The Development of the Foot and Ankle Ability Measure

by

RobRoy Lee Martin

BS Physical Therapy SUNY Health Center at Syracuse, 1991

Submitted to the Graduate Faculty of

University of Pittsburgh in partial fulfillment

of the requirements for the degree of

Doctor of Rehabilitation Science

University of Pittsburgh

2003

UNIVERSITY OF PITTSBURGH

FACULTY OF ARTS AND SCIENCES

This dissertation was presented

by

RobRoy Lee Martin

It was defended on

May 22, 2003

and approved by

Ray Burdett, PhD., PT

Stephen Conti, MD

Jessie VanSwearingen, PhD., PT

James Irrgang, PhD., PT
Dissertation Director

The Development of the Foot and Ankle Ability Measure

RobRoy Lee Martin, PhD

University of Pittsburgh, 2003

The purpose of this project was to develop the Foot and Ankle Ability Measure (FAAM), a self-reported HRQL instrument specific to those with lower leg musculoskeletal disorders. The FAAM consists of the ADL and Sports subscales. Data analysis was done in two stages. Stage I consisted of item selection based on factor loading patterns, inter-item correlations, item to total score correlations, item characteristic curves and the test information functions. 914 subjects participated in the analyses for stage I. Stage II consisted of reliability and validity testing. The data analysis plan consisted of assessing internal consistency, test re-test reliability, responsiveness to change in status, responsiveness compared to general measures of function and validity based on the correlation to concurrent measures of physical and mental functioning. There were 164 subjects in a group expected to change and 79 subjects in a group expected to remain stable. Based on the analyses in state I, 4 items were omitted from the ADL scale. These items were related to pain and sleeping. All items on the Sports subscale were retained. Based on the analyses in stage II, the errors associated with measurement at a single point of time were +/-6.9 and +/-10 points for the ADL and Sports subscales respectively. ICC for test re-test reliability were 0.89 and 0.87 for the ADL and Sports subscales respectively. The minimal detectable change was +/-5.7 and +/-12.3 points for the ADL and Sports subscales respectively. Two-way repeated measures ANOVA and ROC analysis found both the ADL and Sports subscales were responsive to changes in status. The minimal clinically important differences were 8 and 9 points for the ADL and Sports subscales respectively. Guyatt's responsiveness

index and ROC analysis found the ADL subscale was more responsive than general measures of function while the Sports subscale was not. The ADL and Sport subscales had high correlations to SF-36 physical function subscale, physical component summary score and global rating of function and low correlations with the SF-36 mental function subscale and mental component summary score. This study provides evidence of reliability, responsiveness and validity for the FAAM ADL and Sports subscales.

TABLE OF CONTENTS

1 Overview

1.1 The problem

There is a need for an appropriately developed and tested self reported health related quality of life measure for those individuals with foot and ankle impairments. This would be an evaluative measure that could be used by clinicians, researchers, and third party payers to assess self-reported physical performance. The purpose of this research project is to develop such an instrument.

1.2 Introduction

Measuring changes in health status resulting from medical treatment has become increasingly important over the last 20 years[37,44,77]. In the medical community, an important question that is frequently asked is as a result of treatment, what is the final status of a patient compared to his or her initial status in terms of function and quality of life? This type of outcome research, looking at the effect of an intervention on a patient's health related quality of life, has become more important in all medical fields[37,44,77,91]. This has become even more necessary as the need to substantiate the value of intervention to the payers (insurance companies) and consumers (patients) has grown [37,44].

The definition of "outcome" continues to evolve. In recent times, the importance of the patient's perspective in evaluating the outcome of treatment and the success of intervention has been recognized.[26,95]. Consequently, appropriately developed and tested health related quality of life measures in the form of self-reported outcome questionnaires have become a worthy component of outcome assessment[37,55,77].

There are many uses for the information gathered by self-reported health related quality of life measures. In conjunction with other clinical measures, self-reported health related quality of life measures can be used to: 1) justify treatment to third party payers, 2) compare the effectiveness of treatment in research studies, 3) demonstrate the value of treatment to patients, and 4) improve communication between clinicians regarding patients' perceived functional status[37,55]. An appropriately established instrument collects standardized information that is helpful in interpreting the effect of a pathology and subsequent impairment on the patient's functional status and quality of life[55,91]. Score changes on an appropriate instrument during or following various treatments can be monitored over time and compared. Comparisons can be made within the same individuals, within groups, or between groups. Information of this nature can be used to help judge the effectiveness of various treatments to help ascertain the most beneficial treatment methods. This information can also be used to assist in the analysis of cost effectiveness to justify the benefit of treatment to payers [37,55]. A universally accepted instrument would allow easier dissemination of this information between clinicians, researchers and third party payers [37,55].

The importance and usefulness of the information supplied by a self-reported health related quality of life measure focuses around its ability to measure the patient's perceived functional status in a uniform manner[9,55] . Self-reported health related quality of life measures allow information to be collected economically [9] . A great deal of information can be gathered from large numbers of subjects in a quick and efficient manner. The information that is accumulated can be gathered independent of the clinician as the patient should need no assistance to complete the instrument.

The limitations of self reported outcome instruments must also be addressed. This is especially true if the instrument has not gone through appropriate development and testing[9,46,77]. Evidence to support the instrument's reliability validity and responsiveness must be ascertained before the measure can be utilized [46,77]. Clinicians using the instrument must be aware that the patient's motivation while completing the questionnaire will affect the final score and interpretation of the results [9]. Enough information must be provided so that the patients are able to complete the instrument accurately and consistently[9]. Other limitations include the following: 1) data cannot be gathered from a person who is illiterate, 2) test conditions, with respect the amount of assistance given to answer the items, may not be standardized, and 3) subjects may not interpret questions on the instrument in the same manner [9].

The more commonly used measures of HRQL for the foot and ankle are, the American Orthopaedic Foot and Ankle Society Clinical Rating System [48], the Foot Function Index (FFI) [13], the Ankle Osteoarthritis Scale (AOS) [24]and the Lower Extremity Function Scale (LEFS) [8]. Only the FFI, AOS and the LEFS have under gone any reliability and validity testing. This testing was done on a small and limited sample size[8,13,24]. Also, the FFI and AOS use a visual analog format that makes their use in the clinic cumbersome, with respect to scoring, and computerized data collection difficult[13,24]. The SF-36, a generic health related quality of life measure, is also used for patients with impairments of the foot and ankle. Although the SF-36 has evidence to support it's reliability and validity [31],[60-62,78,88] the SF-36 is difficult to score and may not be as sensitive to change as disease or region specific measures [8].

There is a need for an appropriately developed and tested health related quality of life instrument specific to those with foot and ankle disorders. Items need to have appropriate psychometric properties. These psychometric properties include appropriate factoral structure

with items that are all highly related, adequate test re-test reliability, adequate internal consistency, responsiveness to change and evidence of validity based on relations to other variables. The instrument should also provide information across all ability levels, as demonstrated by an appropriate target test information function. Item response theory, such as the graded response model, can be used to assess the items potential to detect change across the spectrum of ability level (potential responsiveness) and target test information function.

1.3 Purpose

The overall purpose of this project is to develop a reliable, valid and responsive self reported health related quality of life instrument specific to those with foot and ankle disorders. This instrument, the Foot and Ankle Ability Measure (FAAM), will consist of two scales. One scale will contain activities related to daily activity. The second scale will contain items related to sports activities. The development of the FAAM will be accomplished in two stages. The purpose of stage I is to develop an instrument that contains items that have appropriate psychometric properties. The purpose of stage II is to assess the instrument's reliability, validity, and responsiveness.

1.4 Research Questions

Specific questions that will be addressed in phase I of the project are:

1) Does the graded response model fit the FAAM ADL and Sports scales?
2) Which items on the instrument are potentially responsive across all ability levels of functional status?
3) Can a target test information function be produced for both the FAAM ADL and Sports scales that maximizes information through a broad range of physical function?

The specific questions that will be addressed in phase II of the project are:

4) What is the factoral structure and how many dimensions are represented by the FAAM ADL and Sports scales?
5) Do the FAAM ADL and Sports scales demonstrate a high level internal consistency?
6) Do the FAAM ADL and Sports scales demonstrate adequate levels of test re-test reliability?
7) Are the FAAM ADL and Sports scales responsive to change in an individual's functional status?
8) Are the FAAM ADL and Sports scales more responsive to changes in physical function than a general measure of health status?
9) Are the FAAM ADL and Sports scales more responsive to changes in physical function than a global rating of self-perceived level of functioning?
10) What is the convergent and divergent evidence to support the interpretation of the final versions of the FAAM ADL and Sports scale?

2 Literature Review

The review of literature will address the following subject matter:

1) disablement models;
2) purposes of health related quality of life measures;
3) types of health related quality of life measures;
4) psychometric properties of health related quality of life measures;
5) currently used health related quality of life measures related to the foot and ankle;
6) development and testing of a health related quality of life instrument;

2.1 Disablement Models

Disablement is a term that attempts to globally identify how human performance is impacted after bodily systems are affected by a pathological condition[40] When describing disablement and the process associated with it, there are two commonly discussed models. One is the Nagi model, developed by Saag Nagi [66]. The second is the International Classification of Impairments, Disabilities and Handicaps model (WHO model) developed for the World Health Organization by Philip Wood [94].

The Nagi model consists of four components: active pathology, impairment, functional limitation and disablement[66]. The WHO model also consists of four components: disease, impairment, disability and handicap[94]. The first two components in both of the models, active pathology and impairment in the Nagi model, and disease and impairment in the WHO model, are essentially the same[40]. Active pathology in the Nagi model and disease in the WHO model are concerned with the pathological condition that affects bodily systems[40]. Impairment, in both models, is concerned with the process of how a pathologic condition directly affects the bodily system[40]. Two examples of impairments could be loss of range of motion and weakness. The final two components in each model, functional limitation and disability in the Nagi model and disability and handicap in the WHO model, deviate in the way they differentiate how changes in bodily systems impact human performance[40].

Nagi describes functional limitation as being at the level of the individual where as disability is at the societal level[40]. Impairments could cause a functional limitation by affecting simple activities of daily living, such as walking and talking. Characteristics or attributes of the individual will determine how performance is altered with respect to functional limitations[40]. Disability, on the other hand, extends beyond the individual and is defined by the roles or tasks society places on that individual. The disability component includes more complex activities,

such as participating in work or school activities that are the combination of a number of simple daily activities[40]. Disability is in part defined by the environmental conditions in which the activities are performed. Because disability is more complex and inter-related, it has been labeled as a "relational concept" that is defined by both individual characteristics as well as environmental conditions[40].

The WHO model defines disability as result of impairment on the individual's capability [40]. As the WHO model defines disability, it is a complex concept, potentially involving the combination and integration of multiple simple activities within the given environmental constraints[40]. The WHO model defines handicap as the consequence of disability interacting with the physical and social environment on an individual's role in society[40]. The WHO model notes that this concept is a classification of circumstances within the individual's environment and society. It is not a direct classification of the individual and is, therefore, a social phenomenon representing social and environmental consequences of impairment or disability[40].

The Nagi model will be used as the model to describe disablement throughout the remainder of the proposal. It has been argued that because the components of the Nagi model can be more clearly differentiated, it is a more desirable model [40]. As the WHO model defines disability, the individual's characteristics and environmental conditions cannot be clearly separated. The WHO model defines handicap as a social phenomenon representing both the social and environmental consequences of impairment or disability and is not a direct classification of the individual [40]. Because of this potential ambiguity, the WHO model may fail to differentiate between limitations in social performance and the causes of these limitations. This may pose problems when trying to define disablement and identify what is being assessed, the individual or the environment[40].

When defining disablement using Nagi's model, the influence of the individual can be identified with functional limitations while the interaction of the person in his or her environment can be identified with disability[40]. A useful health related quality of life measure instrument needs to measure all domains that affect quality of life[46]. As defined by Nagi, areas that can be potentially measured by a health related quality of life instrument include: impairments, functional limitations and disabilities [66]. Defining how someone is performing in every day life is an important function of health related quality of life measures[37]. A main interest of researchers, clinicians, payers and consumers is how pathology and impairment impact performance in every day life[37].

The relationship of impairment to performance in every day life is not direct and improvements in pathology or impairment do not necessarily result in improved functional status[92]. Measures of impairment are therefore felt to be less important when assessing patient change, compared to measures of functional limitation and disability as defined by Nagi[37,92]. Therefore, it has been argued that outcomes research should assess and measure functional limitations and disability [92].

2.2 Purposes of Health Related Quality of Life Measures

Kirshner has outlined three uses of health related quality of life measures: discrimination, prediction, and evaluation[47]. The purpose of a discriminative instrument is to differentiate individuals based on the score they achieve on the measure when no superior gold standard is available[47]. These instruments can also be used to determine the seriousness of the impairments, functional limitation and disability compared to other individuals[47]. Predictive instruments are used in conjunction with a gold standard to categorize individuals [47]. Evaluative instruments measure the patient's change in status over time. Evaluative measures, as opposed to

discriminative and predictive measures, can be used to assess the effectiveness of treatment [47]. One needs to select the appropriate type of instrument as defined by the user's intended purpose. Selecting the type of instrument will be determined by the tester's intended purpose for the instrument and the instrument's associated measurement properties[47]. The primary purpose of outcome research is to evaluate and measure changes in an individual's status over time as a result of treatment. Therefore, evaluative measures are most appropriate for outcome research. Evaluative health related quality of life measures must have items that will relate to a change in health status[47]. Important components of evaluative instruments are as follows: 1) it must be responsive enough to detect clinically important changes in health status; 2) it must be comprehensive but not too tedious to complete; 3) it must be reliable by having reproducible results; and 4) it must be valid by measuring what is was intended to measure[29 47].

2.3 Types of Self Reported Health Related Quality of Life Measures

There are four types of health related quality of life instruments: generic, disease specific, region specific, and patient specific. Generic instruments involve broad questions designed to be appropriate for a wide range of diseases, conditions, and demographics[68,78,88]. Examples include SF-36[83] and the Sickness Impact Profile[7]. The information obtained from generic outcome instruments provide a broad representation of the subject's health. [37,69]. Usually multiple domains are measured. The advantage of a generic measure is that it allows for a comparison of functional limitation and disability among a large number of different disease states and populations [20,69,78]. However, generic outcome instruments may not capture important and unique aspects of a particular disease or population because the items within the instrument are more general in nature[42,69]. Generic instruments may also have limited responsiveness [69,88]

Disease specific instruments are developed for a particular disease or population [65,78,88] and usually measure a single dimension of health status. Examples are the Arthritis Impact Measurement Scale [63] and the Lysholm Knee Scale [53]. The advantage of a disease specific instrument is that it is designed to be more sensitive to the unique characteristics of one distinct disease state or population [69]. Because of this, it maybe more responsive to detect important changes in those patients with the particular disease of interest[37,69]. The obvious shortcoming is that a disease specific measure is only useful for a limited number of individuals as the information gathered cannot be generalized to other diseases or populations [69,91].

Region specific health related quality of life instruments contain elements of both generic and disease specific instruments. Region specific instruments measure self reported functional limitation in subjects or patients with pathology confined to a particular body region and usually measure a single dimension of health status. Examples of these include the Knee Outcome Survey [38], Quebec Disability Index[50] and the DASH (Disabilities of the Arm, Shoulder and Hand) [37]. Region specific instruments are not as all-inclusive as generic instruments because items on region specific instruments pertain to one body region. However, region specific instruments are not as restrictive as disease specific instruments as a number of different pathologies and populations can use the same instrument.

Patient specific instruments were developed to supplement generic, disease specific and region specific instruments[91]. Patient specific instruments use a distinct list of relevant functional activities that each individual patient generates rather than a common predetermined list of activities[91]. In doing so, patient specific instruments attempt to make the scale most meaningful and potentially sensitive to each individual. Examples are the MACTAR[85] and Maximal

Function Measure [54]. The deficiency of patient specific instruments is that comparison across patients is difficult because each patient may generate a unique list of functional activities [91].

2.4 Psychometric Properties: Reliability and Validity

Important properties of self reported outcome instruments are reliability and validity. Reliability refers to the consistency of the measure while validity refers to the accuracy of the measure [39]. In order for an instrument to be useful, it must demonstrate these properties[46].

2.4.1 Reliability

Internal consistency and test-retest reliability are two important forms of reliability. Internal consistency is the degree by which items on the instrument consistently measure the underlying construct[39]. Cronbach's alpha is a popular method to assess internal consistency. Cronbach's alpha is dependent on the number of items and the correlation among the items[39]. It estimates the error associated with inappropriate and/or inadequate sampling of the content domain[39].

Test-retest reliability is used to assess the stability of the score and directly assesses measurement error[47]. When self reported health related quality of life instruments are used, one must make sure that the subject or patient would obtain the same score on the instrument if he or she were to complete the instrument a number of times over a period when his or her status is stable[30,47]. Technically test-retest reliability is the ratio of the variance attributed to true differences among subjects, to the total variance[47]. Test re-test reliability can be assessed using an interclass correlation coefficient[76].

2.4.2 Traditional Validity Theory

There are traditional and contemporary ways of describing validity. Traditional definitions of validity include criterion validity, content validity, and construct validity [81].

Criterion validity refers to the extent to which a health related quality of life instrument will produce the same results as a gold standard or criterion measure [47]. Criterion validity for a measure of functional status is difficult to establish because it is hard to find a gold standard measure of functional limitation[47].

Content validity refers to how well the measure completely covers the important areas of the domain that the measure is attempting to represent[47]. Evaluative instruments need to contain items, which are likely to change as the functional status of the subject or patient changes [47]. In establishing content validity, a pool of items should be generated by a literature review, by experts who work in the domain of interest as well as by subjects or patients who will complete the instrument[47].

Construct validity refers to the relation between the construct being assessed and the particular measure that is being used [47]. Construct validity, as it relates to an evaluative index, attempts to answer the question: How do changes in score on the instrument relate to changes in other related and established clinical measures? [47]. This differs from criterion validity as these clinical measures may not be gold standards. In order for construct validity to be verified, it is imperative that the instruments have good reliability with small measurement error. The score on the instrument must also change in parallel with other related clinical measures [47].

Responsiveness describes an ability to detect change. The issue of responsiveness is controversial. Some believe responsiveness to be an element of validity[81] while others argue that ability to detect change is a separate property[35,93]. Stratford refers to responsiveness as the ability of the instrument to detect clinically significant changes in functional limitations when significant changes have occurred [81]. It is imperative that evaluative instruments demonstrate responsiveness[30]. Stratford feels the greatest challenge in establishing an instrument is to

determine the instrument's capacity to detect clinically meaningful changes [81]. This challenge arises because the definition of meaningful change in functional status is vague[81]. Validated standards of meaningful change in functional status are difficult to find. (disability and handicap)[81].

2.4.3 Contemporary Validity Theory

Related to HRQL measures there are three sources of validity evidence proposed in contemporary validity theory: evidence based on test content, evidence based on internal structure, and evidence based on the relationship to other variables[65]. This evidence of validity is in contrast to the traditional types of validity defined above. These three sources of validity evidence outline how the information collected can contribute to the theoretical justification of validity and do not simply represent different types of validity[65].

2.4.3.1 Evidence Based on Test Content

Evidence based on test content is concerned with how well the test represents the domain of interest[65]. This evidence can come from a theoretical framework, expert judgment, and/or systematic observation of examinees. To obtain this information, the domain of interest must be specified and combined with an account of how the domain is represented by the test. Specific concerns that need to be addressed include construct irrelevance and construct under representation[65]. Construct irrelevance refers to abilities or skills unrelated to the domain of interest that are included on the test. Construct under representation refers to meaningful abilities or skills within the domain of interest that are not assessed by the test. Evidence regarding content will serve to describe the structure and boundaries of the test as well as assist in the interpretation of the test[65].

