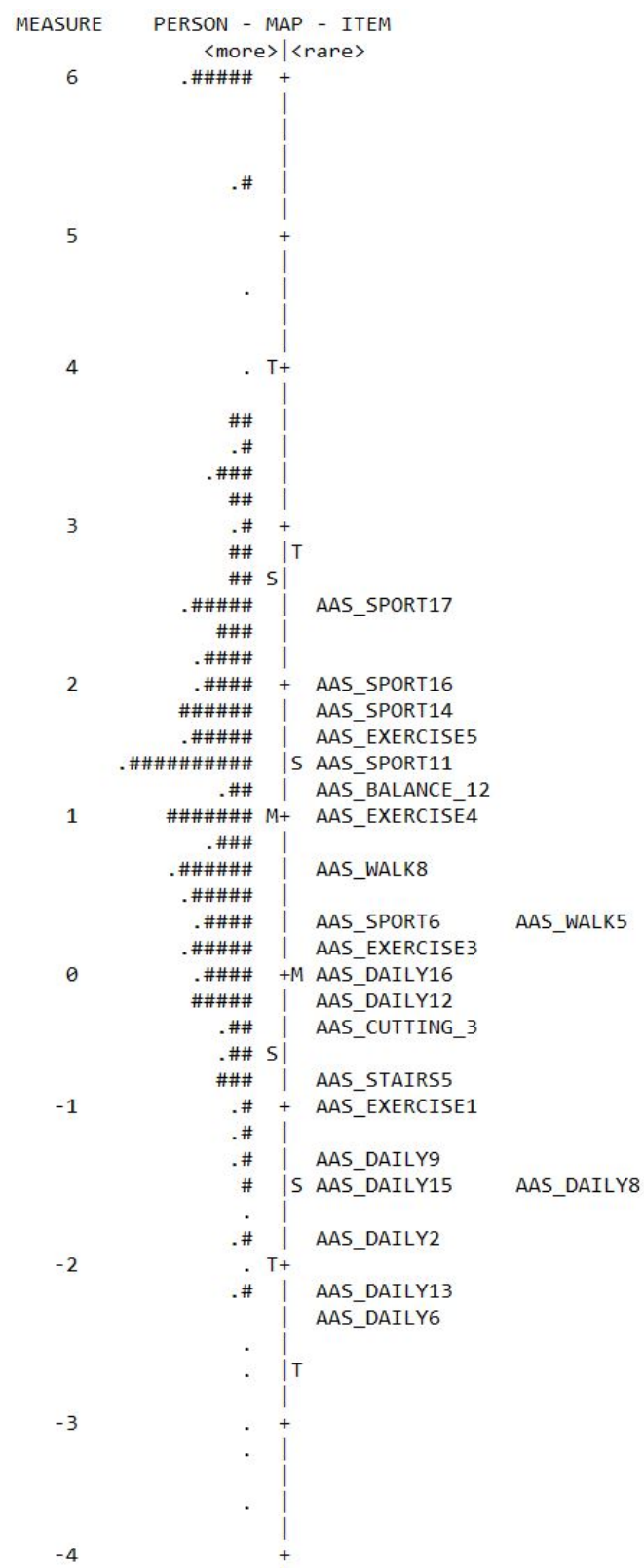
Table 12

Mean-square Infit and Outfit Values: Phase 3 Data

Item	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
AAS_DAILY6	1.28	2.37	1.13	0.67
AAS_SPORT16	1.28	3.93	1.23	2.20
AAS_WALK5	1.20	2.86	1.24	2.74
AAS_SPORT14	1.21	2.95	1.08	0.91
AAS_SPORT17	1.17	2.26	1.13	1.11
AAS_WALK8	1.05	0.82	1.15	1.80
AAS_CUTTING3	0.99	-0.10	1.14	1.37
AAS_DAILY2	1.11	1.13	0.85	-0.79
AAS_BALANCE12	1.07	1.02	1.09	1.15
AAS_STAIRS5	0.98	-0.29	1.06	0.59
AAS_SPORT11	1.06	0.99	0.95	-0.52
AAS_DAILY12	1.05	0.66	0.86	-1.52
AAS_DAILY16	1.01	0.12	1.05	0.60
AAS_DAILY13	1.03	0.28	1.01	0.13
AAS_SPORT6	1.00	-0.05	0.95	-0.54
AAS_DAILY8	0.99	-0.03	0.79	-1.49
AAS_DAILY15	0.99	-0.12	0.99	-0.02
AAS_EXERCISE1	0.99	-0.14	0.86	-1.23
AAS_DAILY9	0.78	-2.88	0.98	-0.11
AAS_EXERCISE5	0.90	-1.55	0.96	-0.43
AAS_EXERCISE3	0.85	-2.22	0.91	-1.01
AAS_EXERCISE4	0.73	-4.52	0.75	-3.46

Person reliability was .92 and person separation was 3.39, which were both above the recommended thresholds. Item reliability was 1.00 and item separation was 20.17, which were

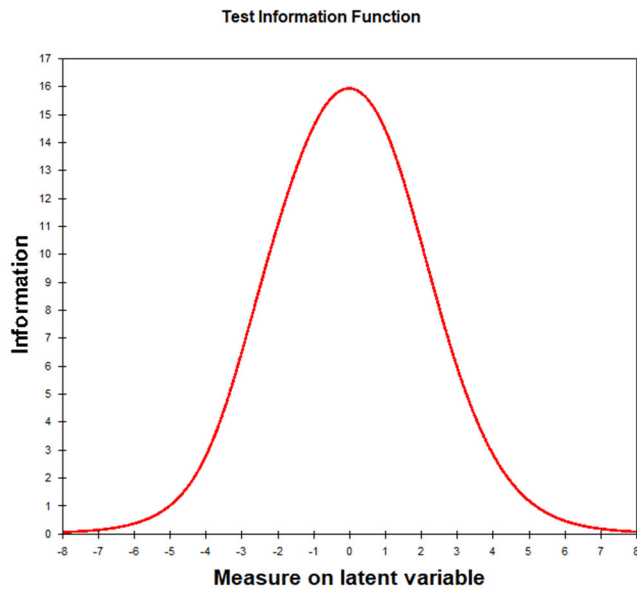
also both above the recommended thresholds. Although there was some indication of redundancy in the instrument as indicated by the Wright item person map (Figure 16), all triplicate redundancy was eliminated, and this was the final step in item reduction since 22 items is considered a reasonable length for a foot and ankle activity instrument (Martin & Irrgang, 2007; Martin et al., 2005). After each step of item reduction, person reliability and person separation decreased. When 22 items were remaining, person reliability was .02 points above the minimum acceptable value, which supported the completion of item reduction at this step. Additionally, 22 items still provide some flexibility to remove items if necessary when tested in a surgical population, which will be discussed further in Chapter 5.

Figure 17*Wright Item Person Map: Phase 3 Data, 22 Items*

Test information function demonstrates that the ankle activity level instrument has high precision, with a value greater than the acceptable threshold of 10 (Figure 18).

Figure 18

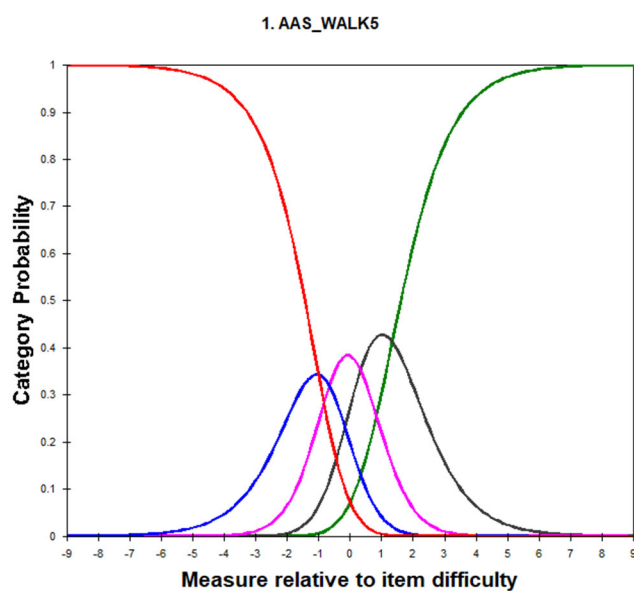
Test Information Function for Phase 3 Data



The category probability plots were similar for all items. Based on step calibrations of Andrich threshold, all categories were ordered and increased monotonically, as demonstrated by item 1 (Figure 19). The category measure, which represents the average ability estimate in logits was -2.46 for persons who chose response category 0, -1.06 for category 1, -0.07 for category 2, 1.02 for category 3, and 2.65 for category 4.

Figure 19

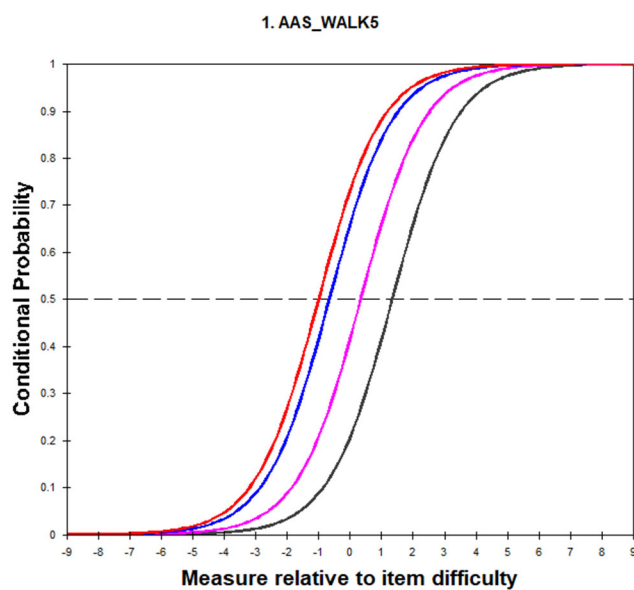
Category Probability Curves for Item 1 of Phase 3 Data



The item characteristic curve demonstrates that as person ability increases, so does the probability of endorsing an item (Figure 20).

Figure 20

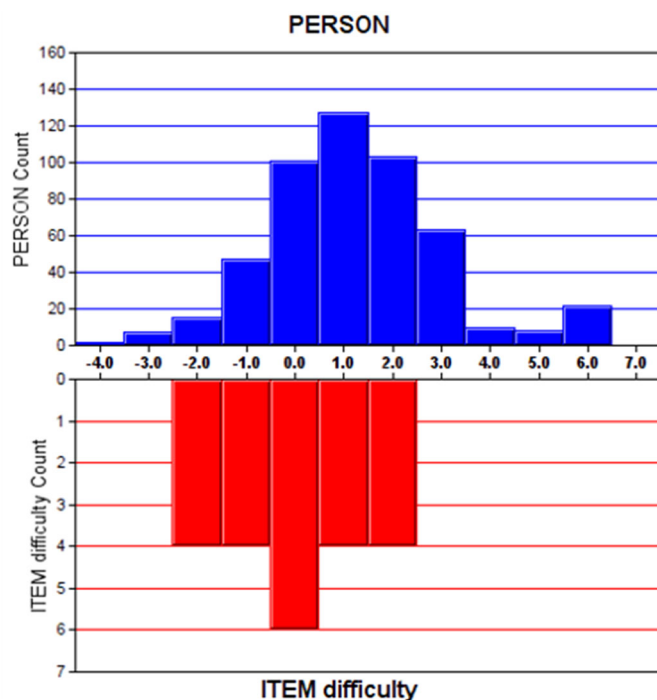
Item Characteristic Curve for Item 1 of Phase 3 Data



The item characteristic curve demonstrates that as person ability increases, so does the probability of endorsing an item. The Wright item person histogram reinforces the gap in items at the upper difficulty level where there are individuals who have no items at their ability level (Figure 21).

Figure 21

Wright Item Person Histogram: Phase 3 Data



Evidence of Convergent and Divergent Validity Analysis

A Pearson correlation was estimated between scores on the foot and ankle activity level instrument and scores from other commonly used outcome measures to determine whether there was evidence of convergent and divergent-related validity (Table 13). The FAAM ADL subscale, FAAM Sport subscale, Tegner activity scale, and the SF-12 PCS scores had correlations with the foot and ankle activity level scale scores above the previously defined threshold of .60, demonstrating evidence of convergent validity. The SF-12 MCS scores had a

low correlation with the FAALS , less than the previously defined threshold of .20, demonstrating evidence of divergent validity (Table 13).

Table 13

Pearson Correlation of FAALS with Common Foot and Ankle Measures

Variable	1	2	3	4	5	6
1. Foot and Ankle Activity Level Scale	-					
2. FAAM ADL	.86**	-				
3. FAAM Sport	.87**	.85**	-			
4. Tegner Activity Scale	.61**	.48**	.51**	-		
5. SF-12 PCS	.77**	.77**	.741**	.48**	-	
6. SF-12 MCS	.12*	.16**	.07	.08	.02	-

* $p < .01$

** $p < .001$

Multiple Linear Regression Analysis

A multiple linear regression analysis was performed to determine whether the ankle activity level instrument was sensitive enough to detect differences in different groups based on body mass index and previous ankle surgery status, which have been well-established to affect function. The dependent variable used in the regression analysis was the foot and ankle activity level raw score. When the ankle activity level scores served as the criterion variable none of the multiple linear regression assumptions were violated including normality, homoscedasticity, linearity, independence, correct model specification, and that all variables are measured without error, based on visualization of a normal probability plot and a histogram of residuals which were included in Appendix F. There was no evidence of collinearity in the regression model as

indicated by the variance inflation factor and tolerance values. There were two cases that were on the cusp of being considered an outlier, with standardized residuals of -3.07 and -3.00. These cases were retained in the analysis since they were extremely close to the threshold of three standard deviations above or below the mean. The overall model was significant, indicating that only 13% of the variance in the ankle activity scale was explained by the model which included BMI and previous ankle surgery status as predictor variables, $F(2, 502) = 38.31, p < .001, R^2 = .13$. Both variables were statistically significant according to the t-statistics (Table 14).

Table 14

Multiple Linear Regression with FAALS as Dependent Variable

	Estimate	<i>SE</i>	95% CI		<i>t</i>	<i>p</i>
			<i>LL</i>	<i>UL</i>		
Fixed effects						
Intercept	90.39	3.71	83.10	97.68	24.36	<.001
BMI	-1.02	.13	-1.28	- .76	-7.69	<.001
Previous Ankle Surgery ^a	-9.01	2.64	-14.19	-3.82	-3.41	.001

Note. ^a 0 = no previous surgery, 1 = previous surgery.

Therefore, as BMI increases, FAALS scores decrease. Specifically, for every one-unit increase in BMI, there was a 1.02 average point decrease in FAALS scores while adjusting for previous ankle surgery status, $t(502) = -7.69, p < .001$. Individuals who had previous ankle surgery had on average 9.01 points fewer than those who did not have previous ankle surgery while adjusting for BMI, $t(502) = -3.41, p < .001$. Overall, the ankle activity level scale demonstrated sensitivity to differences in cohorts based on BMI and previous ankle surgery status. Normative values for BMI, previous ankle surgery, dichotomized age (<40, ≥40), and sex can be found in Table 15.

Table 15*Normative Values Based for FAALS Aggregate Percentage*

Demographic Variable	FAALS	
	<i>M</i>	<i>SD</i>
BMI		
Normal	73.3	18.6
Overweight or Obese	63.2	23.1
Previous Ankle Surgery		
Yes	56.7	24.1
No	68.5	21.3
Sex		
Males	70.3	21.4
Females	64.1	22.1
Age		
<25	76.9	14.9
25 - 39	73.5	19.2
≥40	61.6	23.0

Phase 4

The purpose of this phase was to finalize the order of item difficulty, determine the thresholds for each activity level, and determine the number of clinically relevant activity levels (thresholds). Orthopaedic experts determined the order of difficulty for the 22 items from Phase

3 and identified four activity levels based on the 22 items ankle activity level instrument. (Table 16).

Table 16

Final 22 Items: Highest to Lowest Difficulty

Item	Label	Activity Level
22	Completing a half-marathon	4
21	Rock climbing or bouldering outdoors	4
20	Competitive sports like martial arts or boxing	4
19	Moderate to heavy exercise for 1 hour with no breaks (may include running at a medium to fast pace or heavy use of a fitness machine such as a treadmill or elliptical)	4
18	Competitive sports like basketball, soccer, lacrosse, or tennis	4
17	Standing or walking for an entire 8-hour workday	3
16	Moderate to heavy exercise for 30 minutes with no breaks (may include running at a medium to fast pace or heavy use of a fitness machine such as a treadmill or elliptical)	3
15	Walking up a steep hill	3
14	Moderate cycling for 30 minutes, some hills	3
13	Walking 5 miles at your normal pace (may include treadmill or elliptical use)	3
12	Moderate exercise for 15 minutes with no breaks (may include running at an easy to medium pace or moderate use of a fitness machine such as a treadmill or elliptical)	3
11	Heavy household work (such as mowing the grass, raking leaves, weeding, moving furniture)	2
10	Running a short distance (such as to catch a bus, chase after a child)	2
9	Turning quickly or making quick sideways movements (cutting)	2
8	Squatting to pick up something off of the floor	2
7	Light exercise for 15 minutes with no breaks (may include walking or light use of a fitness machine such as a treadmill or elliptical)	2
6	Putting on pants while standing	1
5	Getting into or out of the bathtub	1
4	Moderate household work (such as vacuuming flat surfaces, mopping, carrying in shopping bags)	1
3	Stepping down from curbs	1
2	Walking to the end of the driveway to get the mail	1
1	Getting on and off the toilet	1

Every orthopaedic expert recommended at least two items to be re-ordered in terms of difficulty; however, after aggregating all results based on majority input, there was no change in item difficulty order as compared to the original order of item difficulty based on the Rasch analysis results. There was one item that did not reach majority, which was Item 11: Heavy household work (such as mowing the grass, raking leaves, weeding, moving furniture). Experts were split in a two versus two decision to place Item 11 one placement above, as Item 12; however, when there was a tie, the tie was broken by referring to the Rasch results. Therefore, no item placement changed based on orthopaedic expert recommendations. Overall, the item difficulty as defined by the Rasch analysis was upheld by the six orthopaedic experts. Based on a 0-point to 4-point Likert-type scale, raw points and percentages were assigned to each activity level (Table 17). Anchor points were no difficulty (4), slight difficulty (3), moderate difficulty (2), extreme difficulty (1), and unable to do (0). Each range of scores is based on the total possible points. For example, for Level 1 there are six items that can have 24 possible total points if an individual chooses “No Difficulty” which is worth 4 points for each item. Percentages of the raw FAALS scores were calculated for easy clinical interpretation.

Table 17

Raw Points and Percentages to Assign Activity Levels

Activity Level	Point Range for Each Activity Level	Percent for Each Activity Level
4	69 - 100 points	78% - 100%
3	45 - 68 points	51% - 77%
2	5 - 44 points	28% - 50%
1	0 - 24 points	0% - 27%

Based on the scoring algorithm described, activity level was assigned based on points for each participant (Table 18).

Table 18

Activity Level Frequencies for FAALS Scores

Activity Level	Frequency	Percent
4	30	6
3	85	17
2	203	40
1	187	37

The majority of participants were assigned to the two lowest levels of activity. This may indicate the need for two additional activity levels towards the lower half of the score to further delineate between activity levels within current levels 1 and 2.

Chapter IV Summary

This study has shown the progression of developing a new instrument to measure foot and ankle activity level in the normal population in the United States. Instrument items were developed and tested through multiple rounds of data collection and analyses using the Rasch measurement model. There were 101 items, initially, for the FAALS. After multiple rounds of data collection and item reduction, the foot and ankle activity level instrument had 22 remaining items. The foot and ankle activity level scores had acceptable levels of person and item reliability and separation, demonstrating evidence of reliability and validity. These values indicate that the foot and ankle activity level instrument is consistent in its ability to measure activity level and is also able to delineate between three or more ability levels.

In addition, there was evidence of convergent and divergent validity for the foot and ankle activity level instrument. In fact, all proposed measures, including the FAAM ADL, FAAM Sport, the Tegner activity scale and the SF-12 PCS were all significantly correlated with the foot and ankle activity level scores. The Tegner activity scale had the lowest correlation with the new instrument's scores, which may be a function of the Tegner being a single item measure and potentially having more measurement error, since instruments with very few items tend to have greater measurement error than instruments with more items. The SF-12 MCS had a correlation less than .2 with the foot and ankle activity level scores, which was expected, and demonstrated evidence of divergent validity.

The multiple linear regression analysis showed that there was a significant difference in foot and ankle activity level scores in BMI groups and people who have had previous ankle surgery and those who have not. People with a normal BMI had higher foot and ankle activity level scores than those who were overweight or obese. People who had previous ankle surgery had significantly lower higher foot and ankle activity level scores than those who did not have ankle surgery. These results demonstrated sensitivity of the foot and ankle activity level instrument. Overall, psychometric properties of the FAALS were very good; however, there seemed to be a lack in extremely difficult items to match the high abilities of some of the participants in the sample. This may be due to the fact that some participants over-reported their abilities.

The purpose of Phase 4 was to finalize the order of item difficulty and the number of activity levels for the foot and ankle activity level instrument. In the end, there were 22 items with four activity levels. The Rasch analysis and orthopaedic expert recommendations for item difficulty did not differ after aggregating expert recommendations, demonstrating agreement

between the data driven procedure of the Rasch analysis and the more subjective assessment by experts in the field.

CHAPTER V

DISCUSSION AND CONCLUSIONS

Through this study I sought to develop a tool that would allow physicians and medical providers an easy and quick way to measure foot and ankle activity level in a clinical setting, since there was not an instrument that has been developed to specifically do so. The first step in developing a foot and ankle activity level instrument was to assess the proposed items in a normal population to establish normative values and assess psychometric properties. There were four phases of this study. The goal of the Phase 1 was to generate items for the FAALS, and operationally define the construct by utilizing expert input. The goal of Phase 2 was to pilot the items for the new instrument and identify poorly functioning items that could be removed in Phase 3. The goal of Phase 3 was to confirm item reduction from the pilot data in Phase 2 and finalize items included for the scale. The purpose of Phase 4 was to develop an aggregate score from the 22 items and define score ranges and activity levels based on expert input. The idea is to produce an activity level scale that can be used as an aggregate score from which an activity level is assigned for easy, quick clinical application. Chapter 5 will include a summary of study findings, as well as a comparative literature discussion of results, score interpretations, issues that arose in the study, limitations, future research, study implications and conclusions.

Study Summary

This study demonstrated acceptable psychometric properties of scores from the FAALS. The Rasch analysis revealed that 66 instrument items had acceptable mean-square infit and outfit values. Therefore, items were removed based on redundancy, which left 22 remaining items that comprised the foot and ankle activity level instrument and supported the assumption of unidimensionality as stated in research question 1. Person reliability was .92 and person separation was 3.39, which were both above the recommended thresholds. Item reliability was 1.00 and item separation was 20.17, which were also considered acceptable. Scores from the new instrument had high internal consistency. These findings support research question one with confirmation. Other patient-reported outcomes were significantly correlated to the FAALS which provided additional evidence of validity and the SF12 MCS had a low correlation with the FAALS which answers research question 2 with confirmation. In addition, the FAALS demonstrated sensitivity, with at least three different ability groups according to person separation that were able to be identified through multiple linear regression analysis which answers research question 4 with an affirmative response. Previously defined demographic and surgical factors, including BMI and previous ankle surgery status were significant predictors of FAALS aggregate scores indicating sensitivity among groups. In Phase 4 of this study, research question 4 was addressed. The order of items for the foot and ankle activity level instrument remained the same as the Rasch analysis recommendations, and experts recommended four activity levels, with the majority of individuals assigned to activity levels 1 and 2.

Comparative Literature Discussion

The final item reduction of the instrument left 22 items. All items had infit and outfit values that were very close to the desired and expected value of 1.0, indicating minimal measurement error and evidence of validity. This seemed to be a good stopping point for item reduction for several

reasons. First, other foot and ankle instruments, such as the FADI and FAAM, are very similar in length to the foot and ankle activity level instrument, and do not seem to be overwhelming to participants in terms of length (Martin & Irrgang, 2007; Martin et al., 2005). In the foot and ankle activity level instrument, each time an item is removed, person reliability is reduced. Person reliability was .92 with 22 items, therefore, removing any further items may decrease person reliability below the acceptable threshold.

However, the Wright item person map revealed a lack of very difficult items to correspond with the upper ability levels of the sample with higher abilities. This may be due to the fact that this study was conducted in a normal population. In addition, perhaps participants overestimated their abilities, or perhaps some individuals have exceptional activity levels. In order to determine whether additional items, with even greater difficulty levels, should be added to the instrument, it was important to understand whether delineating between individuals with activity levels greater than the most difficult item was of clinical relevance. After speaking to various orthopaedic surgeons from the expert panel, it was determined that additional items with difficulty greater than running a half marathon was not very important in the clinical setting. Although there were very difficult items included in the initial total item pool, such as completing a full marathon and competitive sport participation in hockey or powerlifting for example, it was difficult to retain these types of items, based on fit, in a normal population. Perhaps those more difficult items would be more helpful in an elite athlete population, but for every day clinical usage, additional difficult items may not be useful. With 22 items, there is still some flexibility to reduce items, if necessary, for the surgical population. Flexibility in instrument development is imperative for successful finalization of items.

In this study, the foot and ankle activity level scores performed how they were hypothesized to perform, meaning that the new instrument's scores were significantly correlated to scores from other patient reported outcome measures that in some way measure physical function. These findings are similar to a study that examined normative values for the FAAM ADL in a normal population of 284 participants. The correlation between FAAM ADL and SF-12 PCS was significant at .46, and the correlation between the FAAM Sport and SF-12 PCS was significant at $r = .52$ (Matheny et al., 2020). The FAAM ADL and SF-12 MCS, as well as the FAAM Sport and SF-12 MCS were not significantly correlated, with correlations of $r = .07$ and $r = -.01$ respectively (Matheny et al., 2020). In this current study, scores from the foot and ankle activity scale had a correlation of $r = .77$ with the SF-12 PCS and $r = .12$ with the SF-12 MCS. Although there are slight differences in correlation values between the previous study and the current study, the overall correlation trend between the criterion variables (FAAM and FAALS) and the SF-12 component summaries is very similar and supports the hypotheses of this study (Martin et al., 2005; Matheny et al., 2020).

The multiple linear regression analysis revealed that the foot ankle activity level instrument was sensitive enough to decipher between demographic groups that have been established to have distinct differences in function. In this study, the foot and ankle activity level scores were significantly different based on BMI and previous surgery status. In a previous study of normal participants, those who had a normal BMI had significantly higher FAAM ADL and FAAM Sport scores than those who were overweight or obese (Matheny et al., 2020). In addition, those who had previous surgery had approximately 8 points higher for the FAAM ADL and 15 points higher for the FAAM Sport scores (Matheny et al., 2020). In the current study, those who had previous ankle surgery had approximately 10 points less on the FAALS than

those who did not have previous ankle surgery. This difference is in alignment with previous study results (Matheny et al., 2020).

The orthopaedic experts determined that there were four levels of activity for the foot and ankle activity level instrument based on the 22 items. In the current study, person separation was 3.39, which indicated that there were at least three groups of ability for activity level, which supports expert recommendations of four activity levels. The frequencies of activity levels indicated that the majority of participants were assigned to activity Levels 1 and 1 (77%) indicating that there may be a need for further delineation of activity levels in the lower half of the foot and ankle activity level instrument. It will be important to pilot this new instrument in a surgical population to confirm whether or not to increase the number of activity levels towards the lower end of the scale. Although the Rasch analysis identified a lack of extremely high item difficulty, it seems that the upper end of the scale performed quite well in assigning activity levels since only 6% of participants were categorized as a Level 4, which is the highest activity level.

Score Interpretations

Interpretations for the foot and ankle activity level scores are based on aggregate scores solely for ease of interpretability and clinical application per the recommendation of the orthopaedic experts. Although the Rasch model produces Rasch logits, aggregate scores are much better understood in the orthopaedic outcomes. Previous commonly used ankle specific measures, such as the FAAM ADL, the FAAM Sport, and the FADI are recommended to be interpreted as aggregate scores even though item response theory was used to develop those instruments. The foot and ankle activity level instrument was developed to eventually be utilized in a clinical setting; therefore, creating a clear and simple way of interpreting the scores was quite important for clinical application. As the foot and ankle activity level instrument is tested in

the surgical population, it may be necessary to improve or change the process by which thresholds for activity levels are determined, especially if the instrument eventually progresses to administration via computer adaptive testing (CAT).

Study Issues

There were several issues that arose during this study. The first issue that arose was the quality issue with pay-for data collection. The data for this study were collected by a data collection agency, Qualtrics. Qualtrics claimed to have a very short time period required for data collection, while also being able to stratify responses by six different demographic variables that were discussed earlier in Chapter 3. When I first examined the data, I found that some participants had various responses that did not make logical sense. For example, if a participant states that they have no difficulty walking up four flights of stairs, but they have extreme difficulty walking up one flight of stairs, this participant provided non-sensical data and was therefore removed. Although Qualtrics has some quality control measures that are meant to control for participants who may complete the survey too quickly, there were not any quality control measures that accounted for non-sensical answers. There was approximately one-third of the pilot data that was removed and re-collected after implementing nine quality control measures that checked for illogical responses. This whole quality control process also increased the length of time necessary to collect all of the data by approximately eight weeks. Discovering this issue in the pilot data collection phase was paramount in having a smooth data collection process for the main study data collection ($N = 505$). In the future, recommendations would be, that if researchers use a pay-for data collection service such as Qualtrics, it will be important to think about what type of quality control measures can be implemented and to also account for a much longer period of time for data collection.

Another issue that arose is that there were too many well-fitting items, which made item reduction a much longer, more complicated process. There was a large number of items that were originally generated (101); however, there is an expectation that with the first round of item reduction, at least 50% of the items will be removed based on mean-square infit and outfit statistics (Bond & Fox, 2015; Chiu et al., 2020; Christensen et al., 2013; A. M. Keenan et al., 2007; Martin et al., 2005). However, in this study, after items were removed based on expert recommendations and assessment, there were 77 items remaining. From there, item reduction utilizing the Rasch mean-square infit and outfit values only helped to reduce an additional 11 items, leaving 66 remaining. It was necessary to develop a systematic process for item reduction since conventional methods of poor fit assessment could not be used after 66 items. In order to reduce enough items to encourage use of the instrument in a clinical setting, it was necessary to remove a large number of additional items. Redundancy in items in an instrument, although not too detrimental (Linacre, 2002), can create other problems, such as survey fatigue, which, in turn, can lead to missing data (Lafave et al., 2016). In order to reduce items, location on the Wright item person map that indicated two or more items at the same difficulty level was assessed. At ability levels with triplicate items, items were removed based on the mean-square infit values of the three items, and the item with the greatest deviation above or below 1.0 was removed. If all items were the same distance away from 1.0 for infit, outfit was assessed, and the item with the greatest deviation above or below 1.0 was removed. This process also increased the average time it takes for item reduction, which should be accounted for when developing an outcomes instrument.