2.4.3.2 Evidence Based on Internal Structure

Evidence based on internal structure evaluates test items and the relations between items within the construct or domain being assessed by the test[65]. This is accomplished by examining relations of response patterns. Specifically, relations of response patterns include those between individual items, between subdivisions of the test, between items and subdivisions, between items and the entire test, and between subdivisions of the test and the entire test. Analyzing the factoral structure of the test is an important component of evidence based on internal structure. This will allow one to determine how many different domains are represented by the test[65].

2.4.3.3 Evidence Based on Relations to Other Variables

Evidence based on relations to other variables evaluates the level of association between the test score and other measures of the same or related constructs[65]. Evidence based on relations to other variables also evaluates the level of association between the test score and other measures of distinctly different constructs. These other measures can include additional tests that theoretically measure the same construct, related construct, or contrasted construct[63]. Evidence can be convergent in nature, as when a strong relation is established between variables that measure the same or related constructs. Evidence can also be divergent in nature, as when little or no relation is identified between variables that measure distinctly different constructs[65].

Evidence based on relations to other variables also includes evidence related to responsiveness[65]. Demonstration of responsiveness requires evidence that if changes in test scores are related to clinically important changes. Random variability in test scores can confound the ability to identify those who improved from those who did not. Responsiveness can be described in terms of sensitivity and specificity. Sensitivity is defined as the ability to correctly identify those who underwent a change in physical function as demonstrated by a change of

score on the instrument. Specificity is defined as the ability to correctly determine when a change in physical function did not occur as demonstrated by little to no change of score on the instrument[19,20].

2.4.4 Limitations of Traditional Validity Theory

The traditional conceptualization of validity involved in defining, conducting, and reporting validity research has been criticized. The object of validity research is to defend the interpretation of a test score by accounting for the behavior the test score summarizes[65]. Therefore, the interpretation of the test score is validated, not the test itself[65]. The overall process in how this interpretation is defined is the basis of construct related validity. It is argued that construct validity is initiated with test development and continues as the interpretation between test scores and the variables in the domain of interest are more clearly defined[65]. The variables in the domain of interest are identified and defined by the theoretical framework. This process includes defining the meaningfulness and usefulness of the test and will continue through the life of the test. Therefore, it has been suggested that the traditional categories of validity (content, criterion and construct validity) are not distinguishable and that content and criterion validity are two types of evidence of construct validity[65]. Additionally, issues related to validity are never completely satisfied and demonstrated for all uses of the instrument. However, research offers evidence toward validity. This line of thought has lead to more contemporary theories regarding validity[65].

2.5 Health Related Quality of Life Instruments Associated with the Foot and Ankle

A universally accepted self-reported health related quality of life instrument is not available to measure a subject's perceived functional limitation and disability, as defined by Nagi[66], related to impairment of the foot and ankle. An extensive review of the literature was

performed in an attempt to identify all disease specific and region specific health related quality of life instruments that were developed for use with the foot and ankle. Medline was used to search the literature from 1965 to January 1999. Instruments were identified by using the subject "orthopaedics" combined with the following text words: 1) foot, 2) ankle, 3) outcome, 4) index, 5) scale, 6) instrument, 7) clinical rating systems, 8) patient satisfaction, and 9) health status indicators. Thirty-two self reported health related quality of life instruments were identified that included patients' or subjects' perception of their functional status[1,3,6,8,10,13-15,17,24,25,27,36,43,48,49,51,56-58,68,71,73,74,82,87]. Many of these instruments were developed for use in a specific study. Three of the 32 instruments had reliability and validity testing. These include the Foot Function Index (FFI) [13], the Ankle Osteoarthritis Scale (AOS)[24] and the Lower Extremity Function Scale (LEFS) [8]. Five instruments were selected for review. The AOS, FFI, LEFS and SF-36 were chosen because they have research evidence to support their utility. The AOFAS Clinical Rating Systems was chosen for review because it is a commonly found measure in the orthopaedic literature. A brief review of the Foot Function Index (FFI) [13], the Ankle Osteoarthritis Scale (AOS) [24], the Lower Extremity Function Scale (LEFS)[8] and the AOFAS Clinical Rating Systems [48]can be found in Figure 2.1.

2.5.1 American Orthopaedic Foot and Ankle Society Clinical Rating Systems

A commonly used scale in the orthopaedic literature is the American Orthopaedic Foot and Ankle Society clinical rating system[48]. The American Orthopaedic Foot and Ankle Society clinical rating system has four scales that represent four different anatomical regions of the foot. These "clinical rating systems" include the Ankle-Hinfoot, Midfoot, Hallux, and Lesser Toes. Each incorporates items dealing with clinical measures as well as self-report items[48]. To date,

there has been no validity or reliability research done using these clinical rating systems. Also, little description has been given as to how and why items were selected and how the point distribution for each question was determined.

2.5.2 The Foot Function Index

The Foot Function Index (FFI) was developed by Budiman-Mak to measure pain, disability and activity limitation in the elderly population[13]. Item generation was accomplished by a group of clinicians, which included a rheumatologist, physical therapist and podiatrist, with the intent of measuring how foot problems affect pain and function. The FFI consists of 23 items grouped together into three sub-scales including activity limitation (five items), disability (nine items) and pain (nine items). Visual analogue scales that are divided into ten equal segments, with assigned values from zero to nine, are used to score each question[13]. A sub-scale score is obtained by totaling the score for each question, dividing it by the maximum attainable sub-scale score and then multiplying it by 100. If a question is marked not applicable or not answered it is excluded from the total. Sub-scale scores range from zero to 100. A total foot function score is obtained by averaging the sub-scale scores together. A higher score is representative of greater impairment and a lower level of functioning [13].

Factor analysis, internal consistency, reliability, and validity testing were done on 87 subjects with rheumatoid arthritis. [13]. A principle component factor analysis was used to assess internal structure. Items from the pain and disability sub-scales and three items from the activity of daily living sub-scale loaded strongly on one factor. Two items on the activities of daily living sub-scale loaded onto a unique factor. These items questioned the use of an assistive device indoors and outdoors[13].

Internal consistency was examined using Cronbach's alpha[13]. The following results were obtained: 1) total score alpha=0.96, 2) disability sub-scale alpha=0.93, 3) pain sub-scale alpha=0.95 and 4) activity limitation alpha=0.74[13].

Test-retest reliability was done by having subjects complete the FFI during a clinic visit and a second copy was given to the subject to be mailed back within one week[12]. Forty-six percent of the surveys were correctly filled out and returned. Test-retest reliability was estimated for the total score as well as for each of the three sub-scales using intraclass correlation coefficient. The following results were obtained: 1) total score ICC=0.87 (n= 39), 2) activity limitation sub-scale ICC=0.81(n= 40), 3) disability sub-scale ICC=0.84(n= 40), and 4) pain sub-scale ICC=0.70 (n= 39)[13].

Criterion validity was evaluated by assessing the relations between the total score on the FFI and each of the sub-scale scores to the number of painful foot joints, the number of painful hand joints, time to walk fifty feet, and grip strength[13]. The authors hypothesized strong correlations would be found between total score and sub-scale scores to the number of painful foot joints and time it took subjects to walk fifty feet. The authors also hypothesized that weak correlations would be found between total score and sub-scale scores to the number of painful hand joints and to grip strength. Moderate and low correlations were found between these variables[13]. The results are outlined in Figure 2.2.

The ability of the FFI to detect change was assessed over a six-month period by assessing the relationship between a change in an objective measure of disease activity, as defined by the number of painful foot joints to changes in the total FFI score and sub-scale scores[13]. The authors found that changes in the number of painful foot joints correlated moderately with changes in the total FFI score (r=0.45,p=0.003) and changes in the pain sub-scale score (r=0.47, p=0.002).

There was a low correlation between changes in the number of painful foot joints to changes in activity limitation sub-scale (r=0.34, p=0.03). A very low correlation was found between the number of painful foot joints to the change in the disability sub-scale score (r=0.11, p=0.5)[13].

There are a number of limitations with the FFI index. Reliability and validity have been demonstrated, but only in subjects with rheumatoid arthritis. The evidence for validity is not displayed by strong convergent or divergent properties. The correlations between the FFI total score and sub-scale scores to the number of painful foot joint and the time it took subject to walk fifty feet might be expected to be higher. Conversely, the correlations between FFI total score and sub-scale scores to the number of painful hand joints and grip strength might be expected to be lower. However, if a systemic exacerbation of the rheumatoid arthritis is present a stronger correlation may be observed between FFI total score and sub-scale scores to the number of painful hand joints and grip strength. Evidence of validity based on test content may be questionable. The items related to functional activity, in the disability sub-scale of the FFI, are not of a demanding nature as lower level activities such as walking and negotiating stairs are included. There is not any representation for more demanding higher level activities such as running or activities related to sports. Therefore, the FFI exhibits construct under-representation, as all the items represent only lower level activities. Because of this the FFI may exhibit ceiling effects for individuals who function at a high level. Construct irrelevance maybe evident with the activity of daily living sub-scale as demonstrated by the poor internal consistency and inability to load on the common factor. The items on the activities daily of living sub-scale may not be measuring a single construct. The items on this sub-scale may also not be measuring a construct consistent with the other sub-scales. This is particularly true of the items dealing the use of an assistive device. Another potential problem is that the visual analog scale scoring method may be

cumbersome to use in a clinical situation, requiring extra time and the use of a ruler to score. The visual analog scale scoring would also make computerized scanning and automated scoring and data collection difficult. Computerized scoring and collection could be important to automate and chart patient progress. This information could be used by clinicians and others who want to review patient progress. Finally, the contribution of each item with respect to the amount of information it adds to the overall test is unknown.

2.5.3 The Ankle Osteoarthritis Scale

The Ankle Osteoarthritis Scale (AOS) was developed as a modification the Foot Function Index [24]. The objective was to measure patient symptoms and functional limitation as a result of ankle osteoarthritis. To accomplish this, Domsic and Saltzman[23] eliminated the activity limitation subscale and modified anatomical descriptors to apply only to the ankle. Therefore, the AOS consisted of two nine item sub-scales, pain and disability, scored on a 10cm visual analog scale. The responses of the two sub-scale scores are summed[24]. A higher score is representative of greater impairment and a lower level of functioning [13].

Domsic and Saltzman[24] evaluated test-retest reliability and validity of the AOS. Test re-test reliability was assessed by having the 36 subjects with ankle osteoarthritis complete two questionnaires a week apart. Domsic and Saltzman[24] noted "excellent" test-retest reliability of the scale with intraclass correlation coefficients of .97, .95 and .94 for the total score, pain sub-scale and disability subscale respectively (n=28)[24].

Validity was examined by assessing the relation of the pain and disability subscales of the AOS to the Western Ontario McMaster University Osteoarthritis Index (WOMAC) and physical function and bodily pain subscales of the SF-36[24]. The relations were examined using the Pearson Correlation Coefficient in 15 subjects from the test-retest group. Correlations between the WOMAC and AOS pain and disability subscales were r= 0.70 and r=0.65 respectively. The

correlation between the AOS disability score and the SF-36 physical function sub-scale score was r=-0.66 (p=0.005). The correlation between the AOS pain subscale score and the SF-36 pain sub-scale score was r=-0.34 (p<0.2)[24].

The ratio of the number of heel lifts performed by the affected side divided by the number of heel lifts performed by the unaffected side (to a maximum number of 10 heel lifts) was used as a measure of ankle function[24]. The relationship between the total AOS score and the score on each of the AOS subscales to the relative number of heel lifts was assessed using the Spearman rank order correlation coefficient. The authors concluded that the total AOS score, the disability subscale and the pain subscale were "sensitive to the degree of ankle joint dysfunction" with correlation coefficients of 0.88 (p<0.001), 0.90 (p<0.0005) and 0.63 (p<.05) respectively[24].

The authors concluded that the AOS was a reliable and valid outcome instrument to measure patient symptoms and disabilities related to ankle osteoarthritis. However, there are limitations associated with the AOS. A limited population of subjects, those with osteoarthritis, was used when demonstrating reliability and validity of this instrument A greater association should be expected between the SF-36 pain sub-scale and the AOS pain sub-scale. The author's did not give support of divergent evidence. Evidence of validity based on test content may be questionable as described with the FFI. Another potential problem is that the visual analog scale scoring method may be cumbersome to use in a clinical situation, requiring extra time and the use of a ruler to score. The visual analog scale scoring would also make computerized scanning and automated scoring and data collection difficult. Computerized scoring and collection could be important to automate and chart patient progress. This information could be used by clinicians and others who want to review patient progress. Finally, the contribution of each item with respect to the amount of information it adds to the overall test is unknown.

2.5.4 The Lower Extremity Functional Scale

The Lower Extremity Functional Scale (LEFS) was created by Binkley et.al.[8]. Items for the LEFS were generated from review of literature, clinicians, and patient input. Items were intended to assess handicap and disability as defined by the WHO disablement model[40]. Answers to each of the LEFS's 20 items are categorical and scored 0-4 with a total score of 80 indicating a high level of function[8].

The LEFS underwent preliminary statistical testing to assess its factoral structure. Initial factor analysis was done on 57 subjects with 22 interim items[8]. Two items were dropped because of poor factor loading. The remaining 20 items loaded on one factor. The factor loadings ranged from 0.44 (walking between rooms) to 0.86 (performing heavy activities around the house). Coefficient alpha of these 20 items was calculated to be 0.96 (n=107) [8].

Reliability and validity testing was completed on 107 subjects[8]. The site of pathology with these subjects was as follows: hip (n=2), thigh (n=1), knee (n=71), leg (n=8), ankle (n=14), and foot (n=8). Three subjects did not have information with respect to the site of pathology. Test re-test reliability was calculated between the scores on initial evaluation and scores 24-48 hours later. The calculated results using an intraclass correlation coefficient was 0.86(n=107). The 90% confidence interval was calculated to be plus or minus 5.3 points. This implies any score on the LEFS may have up to five points of error associated with it[8].

Construct validity of the LEFS was considered using three separate methods[8]. First, construct validity was assessed by the Pearson correlation coefficient between the initial LEFS and the initial score on the subscales of the SF-36. Convergent evidence was examined by calculating the correlation between the LEFS and the physical function subscale and physical component summary score of the SF-36. These correlation coefficients were 0.80 and 0.64 respectively[8]. Divergent evidence was examined by calculating the correlation between the

LEFS and the mental health subscale and mental component summary score of the SF-36. These correlation coefficients were 0.23 and 0.30 respectively (n=100)[7]. Secondly, construct validity was assessed using a one-way ANOVA to investigate if there were lower LEFS scores in those with recent surgery compared to those without recent surgery. A significant difference was found (P=0.006, n=100). Thirdly, construct validity was assessed by calculating a one-way ANOVA to examine if those subjects with acute conditions had lower LEFS scores than those with chronic conditions. Subjects were given a chronicity rating, based on a three-point scale (1-acute, 2-moderate/unclear, or 3-chronic) by two orthopaedic physical therapists. A significant difference in LEFS scores was found (p=0.027, N=96) between those with acute and those with chronic conditions[8].

A difference in the sensitivity to change between the LEFS and the SF-36 physical function subscale and physical component summary scores was examined at one and three weeks[8]. This was accomplished using a Spearman rank-order correlation coefficient to examine if there was a difference in the relationships between a prognostic rating and the change in LEFS, prognostic rating and the change in SF-36 physical function subscale score and prognostic rating and the change in physical component summary score. Two orthopaedic physical therapists gave the subjects a prognosis rating, from –2 (much worse) to +4 (very large improvement), based on how they felt the subjects would present following one and three weeks of treatment. The results demonstrated: 1) no significant difference between the correlation coefficients of prognosis to the change in score of the physical function subscale and prognosis to the change in score of the LEFS at one week (p=0.106); 2) a significant difference was found between the correlation coefficients of prognosis to the change in the physical component summary score and prognosis to the change in the score of the LEFS at one week (p=0.05); 3) a significant difference was

found between the correlation coefficients of prognosis to the change in the score of the physical function subscale and prognosis to the change in the score of the LEFS at three week.(P=0.002); and 4) a significant difference was found between the correlation coefficients of prognosis to the change in the physical component summary score and prognosis to the change in the score of the LEFS at week three (P=0.019) [8].

The minimally clinically important difference of a change in score was plus or minus 9 points (90% CI) (Binkley). This value was determined using a ROC curve and clinician prognostic ratings. Plus or minus 9 points was associated with a sensitivity of 0.81, a specificity of 0.70 and an area under the curve of 0.76[8].

The authors concluded that the LEFS was reliable, valid, and easy to administer. However, the LEFS does have limitations. The contribution of each item with respect to the amount of information it adds to the overall test is unknown. The items on the LEFS are intended to be relevant to those with hip and knee as well as foot and ankle impairments. Therefore for subjects with foot and ankle disorders, construct irrelevance may be evident, as some of the items may not directly pertain to those with foot and ankle impairments. An example is the item questioning the ability roll over in bed. Construct under-representation may be evident as items particular to foot and ankle impairments may have been excluded. An example would the ability to come up on toes. Because the LEFS is a general HRQL measure, it may not be as sensitive to changes in patient status as a foot and ankle region specific measure. The sample used for testing included only 22 subjects (out of a total of 107 subjects) with impairment of the foot and ankle.

2.5.5 The Short Form Survey (SF-36)

The Short Form Survey (SF-36) was constructed as part of the extensive Medical Outcome Study (MOS) completed in 1986 through 1987, and included over 22,000 subjects [83]. The objective in constructing the MOS SF-36 was to develop a comprehensive, valid, reliable,

precise, and efficient global health survey that could be used in health policy evaluations, general population surveys, clinical research, and clinical practice [88]. Items for the SF-36 were derived from common health related outcome instruments that had been previously used[88]. Since introduction of the SF-36, it has been the principle generic self-reported health related outcome instrument. It has served as a gold standard to which numerous disease specific and region specific health related outcome instruments have been compared.

The SF-36 is a multi-item scale in Likert format that measures eight health concepts, represented by eight different subscales that vary in item number. The eight concepts and the number of items comprising each scale are as follows: 1) physical functioning (10 items); 2) role limitations because of physical health problems (four items); 3) bodily pain (two items); 4) social functioning (two items); 5) mental health (five items); 6) role limitations because of emotional problems (three items); 7) vitality (four items); and 8) general health perceptions (five items). The SF-36 is scored from zero to 100 with higher scores representing better health[88]. Numerous studies have analyzed the psychometric properties of the SF-36[2,11,12,28,31,41,61-63,78, 88].

The test re-test reliability of the SF-36 has been reported to be acceptable in the general medical population [2,11,12,28,41,88]. Patrick et al[70] examined the test re-test reliability in subjects with sciatica. In this orthopaedic population the test re-test reliability results were low. Subjects with sciatica who reported no change in quality of life over a three month period (n=356) were found to have interclass correlations as follows: 1) physical functioning =0.71, role limitations because of physical health problems =0.57, role limitations because of emotional problems=0.73, social functioning=0.32, bodily pain=0.50, mental health=0.50, vitality=0.50, general health perceptions=0.60[70].

McHorney has evaluated the SF-36 for internal consistency, validity and precision. These studies were completed using data from the Medical Outcome Study. The subjects in the Medical Outcome Study had diagnoses of hypertension (HTN), diabetes, congestive heart failure (CHF), recent myocardial infarction(MI)and/or depression (n=3,445) [31,60-63,78] . Internal consistency reliability coefficients were as follows: physical function 0.93, role limitations because of physical health problems 0.84, bodily pain 0.82, social functioning 0.85, mental health 0.90, role limitations because of emotional problems 0.83, vitality 0.87, and general health perceptions0.78[61].

McHorney et al. concluded the SF-36 contained two principal components as extraction of these factors accounted for 70% of the explained variance[61]. The two components were defined as physical and mental health. The pattern of correlation between each of the eight concepts to the two principal components was evaluated. Physical functioning, role limitations because of physical health problems and bodily pain had a strong correlation to the physical health component with correlation coefficients of 0.88, 0.78 and 0.77 respectively. Mental health, role limitations because of emotional problems and social functioning correlated strongly to the mental health component with correlation coefficients of 0.90, 0.81 and 0.71 respectively. Vitality and general health perceptions had moderate correlations to physical health component with correlation coefficients of 0.59 and 0.68 respectively. Vitality and general health also had moderate correlations to the mental health component with correlations of 0.67 and 0.56 respectively[61].

Validity was assessed by evaluating the ability of the SF-36 to distinguish between subjects with serious medical conditions (N=168) from those with minor medical conditions (N=638), as defined by “disease specific severity scales” for HTN, diabetes, CHF and MI [61]. All

of the eight concepts were able to distinguish between those with minor from those with serious medical problems. The concepts of physical function (mean difference=23.18, F=85.9), role limitations because of physical health problems (mean difference=26.35, F=60.6), bodily pain (mean difference=10.96, F=23.6), mental health (mean difference=4.90, F=13.3), social functioning (mean difference=11.59, F=29.9), vitality (mean difference=14.23, F=54.8), general health perceptions (mean difference=17.89, F=84.7) were able to distinguish between those with minor from those with serious medical problems. Role limitation because of emotional problems was able to distinguish between those with minor from those with serious medical problems (mean difference=8.10, F=5.8). [61].

Validity was also assessed by evaluating the ability of the SF-36 to distinguish between those with minor medical problems (N=638) from those with a psychiatric condition or depression (N=163). All concepts except physical function (mean difference=0.09, F=0.0) were able to distinguish between those with minor medical conditions from those with psychiatric conditions at the $p<0.001$ level (mean difference and F-ratios ranged from 29.74, F=294.7 for mental health to 9.11, F=22.9 for general health).[61]

The SF-36 has had comparable results, with respect to responsiveness, when examined against several disease-specific scales for the lower extremity[16,42,45,90,96]. One exception to note was with the LEFS. Binkley et al. compared the LEFS to the SF-36 physical function subscale and the physical component summary score[8]. Specifically, they examined the correlation of the change scores to the clinician prognostic rating. A significant difference in the correlation was found after three weeks of treatment with the change in LEFS score correlating more strongly to the clinician prognostic rating than change scores of both the physical component summary score

(P=0.002) and the physical function subscale (P=0.019). Binkley et al. concluded that the LEFS was "superior to the SF-36 in terms of clinical efficiency and sensitivity to change" in-patients with lower extremity dysfunctions[8].

The SF-36, although it has been shown to be reliable and valid, has a number of flaws. The SF-36, in its entirety, is difficult to use in clinical situations because of the complexity of how it is scored. This may be problematic for those without access to computerized scoring. Also, the SF-36 requires a considerable amount of time to complete. Considering these two issues, compliance in its use by the clinician and the patient may be low. Recently Binkley[8] demonstrated, that the SF-36 is not as sensitive to change as a region specific measure related to the lower extremity. The test re-test reliability in the orthopaedic population may need to be more closely examined. Because the SF-36 is a general HRQL measure, it may not be as sensitive to changes in patient status as a foot and ankle region specific measure. This appears to be supported by Binkley[8]. Psychometric testing has not been completed on subjects with only foot and ankle related impairments. Items on the SF-36, related to functional activities, are general in nature and pertain to lower level activities such as walking and negotiating stairs. Construct under-representation may be present as there is not any representation for more demanding higher level activities such as running or activities related to sports. Because of this ceiling effects maybe exhibited when used with individuals who function at a high level. Construct irrelevance may be evident, as some of the items might not be primarily limited to those with foot and ankle impairments. An example would be the items dealing with bathing or dressing. This activity may be more directly limited by a hip or knee pathology. Construct under-representation may be evident as items particular to foot and ankle impairments may have been excluded. An example would be the ability to come up on toes.

2.5.6 Considerations for the Development of an Instrument

When considering the development of a new self reported health related quality of life instrument one needs to fully examine what is currently being used and ask the question: Why do we need another measure? When evaluating other instruments one must assess and answer the following for each instrument: 1) What is the response burden to the patient or subject required to complete the instrument? 2) What is the burden to the clinician or researcher to score and use the instrument? 3) How was the instrument developed? 4) What is the content of the instrument? 5) How useful is the information the instrument supplies? [9,46,47]. One must utilize the above questions to verify that a new instrument will provide more useful information than that obtained from instruments that have already been developed. At the current time, it is felt that an instrument related to self reported health related quality of life measures for disorders of the foot and ankle can be developed which is superior to that which is presently being utilized.

Perhaps the most significant deficit with the currently used instruments is their failure to use the most precise method psychometric method of test development: namely item response theory [52]. Item response theory allows developers to evaluate how much information each item contributes to the overall score on the instrument[32]. Therefore, items can be more objectively evaluated for their appropriateness and can then be included in or eliminated from the final instrument. Item response theory also allows for the hierarchical ordering of the items on the instrument[32]. Potentially, more information could be obtained from instruments developed using this method.

2.6 Methods Used to Develop a Health Related Quality of Life Instrument

There are two contrasting methods used to develop health related quality of life measures: classic test development and item response theory. Classic test development assesses the reliability and validity of the measure as a whole where as item response theory assesses each item individually.

2.6.1 Classic Test Development

The classic method of test development historically is the most commonly utilized procedure for the development of health related quality of life measures. Typically, when instruments are constructed using classical methods, items are selected based on what the developers deem should be on the instrument. The instrument as a whole is then tested for reliability and validity on a particular sample[32]. Using these classical methods, there is little way to objectively or scientifically assess how each item supplies information about the examinees. Information regarding individual items cannot be disassociated[32]. Classical methods offer no means to select or remove items based on their contribution to the instrument and the examinees overall score. Therefore, examinee characteristics and test characteristics cannot be differentiated. Classical methods offer no means to assess how many items need to be included on the instrument. Testing is sample dependent as conclusions obtained with reliability and validity testing are limited to the population used to complete the testing[32]. Ability estimates are test dependent with harder tests resulting in lower scores and easier tests resulting in higher scores. Also, classical methods do not allow for an ordered continuum of items [32].

2.6.2 Item Response Theory

The basic concept behind item response theory is that the probability of choosing a response for each item is a function the subject's ability and the difficulty level of each item. Calculations can be done and curves can be constructed that represent the probability of choosing

a response for each item based on an examinee's ability. There are a number of different models associated with item response theory. These models can be reduced into three basic models that differ in the number of parameters they incorporate[17,23].

A one-parameter model is able to differentiate items based on how difficult the items are[17,23]. If an item is more difficult, the examinee will have to possess greater ability in order to respond correctly. The difficulty level of each item can be given a numerical value. When the ability level of the examinee is transformed to have a mean ability of zero and a standard deviation of one, the difficulty value of the items usually range from –2 (easier) to +2 (more difficult). A one-parameter model assumes that each item is equal in its ability to rate an examinee based on the examinee's ability[32]. The partial credit model is an example of a one-parameter model [17,23].

A two-parameter model includes an item difficulty parameter and adds a discrimination parameter. In the two-parameter model, each item is assumed to have a different level of discrimination. This discrimination parameter is proportional to the slope of a curve; items that have a steeper slope are better able to separate examinees based on their ability [17,23]. The discrimination factors usually range from 0 to two in value. The graded response model is an example of a two-parameter model [17,2]3.

A three-parameter model includes not only a difficulty and discrimination parameter, but also includes a parameter to reflect success by guessing[17,23]. A three-parameter model is not appropriate in measures of health status because guessing for the correct answer is not a concern[17,23].

Item response theory requires a number of assumptions to be met by the items in order for the results to be valid. These assumptions include unidimensionality, local dependence,

administration of the test is not under time constraints, and guessing for a correct answer is not an option [33]. Unidimensionality implies that the instrument measures a single domain. Local dependence pertains to the examinee's abilities, and it implies that only one factor influences the examinee's response to the items[33]. The assumptions of unidmensionality and local dependence are tested using factor analysis[84]. If it is determined that the items contained on the instrument represent one factor, then the assumptions of unidemensionality and local dependence will be met[33].

In item response theory, the parameters for each item are estimated from the data separately to form the best fitting curves by a maximum likelihood procedure. These curves are called item characteristic curves [33]. This concept is similar to the least squares procedure used in linear regression. The outstanding differences are that item response theory models are not linear in nature, and the regressor or independent variable (ability) is unobservable[33]. However, the concepts are similar because estimates of the parameters are calculated as closely as possible to the expected values in order to fit the best curves for each item[33].

The selection of the IRT model is based on type of data, model fit and the assumptions of the models[23]. There are four models commonly used for instruments with multiple response options; partial credit, graded response, rating scale and nominal response models [17,23]. The partial credit and graded response models are most appropriate when the possible responses can be ordered to represent varying degrees of the ability being measured[23]. The model to be used (one, two, or three-parameter model) will also be determined statistically by the minimal number of parameters that are able to extract a maximal amount of information. A likelihood ratio can

be obtained for each model. The difference between likelihood ratios can be tested statistically, using chi-squared values, to assess if the addition of extra parameters add significantly more information. The simplest model that offers the most information is chosen[23].

Once the model that is most appropriate for the items is determined, the ability to reproduce the results, with respect to model fit and parameter estimates, must be tested. This property is referred to as invariance. It implies that the parameter values determined for each item are unrelated to the characteristics of the examinees answering each item [32]. It also implies that subjects with similar abilities will answer the items the same. The property of invariance should be present as long as a large heterogeneous population is used to calculate the parameter estimates [33]. When the items are determined to be invariant, the same item parameters will be obtained regardless of the population using the instrument[33]. This property of invariance is tested by separating the data into two groups and then comparing the model fit and the parameter estimates of these two groups [33]. The groups can be formed by splitting the data randomly, by age (young and old), and by gender. If the property of invariance is present, then the results of the two groups should be similar[33].

The amount of information supplied by each item and the amount of information supplied by the test as a whole must be examined. The item information function describes the amount of information that each item provides as a function of ability [33]. Item information functions can be summed to provide a test information function. The more information an instrument provides, the more precise the instrument will be with less associated error of estimation [33]. The target test information function for an evaluative instrument should provide a maximal amount of information across all ability ranges[33].

2.6.3 Compare and Contrast Classical Test Development to Item Response Theory

The benefits of classical test theory include smaller sample sizes required for analysis and simpler mathematical analyses compared to item response theory[32]. However, there are significant limitations and they are as follows: 1) the instrument as a whole is tested for reliability and validity on a particular sample; instead of each item as with item response theory; 2) individual items cannot be selected or removed based on their contribution to the instrument and the examinees overall score; 3) examinee characteristics and test characteristics cannot be differentiated as ability estimates are test dependent (harder tests will result in lower scores and easier tests will result in higher scores); 4) there is no means to assess how many items need to be included on the instrument; 5) the testing is sample dependent as conclusions obtained with reliability and validity testing are limited to the population used to complete the testing; and finally 6) an ordered continuum of items based on how difficult the items are cannot be generated [32].

Item response theory offers substantial advantages over classical methods of test construction. One major difference is that item response theory allows one to evaluate the amount of information each item supplies about the examinee's ability. Secondly, the results of item response analysis are invariant. Specifically, the benefits of item response theory over classical testing methods are as follows: 1) item characteristics are not sample dependent; 2) how well a subject performs on the test or instrument is not test dependent; 3) item response theory is item directed while classical test construction methods are test directed; 4) item response theory assesses each item's reliability and not just the test's reliability; 5) selection of items is based on the amount of information the individual item contributes to the total score; and 6) item response theory provides a precise measure that is a function of the subject's ability [32].

There are limitations in using item response theory. One of the major limitations of item response theory is that a large heterogeneous sample of approximately 500 examinees is needed for accurate item parameter calculations and estimates[72]. Also, item response models and calculations are complex, and developers must ensure proper model fit[32].

2.7 Steps to Develop a Health Related Quality of Life Instrument

The construction of a self reported health related quality of life instrument requires six steps: 1) define the purpose of the instrument, 2) item generation, 3) initial item reduction, 4) instrument construction, 5) final item reduction, and 6) reliability and validity testing[37,47,50]. Steps two through five are primarily concerned with issues regarding validity. Specifically, they are concerned with providing evidence based on test content, internal structure and establishing relationships between variables.

2.7.1 Define the Purpose of the Instrument

The first step when developing a health related quality of life instrument is to define the purpose of the instrument[30,47]. Health related quality of life measures maybe used to discriminate between individuals or groups of individuals, to distinguish outcome on a criterion measure or to detect change. The purpose of the instrument will allow the developers to prioritize the properties that the instrument should contain.

2.7.2 Item Generation and Initial Item Reduction

The goal of steps two and three are to establish evidence based on test content. This can be accomplished in two phases: 1) generation of an exhaustive list of all possible items that may be included within the domain of interest and 2) initial item reduction to remove items that are repetitive, complex, too narrow in scope and/or difficult for the subject to interpret. Generation of an exhaustive list of all possible items is done through review of the literature, input from

experts and/or input from a sample of patients or subjects for whom the instrument is intended to be used[47,50,95]. A literature review should examine existing measures to identify what clinical-researchers working in the domain of interest believe are important topics and questions. Experts who work within the domain of interest should have an opportunity to directly contribute ideas, topics and items that they feel need to be included on the instrument in order for it to be successful[95]. Questioning those individuals who will be completing the instrument will also help to make sure that the items address areas they feel should be included on the instrument[95]. Obtaining as much feedback as possible regarding the items to be included in the instrument will help the instrument to be more universally accepted.

Step three, initial item reduction, and can be accomplished by obtaining opinions from experts and patients. Hudac et. al., in developing the DASH, for example, had experts rate individual items in terms of importance, ranging from "very important" to "not important". In doing so he assigned a value to each item, ranging from -2 (not important) to +2 (very important). Descriptive statistics were then obtained to identify the items that the experts cumulatively felt should be included on the index [37]. This method allowed items to be included or excluded based on empirical data.

2.7.3 Instrument Construction

Step four, proper construction of the instrument, includes choosing the most suitable wording and organization of the directions, the items and the possible responses [9,46]. Proper construction of the instrument is implicitly important if useful information is to be collected from it [9,46,47]. Capacity directs the subject or patient to assess what he or she thinks he "could do" or "can do" while performance directs the subject or patient to assess what he or she "did do" or "does do" [36]. The wording of questions to reflect capacity verse performance has been shown to alter scores. Wording in terms of capacity causes potentially inflated scores[59]. However,

wording in terms of capacity should allow patients or subjects to answer all questions based on hypothetical responses if the activity in questions has not been attempted. Wording in terms of performance would allow items to be answered only if the subject has attempted the activity in question within the time frame the instrument is assessing. Answering all items has benefits, as it allows for more equal comparison of scores on the instrument within the same subject as well as between subjects. Therefore, it may be best not to limit the wording of directions to either performance or capacity.

When choosing the wording for potential responses, "difficulty scales" have been utilized successfully [50]. Difficulty scales have questions that read "How much difficulty do you have with?" and responses that range from "no difficulty at all" to "unable to perform". Difficulty scales are also felt to be an appropriate choice because functional limitation and disability are defined in terms of one's difficulty performing various activities [50].

Responses to questions can be set up in a Likert or visual analog format. Likert type response formats use categorical statements for potential responses[9]. These categories are written by the test developer to represent a continuum of response options for each item on the instrument from high to low. Responses set up in a visual analog format have two extreme positive and negative responses, connected by a line, usually 10 centimeters in length[9]. Unlike Likert format, visual analog responses do not force the patient or subject into choosing one category for a response. The patient or subject can choose his or her response to the item in question anywhere along the 10-centimeter line that he/she feels is most representative of his/her status. Use of visual analog scales is sometimes preferred over Likert format because it has the ability to offer more continuous measurement data [50]. However, the use of visual analog scale scoring method may be cumbersome to use in a clinical situation, requiring extra time and the

use of a ruler to score. The visual analog scale scoring would also make computerized scanning and automated scoring and data collection difficult. Computerized scoring and collection could be important to automate and chart patient progress. Also, it would make collection of data by phone interview and computerized scanning and scoring difficult. The disadvantage of a likert format is that the patient if forced to choose a particular response, which may not exactly represent how the patient feels.

Instruments that use a Likert format should: 1) word the item in present tense; 2) avoid statements that are not subject to change; 3) use unambiguous wording; 4) use proper grammar; 5) use simple sentence structure; and 6) consist of easily understandable vocabulary [50].

Item scaling refers to the number of options available to answer each question or item on the instrument[47]. The primary purpose of an evaluative instrument is to register change. Therefore, a score change on the instrument must reflect a clinically important change in the subject [47]. Although the index should be sensitive to change, the subjects' or patients' responses should be consistent. If there are too many choices, the reliability of the subjects' responses will decrease. However, if there are not enough responses, the instrument may not have enough sensitivity to detect change[47]. Likert scales usually allow for a middle or neutral response. Therefore, 5 or 7 point scales are usually suggested[37].

How to score a health related quality of life instrument is also an issue related to scaling. Ability scales imply that a higher score is associated with a higher degree of functioning. Disability scales imply that a lower score is associated with less disability and, therefore, a higher degree of function. Higher scores may be viewed in a more positive light in society while lower scores are usually viewed as more negative. Therefore, ability scales may be easier to interpret.

Construction of the instrument should also include review and pilot testing of the instrument by clinicians and patients or subjects who may use it[37]. At this stage, every effort is made to ensure the instrument is user friendly: easy to administer, easy to complete and easy to score. Being user friendly includes making sure the directions are clear and easy to read.

2.7.4 Final Item Reduction

Step five, final item reduction, involves choosing items that have sound psychometric properties. After the instrument is carefully constructed, items that offer redundant information, those that are not responsive across the complete range of ability, and those that are not reliable or are not contained in the domain of interest need to be eliminated. Factor analysis and item response theory allows items to be selected or eliminated based the characteristics of each individual item. Exploratory factor analysis can be used to assess the appropriateness of a one-factor model. Items that do not fit a one-factor model can be identified and eliminated. Item response theory allows each item to be assessed according to the difficulty level of the item, the ability of each item to discriminate between subjects of varying ability levels, and the amount of information each item provides.

2.7.4.1 Item Reduction Based on Factor Analysis

Factor analysis is a statistical method that can be used to determine the relation among a number of variables[39]. Two commonly used methods are confirmatory factor analysis and exploratory factor analysis. Exploratory factor analysis is a technique that tries to identify relations among a larger number of variables in an effort to condense the number of variables into a smaller subset of variables or factors [39]. Highly related items can be combined into factors and assigned an associated eigenvalue or variance estimate. The amount of variation explained by each factor is indicated by its eigenvalue. The larger the eigenvalue, the more variation that is

explained by the factor. Eigenvalues are divided by the total variance of all the items, and then multiplied by 100 to yield the percent of variance explained by the factor. The eigenvalues can be plotted against the factor number and the curve connecting the points can be examined. This procedure provides test of scree. The test of scree looks for where the "elbow" or curve in the data occurs. The number of points at or to the left of this "elbow" indicates the number of factors that account for largest amount of the variance. A factor rotation procedure can also be done to evaluate how well the items fit into the factors[84]. A factor rotation groups items together based on their inter-item correlation values to allow for more meaningful interpretation. A varimax rotation procedure will produce a factor structure such that each item will load highly on only one factor[75]. Each factor structure should represent a distinct construct[39]. Therefore, items that do load highly on the primary factor may represent a different construct and might be excluded.