Limitations

There are several limitations of this study. After completing item reduction and the final Rasch analysis of the Phase 3 data, there was an over-abundance of well-fitting items. The goal

of phase 1, which included item generation, was to generate 30 to 50 items with the ultimate goal of reducing items by approximately half; however, 101 items were generated from Phase 1. Starting item reduction with nearly double the items may have complicated the process of item reduction; however, it was necessary to include all items from the orthopaedic experts. Since the goal was to develop an instrument that has between 15 and 25 items, it makes logical sense that item reduction would be a longer process. It was unexpected that so many of the items would perform well in terms of mean-square infit and outfit statistics; therefore, items were evaluated not only based on threshold infit and outfit values as previously discussed in Chapters 3 and 4, but also item infit and outfit values were also compared to one another when at the same difficulty level.

Although there were too many well-fitting items, results indicated that there may not be enough items that were very difficult. When looking at the abilities of the sample and the difficulty of the items, there seems to be a gap at the upper end of the items, starting at around three standard deviations above the mean as seen on the Wright item person bar (Figure 21). The ability levels of the sample seem to expand up to 7 standard deviations above the mean. Perhaps this sample was exceptionally active.

A convenience sample was utilized for this study. Quota sampling was implemented for six categories, including age, sex, race, region, income, and educational level, to most represent the “normal” adult US population. However, quotas obtained based on region of the United States were somewhat inaccurate, with an over-representation of the Southern and region by approximately 7%, leaving a slight under-representation of the Midwest and Western regions of approximately 3% each. Even with the addition of quotas based on sociodemographics, this sample was a convenience sample and may introduce some bias.

Future Research

As with most first studies that seek to develop a new instrument, there are future steps that can be taken to further this research and further refine the instrument. This foot and ankle activity level instrument should be assessed in the surgical population. Since this is a tool targeted for use in a foot and ankle surgical population, it would be useful to determine whether the scores from the new instrument exhibit acceptable psychometric properties. Two of the orthopaedic experts have already committed to completing this study for the next step in further testing this instrument in a surgical population.

In order to truly determine whether more difficult items should be included, the foot and ankle activity level instrument should be tested for acceptable psychometric properties in the surgical population. This will allow the FAALS instrument to be re-assessed to confirm that the items selected are the most appropriate for patients who are going to or have undergone foot or ankle surgery. It is also important to recognize whether including more difficult items into the instrument would be useful from a clinical perspective. For example, if a patient is able to complete a half marathon, does it matter to an orthopaedic surgeon whether the patient can complete a full marathon? When two of the orthopaedic experts were asked this question, they both agreed that in terms of clinical treatment and postoperative care, there is no difference in treatment at this level of activity, therefore, it is unnecessary to distinguish between higher activity level. Perhaps including additional, more difficult items would be more useful in an elite or professional population; however, the normal population was the target population for this study.

Once the foot and ankle activity level instrument's scores have been tested in the foot and ankle surgical population, it will be important to perform a differential item functioning analysis that can determine whether measurement bias in the foot and ankle activity level instrument

exists for preoperative and postoperative patients. Although it is expected to see the average scores differ between the two groups, it is also expected that the instrument items will be interpreted in the same way. Eventually, the foot and ankle activity level instrument should be tested for differential item functioning for preoperative and postoperative patients who underwent foot or ankle surgery. This research would aid in providing evidence of reliability and validity of the foot and ankle activity level scores and help to understand whether the activity level instrument is defining activity level differently based on group membership. Results from this research would allow medical providers, specifically foot and ankle orthopaedic practitioners, feel confident in using the foot and ankle activity level instrument to assess patients.

Study Implications

Through this study, I completed the first step in developing a foot and ankle activity level assessment instrument to be used in individuals suffering from foot and ankle injuries and disorders. The first step in this instrument's development entailed developing and examining its use in a normal population. Assessing normative values has become an integral part of the interpretation of patient reported outcomes, with the largest international foot and ankle organization prioritizing the establishment of normative values for a variety of foot and ankle outcome measures in a normal population (Hunsaker et al., 2002). By first understanding normative values of patient reported outcomes in a clinical setting, physicians and medical providers can determine whether patients treated for specific conditions have returned to normative ranges of functioning. This type of information is invaluable for physician-patient communication which facilitates shared decision-making in the patient's medical care and treatment (Henn et al., 2011; Mancuso et al., 2001, 2003). Previous studies have shown that PROs differ based on demographic characteristics (Schneider & Jurenitsch, 2016b). By

determining normative values for PROs based on demographic factors, physicians can truly understand expected values and apply that information in the patient consultation process when describing expected outcomes. The next step would be to assess the instrument in a surgical population, since the foundational work of establishing normative values and assessing instrument performance in a normal population has been completed.

By developing the FAALS instrument, physicians are able to assess activity level for foot and ankle specific activities. Activity level pertaining to the foot and ankle is important for clinical assessment and serves as a benchmark to approximate a patient's position in the continuum of care. Often during clinical assessment, physicians and surgeons utilize short outcome measures that are easily and quickly completed and interpreted. The FAALS aggregate score is comprised of 22 items with a percentage assigned for activity, while an actual level is assigned based on the aggregate score. The activity level instrument will allow physicians to quickly assess a patient's baseline score, as well as their progress following treatment. It is likely that with 22 items, foot and ankle activity level will be more accurately measured with the FAALS as indicated by the psychometric assessment, than compared to the Tegner scale that is only one item and may inherently have greater measurement error. FAALS scores can be rigorously tested for reliability and validity, which is not the case for one item measures such as the Tegner scale.

Conclusions

Through this study I developed a new 22-item foot and ankle activity level scale that will allow practitioners to quickly assess activity level, based on four activity levels, in patients. There was acceptable evidence of psychometric properties, including reliability and validity, of foot and ankle activity level scores in the normal population. In addition, the new activity level instrument demonstrated sensitivity in its ability to measure activity level between groups with

known functional differences, such as BMI and previous ankle surgery. Convergent and divergent validity of foot and ankle activity level scores were also shown. Orthopaedic practitioners can feel confident in using the foot and ankle activity level instrument to assess activity level in the normal population. Further research is necessary to determine clinical application in the foot and ankle surgical population.

REFERENCES

- Abrams, J. S. (2017). Patients Reporting Outcomes: Are we getting it right?: Commentary on an article by Frederick A. Matsen III, MD, et al.: "Relationship between patient-reported assessment of shoulder function and objective range-of-motion measurements". *The Journal of Bone & Joint Surgery*, 99(5), e24.
- Acquadro, C., Berzon, R., Dubois, D., Leidy, N. K., Marquis, P., Revicki, D., Rothman, M., for the, P. R. O. H. G., & Group, P. R. O. H. (2003). Incorporating the patient's perspective into drug development and communication: An ad hoc task force report of the Patient-Reported Outcomes (PRO) Harmonization Group meeting at the Food and Drug Administration, February 16, 2001. *Value in Health*, 6(5), 522-531.
<https://doi.org/10.1046/j.1524-4733.2003.65309.x>
- Alexandrowicz, R. W., & Draxler, C. (2015). Testing the Rasch model with the conditional likelihood ratio test: sample size requirements and bootstrap algorithms. *Journal of Statistical Distributions and Applications*, 3(1), 1-25. <https://doi.org/10.1186/s40488-016-0039-y>
- Alwin, D. F., & Krosnick, J. A. (1985). The Measurement of values in surveys: A comparison of ratings and rankings. *Public Opinion Quarterly*, 49(4), 535-552.
<https://doi.org/10.1086/268949>
- Andersen, E. B. (1972). The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(1), 42-54.
<https://doi.org/10.1111/j.2517-6161.1972.tb00887.x>

- Andrich, D. (1978a). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement*, 2(4), 581-594.
- Andrich, D. (1978b). A rating formulation for ordered response categories. *Psychometrika*, 43(4), 561-573. <https://doi.org/10.1007/BF02293814>
- Andrich, D. (2004). Controversy and the Rasch model: a characteristic of incompatible paradigms? *Medical Care*, 17-116.
- Anselmi, P., Vidotto, G., Bettinardi, O., & Bertolotti, G. (2015, Feb 7). Measurement of change in health status with Rasch models. *Health and Quality of Life Outcomes*, 13, 16. <https://doi.org/10.1186/s12955-014-0197-x>
- Aryadoust, V., Tan, H. A. H., & Ng, L. Y. (2019). A Scientometric Review of Rasch Measurement: The Rise and Progress of a Specialty. *Frontiers in Psychology*, 10, 2197. <https://doi.org/10.3389/fpsyg.2019.02197>
- Balalla, S. K., Medvedev, O. N., Siegert, R. J., & Krägeloh, C. U. (2019). Validation of the WHOQOL-BREF and Shorter Versions Using Rasch Analysis in Traumatic Brain Injury and Orthopedic Populations. *Archives of Physical Medicine and Rehabilitation*, 100(10), 1853-1862. <https://doi.org/10.1016/j.apmr.2019.05.029>
- Balsamo, M., Giampaglia, G., & Saggino, A. (2014). Building a new Rasch-based self-report inventory of depression. *Neuropsychiatric Disease and Treatment*, 10, 153-165. <https://doi.org/10.2147/NDT.S53425>
- Belvedere, S. L., & de Morton, N. A. (2010). Application of Rasch analysis in health care is increasing and is applied for variable reasons in mobility instruments. *Journal of Clinical Epidemiology*, 63(12), 1287-1297. <https://doi.org/10.1016/j.jclinepi.2010.02.012>

- Beretvas, S. N. (2016). Comparison of bookmark difficulty locations under different item response models. *Applied Psychological Measurement, 28*(1), 25-47.
<https://doi.org/10.1177/0146621603259903>
- Bergner, M., Bobbitt, R. A., Carter, W. B., & Gilson, B. S. (1981). The sickness impact profile: development and final revision of a health status measure. *Medical Care, 19*(8), 787-805.
<https://doi.org/10.1097/00005650-198108000-00001>
- Bingham, C. O., Noonan, V. K., Auger, C., Feldman, D. E., Ahmed, S., & Bartlett, S. J. (2017). Montreal accord on patient-reported outcomes (PROs) use series—Paper 4: patient-reported outcomes can inform clinical decision making in chronic care. *Journal of Clinical Epidemiology, 89*, 136-141.
- Birnbaum, A. (1957). Efficient design and use of tests of a mental ability for various decision making problems. *Randolph Air Force Base, Texas: United States Air Force School of Aviation Medicine, series report number 58-16*, (project number 775-23).
- Birnbaum, A. (1958a). Further considerations of efficiency in tests of a mental ability. *Randolph Air Force Base, Texas: United States Air Force School of Aviation Medicine, series report number 17*, (project number 7755-23).
- Birnbaum, A. (1958b). On the estimation of mental ability. *Randolph Air Force Base, Texas: United States Air Force School of Aviation Medicine, (series report number 15, project number 7755-23)*.
- Bond, T., & Fox, C. (2015). Applying the rasch model: Fundamental measurement in the human sciences (3rd ed.). Routledge.
- Boone, W. J. (2016, Winter). Rasch analysis for instrument development: Why, when, and how? *CBE: Life Sciences Education, 15*(4), rm4. <https://doi.org/10.1187/cbe.16-04-0148>

- Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer Science and Business Media. <https://doi.org/10.1007/978-94-007-6857-4>
- Bravini, E., Giordano, A., Sartorio, F., Ferriero, G., & Vercelli, S. (2017). Rasch analysis of the Italian Lower Extremity Functional Scale: insights on dimensionality and suggestions for an improved 15-item version. *Clinical Rehabilitation*, 31(4), 532-543. <https://doi.org/10.1177/0269215516647180>
- Briggs, K. K., Kocher, M. S., Rodkey, W. G., & Steadman, J. R. (2006). Reliability, validity, and responsiveness of the Lysholm knee score and Tegner activity scale for patients with meniscal injury of the knee. *The Journal of Bone & Joint Surgery Am*, 88(4), 698-705.
- Briggs, K. K., Lysholm, J., Tegner, Y., Rodkey, W. G., Kocher, M. S., & Steadman, J. R. (2009). The reliability, validity, and responsiveness of the Lysholm score and Tegner activity scale for anterior cruciate ligament injuries of the knee. *The American Journal of Sports Medicine*, 37(5), 890-897. <https://doi.org/10.1177/0363546508330143>
- Briggs, K. K., Steadman, J. R., Hay, C. J., & Hines, S. L. (2009). Lysholm score and Tegner activity level in individuals with normal knees. *The American Journal of Sports Medicine*, 37(5), 898-901. <https://doi.org/10.1177/0363546508330149>
- Brophy, R. H., Lin, K., & Smith, M. V. (2014). The role of activity level in orthopaedics: an important prognostic and outcome variable. *The Journal of the American Academy of Orthopaedic Surgeons*, 22(7), 430-436.
- Budiman-Mak, E., Conrad, K. J., & Roach, K. E. (1991). The Foot Function Index: a measure of foot pain and disability. *Journal of Clinical Epidemiology*, 44(6), 561-570.

- Budiman-Mak, E., Conrad, K., Stuck, R., & Matters, M. (2006). Theoretical model and rasch analysis to develop a revised Foot Function Index. *Foot & Ankle International*, 27(7), 519-527. <https://doi.org/10.1177/107110070602700707>
- Burton, L. J., & Mazerolle, S. M. (2011). Survey instrument validity part I: Principles of survey instrument development and validation in athletic training education research. *Athletic Training Education Journal*, 6(1), 27-35.
- Button, G., & Pinney, S. (2004). A meta-analysis of outcome rating scales in foot and ankle surgery: is there a valid, reliable, and responsive system? *Foot & Ankle International*, 25(8), 521-525.
- Calderón, J. L., Morales, L. S., Liu, H., & Hays, R. D. (2016). Variation in the readability of items within surveys. *American Journal of Medical Quality*, 21(1), 49-56.
<https://doi.org/10.1177/1062860605283572>
- Camilli, G. (1994). Teacher's corner: origin of the scaling constant $d = 1.7$ in item response theory. *Journal of Educational Statistics*, 19(3), 293-295.
- Cappelleri, J. C., Lundy, J. J., & Hays, R. D. (2014). Overview of classical test theory and item response theory for quantitative assessment of items in developing patient-reported outcome measures. *Clinical Therapeutics*, 36(5), 648-662.
<https://doi.org/10.1016/j.clinthera.2014.04.006>
- Carcia, C. R., Martin, R. L., & Drouin, J. M. (2008). Validity of the Foot and Ankle Ability Measure in athletes with chronic ankle instability. *Journal of Athletic Training*, 43(2), 179-183. <https://doi.org/10.4085/1062-6050-43.2.179>

- Cauffman, E., & MacIntosh, R. (2006). A Rasch Differential Item Functioning Analysis of the Massachusetts Youth Screening Instrument: Identifying Race and Gender Differential Item Functioning Among Juvenile Offenders. *Educational and Psychological Measurement*, 66(3), 502-521. <https://doi.org/10.1177/0013164405282460>
- Cella, D., Choi, S., Garcia, S., Cook, K. F., Rosenbloom, S., Lai, J.-S., Tatum, D. S., & Gershon, R. (2014). Setting standards for severity of common symptoms in oncology using the PROMIS item banks and expert judgment. *Quality of Life Research*, 23(10), 2651-2661. <https://doi.org/10.1007/s11136-014-0732-6>
- Center for Disease Control and Prevention. (2020, September). *Disability and Health Overview*. U.S. Department of Health and Human Services. <https://www.cdc.gov/ncbddd/disabilityandhealth/disability.html>
- Chatterjee, S., & Hadi, A. S. (2015). *Regression analysis by example*. John Wiley & Sons.
- Chen, L., Lyman, S., Do, H., Karlsson, J., Adam, S. P., Young, E., Deland, J. T., & Ellis, S. J. (2012, Dec). Validation of foot and ankle outcome score for hallux valgus. *Foot & Ankle International*, 33(12), 1145-1155. <https://doi.org/DOI: 10.3113/FAI.2012.1145>
- Chiu, M. Y. L., Wong, H. T., & Ho, W. W. N. (2020). A comparative study of confirmatory factor analysis and Rasch Analysis as item reduction strategies for SAMHSA recovery inventory for Chinese (SAMHSA-RIC). *The European journal of psychiatry*, 34(2), 74-81. <https://doi.org/10.1016/j.eipsy.2020.02.002>
- Christensen, K. B., Kreiner, S., & Mesbah, M. (2013). *Rasch models in health*. ISTE Ltd.
- Christensen, K. B., Thorborg, K., Hölmich, P., & Clausen, M. B. (2019). Rasch validation of the Danish version of the shoulder pain and disability index (SPADI) in patients with rotator

- cuff-related disorders. *Quality of Life Research*, 28(3), 795-800.
<https://doi.org/10.1007/s11136-018-2052-8>
- Cizek, G., & Bunch, M. (2007). *Standard Setting*. SAGE Publications, Inc.
- Clark, L. A., & Watson, D. (1995). Constructing Validity: Basic Issues in Objective Scale Development. *Psychological Assessment*, 7(3), 309-319. <https://doi.org/10.1037/1040-3590.7.3.309>
- Comins, J., Brodersen, J., Krogsgaard, M., & Beyer, N. (2008). Rasch analysis of the Knee injury and Osteoarthritis Outcome Score (KOOS): a statistical re-evaluation. *Scandinavian Journal of Medicine & Science in Sports*, 18(3), 336-345.
<https://doi.org/10.1111/j.1600-0838.2007.00724.x>
- Comins, J. D., Siersma, V. D., Lind, M., Jakobsen, B. W., & Krogsgaard, M. R. (2018). KNEES-ACL has superior responsiveness compared to the most commonly used patient-reported outcome measures for anterior cruciate ligament injury. *Knee Surgery, Sports Traumatology, Arthroscopy*, 26(8), 2438-2446. <https://doi.org/10.1007/s00167-018-4961-z>
- Conaghan, P. G., Emerton, M., & Tennant, A. (2007). Internal construct validity of the Oxford Knee Scale: Evidence from Rasch measurement. *Arthritis Care & Research*, 57(8), 1363-1367. <https://doi.org/10.1002/art.23091>
- Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, 10, 7.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297-334. <https://doi.org/10.1007/BF02310555>

- Culpepper, S. A., & Culpepper, S. A. (2016). Revisiting the 4-parameter item response model: Bayesian estimation and application. *Psychometrika*, 81(4), 1142-1163.
<https://doi.org/10.1007/s11336-015-9477-6>
- De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications.
- DeMars, C. (2010). *Item response theory*. Oxford University Press.
- DeVellis, R. F. (2006). Classical test theory. *Medical care*, S50-S59.
- Dillman, D. A., Smyth, J. D., Christian, L. M., & Ebooks, C. (2014). *Internet, phone, mail, and mixed-mode surveys: the tailored design method* (4th ed.). Wiley.
- Dingemans, S. A., Kleipool, S. C., Mulders, M. A. M., Winkelhagen, J., Schep, N. W. L., Goslings, J. C., & Schepers, T. (2017). Normative data for the lower extremity functional scale (LEFS). *Acta Orthopaedica*, 88(4), 422-426.
<https://doi.org/10.1080/17453674.2017.1309886>
- Donnenwerth, M., & Roukis, T. (2012). Outcome of arthroscopic debridement and microfracture as the primary treatment for osteochondral lesions of the talar dome. *Arthroscopy: The Journal of Arthroscopic and Related Surgery*, 28(12), 1902-1907.
<https://doi.org/10.1016/j.arthro.2012.04.055>
- Doostfateme, M., Taghi Ayatollah, S. M., & Jafari, P. (2016). Power and sample size calculations in clinical trials with patient-reported outcomes under equal and unequal group sizes based on graded response model: A simulation study. *Value in Health*, 19(5), 639-647. <https://doi.org/10.1016/j.jval.2016.03.1857>
- Eechaute, C., Vaes, P., Van Aerschot, L., Asman, S., & Duquet, W. (2007). The clinimetric qualities of patient-assessed instruments for measuring chronic ankle instability: a systematic review. *BMC Musculoskeletal Disorders*, 8(1), 6.

- Embretson, S., & Reise, S. (2013). *Item response theory for psychologists*. Lawrence Erlbaum Associates.
- Embretson, S. E. (1985). *Test design: developments in psychology and psychometrics*. Academic Press.
- Farrugia, P., Goldstein, C., & Petrisor, B. A. (2010). Measuring foot and ankle injury outcomes: Common scales and checklists. *Injury*, 42(3), 276-280.
<https://doi.org/10.1016/j.injury.2010.11.051>
- Fisher, W. (1997). Physical disability construct convergence across instruments: towards a universal metric. *Journal of Outcome Measurement*, 1(2), 87-113.
- Franchignoni, F., Ferriero, G., Giordano, A., Sartorio, F., Vercelli, S., & Brigatti, E. (2010). Psychometric properties of QuickDASH – A classical test theory and Rasch analysis study. *Manual Therapy*, 16(2), 177-182. <https://doi.org/10.1016/j.math.2010.10.004>
- Franchignoni, F., Franchignoni, F., Salaffi, F., Salaffi, F., Giordano, A., Giordano, A., Ciapetti, A., Ciapetti, A., Carotti, M., Carotti, M., Ottonello, M., & Ottonello, M. (2012). Psychometric properties of self-administered Lequesne Algofunctional Indexes in patients with hip and knee osteoarthritis: an evaluation using classical test theory and Rasch analysis. *Clinical Rheumatology*, 31(1), 113-121. <https://doi.org/10.1007/s10067-011-1788-0>
- Franchignoni, F., Mora, G., Giordano, A., Volanti, P., & Chiò, A. (2013). Evidence of multidimensionality in the ALSFRS-R Scale: a critical appraisal on its measurement properties using Rasch analysis. *Journal of Neurology, Neurosurgery and Psychiatry*, 84(12), 1340-1345. <https://doi.org/http://dx.doi.org/10.1136/jnnp-2012-304701>

- Frey, B. B. (2018). *The SAGE encyclopedia of educational research, measurement, and evaluation (Vols. 1-4)*. SAGE Publications, Inc. <https://doi.org/10.4135/9781506326139>
- Gable, R. K., & Wolf, M. B. (1993). Instrument development in the affective domain: measuring attitudes and values in corporate and school settings.
- Garratt, A. M., Naumann, M. G., Sigurdson, U., Utvag, S. E., & Stavem, K. (2018). Evaluation of three patient reported outcome measures following operative fixation of closed ankle fractures. *BMC Musculoskeletal Disorders*, 19(1), 134. <https://doi.org/10.1186/s12891-018-2051-5>
- Giesinger, J. M., Behrend, H., Hamilton, D. F., Kuster, M. S., & Giesinger, K. (2019). Normative values for the Forgotten Joint Score-12 for the US General Population. *The Journal of Arthroplasty*, 34(4), 650-655. <https://doi.org/10.1016/j.arth.2018.12.011>
- Goldman, S. H., & Raju, N. S. (1986). Recovery of one-and two-parameter logistic item parameters: An empirical study. *Educational and Psychological Measurement*, 46(1), 11-21.
- Goldstein, C. L., Schemitsch, E., Bhandari, M., Mathew, G., & Petrisor, B. A. (2010). Comparison of different outcome instruments following foot and ankle trauma. *Foot & Ankle International*, 31(12), 1075-1080. <https://doi.org/10.3113/FAI.2010.1075>
- Golightly, Y. M., Devellis, R. F., Nelson, A. E., Hannan, M. T., Lohmander, L. S., Renner, J. B., & Jordan, J. M. (2014). Psychometric properties of the foot and ankle outcome score in a community-based study of adults with and without osteoarthritis. *Arthritis Care & Research*, 66(3), 395-403. <https://doi.org/10.1002/acr.22162>

- Green, K. E., & Frantom, C. G. (2002). *Survey development and validation with the Rasch model*. International Conference on Questionnaire Development, Evaluation, and Testing, Charleston, SC, United States.
- Gulliksen, H. (2013). *Theory of mental tests*. Routledge.
- Gustafsson, J. E. (1980). Testing and obtaining fit of data to the Rasch model. *British Journal of Mathematical and Statistical Psychology*, 33(2), 205-233. <https://doi.org/10.1111/j.2044-8317.1980.tb00609.x>
- Guyer, R., & Thompson, N. (2011). *Item response theory parameter recovery using Xcalibre 4.1*. Assessment Systems Corporation.
<https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1056.8220&rep=rep1&type=pdf>
- Hagell, P., & Westergren, A. (2016). Sample size and statistical conclusions from tests of fit to the Rasch model according to the Rasch Unidimensional Measurement Model (Rumm) program in health outcome measurement. *Journal of Applied Measurement*, 17(4), 416-431.
- Hagquist, C., Välimaa, R., Simonsen, N., & Suominen, S. (2017). Differential item functioning in trend analyses of adolescent mental health – Illustrative examples using HBSC-data from Finland. *Child Indicators Research*, 10(3), 673-691. <https://doi.org/10.1007/s12187-016-9397-8>
- Hale, S. A., & Hertel, J. (2005). Reliability and sensitivity of the Foot and Ankle Disability Index in subjects with chronic ankle instability. *Journal of Athletic Training*, 40(1), 35-40.

- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage.
- Hamilton, C. B., Maly, M. R., Giffin, J. R., Clark, J. M., Speechley, M., Petrella, R. J., & Chesworth, B. M. (2015). Validation of the Questionnaire to Identify Knee Symptoms (QuIKS) using Rasch analysis. *Health and Quality of Life Outcomes*, 13, 157. <https://doi.org/10.1186/s12955-015-0358-6>
- Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care*, 38(9), II28-II42.
- Henn, R. F., III, Ghomrawi, H., Rutledge, J. R., Mazumdar, M., Mancuso, C. A., & Marx, R. G. (2011). Preoperative patient expectations of total shoulder arthroplasty. *The Journal of Bone & Joint Surgery*, 93(22), 2110-2115.
- Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement*, 66(3), 393-416. <https://doi.org/10.1177/0013164405282485>
- Herron, M. L. (2006). A review of outcome measures for the ankle and hindfoot. *Foot and Ankle Surgery*, 12(3), 161-167. <https://doi.org/10.1016/j.fas.2006.03.002>
- Hiller, C. E., Refshauge, K. M., Bundy, A. C., Herbert, R. D., & Kilbreath, S. L. (2006). The Cumberland ankle instability tool: a report of validity and reliability testing. *Archives of physical medicine and rehabilitation*, 87(9), 1235-1241. <https://doi.org/10.1016/j.apmr.2006.05.022>
- Howard, M. C. (2015). A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve? *International Journal of Human-Computer Interaction*, 32(1), 51-62. <https://doi.org/10.1080/10447318.2015.1087664>

- Hung, M., Baumhauer, J. F., Latt, L. D., Saltzman, C. L., SooHoo, N. F., Hunt, K. J., & National Orthopaedic Foot & Ankle Outcomes Research Network. (2013). Validation of PROMIS (R) Physical Function computerized adaptive tests for orthopaedic foot and ankle outcome research. *Clinical Orthopaedics and Related Research*, 471(11), 3466-3474. <https://doi.org/10.1007/s11999-013-3097-1>
- Hung, M., Baumhauer, J. F., Licari, F. W., Voss, M. W., Bounsanga, J., & Saltzman, C. L. (2019). PROMIS and FAAM minimal clinically important differences in foot and ankle orthopedics. *Foot & Ankle International*, 40(1), 65-73. <https://doi.org/10.1177/1071100718800304>
- Hung, M. P., Voss, M. W. M. S., Bounsanga, J. B. S., Crum, A. B. B. S., & Tyser, A. R. M. D. (2016). Examination of the PROMIS upper extremity item bank. *Journal of Hand Therapy*, 30(4), 485-490. <https://doi.org/10.1016/j.jht.2016.10.008>
- Hunnicutt, J. L., Hand, B. N., Gregory, C. M., Slone, H. S., McLeod, M. M., Pietrosimone, B., Kuenze, C., & Velozo, C. A. (2019). KOOS-JR demonstrates psychometric limitations in measuring knee health in individuals after ACL reconstruction. *Sports Health: A Multidisciplinary Approach*, 11(3), 242-246. <https://doi.org/10.1177/1941738118812454>
- Hunsaker, F. G., Cioffi, D. A., Amadio, P. C., Wright, J. G., & Caughlin, B. (2002). The American academy of orthopaedic surgeons outcomes instruments: normative values from the general population. *The Journal of Bone & Joint Surgery*, 84(2), 208-215. <https://doi.org/10.2106/00004623-200202000-00007>

- Hunt, K. J., Alexander, I., Baumhauer, J., Brodsky, J., Chiodo, C., Daniels, T., Davis, W. H., Deland, J., Ellis, S., Hung, M., Ishikawa, S. N., Latt, L. D., Phisitkul, P., SooHoo, N. F., Yang, A., Saltzman, C. L., & Ofar. (2014). The Orthopaedic Foot and Ankle Outcomes Research (OFAR) network: feasibility of a multicenter network for patient outcomes assessment in foot and ankle. *Foot & Ankle International*, 35(9), 847-854.
<https://doi.org/10.1177/1071100714544157>
- Hunt, K. J., & Hurwit, D. (2013). Use of patient-reported outcome measures in foot and ankle research. *The Journal of Bone & Joint Surgery Am*, 95(16), e118(111-119).
<https://doi.org/10.2106/JBJS.L.01476>
- Kallinger, S., Scharm, H., Boecker, M., Forkmann, T., & Baumeister, H. (2019). Calibration of an item bank in 474 orthopedic patients using Rasch analysis for computer-adaptive assessment of anxiety. *Clinical Rehabilitation*, 33(9), 1468-1478.
<https://doi.org/10.1177/0269215519846225>
- Karnofsky, D. A., Abelmann, W. H., Craver, L. F., & Burchenal, J. H. (1948). The use of the nitrogen mustards in the palliative treatment of carcinoma. With particular reference to bronchogenic carcinoma. *Cancer*, 1(4), 634-656.
- Keenan, A. M., Redmond, A. C., Horton, M., Conaghan, P. G., & Tennant, A. (2007). The Foot Posture Index: Rasch analysis of a novel, foot-specific outcome measure. *Archives of Physical Medicine and Rehabilitation*, 88(1), 88-93.
<https://doi.org/10.1016/j.apmr.2006.10.005>

- Kellow, J., & Wilson, V. (2008). Setting standards and establishing cut scores on criterion-referenced assessments some technical and practical considerations. In J. Osborne (Eds.), *Best Practices in Quantitative Methods* (pp. 14). SAGE Publications, Inc.
<https://doi.org/10.4135/9781412995627.d4>
- Kersten, P., White, P. J., & Tennant, A. (2014). Is the pain visual analogue scale linear and responsive to change? An exploration using Rasch analysis. *PLoS One*, 9(6), e99485.
<https://doi.org/10.1371/journal.pone.0099485>
- Khadka, J., Gothwal, V., McAlinden, C., Lamoureux, E., & Pesudovs, K. (2012). The importance of rating scales in measuring patient-reported outcomes. *Health and Quality of Life Outcomes*, 10, 80. <https://doi.org/10.1186/1477-7525-10-80>
- Kirsch, I. S. (1993). *Adult literacy in America: A first look at the results of the National Adult Literacy Survey*. ERIC.
- Ko, Y., Lo, N. N., Yeo, S. J., Yang, K. Y., Yeo, W., Chong, H. C., & Thumboo, J. (2009). Rasch analysis of the Oxford knee score. *Osteoarthritis and cartilage*, 17(9), 1163-1169.
- Korenevskiy, N. A., Shutkin, A. N., Bojcova, E. A., & Dmitrieva, V. V. (2016). Assessment and management of the state of health based on Rasch models. *Biomedical Engineering*, 49(6), 375-379. <https://doi.org/10.1007/s10527-016-9570-x>
- Lafave, M. R., Hiemstra, L., & Kerslake, S. (2016). Factor analysis and item reduction of the Banff Patella Instability Instrument (BPPI): Introduction of BPPI 2.0. *The American Journal of Sports Medicine*, 44(8), 2081-2086.
<https://doi.org/10.1177/0363546516644605>