A conceptually different technique, confirmatory factor analysis, is a method whereby a hypothesized factor model for the variables is tested[75]. When testing the hypothesized factor structure, evidence is used to make a decision to reject or not reject the null hypothesis. The null hypothesis usually involves a statement that the estimated covariance matrix equals the sample covariance matrix[75]. The chi-squared goodness of fit statistic is an example of a statistical test that can be used to test the hypothesis. The chi-squared goodness of fit statistic tests the null hypothesis that the difference between the estimated covariance and the sample covariance matrix is zero. Evidence that the model fits the data will be provided if the null hypothesis is not rejected[75].

2.7.4.2 Item Reduction Based on Item Response Theory

Item difficulty values, item discrimination values, item characteristic curves and item information functions are produced with Item Response Theory. Item characteristic curves are a

plot between the probability of choosing a particular response and the ability level of the examinee. The item characteristic curves describe the relationship between the traits an item assesses and the item's response pattern across ability levels [33]. Items that have four responses, each response describing an ability to perform an activity, should have four distinct and separate curves. Each of the four curves should have one peak and together should span the spectrum of ability for the examinees for which the item is intended[33].

Item information functions describe the amount of information that each item provides across the spectrum of ability of the examinees [33]. Items that provide more information at high and low ability levels are more desirable and are therefore first selected when constructing a test. The amount of information an item provides is inversely related to the standard error of the estimate. Item information functions can be summed to provide a test information function. The more information the test or instrument provides across the complete spectrum of ability the more precise the estimation of ability for evaluative measures [33]. Generally, items are selected or eliminated based on item characteristic curves and item information functions.

2.8 Establishing Reliability and Evidence Based on the Relationship to other Variables

Step six has to do with establishing reliability and providing evidence based on the relationship to other variables. Step six can be initiated once the final list of items that are to be included on the instrument is completed.

2.8.1 Test Re-test Reliability

Test-retest reliability is measured by administering the instrument at least twice over a period in which the construct being measured does not change[47]. When using the instrument for evaluative purposes, the most important aspect of reliability is that with-in person variance is small.[30,47]. Statistical assessment of reliability can be done using Kappa, Pearson correlation

coefficient or the Intraclass correlational coefficient (ICC)[20]. The Kappa statistic is commonly used for categorical data, while the Pearson and ICC are used for continuous data. The Pearson is deficient because it cannot account for systematic error, and it is a measure of association and not agreement. The systematic error associated with Pearson r may make the results and subsequent interpretation inaccurate. The ICC accounts for systematic errors and is felt to be a true reliability coefficient[76]. Therefore ICC has been recommended as the statistic of choice when assessing reliability of interval or ratio level data[20].

2.8.2 Providing Evidence Based on the Relationship to other Variables

Providing evidence based on the relation to other variables can be done by a number of methods. One method involves selecting clinically relevant measures of function that assess the same domain as the self reported outcome instrument[65]. However, finding appropriate clinical measures may be difficult. Clinical measures that are sensitive for the wide spectrum of functional levels, from high function to low function, may not be able to be identified.

A second method of providing evidence based on the relationship to other variables involves using established instruments of the construct for comparison[65]. These established instruments would have already gone through rigorous testing to have their validity and reliability established. Validity evidence can be exhibited through a strong relationship between instruments that measure the same domain, convergent validity. Validity evidence can also be exhibited through a weak relationship between instruments that measure different domains, divergent validity[65]. Statistically, Pearson correlation coefficient can be used to assess these relationships. Statistical testing can also be done to assess if the relationship between the instrument of interest and the divergent measure is statistically different than the relationship between the instrument of interest and the convergent measure.

2.8.3 Responsiveness

When establishing responsiveness, Stratford proposes a number of different designs. These designs include both single and multiple groups [81]. The foremost single group design is one in which measurements are taken at three points in time. The first measurement is an initial baseline measurement. The second measurement is taken a short time after the baseline measurement. There should be no change in the subject's functional status between the first and second measurement and, therefore, no change in the score on the instrument. This allows test re-test reliability and measurement error of the instrument to be estimated [81]. The third measurement is taken at a later time after a change in the subject's functional status is expected. Responsiveness is estimated as the change in score on the instrument that reflects a meaningful change in functional status [81]. This design can be analyzed statistically using effect size, standardized response means, and paired t-value [81]. When using this design there is no way to asses how much variability will occur over time in subjects whose functional status does not change and no way for determining when a true change occurs[81].

Multiple group designs improve on single group designs and involve putting subjects into groups. This allows for comparisons in status between those groups expected to change and those groups not expected to change [81]. There are three multiple group designs . The first design involves randomly putting subjects into two groups. One group will receive a proven treatment and, therefore, will be expected to undergo a clinically significant change in status. The second group will receive a placebo treatment, and no change in status will be expected. The second design involves the separation of subjects into two groups based on their prognosis for change[81]. One group has a good prognosis for clinically significant change while the second group has a

poor prognosis for clinically significant change. The third design involves the separation of groups based on an external standard of change into those that display important changes on this external standard and those that do not [81].

Statistical analyses used with these designs include: Guyatt's Responsiveness Index (GRI), t statistic for independent change scores, analysis of variance of change scores, Norman's S, receiver operating characteristic (ROC) curve and correlation coefficient. GRI is calculated as the change score in a group where change is expected divided by the standard deviation of a group where change is not expected [30]. The GRI can be used to compare competing instruments[81].

A ROC curve is a plot of sensitivity versus 1-specificity. The curve is generated by calculating sensitivity versus 1-specificity for every one-point change one point change on the instrument. Sensitivity is defined as the ability to correctly identify those who underwent a change in physical function as demonstrated by a change of score on the instrument. Specificity is defined as the ability to correctly determine when a change in physical function did not occur as demonstrated by little to no change of score on the instrument[18,19]. A cut-point is chosen as the value when sensitivity is one and 1-specificity is zero. It is the point at which the curve is at the most upper left position on the graph. The ROC curve for a perfect measure would look like a right angle. The curve would ascend and trace the y-axis to 100 (perfect sensitivity) and then form a 90-degree angle and run parallel to the x-axis (prefect specificity). A measure of no use would be represented by a diagonal line that would proceed from the lower left-hand corner to the upper right-hand corner.[19].

Two advantages have been noted for ROC analysis when establishing the ability of an instrument to detect change[81]. The first is that a critical ratio z can be calculated to compare the

area under the ROC curve of two competing instruments. Critical ratio z, as outlined by Hanley and McNeil, make use of the calculated area of two curves, standard errors of each curve and the correlation coefficient between areas to calculate a z value[34]. This z value can be compared to a cutoff value (e.g. z>1.96) to determine if the areas under the two ROC curves are different[34]. The ability to correctly identify subjects is represented by the area under the ROC curve. The value of the area can range from 0.5, no accuracy, to 1.0, perfect accuracy, in identifying those who improved from those who did not improve[34,81].

The second advantage of the ROC curve is that a cut point is generated that can determine who has made a clinically important change in functional status from one who has not[81].

2.9 Summary

Self reported HRQL instruments are commonly used by clinicians and researchers. In order to be useful the instrument needs to be appropriately developed and tested. There is a need for a self-reported HRQL instrument for individuals with impairments of the foot and ankle. The development of the Foot and Ankle Ability Index (FAAM) has been proposed to fill this need. Initial item development and reduction has yielded the interim FAAM. A description of these methods can be found in Appendix A.

The interim FAAM consists of two scales, the Activities of Daily Living (ADL) and Sports sub-scales. The ADL sub-scale contains 27 items pertaining to basic functional activities. The Sports sub-scale contains eight items pertaining to higher level activities, such as those that maybe required in athletics. There are five potential responses in a Likert-type format. These five

responses range from “no difficulty at all” to “unable to perform”. There is also a “non-applicable” category for those activities limited by something other than the foot and ankle. A copy of the interim FAAM can be found in Appendix B.

The final version of the FAAM will be produced through further research. This will include selecting items based on item response theory. Once this final FAAM is produced the usefulness of this measure will be evaluated. This evaluation will include reliability, responsiveness and validity testing.

2.9.1 Purpose of the Project

The overall purpose of this project is to develop a reliable, valid and responsive patient reported health related quality of life instrument specific to those with foot and ankle related impairments. This project will be accomplished in two stages. The purpose of stage I is to develop an instrument that contains items that have appropriate psychometric properties. The purpose of stage II is to assess the instrument’s reliability, validity, and responsiveness.

2.9.2 Research Questions

Specific questions that will be addressed in phase I of the project are:

1) How well does the graded response model fit the FAAM ADL and Sports scales?
2) Are the items responsive across the spectrum of functional status?
3) Can a target test function be produced for the FAAM ADL and Sports scales that maximize information through a broad range of physical function?

Specific questions that will be addressed in phase II of the project are:

4) What is the factoral structure of the FAAM ADL and Sports scales?
5) Do the FAAM ADL and Sports scales demonstrate high levels of internal consistency?
6) Do the FAAM ADL and Sports scales demonstrate adequate levels of test re-test reliability?

7) Are the FAAM ADL and Sports scales responsive to change in an individual's functional status changes?
8) Are the FAAM ADL and Sports scales more responsive to changes in functional status than a general measure of health status?
9) Are the FAAM ADL and Sports scales more responsive to changes in physical function than a global rating of self-perceived level of functioning?
10) What is the convergent and divergent evidence to support interpretation of the final versions of the FAAM ADL and Sports scales?

3 Methods

The overall purpose of this project is to develop a reliable, valid and responsive patient-reported health related quality of life instrument specific to those with foot and ankle disorders. This project will be accomplished in two stages with six proposed steps. The purpose of stage I is to develop an instrument by selecting items that are unidimensional, potentially responsive across ability levels and contribute appropriate information to the test. The purpose of stage II is to assess the instrument's reliability, validity, and responsiveness. Stage I contains five steps: 1) define the purpose of the instrument, 2) item generation, 3) initial item reduction, 4) instrument construction and, 5) final item reduction. Stage II involves a sixth step: 6) reliability and validity testing.

3.1 Methods to Produce Final Version

3.1.1 Research Questions

1) Does the graded response model fit both FAAM ADL and Sports scales?
2) Which items on the instrument are potentially responsive across all ability levels of functional status?

3) Can target test information functions be produced for the FAAM ADL and Sports scales that maximize information through a broad range of physical function?

3.1.2 Item Development and Reduction

The objective of item development and item reduction is to select appropriate items to be included on the instrument. The process which included defining the purpose of the instrument, generating potential items, initial item reduction and instrument construction was completed as preliminary work. These methods can be found in Appendix A. The initial item reduction eliminated items that were felt to be obviously unimportant or repetitive. Final item reduction will be accomplished through analysis of individual items using item response theory.

3.1.3 Procedures for Field Testing to Produce the Final FAAM

Initial item selection was completed April 1997 through December 1997. Once the interim FAAM was completed, its use was implemented on a routine basis in clinics owned and operated by the Centers for Rehab Services. Data was extracted from the outcome database and was used to analyze the psychometric properties of individual items.

The Centers for Rehab Services uses the interim FAAM in everyday treatment as means to assess functional status of patients with foot and ankle disorders undergoing treatment in their clinics. The FAAM is routinely administered to patients during the initial visit, at weekly intervals and again at discharge. The data collected for stage I will include the initial FAAM ADL and Sports scores, initial general measure of health status (SF-36) and demographic information. The demographic information will include a diagnosis (ICD-9) code, surgical procedure, date of surgery, the time from onset of condition to initiation of physical therapy, mechanism of injury, age, race, and gender. This data has already been entered into a computerized database and is available for analysis.

3.1.3.1 Subjects: Inclusion/Exclusion Criteria

The subjects will include patients from the CRS facilities who have a lower leg (below the knee) musculoskeletal disorder (bone, muscle, ligament, and/or joint). Those subjects with pathologies that involve body regions other than the lower leg, as well as subjects with coexisting pathologies will be excluded from the study. Data with more than three missing responses on the FAAM ADL section or one missing response on the FAAM Sports section will be excluded. There will be no discrimination based on age, gender, race or religion. Confidentiality will be maintained, as patient identification information will be separated from the data so that the subject's individual responses will be unknown to the researchers. Responses of approximately 1000 individuals will be used for analysis. Reise and Yu[72]. feel that responses from approximately 500 individuals are needed to fit the item response model. Twice this number will be needed to assess invariance as described in section 3.1.3.2.3. The procedure to assess invariance requires that the data be split in half, forming two groups. The item response model is then fit to each subgroup. Therefore, 500 individuals will be needed in each group, for a total of 1000 individuals.

3.1.3.2 Research Question 1: Does the graded response model fit both FAAM ADL and Sports scales?

The degree of fit of the item response model will depend on the following: 1) the data's ability to meet underlying assumptions, 2) the reproducibility of the results, and 3) the ability of the observed data to fit to the model predictions.

3.1.3.2.1 Item Response Theory Assumptions

Item response theory requires a number of assumptions to be met by the items in order for the results to be appropriate. These assumptions include the following: 1) unidimensionality,

2) local independence, 3) that the administration of the test is not under time constraints, and 4) no guessing for a correct answer [33,84]. Unidimensionality implies that the instrument measures a single domain. Local independence pertains to the examinee's abilities, and implies that only one latent trait influences the examinee's response to the items. The assumptions of unidimensionality and local dependence are tested using factor analysis. If it is determined that the items contained on the instrument represent one factor, then the assumptions of unidimensionality and local dependence will be met[33,84]. The administration of the test is not under time constraints, and therefore this assumption is met. The assumption of guessing for the correct answer is also met, as there are no right or wrong answers with HRQL measures.

NOVAX 1.3 (Poor Professor Software, Davis, CA) will be used to complete the exploratory factor analysis and assess the degree of unidimensionality. This will be done separately for both the ADL and Sports scales. The pattern of correlations between items will be used to identify and extract factors. The amount of variation explained by each factor will be indicated by an eigenvalue. The larger this eigenvalue, the more variation that is explained by the factor. The eigenvalue greater than one rule notes that all eigenvalues greater than one represent a different factor. Each eigenvalue will be divided by the sum of the eigenvalues. This number will then be multiplied by 100 to yield the percentage of variance explained by each factor. The eigenvalues will be plotted against the factor number and the resulting in scree plot will be examined. The test of scree looks for where the "elbow" or curve in the data occurs [84]. The number of points at or to left of this "elbow" indicates the number of factors that should be retained. A varimax rotation will also be done to help evaluate how well the items fit into one factor. A varimax rotation procedure will group items together based on their inter-item correlation values[75]. A varimax rotation procedure will produce a factor structure such that each

item will load highly on only one factor. Each factor structure should represent a distinct construct[75]. Therefore, items that do load highly on the primary factor may represent a different construct and might be excluded.

Since the objective of this analysis is to produce an instrument that is unidimensional (containing one factor) for both the ADL scale and Sports scale, items that do not fit the one factor model will be eliminated. It is hypothesized that the interim ADL and Sports scales will each contain one factor. However, it may be possible that most of the items are contained within the first factor, with a small number of items loading on other factors. Therefore, items contained within the first factor should account for a large portion of the variance and have large eigenvalues relative to the other factors.

If the test of scree and the eigenvalue greater than one rule find more than one factor, in either the ADL scale or the Sports scale, the items within these other factors will be considered for elimination. The exploratory factor analysis will be completed again with the same analysis as described above. At this point, the one factor model should be identified for both scales. If the one factor model is still not achieved, the same procedure of eliminating items will be repeated until a one-factor model is achieved for both scales.

In testing the assumption of item response theory the hypotheses are: 1) there will be one single eigenvalue that is greater than one, 2) the resulting test of scree will have only one point at or to the left of the "elbow", and 3) after factor rotation, all items will load on one factor.

3.1.3.2.2 Assessing model fit

Multilog (Scientific Software Inc., Chicago IL) will be used to calibrate items, separately for both the ADL and Sports scales. Item calibration will be done using both the two-parameter graded response model and the one-parameter partial credit model. Item calibration using the

one-parameter model will provide an estimation of the item difficulty parameters. Item calibration using the two-parameter model will provide an estimation of the difficulty and discrimination parameters for each item[17,23]. A likelihood ratio will also be obtained for each model, for both the ADL and Sports scales. The fit of the one-parameter and two-parameter models will be compared using the difference in the negative twice the log likelihood statistics [33]. The difference in the negative twice the log likelihood statistic is distributed as a chi-squared statistic with the degrees of freedom equal to the difference in the number of parameters between the two models. If the difference in the negative twice the log likelihood statistics is greater than the critical value at the appropriate degree of freedom then it can be assumed that the addition of the extra parameters, with the two-parameter model, adds significantly to the fit of the model to the data over and above the one-parameter model[33].

It is hypothesized that the difference in the negative twice the log likelihood statistics will be greater than the critical value and therefore the two-parameter model will offer a better fit to the data than the one-parameter model for both the ADL and Sports scales.

3.1.3.2.3 Assessing Parameter Invariance

Once the model that is most appropriate for the items is determined, the ability to reproduce the results will be tested. This will be done separately for each scale. This property is referred to as invariance [33]. It implies that the parameter values determined for each item are unrelated to the characteristics of the examinees that are used to calibrate the items [33]. When the items are determined to be invariant, the same item parameters will be obtained regardless of the population using the instrument. If evidence of invariance of the parameter estimates is not provided, one or more of the assumptions underlying the item response model may not be met [33].

The property of invariance will be tested by separating the data into two groups and then comparing the item parameter estimates for these two groups. The two groups will be formed by splitting the data three ways: 1) random assignment, 2) age (young and old), and 3) gender. Multilog will then be re-run three times, once for each of the three sub-groups. Graphs will be plotted comparing the item difficulty and discrimination parameters of each of the three sub-groups: random assignment, young vs. old, and male vs. female. The plots should approximate a regression line that has an intercept of 0.0 and a slope of 1.0 [5]. The degree of scatter for these plots will be evaluated by the averaging the squared distance between the data points and the hypothetical regression line [5]. This will be done separately for both the ADL and Sports scales.

It is hypothesized that the property of invariance for parameter estimates will be demonstrated for both the FAAM ADL and Sports scales. This will be verified by the presence of three linear plots with a slope of 1.00 and intercept of 0.0. The degree of scattering for the plots should be minimal and attributable to sampling error [5].

3.1.3.3 Research Question 2: Which Items are Potentially Responsive Across All Ability Levels of Functional Status?

An important feature of an evaluative index is responsiveness. This means that when the underlying condition that is being measured changes the score on the instrument must change. In order for the FAAM to be responsive, the items must have a wide range of threshold difficulties and high levels of discrimination[33].

The potential responsiveness of the individual items will be assessed by examining item characteristic curves. Item characteristic curves will be constructed for each item as a plot between the probability of choosing a particular response and the ability level of the examinee. This will be done for each item. The item characteristic curves describe the relationship between

the trait an item assesses and the item response pattern across ability levels. Items that have four responses should have four distinct and separate curves[33]. Each of the four curves should have one peak, and together should span the spectrum of ability, as demonstrated in Figure 3.1. Items that do not have four distinct curves or do not span the complete spectrum of ability, as demonstrated in Figure 3.2 will be considered for elimination.

It is hypothesized that items will have curves similar to that represented in Figure 3.1.

3.1.3.4 Research Question 3: Can a Target Test Information Function be Produced that Maximizes Information Through a Broad Range of Physical Function?