- Le, D. T. (2013). *Applying item response theory modeling in educational research*. (Publication No. 13410) [Doctoral dissertation, Iowa State University] Iowa state Digital Repository. <https://pdfs.semanticscholar.org/bbf8/08c2b03b02055d8142387a775eeb9f5eda62.pdf>
- Lewis, D., & Cook, R. (2020). Embedded standard setting: aligning standard-setting methodology with contemporary assessment design principles. *Educational Measurement: Issues and Practice*, 39(1), 8-21. <https://doi.org/10.1111/emip.12318>
- Liao, W. W., Ho, R. G., Yen, Y. C., & Cheng, H. C. (2012). The four-parameter logistic item response theory model as a robust method of estimating ability despite aberrant responses. *Social Behavior and Personality: An International Journal*, 40(10), 1679-1694. <https://doi.org/10.2224/sbp.2012.40.10.1679>
- Lin, C.W. C., Moseley, A. M., Refshauge, K. M., & Bundy, A. C. (2009). The Lower Extremity Functional Scale has good clinimetric properties in people with ankle fracture. *Physical Therapy*, 89(6), 580-588. <https://doi.org/10.2522/ptj.20080290>
- Linacre, J. M. (2000). Comparing "partial credit" and "rating scale" models. *Rasch Measurement Transactions*, 14(3), 768.
- Linacre, J. M. (2005). Rasch dichotomous model vs. one-parameter logistic model. *Rasch Measurement Transactions*, 19(3), 1032.
- Linacre, J. M. (1998). Detecting multidimensionality: which residual data-type works best? *Journal of Outcome Measurement*, 2, 266-283.
- Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, 16, 878.
- Linacre, J. M. (2020). *Reliability and separation of measures*. Winsteps. <https://www.winsteps.com/winman/reliability.htm>

- Linacre, J. M., & Wright, B. D. (2000). Winsteps. <http://www.winsteps.com/index.htm>
- Lo Martire, R., Lis, A., Skillgate, E., & Rasmussen-Barr, E. (2017). Psychometric properties of the Swedish version of the Treatment Outcome Satisfaction Questionnaire. *European Spine Journal*, 26(2), 316-323. <https://doi.org/10.1007/s00586-016-4876-7>
- Loken, E., & Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. *British Journal of Mathematical and Statistical Psychology*, 63(3), 509-525. <https://doi.org/10.1348/000711009X474502>
- Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*, 13(4), 517-549. <https://doi.org/10.1177/001316445301300401>
- Lord, F. M. (1969). Statistical adjustments when comparing preexisting groups. *Psychological Bulletin*, 72(5), 336-337. <https://doi.org/10.1037/h0028108>
- Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika*, 39(2), 247-264. <https://doi.org/10.1007/BF02291471>
- Lord, F. M., Novick, M. R., & Birnbaum, A. (2008). *Statistical theories of mental test scores*. Information Age Pub.
- Lucas, R. E., & Donnellan, M. B. (2012). Estimating the reliability of single-item life satisfaction measures: Results from four national panel studies. *Social Indicators Research*, 105(3), 323-331. <https://doi.org/10.1007/s11205-011-9783-z>
- Mancuso, C. A., Sculco, T. P., & Salvati, E. A. (2003). Patients with poor preoperative functional status have high expectations of total hip arthroplasty. *The Journal of Arthroplasty*, 18(7), 872-878. [https://doi.org/10.1016/S0883-5403\(03\)00276-6](https://doi.org/10.1016/S0883-5403(03)00276-6)

- Mancuso, C. A., Sculco, T. P., Wickiewicz, T. L., Jones, E. C., Robbins, L., Warren, R. F., & Williams-Russo, P. (2001). Patients' expectations of knee surgery. *The Journal of Bone & Joint Surgery*, 83(7), 1005-1012. <https://doi.org/10.2106/00004623-200107000-00005>
- Martin-Löf, P. (1974). The notion of redundancy and its use as a quantitative measure of the discrepancy between a statistical hypothesis and a set of observational data [with discussion]. *Scandinavian Journal of Statistics*, 1(1), 3-18.
- Martin, R. L., & Irrgang, J. J. (2007). A survey of self-reported outcome instruments for the foot and ankle. *Journal of Orthopaedic & Sports Physical Therapy*, 37(2), 72-84.
- Martin, R. L., Irrgang, J. J., Burdett, R. G., Conti, S. F., & Swearingen, J. M. V. (2005). Evidence of validity for the Foot and Ankle Ability Measure (FAAM). *Foot & Ankle International*, 26(11), 968-983.
- Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, 47(2), 149-174. <https://doi.org/10.1007/BF02296272>
- Masters, G. N. (1988). Item discrimination: When more is worse. *Journal of Educational Measurement*, 25(1), 15-29. <https://doi.org/10.1111/j.1745-3984.1988.tb00288.x>
- Matheny, L., Gittner, K., Harding, J., & Clanton, T. (2020). Patient reported outcome measures in the foot and ankle: Normative values do not reflect 100% full function. *Knee Surgery, Sports Traumatology, Arthroscopy*, [epub ahead of print]. doi: 10.1007/s00167-020-06069-3
- Matheny, L. M., & Clanton, T. O. (2020). Rasch analysis of reliability and validity of scores from the Foot and Ankle Ability Measure (FAAM). *Foot & Ankle International*, 41(2), 229-236. <https://doi.org/10.1177/1071100719884554>

- McHorney, C. A. (1999). Health status assessment methods for adults: past accomplishments and future challenges. *Annual Review of Public Health*, 20(1), 309-335.
<https://doi.org/10.1146/annurev.publhealth.20.1.309>
- McKeown, R., Rabiou, A. R., Ellard, D. R., & Kearney, R. S. (2019). Primary outcome measures used in interventional trials for ankle fractures: a systematic review. *BMC Musculoskeletal Disorders*, 20(1), 388. <https://doi.org/10.1186/s12891-019-2770-2>
- McPhail, S. M., Williams, C. M., Schuetz, M., Baxter, B., Tonks, P., & Haines, T. P. (2014). Development and validation of the ankle fracture outcome of rehabilitation measure (A-FORM). *Journal of Orthopaedic & Sports Physical Therapy*, 44(7), 488-499, B481-482.
<https://doi.org/10.2519/jospt.2014.4980>
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *The American Council on Education/Macmillan series on higher education. Educational measurement* (p. 13–103). Macmillan Publishing Co, Inc; American Council on Education.
- Moreton, B. J., Walsh, D. A., Turner, K. V., & Lincoln, N. B. (2015). Rasch analysis of the Chronic Pain Acceptance Questionnaire Revised in people with knee osteoarthritis. *Journal of Rehabilitation Medicine*, 47(7), 655-661. <https://doi.org/10.2340/16501977-1977>
- Moreton, B. J., Wheeler, M., Walsh, D. A., & Lincoln, N. B. (2012). Rasch analysis of the intermittent and constant osteoarthritis pain (ICOAP) scale. *Osteoarthritis and Cartilage*, 20(10), 1109-1115. <https://doi.org/10.1016/j.joca.2012.06.011>
- Muller, S., & Roddy, E. (2009). A Rasch analysis of the Manchester foot pain and disability index. *Journal of Foot and Ankle Research*, 2(1), 29. <https://doi.org/10.1186/1757-1146-2-29>

- Nguyen, T. H., Han, H.-R., Kim, M. T., & Chan, K. S. (2014). An introduction to item response theory for patient-reported outcome measurement. *The Patient - Patient-Centered Outcomes Research*, 7(1), 23-35. <https://doi.org/10.1007/s40271-013-0041-0>
- Niama Natta, D. D., Thienpont, E., Bredin, A., Salaun, G., & Detrembleur, C. (2019). Rasch analysis of the Forgotten Joint Score in patients undergoing knee arthroplasty. *Knee Surgery, Sports Traumatology, Arthroscopy*, 27(6), 1984-1991. <https://doi.org/10.1007/s00167-018-5109-x>
- Niki, H., Aoki, H., Inokuchi, S., Ozeki, S., Kinoshita, M., Kura, H., Tanaka, Y., Noguchi, M., Nomura, S., Hatori, M., & Tatsunami, S. (2005). Development and reliability of a standard rating system for outcome measurement of foot and ankle disorders I: development of standard rating system. *Journal of Orthopaedic Science*, 10(5), 457-465. <https://doi.org/10.1007/s00776-005-0936-2>
- Nishigami, T., Mibu, A., Tanaka, K., Yamashita, Y., Yamada, E., Wand, B. M., Catley, M. J., Stanton, T. R., & Moseley, G. L. (2017). Development and psychometric properties of knee-specific body-perception questionnaire in people with knee osteoarthritis: The Fremantle Knee Awareness Questionnaire. *PLoS One*, 12(6), e0179225. <https://doi.org/10.1371/journal.pone.0179225>
- Ortega-Avila, A. B., Ramos-Petersen, L., Cervera-Garvi, P., Nester, C. J., Morales-Asencio, J. M., & Gijon-Nogueron, G. (2019). Clinical rehabilitation. *Clinical Rehabilitation*, 33(11), 1788-1799. <https://doi.org/10.1177/0269215519862328>

- Oude Voshaar, M. A. H., ten Klooster, P. M., Vonkeman, H. E., & van de Laar, M. A. F. J. (2017). Measuring everyday functional competence using the Rasch assessment of everyday activity limitations (REAL) item bank. *Quality of Life Research*, 26(11), 2949-2959. <https://doi.org/10.1007/s11136-017-1627-0>
- Perera, S., VanSwearingen, J., Shuman, V., & Brach, J. S. (2020). Assessing gait efficacy in older adults: An analysis using item response theory. *Gait & Posture*, 77, 118-124. <https://doi.org/10.1016/j.gaitpost.2020.01.028>
- Perruccio, A. V., Stefan Lohmander, L. M., Canizares, M. M., Tennant, A., Hawker, G. A., Conaghan, P. G., Roos, E. M., Jordan, J. M., Maillefert, J. F., Dougados, M. M., & Davis, A. M. (2008). The development of a short measure of physical function for knee OA KOOS-Physical Function Shortform (KOOS-PS) – an OARSI/OMERACT initiative. *Osteoarthritis and Cartilage*, 16(5), 542-550. <https://doi.org/10.1016/j.joca.2007.12.014>
- Pinsker, E., & Daniels, T. R. (2011). AOFAS position statement regarding the future of the AOFAS Clinical Rating Systems. *Foot & Ankle International*, 32(9), 841-842. <https://doi.org/10.3113/FAI.2011.0841>
- Pogorzelski, J., & Millett, P. J. (2017). Editorial commentary: Postoperative outcomes—Are we asking the right questions? Shoulder arthroscopy patient quality of life correlates with joint-specific outcome and is predicated on patient expectation. *Arthroscopy: The Journal of Arthroscopic and Related Surgery*, 33(10), 1786-1787. <https://doi.org/10.1016/j.arthro.2017.07.014>
- Ponkilainen, V. T., Tukiainen, E. J., Uimonen, M. M., Hakkinen, A. H., & Repo, J. P. (2020). Assessment of the structural validity of three foot and ankle specific patient-reported

outcome measures. *Foot and Ankle Surgery*, 26(2), 169-174.

<https://doi.org/10.1016/j.fas.2019.01.009>

Prior, M. E., Hamzah, J. C., Francis, J. J., Ramsay, C. R., Castillo, M. M., Campbell, S. E.,

Azuara-Blanco, A., & Burr, J. M. (2011). Pre-validation methods for developing a patient reported outcome instrument. *BMC medical research methodology*, 11(1), 112-112.

<https://doi.org/10.1186/1471-2288-11-112>

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. University of Chicago Press.

Reise, S. P., & Yu, J. (1990). Parameter Recovery in the Graded Response Model Using MULTILOG. *Journal of Educational Measurement*, 27(2), 133-144.

<https://doi.org/10.1111/j.1745-3984.1990.tb00738.x>

Repo, J. P., Tukiainen, E. J., Roine, R. P., Sampo, M., & Häkkinen, A. (2017). Rasch measurement analysis of the Lower Extremity Functional Scale for foot and ankle patients. *Value in Health*, 20(9), A680. <https://doi.org/10.1016/j.jval.2017.08.1696>

Richter, M., Agren, P. H., Besse, J. L., Coster, M., Kofoed, H., Maffulli, N., Rosenbaum, D., Steultjens, M., Alvarez, F., Boszczyk, A., Buedts, K., Guelfi, M., Liska, H., Louwerens, J. W., Repo, J. P., Samaila, E., Stephens, M., & Witteveen, A. G. H. (2018). EFAS Score - Multilingual development and validation of a patient-reported outcome measure (PROM) by the score committee of the European Foot and Ankle Society (EFAS). *Foot and Ankle Surgery*, 24(3), 185-204. <https://doi.org/10.1016/j.fas.2018.05.004>

- Richter, M., Zech, S., Geerling, J., Frink, M., Knobloch, K., & Krettek, C. (2006). A new foot and ankle outcome score: Questionnaire based, subjective, Visual-Analogue-Scale, validated and computerized. *Foot and Ankle Surgery*, 12(4), 191-199.
<https://doi.org/10.1016/j.fas.2006.04.001>
- Riskowski, J. L., Hagedorn, T. J., & Hannan, M. T. (2011). Measures of foot function, foot health, and foot pain: American Academy of Orthopedic Surgeons Lower Limb Outcomes Assessment: Foot and Ankle Module (AAOS-FAM), Bristol Foot Score (BFS), Revised Foot Function Index (FFI-R), Foot Health Status Questionnaire (FHSQ), Manchester Foot Pain and Disability Index (MFPDI), Podiatric Health Questionnaire (PHQ), and Rowan Foot Pain Assessment (ROFPAQ). *Arthritis Care & Research*, 63 Suppl 11, S229-239. <https://doi.org/10.1002/acr.20554>
- Roos, E. M., Brandsson, S., & Karlsson, J. (2001). Validation of the Foot and Ankle Outcome Score for ankle ligament reconstruction. *Foot & Ankle International*, 22(10), 788-794.
<https://doi.org/10.1177/107110070102201004>
- Rothrock, N. E., Cook, K. F., O'Connor, M., Cella, D., Smith, A. W., & Yount, S. E. (2019). Establishing clinically-relevant terms and severity thresholds for Patient-Reported Outcomes Measurement Information System® (PROMIS®) measures of physical function, cognitive function, and sleep disturbance in people with cancer using standard setting. *Quality of Life Research*, 28(12), 3355-3362. <https://doi.org/10.1007/s11136-019-02261-2>
- Rousson, V., Gasser, T., & Seifert, B. (2002). Assessing intrarater, interrater and test-retest reliability of continuous measurements. *Statistics in Medicine*, 21(22), 3431-3446.
<https://doi.org/10.1002/sim.1253>