The item information function describes the amount of information that each item provides as a function of ability[33]. Item information functions can be summed to provide a test information function. The more information an instrument provides, the more precise the instrument will be with less associated error of estimation. The target test information function for an evaluative instrument should provide a maximal amount of information across all ability ranges[33]. Therefore, the target test information function should be flat throughout the range of ability, as demonstrated in Figure 3.3. Test information function will be constructed separately for the ADL scale and the Sports scale. The amount of information each item supplies is given by Multilog. Items are chosen one at a time, starting with hard to fill areas. These hard to fill areas are located at the high and low ability extremes. After each item is chosen, the test information function is recalculated until it matches target test function.

It is hypothesized the test information function will match the target test information function demonstrated in Figure 3.3 for both the FAAM ADL and Sports scales.

3.1.4 Summary of methods

The intake data from approximately 1000 subjects contained in the CRS database will be used for these analyses. The object of these analyses is to produce a final instrument that is

unidimensional, potentially responsive across ability levels, and contributes appropriate information to the test. The specific research questions and methods of analysis are outlined in Figure 3.4.

3.2 Methods to Provide Evidence of the Usefulness of the Final Version of the FAAM

3.2.1 Research Questions

The research questions that will be attended to in this phase are:

4) What is the factoral structure of the FAAM ADL and Sports scales?
5) Do the FAAM ADL and Sports subscales demonstrate high levels of internal consistency?
6) Do the FAAM ADL and Sports subscales demonstrate adequate levels of test re-test reliability?
7) Are the FAAM ADL and Sports subscales responsive to changes in an individual's level of physical function?
8) Are the FAAM ADL and Sports subscales more responsive to changes in physical function than a general measure of health status?
9) Are the FAAM ADL and Sports subscales more responsive to changes in physical function than a global rating of self-perceived level of functioning?
10) What is the convergent and divergent evidence to support the interpretation of the final version of the FAAM ADL and Sports scales?

3.2.2 Procedures to Provide Evidence of the Usefulness of the Final Version of the FAAM

3.2.2.1 Subjects

Data for analysis will be extracted from the Centers for Rehab Services (CRS) outpatient facilities within the Southwestern Pennsylvania area. The data will be extracted from the CRS

clinical outcomes database and have been normally collected as part of routine care. Subjects who have been referred by a physician and have been scheduled for physical therapy evaluation of a lower leg disorder will be identified. Subjects will include anyone with a lower leg (below the knee) musculoskeletal disorder (bone, muscle, ligament, and/or joint). Subjects with pathologies that involve regions other than the lower leg, as well as subjects with coexisting pathologies will be excluded from the study. Identifying appropriate subjects will be accomplished based on the information from the ICD-9 codes. There will be no discrimination based on age, gender, race or religion. Two groups will be formed for comparison. One group will consist of subjects whose functional status should change. This group will consist of subjects who are involved in physical therapy treatment. The second group will consist of subjects whose functional status should be stable and not change. The stable group will consist of subjects who received physical therapy treatment more than one year ago. In order to be included, the data must meet the following criteria: 1) no more than three missing scores for the FAAM ADL scale and 2) no more than one missing score for the FAAM Sports scale. An estimation of sample size required for this analysis will be discussed in section 3.2.2.4.

3.2.2.2 Procedure for Data Collection

3.2.2.2.1 Subjects Expected to Change

Demographic data that will be extracted will include the subject's diagnosis (ICD-9) code, age, race, gender, date of injury, date of initial evaluation, mechanism of injury, date of onset of lower leg disorder, surgical procedure (if applicable), initial SF-36 score, initial FAAM ADL score, initial FAAM Sports score, pain levels (best, worst and present), and global rating. A global rating will be obtained by asking subjects to rate their current level of function from zero to 100. Zero will be defined as a complete loss of function and 100 will be defined as the

individual's level of function prior to the onset of the problem. After four weeks or discharge from physical therapy, which ever comes first, data will again be collected. This will include SF-36, FAAM ADL and Sports scores, pain levels and global rating. Added to this will be a rating scale that asks "Over the past four weeks, how would you rate your overall physical ability?". This question will have seven potential responses: "much worse, worse, slightly worse, no change, slightly improved, improved, much improved." This information is routinely collected as part of the normal and standard physical therapy evaluation and treatment across all of the CRS facilities. The data is collected and stored via an established TELE*form* (Cardiff Software Incorporated, San Marcos, CA) format. Confidentiality will be maintained, as each subject's social security number will be separated from the data so that individual responses will be unknown to the researchers.

3.2.2.2.2 Subjects Expected to Remain Stable

Potential subjects for the stable group will be obtained from the CRS database. The CRS database will be used to identify patients who were treated approximately one year ago (or longer) at one of the facilities with a lower leg musculoskeletal disorder (bone, muscle, ligament, and/or joint). These patients will be asked to participate in the study via mail. If they agree to participate, they will complete the following seven items: SF-36, FAAM ADL and Sports scales, pain scales (best, worst and present) and global rating. These items will be completed on two separate occasions, approximately four weeks apart. At the four-week period a rating scale that asks "Over the past four weeks, how would you rate your overall physical ability?" will be included. This question will have seven potential responses (ranging from-3 to +3): "much worse (-3), worse (-2), slightly worse (-1), no change (0), slightly improved (+1), improved (+2), much improved (+3)." This preceding information will be collected via mail. These subjects will also

have demographic information extracted from the CRS database. This will include each patient's, diagnosis (ICD-9) code, age, race, gender, date of injury, date of initial evaluation, mechanism of injury, date of onset of their lower leg disorder, and surgical procedure (if applicable). Once the data is obtained the social security number will be separated from the data so that each subject's individual responses will be unknown and confidentiality will be maintained.

The mailing process for this stable group has been described in detail by Dillman and consists of four separate mailings [22]. The first mailing will consist of the seven items, as outlined above, a back cover along with a cover letter. The back cover will not only allow the subject to make comments about the study or treatment they received but will also allow them to indicate if they do not wish to be involved in the study. A return business envelope fully addressed with first class postage will also be included so that they may return either the completed items or just the back cover indicating they do not wish to be involved in the study. The subjects who indicate they do not want to be involved in the study will not be contacted in the future. One week after the survey is sent a follow up postcard will be mailed out to the subjects. This is done in an effort to thank those who have returned the completed forms and to remind those who have not completed the forms to please do so if they desire. Three weeks later a third correspondence will be mailed out to those who have not responded. This third mailing will contain a new cover letter, the seven items, back cover, and the business envelope with paid return postage. Seven weeks after the first mailing, a fourth and final mailing will be sent out using certified mail. This will include a new cover letter, the seven items, a back cover and an envelope with paid return postage.

The physical therapists that have delivered treatment to the subjects will participate in the study by signing the individual cover letters. This will be done in an effort to respect the subjects' confidentially.

3.2.2.3 Data Analysis Plan

3.2.2.3.1 Research Question 4: What is the Factoral Structure of the FAAM ADL and Sports Subscales?

NOVAX (Poor Professor Software, Davis, CA) will be used to complete exploratory factor analysis for both the ADL scale and Sports scale. This exploratory factor analysis will be done using the data from the stable group and the group expected to undergo change separately. The pattern of correlations between items will be used to identify and extract factors. Eigenvalues will be generated. Each eigenvalue will be divided by the sum of the generated eigenvalues. This number will then be multiplied by 100 to yield the percentage of variance explained by each factor. The eigenvalues will be plotted against the factor number resulting in scree plot. A Varimax rotation will be done to evaluate how well the items fit this first factor.

The hypothesis for the above analysis are as follows: 1) there will be one single eigenvalue that is greater than one, 2) after factor rotation, all items will load on one factor, and 3) the resulting test of scree will indicate a one factor as demonstrated by only one point being at or to the left of the "elbow".

3.2.2.3.2 Research Question 5: Do the FAAM ADL and Sports Subscales Demonstrate a High Level of Internal Consistency?

Internal consistency assesses how well a subject's responses to each of the items relate to one another[39]. If the individual items measure the same domain, then each of the responses for the individual items should be highly related.

Cronbach's alpha will be used to assess internal consistency of the FAAM ADL and Sports scales. Cronbach's alpha calculates the correlation among items [39]. As a result of this process, the amount of error accounted for by inappropriate and/or inadequate sampling of the content domain can be obtained[39]. If the factoral structure of the change group and stable group are the same, then the groups can be combined for the analysis of internal consistency. The assessment of internal consistency will be done using the initial administration of the FAAM for both groups. If the factoral structure is different between the stable and change groups, then Cronbach's alpha will be calculated for both groups separately.

It is hypothesized that the Cronbach's coefficient alpha will be greater than 0.90 for the FAAM ADL and Sports scales.

3.2.2.3.3 Research Question 6: Do the FAAM ADL and Sports Subscales Demonstrate Adequate Levels of Test Re-test Reliability?

Test re-test reliability measures the stability of test scores and directly assesses measurement error[47]. Repeated administration of the same test should give the same score if there is no change in the construct that is being measured. Test re-test reliability of the FAAM ADL and Sports subscales will be assessed using an ICC (2,1)[76] with the stable group of subjects. Minimal detectable change will be obtained using the standard error of measure for the FAAM ADL and Sports scales. This will be accomplished by calculating the standard deviation (square root of 2 times the mean standard error) of the stable group. It is assumed that because the underlying condition should not change in the stable group, the level of physical function should not change, and concurrently, the scores of the FAAM ADL and Sports subscales should not change over the four-week period.

It is hypothesized that the test-re test reliability of the FAAM ADL and Sports scales will be good or excellent (>.90).

3.2.2.3.4 Research Question 7: Are the FAAM ADL and Sports Subscales Responsive to Change in the Functional Status of the Individual?

Responsiveness is the ability of the instrument to detect clinically significant changes in functional status when significant changes have occurred[81]. Three analyses will be done to assess responsiveness of the FAAM ADL and Sports scales: 1) two-way ANOVA with repeated measures, 2) Guyatt's responses index (GRI) and 3) construction of ROC curves.

The first of these three analyses will consist of comparing the scores of the change group to the stable group. Two-way analysis of variance with repeated measures of the initial and discharge FAAM ADL scale scores will be completed. This process will be repeated for the FAAM Sports score. It is hypothesized that the change in score in the change group will be greater than the change in score in the stable group, and therefore a significant interaction should be found with both scales. The a priori alpha for type I error will be set at 0.05.

The second analysis will involve calculating Gyatt's responsiveness index (GRI) for both the FAAM ADL and Sports subscales. GRI is calculated by dividing the average change in score of the change group by the standard deviation of the stable group[30,86]. A 95% confidence interval will be calculated to test if the change in score is significantly different than zero[86]. It hypothesized that the 95% confidence intervals will not contain zero for either the FAAM ADL or Sports Subscales.

The third analysis will consist of constructing ROC curves for the FAAM ADL and Sports subscales. The ROC curve will be constructed as a plot of sensitivity verse 1-specificity. The ROC curve will be constructed using all subjects, the stable group and change group. The

rating scale, obtained from all subjects, at the four-week period will be used to define those that changed from those that did not change. This rating scale is obtained by asking the subjects "How has your physical ability changed?" The seven potential responses (ranging from –3 to +3) are as follows: much worse (-3), worse (-2), slightly worse (-1), no change (0), slightly improved (+1), improved (+2), much improved (+3). The responses slightly worse, no change and slightly improved (-1 to +1) will be used for analysis as group that remains stable. Sensitivity, specificity and 1- specificity will be calculated and plotted for each one-point change in each of the two scales. Sensitivity is defined as the ability to correctly identify those who underwent a change in physical function[19]. Specificity is defined as the ability to correctly identify those who did not undergo a change in physical function[19]. The cut-point for each of the two scales will be chosen as the values when the sensitivity is closet to one and specificity is closest to zero, or the point at which the ROC curve is at the most upper left position on the graph. The ROC curve for a perfect measure would look like a right angle. The curve would ascend and trace the y-axis to 100 (perfect sensitivity) and then form a 90-degree angle and run parallel to the x-axis. A measure of no use would be represented by a diagonal line that would proceed from the lower left hand corner to the upper right hand corner[19].

A score change of one point might be highly sensitive, as most subjects who actually improve their physical function level will have a score that improves by one point. However, a one-point change might also have a low specificity, implying that many subjects who do not actually improve might have a change in score of one point due to measurement error. On the contrary, a score may change by 10 points. A change score of 10 points may be highly specific, as few subjects who do not actually improve in their physical function level would have a score

change to this degree. However, a 10-point change may have low sensitivity, implying many subjects who actually did improve their physical function level would not have a score change to this degree.

It is hypothesized that the ROC curves for the FAAM ADL and Sports Subscales will resemble a right angle with the angle occurring in the upper most left hand corner (high specificity and sensitivity) and not a diagonal line (low specificity and sensitivity).

3.2.2.3.5 Research Question 8: Are the final FAAM ADL and Sports Scales More Responsive to Changes in Physical Function than General Measure of Health Status?

Comparing the responsiveness of the FAAM ADL and Sports scales to the SF-36 physical function subscale and the SF-36 physical component summary score will be done using Gyatt's response index (GRI) and the area under the ROC curves.

GRI is calculated by dividing the change in score of the change group by the standard deviation of the stable group[30]. The GRI will be calculated for the FAAM ADL and Sports subscales, the physical function subscale of the SF-36 and the physical component summary score of the SF-36. To assess GRI values, four separate 95% confidence intervals will be constructed for the difference in GRI values between the: 1) FAAM ADL scale and SF-36 physical function subscale, 2) FAAM Sports scale and SF-36 physical function subscale, 3) FAAM ADL scale and physical component summary score of the SF-36, and 4) FAAM Sports scale and physical component summary score of the SF-36.

It is hypothesized that the FAAM ADL and Sports scales will be more responsive then the SF-36 physical function subscale and physical component summary scores. If the FAAM

ADL and Sports scales are more responsive then the SF-36 physical function subscale and physical component summary scores, then none of the confidence intervals will contain zero.

ROC curves will be constructed (using the process described below) for the FAAM ADL and Sports subscales, SF- 36 physical function sub-scale score and the SF-36 physical component summary score. A comparison of the area underneath each ROC curve will be made, as described by Hanley and McNeil[33], between: 1) FAAM ADL scale and SF-36 physical function subscale, 2) FAAM Sports scale and SF-36 physical function subscale, 3) FAAM ADL scale and physical component summary score of the SF-36, and 4) FAAM Sports scale and physical component summary score of the SF-36.

The ROC curve will be constructed as a plot of sensitivity versus 1-specificity. The ROC curve will be constructed using all subjects, the stable group and change group. The rating scale, obtained from all subjects, at the four-week period will be used to define those that changed from those that did not change. This rating scale is obtained by asking the subjects "How has your physical ability changed?" The seven potential responses (ranging from –3 to +3) are as follows: much worse (-3), worse (-2), slightly worse (-1), no change (0), slightly improved (+1), improved (+2), much improved (+3). The responses slightly worse, no change and slightly improved (-1 to +1) will not be used for analysis as these subjects might have had only a questionable change in physical ability. Sensitivity, specificity and 1- specificity will be calculated and plotted for each one-point change in each of the two scales. The cut points, when the sensitivity is closest to one and 1-specificity is closest to zero, will be determined for the SF-36 physical function subscale and the SF-36 physical component summary score.

The area under each of the four curves, standard errors associated with these areas and correlation coefficient between areas will be determined to allow for calculating the critical ratio

z as described by Hanley and McNeil[34]. To assess for difference in the areas under the ROC curves, critical ratio z will be calculated, using the formula described by Hanley and McNeil[34]. The differences between the following ROC curve areas will be assessed: 1) FAAI ADL scale and SF-36 physical function subscale, 2) FAAI Sports scale and SF-36 physical function subscale, 3) FAAI ADL scale and physical component summary score of the SF-36, and 4) FAAI Sports scale and physical component summary score of the SF-36. Critical ratio z will be compared to a z value 1.96.

It is hypothesized that the FAAM ADL and Sports subscales will be more responsive then the SF-36 physical function subscale and the physical component summary scores and therefore the critical z ratio will be greater or equal to 1.96 for all four comparisons.

3.2.2.3.6 Research Question 9: Are the Final FAAM ADL and Sports Subscales More Responsive to Changes in Physical Function than a Global Rating of Self-perceived Level of Functioning?

Comparing the responsiveness of the FAAM ADL and Sports scales to the global rating of self-perceived level of functioning will be done using Gyatt's response index (GRI) and the area under the ROC curves.

The GRI will be calculated for the FAAM ADL scale, FAAM Sports scale, and the global rating. To assess GRI values, two separate 95% confidence intervals will be constructed for the difference in GRI values between the: 1) FAAM ADL scale and the global rating, 2) FAAM Sports scale and global rating.

It is hypothesized that the FAAM ADL and Sports scales will be more responsive then the global rating of self-perceived level of functioning. If the FAAM ADL and Sports scales are

more responsive then the global rating of self-perceived level of functioning then none of the confidence intervals will contain zero.

ROC curves will be constructed (using the process described below) for the FAAM ADL and Sports scales and the global rating of self-perceived level of functioning. A comparison of the area underneath each ROC curve will be made, as described by Hanley and McNeil[34], between: 1) FAAM ADL scale and global rating, 2) FAAM Sports scale and global rating.

The ROC curve will be constructed as a plot of sensitivity verse 1-specificity. The ROC curve will be constructed using all subjects, the stable group and change group. The rating scale, obtained from all subjects, at the four-week period will be used to define those that changed from those that did not change. This rating scale is obtained by asking the subjects "How has your physical ability changed?" The seven potential responses (ranging from –3 to +3) are as follows: much worse (-3), worse (-2), slightly worse (-1), no change (0), slightly improved (+1), improved (+2), much improved (+3). The responses slightly worse, no change and slightly improved (-1 to +1) will be used for analysis as group that remains stable. Sensitivity, specificity and 1- specificity will be calculated for each one-point change in the FAAM ADL and Sports scales and the global rating. The cut-point for the global rating will be chosen as the values when the sensitivity is closest to one and 1-specificity is closest to zero, or the point at which the ROC curve is at the most upper left position on the graph.

The area under each of the three curves, standard errors associated with these areas and correlation coefficient between areas will be determined in methods described by Hanley and McNeil[34]. The ability to correctly identify subjects is represented by the area under the ROC curve. The value of the area ranges from 0.5, no accuracy, to 1.0, perfect accuracy, in identifying those who improved from those who did not improve. To assess for difference in the

areas under the ROC curves between: 1) FAAM ADL scale and global rating and 2) FAAM ADL scale and global rating, two critical ratio z values will be calculated, using the formula described by Hanley and McNeil[34]. The two critical z values will be compared to a z value 1.96.

It is hypothesized that the FAAM ADL and Sports scales will be more responsive then the global rating of self-perceived level of functioning and therefore the critical z ratio will be greater or equal to 1.96 for the two comparisons.

3.2.2.3.7 Research Question 10: What is the Convergent and Divergent Evidence to Support the Interpretation of the FAAM ADL and Sports Subscales?

Convergent and divergent evidence will be provided based on the association between the FAAM ADL and Sports subscales and concurrent measures of physical and emotional function using Pearson correlation coefficient. Convergent evidence is provided when a strong association is found between variables that measure the same or related constructs[65]. Convergent evidence will be examined by assessing the associations between the FAAM ADL and Sports scales with concurrent measures of physical function including the SF-36: 1) physical function and 2) physical components summary scores. Convergent evidence will also be examined by assessing the associations between the FAAM ADL and FAAM Sports scales to the subject's global rating. A global rating will be obtained by asking subjects to rate their current level of function from zero to 100. Zero will be defined as a complete loss of function and 100 will be defined as the individual's level of function prior to the onset of the problem.

It is hypothesized that there will be moderate to strong correlations ($r >= 0.6$) between FAAM ADL and Sports scores and concurrent measures of physical function.

Divergent evidence is provided when little or no association is identified between variables that measure distinctly different constructs[65]. Divergent evidence will be examined by

assessing the associations between the FAAM ADL and Sports scales to concurrent measures of emotional function including the SF-36: 1) mental health and 2) mental components summary scores.

It is hypothesized that there will be low correlations (r <=0.3) between FAAM ADL and Sports scores and concurrent measures of emotional function

Differences in the level of association between the variables that measure similar constructs and the variables that measure different constructs will be examined. It is hypothesized that the level of association between the FAAM ADL scale and concurrent measures of physical function will be greater than the association between the FAAM ADL scale and concurrent measures of emotional function. It is also hypothesized that the level of association between the FAAM Sports scale and concurrent measures of physical function will be greater than the association between the FAAM Sports scale and concurrent measures of emotional function.

Testing for difference in correlation coefficients between the FAAM ADL and Sports subscales to concurrent measures of physical and emotional function was done based on the equation by Meng et al[64]. These calculated values were compared to a critical t value. The a priori type I error rate will be set at .001 to account for the multiple comparisons. This was calculated by dividing 0.05 by the number of comparisons.

3.2.2.4 Sample Size

An a priori power analysis was done to estimate the sample size required to have an acceptable probability of rejecting a false null hypothesis. Sample size estimation was done taking into account the two-way repeated measure analysis of variance and also assessing for

differences between two sample correlations. SPSS (SPSS inc., Chicago, IL.) was used to estimate sample size based on our known values for alpha, sample variance, effect size and desired power level.

A two-way repeated measure analysis of variance will be used to assess responsiveness of the FAAM ADL and Sports subscales. This will be accomplished by comparing the change in scores in the stable group (those who received physical therapy greater than one year ago) to the change in scores in the group receiving treatment over a four week period. Sample size estimates for an independent t-test of the pre- to post- change scores can be used to approximate sample size under these conditions.

A random group of 30 subjects in the CRS data base was used for preliminary estimates of mean score changes of the FAAM ADL and Sports scales. The mean difference between the FAAM ADL scores was 18.1 with a standard deviation of 25.1. The mean difference of the FAAM Sports score was 19.1 with a standard deviation of 31.8. Effect size is calculated as the difference between means divided by their standard deviation. Given these values the effect size for the FAAM ADL scale was 0.72 and the effect size for the FAAM Sports scale was 0.60. The stable group, those not receiving treatment, is expected to have an effect size of zero on both of the scales, as no change in their status should occur. Alpha was set at 0.05 for a one tailed test. To obtain a desired power of 80%, a sample size of 50, 25 subjects in each group, will be required to test the FAAM ADL scale. A sample size of 70, 35 in each group, will be required to test the FAAM Sports scale.

Assessing the correlation between the FAAM ADL and Sports scores with the SF- 36 physical function sub-scale score will provide convergent evidence of validity. Assessing the correlation between the FAAM ADL and Sports scores with the SF-36 mental health sub-scale

score will provide divergent evidence. It is hypothesized that the association between the FAAM scores and the concurrent measures of physical function will be greater than the association between the FAAM scores and the concurrent measures of emotional function. A t-test using the equation by Meng etal.[64] will be used to test for differences in the strength of association.

We had a test sample of 1027 subjects that was used to measure the correlation between the FAAM scales and the SF-36 sub-scales. The correlations between the FAAM ADL score and the physical function and mental health sub-scale scores were 0.70 and 0.30 respectively. The correlations between the FAAI Sports score and the physical function and mental health sub-scale scores were 0.63 and 0.20 respectively.

Sixteen comparisons will be required in the analysis for convergent and divergent evidence. These 16 comparisons will include comparing the FAAM scales with four components of the SF-36, physical function sub-scale, physical component summary, mental health and mental component summary sub-scale scores. In an effort to be conservative with the analysis, 16 was rounded to 20. To account for the number of comparisons, 0.05 was divided by 20 to calculate an alpha level that would reduce the chance of a type I error (probability of false rejection of the null hypothesis). Therefore, alpha level will be set at 0.001 using a one tailed test. The desired power level will be chosen to be 80%. Two hundred subjects are required to detect a significant difference between the correlation values of 0.70 and 0.30 for the FAAM ADL Scale. Two hundred and twenty subjects will be required to detect a significant difference between the correlation values of 0.63 and 0.20 for the FAAM Sports Scale. Because there is no reason to believe the association between the FAAM scores and the SF-36 sub-scale scores will be different between the stable group and the group receiving treatment the groups can be combined.

In summary a total of 220 subjects, 110 subjects in stable group and 110 subjects in the change group will be recruited for this study.

3.2.3 Summary of Methods

To provide evidence for the usefulness of the final version of the FAAM ADL and Sports scales analyses will be done to assess the reliability, validity and responsiveness. Two groups of subjects, a stable group and a change group will be used. The stable group will consist of 110 subjects whom completed physical therapy greater than one year ago. Initial and four-week information will be obtained using mailed responses. The change group will consist of 110 subjects currently receiving therapy. Initial and four-week information will be obtained from the CRS database. The information to be used in the analysis will be initial and four-week FAAM, initial and four-week SF-36 scores, initial and four-week global ratings and a question regarding the change in status at the four week period. A summary of the research questions, samples to be used, methods of analyses and the hypotheses of the analyses are summarized in Figure 3.5.

4 Results

4.1 Results to Produce the Final Version of the FAAM ADL and Sports Scales

4.1.1 Description of Subjects

The subjects consisted of patients who received treatment at CRS facilities and completed an intake FAAM. The initial potential sample consisted of 1027 subjects. With respect to the ADL subscale, there were 914 subjects who had three or fewer missing responses. Of these 914 subjects, 659 (64.2%) had no missing responses, 148 (14.4%) had one missing response, 76

(7.4%) had two missing responses, and 31(3.0%) had three missing responses. For the Sports subscale, there were 796 subjects who had zero or one missing response. Of these 796 subjects, 659 (64.2%) had no missing responses and 137 (13.3%) had one missing response.

The average age of the total sample was 42.0 years (SD 17.39 median 42.81 range 8-83years). Age could not be determined for nine (0.88%) individuals. Six hundred twenty nine (61.2%) were females and 391(38.1%) were males. Gender was not recorded for seven (0.68%) individuals. Mechanism of injury was related to activities of daily living for 203 (19.8%) subjects; work for 98 (9.5%); sports for 72 (7.0%); post surgery for 61 (5.9%), and motor vehicle accident for 15 (1.5%). The mechanism of injury was not recorded for 322 (36.2%) individuals. Duration of the symptoms was defined as the time from the onset of symptoms to the initiation of treatment at CRS. The duration symptoms averaged 3.7 months (SD 8.55 months, median 1.45 months, range 1 day to 7.88 years). Duration of symptoms could not be determined for 85 (8.3%) individuals.

Diagnosis was determined using ICD-9 codes. The ICD-9 codes for individual diagnoses were organized into six categories: ankle joint pathology, sprains and strains, heel pathologies, fractures, forefoot pathology and non-specific leg pain. Diagnoses were as follows: 193 (18%) subjects had ankle joint pathology, 321 (31.3%) had sprains and strains, 113 (11.9%) had heel pathology, 151 (14.7%) had fractures, 37 (3.6%) had forefoot pathology and 87 (8.5%) had nonspecific leg pain. Diagnosis was not accurately recorded for 125 (12.2%) subjects.

4.1.2 Descriptive Statistics for Items on the Preliminary Version of the FAAM

The number of missing responses for individual items on the ADL subscale was relatively uniform across items, ranging from 4.3% to 11% missing. The average number of missing responses was 6.9%. Items 10 (squatting), 12 (coming up on your toes) and 22 (recreational activities) had the highest percentage of missing responses, with 11%, 9.9%, and

9% missing values respectively. All items had median score values of two or three. The two exceptions to this were items 11 (sleeping) and 19 (personal care). The median values for these items were four. The degree of kurtosis was assessed by dividing kurtosis by the standard error of kurtosis. This value was then compared to a critical value of 1.96. All items had significant kurtosis values with the exception of 8 (Walking on uneven ground), 23 (General level of pain) and 24 (Pain at Rest). The items with significant kurtosis values were platykurtic with the exception of 11 (Sleeping), and 19 Personal care), which were leptokurtic. A test of significance was also done to assess the degree of skewness. All items were significantly skewed with the exception of items 5 (Walking down hills), 8 (Walking on uneven ground), 21 (Heavy work) and 25 (Pain during normal activity). The items with significant skewness were negatively skewed with the exception of items 8 (Walking on uneven ground), 16 (Walking 15 minutes or greater), and 22 (Recreational activities) which were positively skewed.

The items on the Sports subscale had a noticeably higher percentage of missing responses, ranging from 13% to 20%. The average number of missing responses was 16.5%. The median scores for the Sports subscale items were also noticeably lower then the median scores of the ADL subscale items. The median scores of the items on the Sports subscale were either zero or one. The higher number of missing responses may be related to individuals having not attempted these more challenging activities. All items had significant kurtosis values with the exception of item 8 (Ability to participate in your desired sports as long as you would like). Items 1 (Running), 2 (Jumping), and 3 (Landing) were leptokurtic, while items 4 (Starting and stopping quickly), 5 (Cutting and lateral movements), 6 (Low impact activities), and 7 (Ability to perform activity with your normal technique) were platykurtic. All items had significant skewness values with the exception of item 7 (Ability to perform activity with your normal

technique). All items were positively skewed. The positive skewness and lower scores of the items reflect the more challenging content of the items on the Sports subscale.

Descriptive statistics for FAAM ADL and Sports subscales can be found in Table 4.1 and Table 4.2.

4.1.3 Analysis of Missing Data

The subject's responses for each item were categorized as present, missing or not applicable. A response that was able to be scored was assigned the category "present". Chi-squared analyses were used to evaluate the relationship between the presence of missing data and gender, age (younger or older) and diagnostic category.

4.1.3.1 Missing Data versus Gender

There was no relationship between gender and missing data on the ADL subscale. Items 5 (Cutting/lateral movements), 7 (Ability to perform activity with your normal technique) and 8 (Ability to participate in your desired sport as long as you would like) on the Sports subscale demonstrated a significant relationship between gender and the presence of missing data. These results demonstrate that males were more likely to respond to items 5, 7 and 8 compared to females. There was no significant relationship between gender and presence of missing data for the remaining items on the Sports subscale.

4.1.3.2 Missing Data versus Age

The sample was dichotomized by the mean age. A significant relationship was found between age and presence of missing data for items 3 (Walking on even ground without shoes), 6 (Going up stairs), 8 (Walking on uneven ground), 10 (Squatting), 19 (Personal care), and 22

(Recreational activities) on the ADL subscale. Younger individuals were more likely to respond to these items than older individuals. There was no significant relationship between the presence of missing data and age for the remaining ADL subscale items.

A significant relationship was found between age and the presence of missing data for all of the Sports subscale items ($p<.001$). Younger individuals were more likely to respond to items on the Sports subscale than older individuals.

4.1.3.3 Missing Data versus Diagnosis

A significant relationship was found between the diagnostic category and the presence of missing data for items 10 (Squatting), 17 (Home responsibilities), 18 (Activities of daily living), 19 (Personal care), 20 (Light to moderate work), 24 (Pain at rest), and 25 (Pain during your normal activity) on the ADL subscale. For item 10, diagnosis category 1 (ankle pathology) was more likely to have missing data. For items 17, 18, 19, 24 and 25, diagnosis category 2 (sprains and strains) was more likely to have a response. For item 20, diagnosis category 4 (fractures) was more likely to have not applicable marked.

A significant relationship was found between diagnostic category and the presence of missing data for items 1 (Running), 2 (Jumping), 3 (Landing), 4 (Starting and stopping quickly), 5 (Cutting/lateral movements) and 6 (Low impact activities) on the Sports subscale. For these item diagnosis category 2 (sprains and strains) was more likely to have a response.

4.1.4 Evaluating the Assumptions of Item Response Theory

PRELIS (Scientific Software International, Chicago, IL) was used to perform exploratory factor analysis to assess for unidimensionality. Specifically, eigenvalues, scree plots, inter-item

correlations and item to principal component correlations were evaluated. PRELIS requires the use of complete data with no missing responses. Therefore, 659 (64.2%) subjects were used to evaluate both the ADL and Sports subscales

The polychoric correlation matrix for items on the interim 26-item ADL subscale is reported in Table 4.3 items had consistently high inter-item correlations except for items 23 through 26. These items did not strongly correlate to items 1 through 22 but did correlate strongly to one another. The content of items 23 through 26 were as follows: item 23- general level of pain, item 24- pain at rest, item 25- pain during normal activity, and item 26- pain first thing in the morning.

Factor analysis Principal component analysis of the 26 items on the interim FAAM ADL subscale indicated that the items loaded on two factors. These two factors accounted for 75.0% of the variance. Factor one accounted for 66.24% of the variance and had an eigenvalue of 17.22. Factor two accounted for 8.81% of the variance and had an eigenvalue of 2.29. The third factor accounted 3.35% of the variance and had an eigenvalue of 0.87. The scree plot can be found in Figure 4.1. The factor loadings of each item on the first two principal components are reported in Table 4.4. Items 23 through 26 had higher factor loadings on the second principal component than the first principal component.

Review of the scree plot and the eigenvalue greater than one rule, lead to the conclusion that the 26 item ADL subscale best fit a two factor model. This was consistent with the inter-item correlations and factor loadings. Items 23 through 26 were therefore considered for elimination from the ADL subscale so that it would conform to a one-factor model. A one-factor model is necessary to meet the assumption of unidimensionality for item response theory.

The factor analysis was repeated with items 23 through 26 omitted. When these four items were omitted the number of subjects without a missing response increased to 678 (66%). The polychoric correlation matrix for the 22-item interim ADL subscale is reported in Table 4.5. All of the 22 items had strong inter-item correlation values. Item 11 (sleeping) had the lowest inter-item correlation values.

The 22 items on the interim ADL subscale loaded on one factor. This one factor accounted for 74.09% of the variance and had an eigenvalue of 16.30. The second factor accounted for 3.88% of the variance and had an eigenvalue of 0.85. The scree plot for these 22 items can be found in Figure 4.2. The factor loading of each item to the first principal component is found in Table 4.6. Item 11 (Sleeping) had the lowest factor loading.

After examining the scree plot and eigenvlaue greater than one rule, a one-factor model was determined to be appropriate for the 22- item ADL subscale. This is consistent with the assessment of the inter-item correlations, and factor loadings to the first principal component.

The polychoric correlation matrix for the items on the Sports subscale is reported in Table 4.7. The eight items on the interim Sports subscale loaded on one factor. This one factor accounted for 86.33% of the variance and had an eigenvalue of 6.91. The second factor accounted for 4.24% of the variance and had an eigenvalue of 0.34. The scree plot can be found in Figure 4.3. The factor loadings to the first principal component are found in Table 4.8.

After examining the scree plot and eigenvlaue greater than one rule, a one factor model was found to be appropriate for the eight item Sports subscale. This is consistent with the inter-item correlations and factor loadings to the principal component.

4.1.5 Assessment of Model Fit

Multilog was used to calibrate items separately for the ADL and Sports subscales. Subjects were included if they responded to at least 19 of the 22 items on the ADL subscale and

if they responded to at least 7 of the 8 items on the Sports subscale. Therefore, 914 (90.0%) subjects were included in the analysis of the ADL subscale and 796 (77.5%) subjects were included in the analysis of the Sports subscale. Item calibration was done using both the one- and two-parameter models. Fit of the one and two parameter models was compared using the difference in the negative twice the log likelihood statistic.

The negative twice log likelihood statistics for the 22-item ADL subscale were -29670.6 and -28777.5 for the one- and two- parameter models respectively. The observed difference of 893.1 is greater than critical value of 32.67 for p=0.05 and 21 degrees of freedom. This indicates the additional parameters estimated in the two-parameter model contribute significantly to the fit of the model to the data. The parameter estimates for the 22-item ADL subscale using the two-parameter graded response model are reported in Table 4.9.

The negative twice log likelihood statistics for Sports subscale were -977.6 and -860.8 for the one- and two- parameter models respectively. The observed difference of 116.8 is greater than the critical value of 14.07 for p=0.05 and 7 degrees of freedom. This indicates that the additional parameters estimated in the two-parameter model contributed significantly to the fit of the model to the data. The parameter estimates for the eight item Sports subscale using the two-parameter graded response model are reported in Table 4.10.

4.1.6 Assessment of Parameter Invariance

The property of invariance refers to the ability to reproduce results with respect to model fit and parameter estimates[33]. Two methods were used to assess the property of invariance. The first method involved separating the data into two groups and comparing the results of these two groups, with respect to the item difficulty and discrimination parameters. The second method involved comparing the negative twice log likelihood statistic between a restricted and unrestricted model. The unrestricted model was the model where the item parameter estimates

were determined separately for each subgroup. The restricted model was the model where the item parameter estimates were constrained and set equal for each subgroup. In a restricted model there were only one set of parameter estimates for each item from the entire sample. In the unrestricted model with two subgroups there were two sets of parameter estimates for each item, one for each subgroup.

The data was split into two subgroups randomly, by age (young versus old) and by gender. If the property of invariance is met then the plots of item difficulty and discrimination parameters estimated separately for each subgroup should approximate a theoretical regression line with a slope of 1.0 and an intercept of zero[5]. Also, the negative twice log likelihood statistic between the restricted model, where the item parameter estimates are set equal for each subgroup, and unrestricted model, where the item parameter estimates are determined separately for each subgroup, should not be significantly different. Multilog was used to generate item parameter estimates for each subgroup as well as the negative twice log likelihood statistics for each of the restricted and unrestricted models.

4.1.6.1 Assessment of Parameter Invariance in the Randomly Generated Sample

4.1.6.1.1 Assessment of Parameter Invariance for the Randomly Generated Sample for the ADL Subscale

Subjects were randomly split into two subgroups. This was accomplished by using SPSS to create a random variable from a binomial distribution using a probability of 0.50. For analysis of the 22-item ADL subscale there were 470 subjects in group one and 453 subjects in group two. There were no significant differences between these two groups with respect to age, gender, duration of symptoms or subscale scores.

The item discrimination parameters for the 22-item ADL subscale for the two randomly split subgroups were plotted in Figure 4.4. A best-fit regression line through these points had a slope of 0.84 and an intercept of 0.41. The mean squared distance between each point and the theoretical regression line was 0.24. The mean squared distance between each point and the actual regression line was 0.22. The Pearson correlation between the paired item discrimination parameters was 0.94

The item difficulty parameters were plotted in Figure 4.5. The resulting regression line had a slope of 0.93 and an intercept of 0.069. The mean squared distance between each point and the theoretical regression line was 0.028. The mean squared distance between each point and the actual regression line was 0.022. The correlation between the paired item difficulty parameters was 0.99.

The negative twice log likelihood values for the 22 -item ADL subscale for the restricted and unrestricted models were -28777.5 and -37897.2. The difference between the two models was 9117.7. This value was greater than the critical value of 67.51. This indicated that some of the items demonstrate differential item functioning.

4.1.6.1.2 Assessment of Parameter Invariance for the Randomly Generated Sample for the Sports Subscale

For analysis of the Sports subscale there were 398 subjects in each group. There were no significant differences between these two grouped with respect to age, gender, duration of symptoms or subscale scores.

The item discrimination parameters for the Sports subscale for the two randomly split subgroups are plotted in Figure 4.6. The regression line had a slope of 0.30 and an intercept of

2.1. The mean squared distance between each point and the theoretical regression line was 0.99. The mean squared distance between each point and the actual regression line was 4.1. The correlation between the paired item discrimination parameters was 0.39.