- Ruiz-Menjivar, J. (2016). *Using Rasch Measurement Theory to evaluate the psychometric quality of a financial risk tolerance scale* [Doctoral dissertation, University of Georgia] Athenaeum@UGA. http://purl.galileo.usg.edu/uga_etd/ruiz-menjivar_jorge_201605_phd
- Sadjadi, R., Reilly, M. M., Shy, M. E., Pareyson, D., Laura, M., Murphy, S., Feely, S. M. E., Grider, T., Bacon, C., Piscosquito, G., Calabrese, D., & Burns, T. M. (2014). Psychometrics evaluation of Charcot-Marie-Tooth Neuropathy Score (CMTNSv2) second version, using Rasch analysis. *Journal of the Peripheral Nervous System*, 19(3), 192-196. <https://doi.org/10.1111/jns.12084>
- Sahin, A., & Anil, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Education Theory and SciencesL Theory & Practice*, 17(1), 321-335.
- Sallay, P. I., & Reed, L. (2003). The measurement of normative American Shoulder and Elbow Surgeons scores. *Journal of Shoulder and Elbow Surgery*, 12(6), 622-627. [https://doi.org/10.1016/s1058-2746\(03\)00209-x](https://doi.org/10.1016/s1058-2746(03)00209-x)
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, 34(S1), 1-97. <https://doi.org/10.1007/BF03372160>
- Samejima, F., van der Liden, W., & Hambleton, R. (1997). *Handbook of modern item response theory*. New York, NY: Springer.
- Schneider, W., & Jurenitsch, S. (2016a). Age- and sex-related normative data for the Foot Function Index in a German-speaking cohort. *Foot & Ankle International*, 37(11), 1238-1242. <https://doi.org/10.1177/1071100716659747>

- Schneider, W., & Jurenitsch, S. (2016b). Normative data for the American Orthopedic Foot and Ankle Society ankle-hindfoot, midfoot, hallux and lesser toes clinical rating system. *International Orthopaedics*, 40(2), 301-306. <https://doi.org/10.1007/s00264-015-3066-2>
- Sick, J. (2008). Rasch measurement in language education: Part 1. *Shiken: JALT Testing & Evaluation SIG Newsletter*, 12(1), 1-6.
- Sierevelt, I. N., Zwiers, R., Schats, W., Haverkamp, D., Terwee, C. B., Nolte, P. A., & Kerkhoffs, G. (2018, Jul). Measurement properties of the most commonly used Foot- and Ankle-Specific Questionnaires: the FFI, FAOS and FAAM. A systematic review. *Knee Surgery, Sports Traumatology, Arthroscopy*, 26(7), 2059-2073. <https://doi.org/10.1007/s00167-017-4748-7>
- Smith, A. B., Fallowfield, L. J., Stark, D. P., Velikova, G., & Jenkins, V. (2010). A Rasch and confirmatory factor analysis of the general health questionnaire (GHQ)--12. *Health and quality of life outcomes*, 8(1), 45-45. <https://doi.org/10.1186/1477-7525-8-45>
- Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G., & Sharpe, M. (2008a). Rasch fit statistics and sample size considerations for polytomous data. *BMC medical research methodology*, 8(1), 33-33. <https://doi.org/10.1186/1471-2288-8-33>
- Smith, H. J., Richardson, J. B., & Tennant, A. (2008b). Modification and validation of the Lysholm Knee Scale to assess articular cartilage damage. *Osteoarthritis and cartilage*, 17(1), 53-58. <https://doi.org/10.1016/j.joca.2008.05.002>
- Smith, J. E. V. (2001). Evidence for the reliability of measures and validity of measure interpretation: a Rasch measurement perspective. *Journal of applied measurement*, 2(3), 281.

- Smith, J. E. V. (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.
- Smyth, J. D., Olson, K., & Burke, A. (2018). Comparing survey ranking question formats in mail surveys. *International Journal of Market Research*, 60(5), 502-516.
<https://doi.org/10.1177/1470785318767286>
- Steinbrocker, O., Traeger, C. H., & Batterman, R. C. (1949). Therapeutic criteria in rheumatoid arthritis. *Journal of the American Medical Association*, 140(8), 659-662.
<https://doi.org/10.1001/jama.1949.02900430001001>
- Stuber, J., Zech, S., Bay, R., Qazzaz, A., & Richter, M. (2011, Sep). Normative data of the Visual Analogue Scale Foot and Ankle (VAS FA) for pathological conditions. *Foot and Ankle Surgery*, 17(3), 166-172. <https://doi.org/10.1016/j.fas.2010.05.005>
- Stüber, J. M. D., Zech, S. M. D., Bay, R. M. D., Qazzaz, A. M. D., & Richter, M. M. D. P. (2010). Normative data of the Visual Analogue Scale Foot and Ankle (VAS FA) for pathological conditions. *Foot and Ankle Surgery*, 17(3), 166-172.
<https://doi.org/10.1016/j.fas.2010.05.005>
- Tegner, Y., & Lysholm, J. (1985). Rating systems in the evaluation of knee ligament injuries. *Clinical Orthopaedics and Related Research*(198), 43.
- Tegner, Y., Lysholm, J., Lysholm, M., & Gillquist, J. (1986). A performance test to monitor rehabilitation and evaluate anterior cruciate ligament injuries. *The American Journal of Sports Medicine*, 14(2), 156-159.
- Tegner, Y., Lysholm, J., Odensten, M., & Gillquist, J. (1988). Evaluation of cruciate ligament injuries: A review. *Acta Orthopaedica Scandinavica*, 59(3), 336-341.

- Tennant, A., McKenna, S. P., & Hagell, P. (2004). Application of Rasch analysis in the development and application of quality of life instruments. *Value in Health, 7 Suppl 1*, S22-26. <https://doi.org/10.1111/j.1524-4733.2004.7s106.x>
- Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika, 47*(4), 397-412.
- Tilden, V. P., Nelson, C. A., & May, B. A. (1990). Use of qualitative methods to enhance content validity. *Nursing Research, 39*(3), 172-175. <https://doi.org/10.1097/00006199-199005000-00015>
- Tsutakawa, R. K., & Johnson, J. C. (1990). The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika, 55*(2), 371-390. <https://doi.org/10.1007/BF02295293>
- Tsutakawa, R. K., & Soltys, M. J. (1988). Approximation for Bayesian ability estimation. *Journal of Educational Statistics, 13*(2), 117-130. <https://doi.org/10.2307/1164749>
- Tucker, G., Adams, R., & Wilson, D. (2016). The case for using country-specific scoring coefficients for scoring the SF-12, with scoring implications for the SF-36. *Quality of Life Research, 25*(2), 267-274. <https://doi.org/10.1007/s11136-015-1083-7>
- Turner, K. V., Moreton, B. M., Walsh, D. A., & Lincoln, N. B. (2017). Reliability and responsiveness of measures of pain in people with osteoarthritis of the knee: a psychometric evaluation. *Disability and Rehabilitation, 39*(8), 822-829. <https://doi.org/10.3109/09638288.2016.1161840>
- van Cranenburgh, O. D., Prinsen, C. A. C., Sprangers, M. A. G., Spuls, P. I., & de Korte, J. (2012). Health-related quality-of-life assessment in dermatologic practice: relevance and application. *Dermatologic clinics, 30*(2), 323.

- Van Someren, M., Barnard, Y., & Sandberg, J. (1994). The think aloud method: a practical approach to modelling cognitive. *London: AcademicPress*.
- Veltman, E. S., Hofstad, C. J., & Witteveen, A. G. H. (2017). Are current foot- and ankle outcome measures appropriate for the evaluation of treatment for osteoarthritis of the ankle?: Evaluation of ceiling effects in foot- and ankle outcome measures. *Foot and Ankle Surgery*, 23(3), 168-172. <https://doi.org/10.1016/j.fas.2016.02.006>
- Verbrugge, L. M., & Jette, A. M. (1994). The disablement process. *Social Science & Medicine*, 38(1), 1-14. [https://doi.org/https://doi.org/10.1016/0277-9536\(94\)90294-1](https://doi.org/https://doi.org/10.1016/0277-9536(94)90294-1)
- Vrotsou, K., Cuellar, R., Silio, F., Rodriguez, M. A., Garay, D., Busto, G., Trancho, Z., & Escobar, A. (2016). Patient self-report section of the ASES questionnaire: a Spanish validation study using classical test theory and the Rasch model. *Health and Quality of Life Outcomes*, 14(1), 147. <https://doi.org/10.1186/s12955-016-0552-1>
- Wang, I., Kapellusch, J., Rahman, M. H., Lehman, L., Liu, C. J., & Chang, P. F. (2020). Psychometric evaluation of the disabilities of the arm, shoulder and hand (DASH) in patients with orthopedic shoulder impairments seeking outpatient rehabilitation. *Journal of Hand Therapy*. <https://doi.org/10.1016/j.jht.2020.01.002>
- Wang, W., Guedj, M., Bertrand, V., Fouquier, J., Jouve, E., Commenges, D., Proust-Lima, C., Murphy, N. P., Blin, O., Magy, L., Cohen, D., & Attarian, S. (2017). A Rasch analysis of the Charcot-Marie-Tooth Neuropathy Score (CMTNS) in a cohort of Charcot-Marie-Tooth type 1A patients. *PLoS One*, 12(1), e0169878. <https://doi.org/10.1371/journal.pone.0169878>

- Wang, Y. C., Sindhu, B., Lehman, L., Li, X., Yen, S. C., & Kapellusch, J. (2018). Rasch analysis of the activities-specific balance confidence scale in older adults seeking outpatient rehabilitation services. *The Journal of Orthopaedic and Sports Physical Therapy*, 48(7), 574-583. <https://doi.org/10.2519/jospt.2018.8023>
- Ware, J., Kosinski, M., & Keller, S. (1995). *How to score the SF-12 physical and mental health summaries: a user's manual*. The Health Institute, New England Medical Centre, Boston, MA.
- Ware, J. E., Kosinski, M., & Keller, S. D. (1996). A 12-item Short-Form Health survey: construction of scales and preliminary tests of reliability and validity. *Medical Care*, 34(3), 220-233. <https://doi.org/10.1097/00005650-199603000-00003>
- Ware, J. E., Jr., Kosinski, M., Bjorner, J. B., Turner-Bowker, D. M., Gandek, B., & Maruish, M. E. (2008). *SF-36v2® health survey: Administration guide for clinical trial investigators*. QualityMetric Incorporated, 1-34.
- Warholak, T. L., Menke, J. M., Hines, L. E., Murphy, J. E., Reel, S., & Malone, D. C. (2011). A drug-drug interaction knowledge assessment instrument for health professional students: a Rasch analysis of validity evidence. *Research in Social and Administrative Pharmacy*, 7(1), 16-26. <https://doi.org/10.1016/j.sapharm.2010.01.001>
- Wolfe, E., & Smith, E. (2007). Instrument development tools and activities for measure validation using rasch models: Part II - Validation activities. *Journal of Applied Measurement*, 8, 204-234.

- Woodburn, J., Turner, D. E., Rosenbaum, D., Balint, G., Korda, J., Ormos, G., Szabo, A., Vliet Vlieland, T. P., van der Leeden, M., & Steultjens, M. P. M. (2012). Adaptation and crosscultural validation of the foot impact scale for rheumatoid arthritis using rasch analysis. *Arthritis Care & Research*, 64(7), 986-992. <https://doi.org/10.1002/acr.21635>
- World Health Organization. (2002). Towards a Common Language for Functioning, Disability and Health International Classification of Functioning, Disability and Health (ICF). <https://www.who.int/classifications/icf/icfbeginnersguide.pdf>
- Wright, B. (1992). IRT in the 1990s: Which models work best. *Rasch Measurement Transactions*, 6(1), 196-200.
- Wright, B., Linacre, J., Gustafson, J., & Martin-Löf, P. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8(3), 370.
- Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. MESA press.
- Wu, A. W., Kharrazi, H., Boulware, L. E., & Snyder, C. F. (2013). Measure once, cut twice--adding patient-reported outcome measures to the electronic health record for comparative effectiveness research. *Journal of Clinical Epidemiology*, 66(8 Suppl), S12-20. <https://doi.org/10.1016/j.jclinepi.2013.04.005>
- Yen, Y. C., Ho, R. G., Laio, W. W., Chen, L. J., & Kuo, C. C. (2012). An empirical evaluation of the slip correction in the four parameter logistic models with computerized adaptive testing. *Applied Psychological Measurement*, 36(2), 75-87. <https://doi.org/10.1177/0146621611432862>
- Yorke, J., Horton, M., & Jones, P. W. (2012). A critique of Rasch analysis using the Dyspnoea-12 as an illustrative example. *Journal of Advanced Nursing*, 68(1), 191-198. <https://doi.org/10.1111/j.1365-2648.2011.05723.x>

- Yuksel, S., Elhan, A. H., Gokmen, D., Kucukdeveci, A. A., & Kutlay, S. (2018). Analyzing differential item functioning of the Nottingham Health Profile by mixed Rasch model. *Turkish Journal of Physical Medicine and Rehabilitation*, 64(4), 300-307.
<https://doi.org/10.5606/tftrd.2018.2796>
- Yusuf, F., Liu, G., Wing, K., Crump, T., Penner, M., Younger, A., Veljkovic, A., & Sutherland, J. M. (2019). Validating the Foot and Ankle Outcome score for measuring foot dysfunction among hallux valgus surgery patients using item response theory. *Foot and Ankle Surgery*, 26(8), 864-870. <https://doi.org/10.1016/j.fas.2019.11.002>
- Zamanzadeh, V., Ghahramanian, A., Rassouli, M., Abbaszadeh, A., Alavi-Majd, H., & Nikanfar, A. R. (2015). Design and implementation content validity study: Development of an instrument for measuring patient-centered communication. *Journal of Caring Sciences*, 4(2), 165-178. <https://doi.org/10.15171/jcs.2015.017>
- Zanon, C., Hutz, C. S., Yoo, H., & Hambleton, R. K. (2016). An application of item response theory to psychological test development. *Psicologia: Reflexão e Crítica*, 29(1), 1-10.
<https://doi.org/10.1186/s41155-016-0040-x>
- Zwiers, R., Weel, H., Mallee, W. H., Kerkhoffs, G., van Dijk, C. N., & Ankle Platform Study Collaborative - Science of Variation, G. (2018). Large variation in use of patient-reported outcome measures: A survey of 188 foot and ankle surgeons. *Foot and Ankle Surgery*, 24(3), 246-251. <https://doi.org/10.1016/j.fas.2017.02.013>

APPENDIX A

INSTITUTIONAL REVIEW BOARD APPROVAL



Date: 08/04/2020

Principal Investigator: Lauren Matheny

Committee Action: **IRB EXEMPT DETERMINATION – New Protocol**

Action Date: 08/04/2020

Protocol Number: [2007007178](#)

Protocol Title: DEVELOPMENT OF A NEW ANKLE ACTIVITY LEVEL INSTRUMENT USING THE RASCH MODEL FOR ORTHOPAEDIC CLINICAL APPLICATION

Expiration Date:

The University of Northern Colorado Institutional Review Board has reviewed your protocol and determined your project to be exempt under 45 CFR 46.104(d)(7)(2) for research involving

Category 2 (2018): EDUCATIONAL TESTS, SURVEYS, INTERVIEWS, OR OBSERVATIONS OF PUBLIC BEHAVIOR. Research that only includes interactions involving educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures, or observation of public behavior (including visual or auditory recording) if at least one of the following criteria is met: (i) The information obtained is recorded by the investigator in such a manner that the identity of the human subjects cannot readily be ascertained, directly or through identifiers linked to the subjects; (ii) Any disclosure of the human subjects' responses outside the research would not reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, educational advancement, or reputation; or (iii) The information obtained is recorded by the investigator in such a manner that the identity of the human subjects can readily be ascertained, directly or through identifiers linked to the subjects, and an IRB conducts a limited IRB review to make the determination required by 45 CFR 46.111(a)(7).

You may begin conducting your research as outlined in your protocol. Your study does not require further review from the IRB, unless changes need to be made to your approved protocol.

As the Principal Investigator (PI), you are still responsible for contacting the UNC IRB office if and when:



- You wish to deviate from the described protocol and would like to formally submit a modification request. Prior IRB approval must be obtained before any changes can be implemented (except to eliminate an immediate hazard to research participants).
- You make changes to the research personnel working on this study (add or drop research staff on this protocol).
- At the end of the study or before you leave The University of Northern Colorado and are no longer a student or employee, to request your protocol be closed. *You cannot continue to reference UNC on any documents (including the informed consent form) or conduct the study under the auspices of UNC if you are no longer a student/employee of this university.
- You have received or have been made aware of any complaints, problems, or adverse events that are related or possibly related to participation in the research.