The item difficulty parameters are plotted in Figure 4.7. The regression line had a slope of 0.94 and an intercept of -.070. The mean squared distance between each point and the theoretical regression line was 0.39. The mean squared distance between each point and the actual regression line was 0.088. The correlation between the paired item difficulty parameters was 0.92.

The negative twice log likelihood values for the Sports subscale for the restricted and unrestricted models were –860.8 and –1916.8 respectively. The difference between the two models was 1056.0. This value was greater than the critical value of 17.71. This indicated that some of the items demonstrate differential item functioning.

4.1.6.1.3 Summary of the Assessment of the Property of Invariance for the Randomly Generated Samples

The item discrimination plots for the ADL subscale revealed items 14 (Walking 5 minutes or less) and 15 (Walking approximately 10 minutes) were furthest from the theoretical and actual regression lines. The item difficulty plots for the ADL subscale revealed item 11 (Sleeping) had the largest deviation from the theoretical and actual regression lines. Deviations from the theoretical regression line imply that the paired item parameters are not equal, while deviations from the actual regression line imply that the paired item parameters are not consistent with the other items. The difference in the negative twice the log likelihood statistics between the restricted and unrestricted models indicates that some of the items demonstrate differential item functioning. This may be explained by the functioning of items 11, 14 and 15.

The item discrimination plots for the Sports subscale revealed that items 3 (Landing) and 8 (Ability to participate in your desired sport as long as you would like) had large deviations from the theoretical and actual regression lines. The actual regression line poorly approximated the theoretical regression line. The actual regression line for the pairs of item difficulty parameters was below the theoretical regression line. A difference between the intercepts is not as serious as a difference in slope. A difference in intercept indicates that the two sets of estimates deviate by a constant that is equal at all levels of ability[5]. The difference in the negative twice the log likelihood statistic between the restricted and unrestricted models indicated that some of the items demonstrated differential item functioning. This may be explained by the discrimination parameters for items 3 and 8. This may also be explained by the constant difference that the items had at all ability levels with the paired item difficulty plots**.**

4.1.6.2 Assessment of Parameter Invariance by Age

4.1.6.2.1 Assessment of Parameter Invariance by Age for the ADL Subscale

The sample was dichotomized into younger and older subgroups according to mean age of the sample. For analysis of the 22- item ADL subscale there were 450 subjects (mean age 27.01 SD 9.63) in the young group and 470 subjects (mean age 55.16 SD 9.85) in the old group. There were no significant differences between these two groups with respect to gender and duration of symptoms. The younger group had significantly higher ADL subscale scores than the older group (p=0.001).

The item discrimination parameters for the 22- item ADL subscale for the young and old subgroups are plotted in Figure 4.8. The best-fit regression line through these points had a slope of 0.89 and an intercept of 0.37. The mean squared distance between each point and the

theoretical regression line was 0.098. The mean squared distance between each point and the actual regression line was 0.068. The correlation between the paired item discrimination parameters was 0.86.

The pairs of item difficulty values are plotted in Figure 4.9. The regression line had a slope of 1.01 and an intercept of –0.15. The mean squared distance between each point and the theoretical regression line was 0.052. The mean squared distance between each point and the actual regression line was 0.021. The correlation between the paired item difficulty parameters was 0.98.

The negative twice log likelihood values for the 22- item ADL subscale for the restricted and unrestricted models were–28777.5 and –27722.7 respectively. The difference between the two models was 1054.8. This value was greater than the critical value of 67.51. This indicates some of the items demonstrated differential item functioning.

4.1.6.2.2 Assessment of Parameter Invariance by Age for the Sports Subscale

For analysis of the Sports subscale there were 413 subjects (mean age 26.66 years SD 9.66) in the young group and 381 subjects (mean age 54.42 years SD 9.60) in the old group. There were no significant differences between these two groups with respect to gender and duration of symptoms. There was a significant difference with respect to the Sports subscale score with the younger group scoring significantly higher than the older group (p=0.001).

The item discrimination parameters estimated for the young and old sub-sample are plotted in Figure 4.10. The regression line had a slope of 1.20 and an intercept of –0.30. The mean squared distance between each point and the theoretical regression line was 0.22. The mean squared distance between each point and the actual regression line was 0.39. The correlation between the paired item discrimination parameters was 0.92.

The item difficulty parameters for the sub-samples defined by age are plotted in Figure 4.11. The regression line had a slope of 0.96 and an intercept of 1.46. The mean squared distance between each point and the theoretical regression line was 1.1. The mean squared distance between each point and the actual regression line was 0.014. The correlation between the paired item difficulty parameters was 0.99.

The negative twice log likelihood values for the restricted and unrestricted models were –860.8 and –656.8 respectively. The difference between the two models was 204.0. This value was greater than the critical value of 55.76. This indicates that some of the items demonstrate differential item functioning.

4.1.6.2.3 Summary of the Assessment of the Property of Invariance by Age

The item discrimination plots for the ADL subscale revealed items 6 (Going up stairs), 7 (Going down stairs), 20 (Light to moderate work) and 22 (Recreational activities) had the largest deviations from the theoretical and actual regression lines The item difficulty plots for the ADL subscale found that all item pairs, with the exception of items 11 (Sleeping) and 19 (Personal care), closely approximated the theoretical and actual regression lines However, the calculated difference in the negative twice the log likelihood statistic between the restricted and unrestricted models indicated that the item parameters for the ADL subscale may not be invariant with respect to age**.** This may be explained by the difference in discrimination parameters noted with items 6,7,20 and 22 and the difference in the difficulty parameters of item 11 and 19.

The item discrimination plots for the Sports subscale revealed item 3 (Landing) had the largest deviation from the theoretical and actual regression lines. The regression line item difficulty parameters plot for the Sports subscale was above the theoretical regression line indicating intercept indicates that the two sets of estimates deviate by a constant that is equal at

all levels of ability, The difference in the negative twice the log likelihood statistic between the restricted and unrestricted models indicated differential item functioning. This discrepancy may be explained by the difference in discrimination parameters with item 3 and by the constant differences in the item difficulty parameters that the items had across ability levels.

4.1.6.3 Assessment of Parameter Invariance by Gender

4.1.6.3.1 Assessment of Parameter Invariance by Gender for the ADL Subscale

For analysis of the 22-item ADL subscale, the groups consisted of 564 females and 354 males. There were no significant differences between these two groups with respect to age, duration of symptoms or ADL subscale score

The item discrimination parameters for the 22-item ADL subscale by gender are plotted in Figure 4.12. The regression line had a slope of 1.13 and an intercept of 0.13. The mean squared distance between each point and the theoretical regression line was 0.24. The mean squared distance between each point and the actual regression line was 0.22. The correlation between the item discrimination parameters estimated for each sample was 0.85

The item difficulty parameters estimated separately for males and females are plotted in Figure 4.13. The regression line had a slope of 0.75 and an intercept of -0.0051. The mean squared distance between each point and the theoretical regression line was 0.15. The mean squared distance between each point and the actual regression line was 0.11. The correlation between the item difficulty parameters estimated separately for males and females was 0.96.

The negative twice log likelihood values for the 22 item ADL subscale for the restricted and unrestricted models were –28777.5 and -27648.6 respectively. The difference between the two models was 1128.9. This value was greater than the critical value of 67.51. This indicates some of the items demonstrate differential item functioning.

4.1.6.3.2 Assessment of Parameter Invariance by Gender for the Sports Subscale

For analysis of the Sports subscale, the groups consisted of 463 females and 327 males. There were no significant differences between these two groups with respect to age, duration of symptoms and subscale score.

The item discrimination parameters that were estimated separately for males and females are plotted in Figure 4.14. The regression line had a slope of 1.34 and an intercept of 1.46. The mean squared distance between each point and the theoretical regression line was 0.28. The mean squared distance between each point and the actual regression line was 2.36. The correlation between the item discrimination parameters was 0.82 .

The item difficulty parameters are plotted in Figure 4.15. The regression line had a slope of 0.90 and an intercept of 0.44. The mean squared distance between each point and the theoretical line was 0.096. The mean squared distance between each point and the actual regression line was 0.018. The correlation between the difficulty parameters for the Sports subscale estimated separately for males and females was 0.99.

The negative twice log likelihood values for the Sports subscale for the restricted and unrestricted models were –860.8 and –627.4 respectively. The difference between the two models was 233.4. This value was greater than the critical value of 55.76. This indicates some of the items demonstrate differential item functioning.

4.1.6.3.3 Summary of the Assessment of Invariance by Gender

The item discrimination plots for the ADL subscale revealed that items 2 (Walking on even ground) and 4 (Walking up hills) had largest deviations from the theoretical and actual regression lines. The item difficulty plots for the ADL subscale revealed all points, except for items 11 (Squatting) and 19 (Personal care) closely approximated the theoretical and actual

regression lines. The calculated difference in the negative twice the log likelihood statistic between the restricted and unrestricted models indicates some of the items demonstrate differential item functioning. This may be due to the difference in the discrimination parameters of items 2 and 4.This difference in item functioning may also be due to the difference in the difficulty parameters noted by items 11 and 19.

The item discrimination plots for the Sports subscale revealed that item 6 (Low impact activities) had the largest deviation from the theoretical and actual regression lines The item difficulty plots for the Sports subscale revealed that all points were above the theoretical regression line The difference in the negative twice the log likelihood statistic between the restricted and unrestricted models indicates some of the items demonstrate differential item functioning. This may be due to the difference in difficulty parameters noted by item 6. This may also be explained by the constant differences between all of the item difficulty parameters for males and females.

4.1.7 Assessing the Potential Responsiveness Across Ability Levels for each Item

Item characteristic curves were constructed using the item difficulty and item discrimination parameters produced from Multilog. Item characteristic curves that were plotted for the 22 items on the interim ADL subscale can be found in Appendix C. Item characteristic curves for an item that has five response categories should have five distinct and separate curves. Each curve should have one peak and together the curves should span the spectrum of ability. This is represented by the item characteristic curve in Figure 4.16. All items, except item 11 (sleeping) and 19 (personal care), had well fitting item characteristic curves. The item characteristic curves for item 11 and 19 are displayed in Figure 4.17 and Figure 4.18. These items provided the most information at the lower end of the ability spectrum. Item 11 functioned as an item with two potential responses as most of the subjects reported having either no

difficulty or slight difficulty sleeping as a result of their foot or ankle problem. Item 19 functioned as an item with three potential responses as most subjects reported having moderate difficulty, slight difficulty or no difficulty with personal care as a result of their foot or ankle problem. The discrimination parameters for items 11 and 19 were 0.77 and 1.28 respectively. The other 20 items had discrimination parameters that averaged 2.13 (range 1.26 to 3.27)

Item characteristic curves were also plotted for eight items on the interim Sports subscale. These curves can be found in Appendix D. All eight had well fitting curves similar to that displayed in Figure 4.19.

4.1.8 Assessing the Target Test Information Function

The amount of information each item provided at nine ability intervals, ranging from -2.0 to 2.0 was provided by Multilog for the two-parameter model. The item information function for each item at each ability level was summed to produce the test information function. The test information is related to precision of measurement and as the information function increases the precision of measurement increases. A target test information function for an evaluative instrument should provide maximal information across all ability ranges[33]. The test information function for the 22-item ADL subscale can be found in Figure 4.20. Most information was supplied at the lower end of ability. Items 11 (Sleeping) and 19 (Personal care) were noted to give the most information at the lower end of ability. Item 11 was deleted first because it functioned as an item with only two potential responses as demonstrated by the item characteristic curve. It also had the lowest inter-item correlation, item to score correlation, factor loading to the first principle component and the lowest discriminative parameter. The test information function, after deleting item 11 did not substantially change (Figure 4.21). Item 19 was than deleted. The test information function was recalculated and a decrease in information

was noted throughout the range of ability (Figure 4.22). Therefore, item 19 was retained to maintain the instrument's precision.

The test information function for the eight-item Sports subscale can be found in Figure 4.23. The test information function provided most information in the higher ability ranges.

4.1.9 Selection of Items for the Final Version

Items for the final version of the FAAM were selected based on the item character curves, inter-item correlations, item to score correlations, and factor loadings to the first principal component. The purpose of the FAAM is to assess self-reported physical performance. Therefore, items 23 to 26, which were related to pain and could be eliminated without jeopardizing content of the ADL subscale. These items had lower inter-item and item to total score correlations.

Item 11 (Sleeping) was also deleted. This was done because of its item character curve, low inter-item and item to score correlations, and the factor loading to the first principal component. Therefore, the final version of the FAAM ADL subscale has 21 items. The final version of the FAAM can be found in Appendix E.

All eight items on the FAAM Sports scale were appropriate based on the item character curves, inter-item and item to total score correlations and the factor loadings to the first principal component. The final version of the Sports Subscale can be found in Appendix E.

4.2 Evidence of the Usefulness of the Final Version of the FAAM

4.2.1 Description of the Subjects

4.2.1.1 Description of the Subjects in the Group Expected to Change

The sample expected to change was obtained from individuals receiving physical therapy treatment between July 2002 and January 2003. These subjects were participating in treatment at one of the Centers for Rehab Services (CRS) clinics in the southwestern Pennsylvania area. Subjects were included in the study if they had received treatment for a lower leg (below the knee) musculoskeletal disorder (bone, muscle, ligament, and/or joint) for greater than four weeks. Information contained within the CRS database was collected. This database included demographics, mechanism of injury, duration of symptoms, diagnosis, SF-36 scores, global rating of function and perceived change in status. Subjects with coexisting pathologies were excluded. Subjects were also excluded if they had not received at least four weeks of physical therapy treatment. Two hundred and sixty six potential subjects that met these criteria were identified. Evaluation and discharge information could be obtained from 164 (62%) of these individuals. These 164 individuals were included in the group expected to change. Subjects were included in the analysis if they responded to at least 19 of the 21 Activities of Daily Living subscale items and if they responded to at least 7 of the 8 Sports subscale items. For the Activities of Daily Living Subscale 112 (68.3%) subjects had no missing responses, 31(18.9%) had one missing response, and eight (4.9%) subjects had two missing responses on the initial ADL subscale. The remaining 13 (7.9%) subjects had three or more missing responses. For the Sports subscale, 106 (64.6%) subjects had no missing responses and 24 (14.6%) had one missing response. The remaining 34 (20.8%) subjects had two or more missing responses. Thus the

analyses for the Activities of Daily Living subscale included 151 patients and the analyses for the Sports subscale included 130 patients.

The average age of the group expected to change was 41.2 (SD16.3 median 45.6 range 9-75). There was no information regarding age for two (1.2%) individuals. There were 97 (59.1%) females and 67 (40.9%) males. Mechanism of injury was related to the following: activities of daily living for 52 (31.7%) patients, sports for 40 (24.4%), work for 13 (7.9%), post surgery for four (2.4%), motor vehicle accident for three (1.8%) and "other" for 29 (17.7%) subjects. There was no information regarding mechanism of injury for 23 (14%) subjects. Duration of symptoms was defined as the time from onset of symptoms to the initiation of treatment at CRS. The duration of symptoms averaged 4.7 months (SD 0.17 yr, median 2.1 months, range1day-3.8yr). The duration of symptoms could not be determined for 26 (15.9%) individuals.

ICD-9 codes and when possible the physical therapy chart were used to determine diagnosis. The ICD-9 codes were organized into seven categories: joint/limb pain, sprains/strains, fractures, plantar fasciitis, bunion, Achilles tendon rupture and "other". For the group expected change 55 (33.5%) subjects had joint/limb pain, 47(28.7%) had sprains/strains, 28 (17.1%) had fractures, 22 (13.4%) had plantar fasciitis, three (1.8%) had bunion, and two (1.2%) had Achilles tendon rupture. Four (2.4%) subjects had a diagnosis that did not fit these categories and were put in the "other" category. There was no information regarding diagnosis for three (1.8%) subjects.

The FAAM ADL and Sports subscales were scored separately. Responses to individual items were assigned a value of value zero (unable to do) to four (no difficulty). The response "not applicable" was recorded when the activity was limited by something other than the lower leg disorder. The scores for the items responses were summed. Items that did not have a

response or were marked “not applicable” could not be scored. When an item could not be scored the average item score for that individual was calculated. This value was substituted for the missing response when summing items to obtain a total score. This total score was then divided by 84 for the ADL subscale, and by 32 for the Sports subscale and then multiplied by 100 to transform the score to a 0 to 100 scale such that a higher score represents a higher level of function.

The average score for initial FAAM ADL subscale for the subjects in the group expected to change was 58.0 (SD 24.8 median 59.5 range 5-98). The distribution was approximately normal (skewness -0.26, SE skewness 0.20, kurtosis -0.72, SE kurtosis 0.39). The average initial FAAM Sports subscale for the subjects in the group expected to change was 25.2 (SD 26.7 median 15.0 range 0-94). The distribution was positively skewed (skewness 0.86, SE skewness 0.21) but was normal with respect to kurtosis (kurtosis -0.42, SE kurtosis 0.42)

The final FAAM ADL and Sports subscales were administered approximately four weeks after initiation of physical therapy. The average time between completing the initial and final surveys was 32.3 days (SD 12.1 median 28 range 23-106). The average final FAAM ADL subscale score for the subjects in the group expected to change was 74.9 (SD 20.0 median 77.5 range 13-100). The distribution was approximately normal (skewness -0.71, SE skewness 0.19, kurtosis -0.33, SE kurtosis 0.38). The average final FAAM Sports subscale score for the subjects in the group expected to change was 43.9 (SD 30.0 median 45.1 range 0-100). The distribution was approximately normally skewed (skewness 0.098, SE skewness 0.21) but platykurtic (kurtosis -1.147, SE kurtosis .410).

The perceived global rating of change reported by patients in the group expected to change revealed 75 patients (45.7%) perceived themselves to be much better, 42 (25.6%) were

somewhat better, 24 (14.6%) were slightly better, five (3.0%) were unchanged and two (1.2%) were slightly worse. Sixteen (9.8%) subjects did not provide a perceived global rating of change. No subjects reported being somewhat worse or much worse. It was hypothesized that subjects undergoing physical therapy would report being much better or somewhat better from the initial to final administration of the FAAM. One-hundred and seventeen (71.3%) subjects met this hypothesis and are considered to have improved over the 4 week duration of physical therapy. The average change scores for the ADL and Sports subscale scores, those that were improved were respectively. 21.01 (SD 19.60 median 15.48 range -19 to 81), 19.34 (SD 21.94 median 15.00 range -45 to 40) and 20.97 (SD 25.71 median 14.29 range -34 to 97) and 28.30 (SD 28.31 median 25.00 range -20 to 90) respectively. Thirty-one (18.9%) subjects reported they were slightly better, unchanged or slightly worse. The average change ADL and Sports subscale scores for these 31 subject were 3.44 (SD 12.03 median -0.60 range -18 to 40) and 2.73 (SD 21.31 median 0.50 range -45 to 40).

An analysis was done to compare those who were included in group expected to change to those who could not be included because of missing FAAM information. Chi-squared analysis revealed that there was no difference with respect to gender ($p=0.63$), diagnosis ($p=0.37$), or mechanism of injury ($p=0.40$). Independent t-test revealed that there were no differences with respect to age ($p=0.17$). Mann-Whitney test revealed that there was a difference with respect to duration of symptoms ($p=0.028$). Subjects that had complete information had a longer duration of symptoms compared to those that did not have complete information.