If you have any questions, please contact the Research Compliance Manager, Nicole Morse, at 970-351-1910 or via e-mail at nicole.morse@unco.edu. Additional information concerning the requirements for the protection of human subjects may be found at the Office of Human Research Protection website - <http://hhs.gov/ohrp/> and <https://www.unco.edu/research/research-integrity-and-compliance/institutional-review-board/>.

Sincerely,

A handwritten signature in black ink that reads "Nicole Morse".

Nicole Morse
Research Compliance Manager

University of Northern Colorado: FWA00000784

APPENDIX B

INTERVIEW QUESTIONS AND PROMPTS

1. What do you look for in terms of activities and ability when conducting a physical examine on a patient?
2. How do you judge a patient's activity level? Are there any activities that they must be able to perform to be considered high functioning? Moderately functioning? Low functioning?
3. What are the minimal activity requirements for patients to be considered sedentary with minimal function?
4. What type of questions do ask patients when questioning them about their activities regarding their ankle?
 - a. Are these questions more related to activities of daily living? Sport participation?
5. When you ask patients about activities that they are able to perform, do you ask about frequency?
6. Does the amount of pain a patient has when performing an activity play a role in your assessment of their activity level, or is it solely based on activities performed?

APPENDIX C

PANEL OF ORTHOPAEDIC EXPERTS' QUESTIONNAIRE

Q1. There are many different activities that represent ankle activity level. The goal of this question is to determine whether the bolded item effectively represents ankle activity level. In evaluating each item, please ask yourself: *How well does this item represent ankle activity level?*

	Very Poor	Poor	Average	Good	Excellent
	1	2	3	4	5
1. Walking (1 mile) ○	○	○	○	○	○
2. Running (1 mile) ○	○	○	○	○	○

Q2. The next set of questions is related to different activities that can be affected by ankle function. The goal is to have a variety of ankle activities that are representative of a wide range of ankle activity levels. *Please rate the level of difficulty it would take to perform each of the following activities, with 1 representing the easiest level and 5 representing the most difficult level.*

	Very Easy	Easy	Moderate	Difficult	Very Difficult
	1	2	3	4	5
1. Walking (1 mile) ○	○	○	○	○	○
2. Running (1 mile) ○	○	○	○	○	○

Q3. Do you have any other comments or suggestions?

APPENDIX D

READING EXPERTS' QUESTIONNAIRE

Q1. This set of questions is related to reading level. The goal is to have items that are able to be understood by all adults who would complete the survey. *Please rate the level of reading difficulty, with 1 representing the easiest reading level and 5 representing the most difficult reading level.*

	Very Easy	Easy	Moderate	Difficult	Very Difficult
	1	2	3	4	5
1. Walking (1 mile)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
2. Running (1 mile)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Q2. Do you have any other comments or suggestions?

APPENDIX E

ANKLE ACTIVITY LEVEL QUESTIONNAIRE

CONSENT Project Title: Development of a New Ankle Activity Level Instrument Using the Rasch Model for Orthopaedic Clinical Application Purpose and Description: **The primary purpose of this research study is to collect responses for questions that will make up a new ankle activity level scale and determine which questions are the best suited for this new scale. Some questions are very similar or repetitive, but please answer all of the questions.** As a participant you will participate in a 10 to 15 minute survey. Survey questions will be asked about your physical activities that involve the use of your foot/ankle in daily living, as well as at work, and in sport. Demographic questions will be asked about you and your previous foot, ankle, and knee injuries. All attempts to maintain confidentiality will be made by keeping all data stored on a password-protected computer with no identifying information being recorded. IP addresses of participants will not be collected when participating in the online survey, thus there will be no linking of participants to their answers. Confidentiality of data collected online can only be offered to the point of Qualtrics data storage and safety policies. The researcher will be the only person who has access to the login information to Qualtrics. Digital responses will only record your responses. The results of the survey will not be reported individually but compiled as grouped data. There are no foreseeable risks to participating in this study beyond those that occur in natural conversation about your current ankle and foot ability. You may feel anxious or frustrated responding to how you feel about your current ankle and foot ability. There is no benefit to you other than the potential to allow researchers to determine a baseline measure of ankle activity level. There is no cost to participate beyond the time you spend completing this survey. Participation is voluntary. You may decide not to participate in this study and if you begin participation, you may still decide to stop and withdraw at any time. Your decision will be respected and will not result in loss of entitled benefits. Having read the above and having had an opportunity to ask any questions, please complete the survey if you would like to participate in this research. By completing the survey you give your permission to be included as a participant in this study. You may keep this form for future reference. If you have any concerns about your selection or treatment as a research participant, please contact the Nicole Morse, UNC Research Compliance Manager, Office of Sponsored Programs, 25 Kepner Hall, University of Northern Colorado Greeley, CO 80639; 970-351-1910, nicole.morse@unco.edu. You may keep this form for future reference.

Researcher: Lauren Matheny, M.P.H., Principal Investigator, Department of Applied Statistics and Research Methods Phone: 419-410-9443 E-mail: lauren.matheny@unco.edu
 Research Advisor: Dr. Susan Hutchinson, faculty, Department of Applied Statistics and Research Methods, phone: 970-351-1643 e-mail: susan.hutchinson@unco.edu

- ☐ Agree to participate
- ☐ Don't agree

AGE What is your age?

GENDER What is your sex?

- ☐ Male
- ☐ Female

Q48 What is the highest level of school you have completed or the highest degree you have received?

- ☐ Less than high school degree
- ☐ High school graduate (high school diploma or equivalent including GED)
- ☐ Some college but no degree
- ☐ Associate degree in college (2-year)
- ☐ Bachelor's degree in college (4-year)
- ☐ Master's degree
- ☐ Doctoral degree
- ☐ Professional degree (JD, MD)

Q50 Are you Spanish, Hispanic, or Latino or none of these?

- ☐ Yes
- ☐ None of these

Q52 Choose one or more races that you consider yourself to be:

- ☐ White
- ☐ Black or African American
- ☐ American Indian or Alaska Native
- ☐ Asian
- ☐ Native Hawaiian or Pacific Islander
- ☐ Other _____

Q54 Information about income is very important to understand. Would you please give your best guess? Please indicate the answer that includes your entire household income in (previous year) before taxes.

- ☐ Less than \$10,000
- ☐ \$10,000 to \$19,999
- ☐ \$20,000 to \$29,999
- ☐ \$30,000 to \$39,999
- ☐ \$40,000 to \$49,999
- ☐ \$50,000 to \$59,999
- ☐ \$60,000 to \$69,999
- ☐ \$70,000 to \$79,999
- ☐ \$80,000 to \$89,999
- ☐ \$90,000 to \$99,999
- ☐ \$100,000 to \$149,999
- ☐ \$150,000 or more

Q56 In which state do you currently reside?

▼ Alabama ... I do not reside in the United States

Height and weight are very important to our study so please report as accurately as possible. Thank You.

How would you like to answer questions about height and weight?

- ☐ Feet, Inches and Pounds
- ☐ Centimeters and kilograms

What is your height in feet and inches?

Feet

Inches

What is your weight in pounds?

What is your height in centimeters?

What is your weight in kilograms?

Do you **currently** have any **foot or ankle injuries**?

- ☐ Yes
- ☐ No

Have you had any **previous foot or ankle injuries**?

- ☐ Yes
- ☐ No

What YEAR did your **most recent ankle injury** occur? If you can't remember exactly, please estimate.

Year

Have you had any **previous foot or ankle surgeries**?

- ☐ Yes
- ☐ No

What YEAR did your **most recent ankle surgery** occur? If you can't remember exactly, please estimate.

Year

Q150 Do you currently have any knee injuries?

- ☐ Yes
- ☐ No

Have you ever had knee surgery?

- ☐ Yes
- ☐ No

What YEAR did your **most recent knee surgery** occur? If you can't remember exactly, please estimate.

Year

Have you ever been diagnosed with Covid-19?

- ☐ Yes
- ☐ No

Do you currently have Covid-19?

- ☐ Yes
- ☐ No

Are you currently pregnant?

- ☐ Yes
- ☐ No
- ☐ Not Applicable

These questions ask about your **PHYSICAL FUNCTIONING** in relation to your FOOT and ANKLE. Some questions are **intentionally similar to others**, but please answer **ALL** questions to the best of your ability. If you do not participate in some activities, please answer the question based on how much difficulty you believe you would have in performing the activities.

In the last week how much difficulty have you had in performing the following activities due to your foot or ankle?

WALKING

	No Difficulty	Slight Difficulty	Moderate Difficulty	Extreme Difficulty	Unable to do
Walking on flat, even surfaces	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Walking on uneven surfaces (i.e. your yard, a dirt trail)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Walking 1 block at your normal pace	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Walking 1 mile at your normal pace	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Walking 5 miles at your normal pace	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Walking short distances uphill	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Walking long distances uphill	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Walking up a steep hill	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Same instructions as previous set of questions.

These questions ask about your **PHYSICAL FUNCTIONING** in relation to your FOOT and ANKLE. Some **questions are intentionally similar** to others, but please answer **ALL** questions to the best of your ability. If you do not participate in some activities, please answer the question based on how much difficulty you believe you would have in performing the activities.

In the last week how much difficulty have you had in performing the following activities due to

your foot or ankle?

RUNNING

	No Difficulty	Slight Difficulty	Moderate Difficulty	Extreme Difficulty	Unable to do
Running 1 mile at a comfortable pace	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Please select Slight Difficulty for this question	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Running 5 miles at a comfortable pace	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Running 10 miles at a comfortable pace	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Running 100 meters as fast as you can (109 yards)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Running 400 meters as fast as you can (about a quarter mile)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Same instructions as previous set of questions.

These questions ask about your **PHYSICAL FUNCTIONING** in relation to your **FOOT** and **ANKLE**. Some questions are intentionally similar to others, but please answer **ALL** of the questions to the best of your ability. If you do not participate in some of the activities, please answer the question based on how much difficulty you believe you would have in performing the activities.

In the last week how much difficulty have you had in performing the following activities due to your foot or ankle?

BALANCE

	No Difficulty	Slight Difficulty	Moderate Difficulty	Extreme Difficulty	Unable to do
Standing for 15 minutes without leaning on anything	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Standing for 1 hour without leaning on anything	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Going up on your tip toes on <u>both</u> feet <u>without</u> <u>holding</u> on to anything	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Going up on your tip toes on <u>one</u> foot <u>without</u> <u>holding</u> on to anything	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Going up on your tip toes on <u>one</u> foot <u>while holding</u> on to something	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Standing with your knees straight (locked, not bent)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Standing
without
losing
balance for
10 minutes
without
holding on to
or leaning on
anything

☐ ☐ ☐ ☐ ☐

Standing for
20 minutes
without
holding on to
or leaning on
anything

☐ ☐ ☐ ☐ ☐

Please select
Moderate
Difficulty for
this question

☐ ☐ ☐ ☐ ☐

Standing for
an hour
without
holding on to
or leaning on
anything

☐ ☐ ☐ ☐ ☐

Balancing on
one foot for
10 seconds
without
holding on to
anything

☐ ☐ ☐ ☐ ☐

Balancing on
one foot for
60 seconds
without
holding on to
anything

☐ ☐ ☐ ☐ ☐

Standing or
walking for
an entire 8-
hour
workday

☐ ☐ ☐ ☐ ☐

Same instructions as previous set of questions.

These questions ask you about your **PHYSICAL FUNCTIONING** in relation to your FOOT and ANKLE. Some **questions are intentionally similar** to others, but please answer **ALL** questions to the best of your ability. If you do not participate in some activities, please answer the question based on how much difficulty you believe you would have in performing the activities.

In the last week how much difficulty have you had in performing the following activities due to your foot or ankle?

STAIRS, CLIMBING, SQUATTING

	No Difficulty	Slight Difficulty	Moderate Difficulty	Extreme Difficulty	Unable to do
Walking up 1 flight of stairs at your normal pace	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Walking down 1 flight of stairs at your normal pace	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Walking up 4 flights of stairs at your normal pace	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Walking down 4 flights of stairs at your normal pace	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Squatting to pick up something off of the floor	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Standing up from a kneeling position	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Standing up from a sitting position in a chair	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Climbing a ladder	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Same instructions as previous set of questions.

These questions ask you about your **PHYSICAL FUNCTIONING** in relation to your FOOT and ANKLE. Some questions are intentionally similar to others, but please answer ALL questions to the best of your ability. If you do not participate in some activities, please answer based on how much difficulty you believe you would have in performing the activities.

In the last week how much difficulty have you had in performing the following activities due to your foot or ankle?

SIDE TO SIDE MOVEMENTS, CUTTING, PIVOTING

	No Difficulty	Slight Difficulty	Moderate Difficulty	Extreme Difficulty	Unable to do
Stopping quickly while running or jogging	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Jumping up and down on both feet	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Turning quickly or making quick sideways movements (cutting)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Stopping quickly	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Jumping in the air and landing on both feet	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Same instructions as previous set of questions. These questions ask you about your PHYSICAL FUNCTIONING in relation to your FOOT and ANKLE. Some questions are intentionally similar to others, but please answer ALL questions to the best of your ability. If you do not participate in some activities, please answer based on how much difficulty you believe you would have in performing the following activities. In the last week how much difficulty have you had in performing the following activities due to your foot or ankle?

DAILY ACTIVITIES

	No Difficulty	Slight Difficulty	Moderate Difficulty	Extreme Difficulty	Unable to do
Stepping <u>up</u> on curbs	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Stepping <u>down</u> from curbs	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Getting out of a car	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Putting on shoes while sitting (not slip-ons)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Putting on shoes while standing (not slip-ons)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Getting on and off the toilet	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Getting into or out of the shower	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Getting into or out of the bathtub	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Putting on pants while standing	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Bending over and picking up 5 pounds	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Carrying a laundry basket up a flight of stairs	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Running a
short distance
(such as to
catch a bus,
chase after a
child)

☐ ☐ ☐ ☐ ☐

Walking to
the end of the
driveway to
get the mail

☐ ☐ ☐ ☐ ☐

Light
household
work (such as
dusting)

☐ ☐ ☐ ☐ ☐

Please select
Extreme
Difficulty for
this question

☐ ☐ ☐ ☐ ☐

Moderate
household
work (such as
vacuuming
flat surfaces,
mopping,
carrying in
shopping
bags)

☐ ☐ ☐ ☐ ☐

Heavy
household
work (such as
mowing the
grass, raking
leaves,
weeding,
moving
furniture)

☐ ☐ ☐ ☐ ☐

Same instructions as previous set of questions. These questions ask you about your PHYSICAL FUNCTIONING in relation to your FOOT and ANKLE. Some questions are intentionally similar to others, but please answer ALL questions to the best of your ability. If you do not

participate in some activities, please answer based on how much difficulty you believe you would have in performing the following activities.

In the last week how much difficulty have you had in performing the following activities due to your foot or ankle?

EXERCISE

	No Difficulty	Slight Difficulty	Moderate Difficulty	Extreme Difficulty	Unable to do
Light exercise for 15 minutes with no breaks (may include walking or light use of a fitness machine such as a treadmill or elliptical)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Light exercise for 30 minutes with no breaks (may include walking or light use of a fitness machine such as a treadmill or elliptical)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Moderate exercise for 15 minutes with no breaks (may include running at an easy to medium pace or moderate use of a fitness machine such as a treadmill or elliptical)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Moderate to heavy exercise for 30 minutes with no breaks (may include running at a medium to fast pace or heavy use of a fitness machine such as a treadmill or elliptical)



Moderate to heavy exercise for 1 hour with no breaks (may include running at a medium to fast pace or heavy use of a fitness machine such as a treadmill or elliptical)



Same instructions as previous set of questions. These questions ask you about your PHYSICAL FUNCTIONING in relation to your FOOT and ANKLE. Some questions are intentionally similar to others, but please answer ALL questions to the best of your ability. If you do not participate in some activities, please answer based on how much difficulty you believe you would have in performing the activities. In the last week how much difficulty have you had in performing the following activities due to your foot or ankle?

SPORTS, PHYSICAL AND RECREATIONAL ACTIVITIES

	No Difficulty	Slight Difficulty	Moderate Difficulty	Extreme Difficulty	Unable to do
<u>Recreational</u> activities or sports like throwing bean bags, corn hole or horseshoes	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<u>Recreational</u> activities or sports like bowling or shuffleboard	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<u>Recreational</u> activities like hiking	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<u>Recreational</u> sports or activities like golf or swimming	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Light cycling for 15 minutes or less, no hills	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Moderate cycling for 30 minutes, some hills	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Recreational
sports or
activities like
basketball,
soccer,
lacrosse,
tennis,
badminton,
throwing and
catching a
frisbee,
racquetball, or
pickleball

☐ ☐ ☐ ☐ ☐

Recreational
sports or
activities like
hockey, alpine
skiing,
snowboarding,
snowshoeing,
waterskiing,
wakeboarding

☐ ☐ ☐ ☐ ☐

Please select
No Difficulty
for this
question

☐ ☐ ☐ ☐ ☐

Recreational
sports or
activities like
martial arts,
mixed-martial
arts, boxing,
kickboxing, or
CrossFit

☐ ☐ ☐ ☐ ☐

Recreational
sports or
activities like
yoga or
Pilates

☐ ☐ ☐ ☐ ☐

Competitive
sports like
basketball,
soccer,
lacrosse, or
tennis

☐ ☐ ☐ ☐ ☐

Competitive
sports like
hockey, alpine
skiing,
snowboarding,
snowshoeing,
or cross-
country skiing

☐ ☐ ☐ ☐ ☐

Competitive
sports like
baseball or
softball

☐ ☐ ☐ ☐ ☐

Competitive
sports like
martial arts or
boxing

☐ ☐ ☐ ☐ ☐

Competitive
sports like
powerlifting
or CrossFit

☐ ☐ ☐ ☐ ☐

Rock climbing
or bouldering
outdoors

☐ ☐ ☐ ☐ ☐

Completing a
half-marathon

☐ ☐ ☐ ☐ ☐

Completing a
half-marathon
in 2 hours or
less (9.2
minutes per
miles)

☐ ☐ ☐ ☐ ☐

TEGNER Please choose **ONE** of the following that best describes your current activity level:

- ☐ **Level 10:** Competitive sports (Soccer, Football, Rugby (national elite))
- ☐ **Level 9:** Competitive sports (Soccer, Football, Rugby (lower divisions), Hockey, Wrestling, Gymnastics)
- ☐ **Level 8:** Competitive sports (Racquetball, Squash, Track and Field, Alpine Skiing)
- ☐ **Level 7:** Competitive sports (Tennis, Athletics (running), Handball, Basketball, Motocross, Cross Country Track), Recreational sports (Soccer, Football, Squash, Athletics (jumping), Track)
- ☐ **Level 6:** Recreational sports (Tennis, Handball, Basketball, Alpine Skiing, Jogging 5x/week)
- ☐ **Level 5:** Work (heavy labor), Competitive sports (Cycling, Cross-Country Skiing) Recreational sports (Jogging on uneven ground 2x/week)
- ☐ **Level 4:** Work (moderately heavy labor (truck driving, etc.)) Recreational sports (Cycling, Cross-Country Skiing, Jogging on even ground 2x/week)
- ☐ **Level 3:** Work (light labor), Comp & Rec sports (Swimming), Hiking, Backpacking
- ☐ **Level 2:** Work (light labor), Walking on uneven ground possible but impossible to backpack or hike
- ☐ **Level 1:** Work (light labor), Walking on even ground possible
- ☐ **Level 0:** Sick leave or disability pension because of ankle problems

Please answer every question with one response that most closely describes your physical ability within the past week. We realize some questions are repetitive, but please answer every question.

Because of your foot and ankle how much **difficulty do you have with:**

	No Difficulty	Slight Difficulty	Moderate Difficulty	Extreme Difficulty	Unable to Do
Standing	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Walking on even ground	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Walking on even ground without shoes	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Walking up hills	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Walking down hills	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Going up stairs	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Going down stairs	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Walking on uneven ground	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Stepping up and down curbs	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Squatting	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Coming up on your toes	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Walking Initially	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Walking 5 minutes or less	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Select Unable to Do for this question	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Walking approximately 10 minutes	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Walking 15 minutes or more	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Home responsibilities	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Activities of daily living	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Personal care	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Light to moderate work (standing and walking)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Heavy work (pushing/pulling, climbing, carrying)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Recreational activities	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Please answer **every question** with **one response** that most closely describes your physical ability within the past week. We realize some questions are repetitive, but please answer every question.

Because of your **foot and ankle** how much **difficulty do you have with:**

	No Difficulty	Slight Difficulty	Moderate Difficulty	Extreme Difficulty	Unable to do
Running	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Jumping	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Landing	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Squatting and stopping quickly (i.e. stopping quickly during running)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Cutting or lateral or side to side movements	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Low-impact activities	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ability to perform activities with your normal technique	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ability to participate in your desired sport as long as you would like	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

These next questions ask for your views about your health, how you feel and how well you are able to do your usual activities. Answer questions by selecting the one response that most closely describes your condition. If you are unsure about how to answer a question, please give the best answer you can.

In general, would you say your health is:

- ☐ Excellent
- ☐ Very Good
- ☐ Good
- ☐ Fair
- ☐ Poor

The following two questions are about activities you might do during a typical day. Does your **health** now limit you in these activities? If so, how much?

Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf:

- ☐ Yes, limited a lot
- ☐ Yes, limited a little
- ☐ No, not limited at all

Climbing **several** flights of stairs:

- ☐ Yes, limited a lot
- ☐ Yes, limited a little
- ☐ No, not limited at all

During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities as a result of your **PHYSICAL** health?

Accomplished less than you would like:

- ☐ Yes
- ☐ No

Were limited in the **kind** of work or other activities:

- ☐ Yes
- ☐ No

During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities as a result of any EMOTIONAL problems (such as feeling depressed or anxious)?

Accomplished less than you would like:

- ☐ Yes
- ☐ No

Didn't do work or other activities as **carefully** as usual:

- ☐ Yes
- ☐ No

During the **past 4 weeks**, how much did **PAIN** interfere with your normal work (including both work outside the home and housework)?

- ☐ Not at all
- ☐ A little bit
- ☐ Moderately
- ☐ Quite a bit
- ☐ Extremely

The next three questions are about how you feel and how things have been with you during the past 4 weeks. For each question, please give the one answer that comes closest to the way you have been feeling.

How much of the time during the *past 4 weeks*....

	All of the time	Most of the time	A good bit of the time	Some of the time	A little of the time	None of the time
Have you felt calm and peaceful?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Did you have a lot of energy?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Have you felt downhearted and blue?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

During the *past 4 weeks*, how much of the time has your **physical health or emotional problems** interfered with your social activities (like visiting friends, relatives, etc.)?

- ☐ All of the time
- ☐ Most of the time
- ☐ Some of the time
- ☐ A little of the time
- ☐ None of the time

BORN What year were you born?

ZIP What is your zip code?

DIFFICULT_QUESTIONS Were there any questions that were unclear or difficult to understand? If so, which ones and why?

COMMENTS Do you have any other comments or feedback about the survey?

APPENDIX F

REMAINING 77 ITEMS WITH LABELS

Table Appendix E*Remaining Items for Rasch Analysis (77 items)*

Item Label	Item
AAS Balance 1	Standing for 15 minutes without leaning on anything
AAS Balance 2	Standing for 1 hour without leaning on anything
AAS Balance 3	Going up on your tip toes on both feet without holding on to anything
AAS Balance 4	Going up on your tip toes on one foot without holding on to anything
AAS Balance 5	Going up on your tip toes on one foot while holding on to something
AAS Balance 6	Standing with your knees straight (locked, not bent)
AAS Balance 7	Standing without losing balance for 10 minutes without holding on to or leaning on anything
AAS Balance 8	Standing for 20 minutes without holding on to or leaning on anything
AAS Balance 9	Standing for an hour without holding on to or leaning on anything
AAS Balance 10	Balancing on one foot for 10 seconds without holding on to anything
AAS Balance 11	Balancing on one foot for 60 seconds without holding on to anything
AAS Balance 12	Standing or walking for an entire 8-hour workday
AAS Cutting 1	Stopping quickly while running or jogging
AAS Cutting 2	Jumping up and down on both feet
AAS Cutting 3	Turning quickly or making quick sideways movements (cutting)
AAS Cutting 4	Stopping quickly
AAS Cutting 5	Jumping in the air and landing on both feet
AAS Daily 1	Stepping up on curbs
AAS Daily 2	Stepping down from curbs
AAS Daily 3	Getting out of a car
AAS Daily 4	Putting on shoes while sitting (not slip-ons)
AAS Daily 5	Putting on shoes while standing (not slip-ons)
AAS Daily 6	Getting on and off the toilet
AAS Daily 7	Getting into or out of the shower
AAS Daily 8	Getting into or out of the bathtub
AAS Daily 9	Putting on pants while standing
AAS Daily 10	Bending over and picking up 5 pounds
AAS Daily 11	Carrying a laundry basket up a flight of stairs

AAS Daily 12	Running a short distance (such as to catch a bus, chase after a child)
AAS Daily 13	Walking to the end of the driveway to get the mail
AAS Daily 14	Light household work (such as dusting)
AAS Daily 15	Moderate household work (such as vacuuming flat surfaces, mopping, carrying in shopping bags)
AAS Daily 16	Heavy household work (such as mowing the grass, raking leaves, weeding, moving furniture)
AAS Exercise 1	Light exercise for 15 minutes with no breaks (may include walking or light use of a fitness machine such as a treadmill or elliptical)
AAS Exercise 2	Light exercise for 30 minutes with no breaks (may include walking or light use of a fitness machine such as a treadmill or elliptical)
AAS Exercise 3	Moderate exercise for 15 minutes with no breaks (may include running at an easy to medium pace or moderate use of a fitness machine such as a treadmill or elliptical)
AAS Exercise 4	Moderate to heavy exercise for 30 minutes with no breaks (may include running at a medium to fast pace or heavy use of a fitness machine such as a treadmill or elliptical)
AAS Exercise 5	Moderate to heavy exercise for 1 hour with no breaks (may include running at a medium to fast pace or heavy use of a fitness machine such as a treadmill or elliptical)
AAS Run 1	Running 1 mile at a comfortable pace
AAS Run 2	Running 5 miles at a comfortable pace
AAS Run 3	Running 10 miles at a comfortable pace
AAS Run 4	Running 100 meters as fast as you can (109 yards)
AAS Run 5	Running 400 meters as fast as you can (about a quarter mile)
AAS Sport 1	Recreational activities or sports like throwing bean bags, corn hole or horseshoes
AAS Sport 2	Recreational activities or sports like bowling or shuffleboard
AAS Sport 3	Recreational activities like hiking
AAS Sport 4	Recreational sports or activities like golf or swimming
AAS Sport 5	Light cycling for 15 minutes or less, no hills
AAS Sport 6	Moderate cycling for 30 minutes, some hills
AAS Sport 7	Recreational sports or activities like basketball, soccer, lacrosse, tennis, badminton, throwing and catching a frisbee, racquetball, or pickleball
AAS Sport 8	Recreational sports or activities like hockey, alpine skiing, snowboarding, snowshoeing, waterskiing, wakeboarding
AAS Sport 9	Recreational sports or activities like martial arts, mixed-martial arts, boxing, kickboxing, or CrossFit
AAS Sport 10	Recreational sports or activities like yoga or Pilates
AAS Sport 11	Competitive sports like basketball, soccer, lacrosse, or tennis

AAS Sport 12	Competitive sports like hockey, alpine skiing, snowboarding, snowshoeing, or cross-country skiing
AAS Sport 13	Competitive sports like baseball or softball
AAS Sport 14	Competitive sports like martial arts or boxing
AAS Sport 15	Competitive sports like powerlifting or CrossFit
AAS Sport 16	Rock climbing or bouldering outdoors
AAS Sport 17	Completing a half-marathon
AAS Sport 18	Completing a half-marathon in 2 hours or less (9.2 minutes per miles)
AAS Stairs 1	Walking up 1 flight of stairs at your normal pace
AAS Stairs 2	Walking down 1 flight of stairs at your normal pace
AAS Stairs 3	Walking up 4 flights of stairs at your normal pace
AAS Stairs 4	Walking down 4 flights of stairs at your normal pace
AAS Stairs 5	Squatting to pick up something off of the floor
AAS Stairs 6	Standing up from a kneeling position
AAS Stairs 7	Standing up from a sitting position in a chair
AAS Stairs 8	Climbing a ladder
AAS Walk 1	Walking on flat, even surfaces
AAS Walk 2	Walking on uneven surfaces (i.e. your yard, a dirt trail)
AAS Walk 3	Walking 1 block at your normal pace
AAS Walk 4	Walking 1 mile at your normal pace
AAS Walk 5	Walking 5 miles at your normal pace
AAS Walk 6	Walking short distances uphill
AAS Walk 7	Walking long distances uphill
AAS Walk 8	Walking up a steep hill

APPENDIX G

MULTIPLE LINEAR REGRESSION ASSUMPTIONS

Figure A.

Histogram of Foot and Ankle Activity Level Scores.

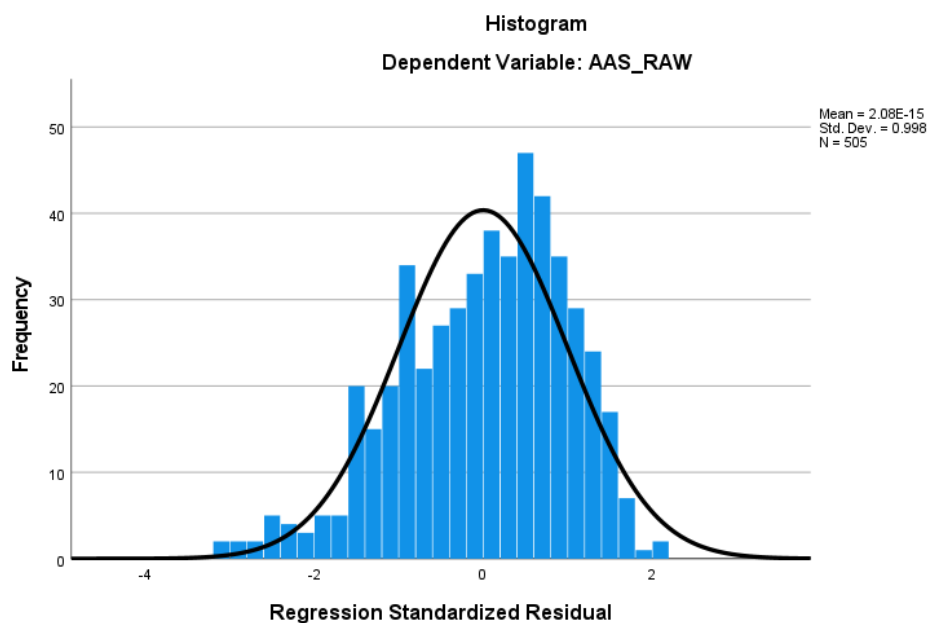


Figure B.

Normal Probability Plot of Foot and Ankle Activity Level Scores.