4.2.1.2 Description of the Subjects in the Group Expected to Remain Stable

Potential subjects for the group expected to remain stable were obtained from the CRS database. This data base was used to identify patients who were treated at least one year ago for

a lower leg musculoskeletal disorder. One hundred and eighty potential subjects who were treated between the time period between January 3, 2000 and October 3, 2001 were identified. The mailing process outlined by Dillman[21] was used to obtain initial and final FAAM, SF-36, global rating and perceived change in status.

Initial baseline survey information was obtained from 79 (42%) subjects. Of the potential subjects that initial information was not obtained from, 14 (16.9%) had the wrong address, 11 (5.9%) refused to participate, three (2%) noted having a different problem, and 72 (42%) did not respond.

Of the 79 subjects from whom initial baseline information was obtained, 61 (77.2%) had no missing responses, 10 (12.7%) had one missing response, and three (3.8%) had two responses on the initial ADL subscale. The remaining five (6.4%) subjects had three or more missing responses. For the Sports subscale, 61 (77.2%) subjects had no missing responses and nine (11.4%) had one missing response. The remaining nine (11.4%) subjects had two or more missing responses.

Analysis of demographic information of the group expected to remain stable revealed the average age was 45.2 (SD15.0 median 44.5 range19-86). Forty-seven (59.5%) subjects were female and 32 (40.5%) subjects were male. Mechanism of injury was related to activities of daily living for 13 (16.5%) subjects, sports for 22 (27.8%) subjects, and post surgery for one (1.3%) subject. There was no information regarding mechanism of injury for 43 (54.4%) individuals. The duration of symptoms, defined as the time from the onset of symptoms to the initiation of treatment at CRS averaged 4.0 months (SD 0.33yr, median 43 days, range 2 days-2.2 yr). There was no information regarding duration of symptoms for 22 (27.8%) individuals.

ICD-9 codes were used to determine diagnosis. Thirty-nine (49.4%) subjects had joint/limb pain, 24 (30.4%) had sprains/strains, 5 (6.3%) had fractures, 5 (6.3%) had plantar fasciitis, 1 (1.3%) had a bunion, and 1 (1.3%) was in the category "other".

The average score for initial FAAM ADL subscale for the group expected to remain stable was 91.5 (SD 13.6 median 97.5 range 37-100). The distribution was negatively skewed (skewness -2.5, SE skewness 0.28) and leptokurtic (kurtosis 6.1, SE kurtosis 0.56). The average initial FAAM Sports subscale score was 78.6 (SD 23.8 median 88.4 range 13-100). The distribution was approximately normal (skewness -0.97, SE skewness 0.29, kurtosis 0.027, SE kurtosis 0.55).

The final administration of the FAAM, SF-36 and global rating of function was mailed to the participating subjects. The time period between the initial and follow up surveys averaged 65.6 days (SD 19.8 median 67.0 range 31-101). Of the 79 subjects from whom initial baseline information was obtained, the final survey information was obtained from 71(90.0%). The reason that eight subjects did not return the final survey is unknown. For the final FAAM ADL subscale, 53 (74.6%) subjects had no missing responses, 6 (8.5%) had 1 missing response, and 4 (5.6%) had two missing responses. The remaining eight (11.3%) subjects had three or more missing responses. For the initial Sports subscale, 46 (64.8%) subjects had no missing responses and nine (12.7%) had one missing response. The remaining 16 (22.5%) subjects had two or more missing responses.

The average score for final FAAM ADL subscale for the group expected to remain stable was 92.6 (SD 13.2 median 97.6 range 27-100). The distribution was negatively skewed (skewness -2.95, SE skewness 0.302) and leptokurtic (kurtosis 10.37, SE kurtosis 0.60). The

average final FAAM Sports subscale score was 81.9 (SD 23.3 median 93.8 range 13-100). The distribution was negatively skewed (skewness -1.21 SE skewness 0.32) but had normal kurtosis (kurtosis 0.44, SE kurtosis 0.64).

Fifty-two (65.8%) of the subjects reported their perceived change in status to be unchanged. Three (3.8%) subjects reported they were much better, three (3.8%) were somewhat better, two (2.5%) were slightly better, one (1.3%) was slightly worse, and one (1.3%) was much worse. This information was no reported by 13 (16.5%) subjects. It was hypothesized that subjects who had not received physical therapy within the last year would not change from the initial to final administration of the FAAM. Fifty-two (65.8%) subjects met this hypothesis. Fourteen (17.7%) subjects did not meet this hypothesis. These 14 subjects had changes in ADL and Sports subscales scores that were -0.36 (SD 9.2 median 0.00 range -13 to 24) and -3.98 (SD 16.1 median 0.00 range -28 to 31) respectively.

An analysis was done to determine if there were differences between those in the group expected to remain stable that did and did not respond to the surveys. Chi-squared analysis revealed no differences in gender (p=0.67), however there was a significant difference in diagnosis (p=.005) between those that returned a survey and those that did not. The individuals that did not return the surveys more frequently had a diagnosis of joint/limb pain than those that returned the surveys. Additionally, subjects that returned the surveys more frequently had a diagnosis of sprain/strain. Independent t-test revealed that there was no difference (p=0.46) with respect to age. Mann-Whitney test revealed no difference (p=0.84) with respect to duration of symptoms between those that did and did not respond to the surveys.

When comparing the group expected to change to the group expected to remain stable, there were no differences in gender (p=0.77), diagnoses (p=0.63) or age (p=.10), however there

was a significant duration in symptoms (p=0.004). Subjects in the group expected to remain stable had a longer duration of symptoms prior to treatment than subjects in the group expected to change.

4.2.2 Factoral Structure of the FAAM ADL and Sports subscale

PRELIS (Scientific Software International, Chicago, IL) was used to perform exploratory factor analysis to assess dimensionality of the FAAM ADL and Sports subscales. Specifically, eigenvalues, scree plots, inter-item correlations and factor loadings were evaluated. Data from the initial FAAM ADL and Sports subscales were used for these analyses. The group expected to change and the group expected to remain stable were initially analyzed separately. PRELIS requires use of complete data with no missing responses. Therefore, responses from 112 (68.3%) of the 164 in the group expected to change were used to evaluate the ADL subscale and 106 (6.4.6%) were used to evaluate the Sports subscale. In the group expected to remain stable, responses from 61 (77.2%) of the 79 subjects were used to evaluate the ADL and Sports subscales.

Principal component analysis found the items on the ADL subscale loaded on one factor in the group expected to change. Factor one accounted for 80.46% of the variance and had an eigenvalue of 16.90. The second factor accounted for 4.39% of the variance and had an eigenvalue of 0.92. The scree plot can be found in Figure 4.24. The factor loadings for each item to the first principal component are reported in Table 4.11 for the group expected to change.

In the group expected to remain stable the principal component analysis revealed two factors. Factor one accounted for 78.37% of the variance and had an eigenvalue of 16.46. The second factor accounted for 12.28% of the variance and had an eigenvalue of 2.58. The scree plot can be found in Figure 4.25. The factor loadings of each item to the first principal component are reported in Table 4.12 for the group expected to remain stable. With the

exception of item 18, all of the items loaded on the first factor. Item 18 had a low factor loading on the first factor but a high factor loading to the second factor. No attempt was made to perform a factor analysis on the combined sample because the factorial structure was different between for each group.

The principal component analysis of the Sport subscale in the group expected to change revealed all items loaded on one factor. Factor one accounted for 86.7% of the variance and had an eigenvalue of 6.94. The second factor accounted for 4.13% of the variance and had an eigenvalue of 0.33. The scree plot can be found in Figure 4.26. The principal component analysis of the Sports subscale in the group expected to remain stable revealed all items loaded on one factor. Factor one accounted for 86.42% of the variance and had an eigenvalue of 6.91. The second factor accounted for 4.79% of the variance and had an eigenvalue of 0.38. The scree plot can be found in Figure 4.27. The factorial structure of the Sports subscale was similar in both groups, therefore we performed a factor analysis on the combined samples. The effective sample size for this analysis was 167 (83.5%) of the total 200 subjects. The resulting principal component analysis found the items loaded on one factor. Factor one accounted for 93.48% of the variance and had an eigenvalue of 7.48. The second factor accounted for 1.85% of the variance and had an eigenvalue of 0.15. The scree plot can be found in Figure 4.28. The factor loadings of each item to the first principal component for the group expected to change, group expected to remain stable and the combined groups can be found in Table 4.13.

4.2.3 Internal Consistency of the FAAM ADL and Sports subscale.

The factorial structure of the FAAM ADL subscale was not the same in the group expected to change as it was in the group that was expected to remain stable, therefore, the assessment of internal consistency was done for each sample separately. In the group expected to change coefficient alpha was 0.98 with a standard error of measurement (SEM) of 3.5 and

95% confidence interval of plus or minus 6.9 points. In the group expected to remain stable coefficient alpha was 0.96 with a SEM of 2.7 and a 95% confidence interval of plus or minus 5.3. The item to total score correlations and coefficient alpha with each item deleted are presented in Table 4.14 for the group expected change and the group expect to remain stable.

The result of factor analysis for the FAAM Sports subscale found a one-factor model was appropriate for the group expected to change as well as for the group expected to remain stable. Therefore, the assessment of internal consistency was done with the combined samples yielding a coefficient alpha of 0.98. The SEM was 5.1 with a 95% confidence interval of pus or minus 10.0.

4.2.4 Test Re-test Reliability of the Final FAAM ADL and Sports subscale

Test Re-test reliability measures the stability of test scores over time. The initial and final FAAM ADL and Sports subscale scores for the group expected to remain stable were used for this analysis. Test re-test reliability was evaluated using an Intra-class Correlation Coefficient (ICC) (formula 2,1)[76]. The resulting ICC was 0.89 for the ADL subscale and 0.87 for the Sports subscale. The SEM calculated using the test re-test reliability coefficient and the standard deviation of the group expected to remain stable was 2.1 with a corresponding 95% confidence interval of plus or minus 4.0 points for the ADL subscale. The minimal detectable change for the ADL subscale using a 95% confidence interval was plus or minus 5.7 points. For the Sports subscale the SEM was 4.5 with a corresponding 95% confidence interval of plus or minus 8.8 points. The minimal detectable change for the Sports subscale using a 95 % confidence was plus or minus 12.5.

It was hypothesized that subjects in the group expected to remain stable would report their perceived change in status to be unchanged. Fifty-two (65.8%) subjects met this hypothesis. Fourteen (17.7%) subjects did not meet this hypothesis. These subjects were examined to

determine if their status may have changed. This determination was based two measures external to the FAAM, change in perceived status and change in global rating for activities of daily living. One subject reported being somewhat better and had an improvement in global rating of 10 points. Two subjects reported being much worse and slightly worse and had changes in global ratings of -21 and -35 respectively. When these subjects were deleted from the analysis, the ICC for test re-test reliability for the ADL and Sports subscales increased to 0.92. Test re-test reliability for the Sports subscale remained the same. For the ADL subscale the SEM decreased to 1.5 with a 95% confidence interval of plus or minus 2.9 points. The minimal detectable change using a 95% confidence interval decreased to plus or minus 4.1. For the Sports subscale the SEM decreased to 3.4 with a 95% confidence interval of plus or minus 6.7 points. The minimal detectable change using a 95% confidence interval decreased to plus or minus 9.5.

4.2.5 Responsiveness of the FAAM ADL and Sports Subscales to Change in Functional Status

The ability of an instrument to detect a clinically significant change in functional status when a change has occurred is termed responsiveness. Responsiveness was assessed using three analyses: 1) two-way analysis ANOVA with repeated measures, 2) Guyatt's response index, and 3) the construction of receiver operating characteristic (ROC) curves. The following hypotheses were tested: 1) the change in the ADL and Sports subscales scores for the group expected to change would be greater than the change scores for the group expected to remain stable, 2) the 95% confidence interval of the Gyuatt's response index for the ADL and Sports subscales will not contain zero and, 3) the 95% confidence intervals for area under the ROC curve for both the ADL and Sports subscales will be greater than 0.5.

The two way repeated measures ANOVA was done to compare the change from the initial to final FAAM ADL and Sports subscale scores, between the group expected to change and the group expected to remain stable. The average difference between the initial and final ADL subscale scores in the group expect to change was 17.1 (SD 19.88 median 12.59 range -25 to 81). The average difference between the initial and final ADL subscale scores in the group expected to remain stable was -0.2 (SD 6.21 median 0.00 range -19 to 24). The initial and final FAAM ADL scores for the two groups are presented in Figure 4.29. The analysis of variance summary table is present in Table 4.15. The group by time interaction was significant ($F(1,202)= 42.562$ $p<.001$). This supports the hypothesis that the difference between the initial and final ADL scores in the group expected to change would be greater than the difference between the initial and final ADL subscale scores in the group expected to remain stable.

The two-way repeated measures ANOVA was also done with the Sports subscale scores. The average change score on the Sports subscale in the group expect to change was 17.2 (SD 24.8 median 11.6 range -34 to 97). The average change in the group expected to remain stable was 0.0 (SD 12.3 median 0.00 range -28 to 33). The initial and final FAAM Sports scores for the two groups are presented in Figure 4.30. The analysis of variance summary table is present in Table 4.16. The group by time interaction was significant ($F(1,165)=22.466$ $p<.010$). This supports the hypothesis that the difference between the initial and final Sports scores in the group expected to change would be greater than the difference between initial and final Sports subscale scores in the group expected to remain stable.

Gyatt's responsiveness index (GRI) was calculated for the ADL subscale by dividing the average change in score of the group expected to change, 17.1 points, by the standard deviation

of the change scores for stable group, 6.2. The resulting GRI for the ADL subscale was 2.75 with a 95% confidence interval ranging from 2.02 to 3.48. Since this confidence interval did not contain zero, the GRI was significantly different than zero.

GRI was also calculated for the Sports subscale. The average change score for the group expected to change was 17.2 points. The standard deviation of the change score for the group expected to remain stable was 12.3. The resulting GRI for the Sports subscale was 1.40 with a 95% confidence interval ranging from 0.93 to 1.86. This confidence interval did not contain zero, indicating the GRI for the Sports subscale is significantly different from zero.

A ROC curve, which is a plot of sensitivity versus 1 – specificity for each one-point change on the ADL subscale was constructed using subjects in both the group expected to change and the group expected to remain stable. Sensitivity and specificity was calculated for each one-point change on the ADL subscale. The criterion measure to calculate sensitivity and specificity was whether the subject was in the group expected to change or the group expected to remain stable. The resulting ROC is presented in Figure 4.31. The area under the curve was 0.80 with a 95% confidence interval ranging from 0.84 to 0.74. The results indicate the area under the curve was greater than 0.5. The change score that best differentiates an individual in the group expected to change from an individual in the group expected to remain stable is that which lies closest to the upper left hand corner of the ROC curve. The change score that was closest to the upper left hand corner of the graph was 4 points which had sensitivity and specificity of change values of 0.76 and 0.88 respectively. This change score had the highest sensitivity and specificity values. The positive and negative likelihood ratios were 6.39 and 0.27 respectively for this 4-point change score.

A ROC curve was also constructed for the Sports subscale Figure 4.32. The area under the curve was 0.76 with a 95% confidence interval ranging from 0.84 to 0.68. The results indicate the area under the curve was greater than 0.5.The change score that was closest to the upper left hand corner of the ROC curve was 5 points with sensitivity and specificity of 0.63 and 0.81 respectively. This change score had the highest sensitivity and specificity values. The positive and negative likelihood ratios were 3.36 and 0.45 respectively.

The above ROC analyses established the change score that best differentiated between a subject in the group expected to change and the group expected to remain stable. To determine the best change score to differentiate a patient that perceives him/herself to be improved after four weeks of physical therapy from a patient that does not perceive themselves to be improved, we also constructed a ROC for subjects within the group receiving treatment. For this analysis the group expected to change was dichotomized based on how subjects perceived their change in status between the initial and final administration of the FAAM. When this study was planned, we hypothesized that because subjects in the group expected to change were receiving treatment, they would perceive their condition to be either somewhat or much improved between the initial and final administration of the FAAM. One hundred and seventeen (73.3%) subjects in the group expected to change described their perceived change in status as much better (75 or 45.7%) or somewhat better (42 or 25.6%). Sixteen (9.8%) subjects did not provide a global rating of change. The remaining 31 (18.9%) subjects described their change in functional status to be slightly better (24 or 14.6%), unchanged (five or 3.0%) or slightly worse (2 or 1.2%). No subjects reported being somewhat worse or much worse. We used the global rating of change to dichotomize the group that was expected to change into those that were improved (n=117) and those that were not improve (n=31). The ROC analyses for the ADL and Sports scales were

repeated using this dichotomy (i.e. improved vs. not improved after 4 weeks of physical therapy) as the criterion measures to calculate sensitivity and specificity for multiple change scores of the ADL and Sports scale.

The ROC for the ADL subscale is presented in Figure 4.33. The area under the ROC curve was 0.80 SE with a 95% confidence interval ranging from 0.89 to 0.71. The results indicate the area under the curve was greater than 0.5. The change score that best distinguished between a patient that perceive him/herself to be improved from a patient that does not perceive him/herself to be improved was 8 points which had a sensitivity and specificity of 0.77 and 0.75 respectively. The positive and negative likelihood ratios were 3.09 and 0.30 for this 8-point change in the ADL subscale score.

The ROC curve for the Sports subscale is presented in Figure 4.34. The area under the ROC curve was 0.72 with a 95% confidence interval ranging from 0.78 to 0.66. The results indicate the area under the curve was greater than 0.5. The change score the best distinguished an improved from an unimproved patient after 4 weeks of physical therapy was 9 points which had a sensitivity and specificity of 0.64 and 0.75 respectively. The positive and negative likelihood ratios were 2.57 and 0.48 respectively for this 9-point change score.

4.2.6 Responsiveness to Change in Functional Status of FAAM ADL and Sports Subscales Compared to a General Measure of Functional Status

Responsiveness of the FAAM ADL and Sports subscales were compared to the physical function subscale and physical component summary score of the SF-36 using Guyatt's Responsiveness Index and ROC Curves. It was hypothesized that the FAAM ADL and Sports subscale would be more responsive to a change in function than the general measures of function.

Guyatt's responsiveness indices were calculated for the FAAM ADL and Sports subscales and physical function and physical component summary scores for the SF-36. The GRI and 95% confidence intervals for these instruments were as follows: ADL subscale 2.75(2.02, 3.48), Sports subscale 1.40(0.95,1.86), physical function subscale 1.77(1.15,2.39), and physical component summary score 1.12(0.67,1.58) Differences and 95% confidence intervals for the differences between GRIs for the ADL and Sports subscales and physical function subscale and the physical component summary score of the SF-36 were calculated as follows: 1) ADL subscale and physical function subscale 0.98 (0.27,1.69), 2) ADL subscale and physical component summary score 1.63 (1.02,2,24), 3) Sports subscale and physical function subscale -0.38 (-1.00,0.25) and 4) Sports subscale and physical component summary score 0.27 (-.022,0.77). These results indicate that the ADL subscale was more responsive than either the physical function subscale or physical components summary score of the SF-36, however the Sports subscale was not found to be more responsive than the physical function subscale or physical components summary score of the SF-36.

ROC curves were constructed to compare the FAAM ADL and Sports subscales to the physical function and physical component summary scores of the SF-36. Four ROC curves were constructed to compare the areas under the curves between the: 1) ADL subscale and physical function subscale (Figure 4.35) 2) ADL subscale and physical component summary score (Figure 4.36) 3) Sports subscale and physical function subscale (Figure 4.37) and 4) FAAM Sports subscale and SF-36 physical component summary score (Figure 4.38) The criterion measure to calculate sensitivity and specificity was whether the subject was in the group expected to change or the group expected to remain stable. The correlation between the areas under the curves and the critical ratio z were calculated as outlined by Haney and McNeil[33]. The

results are reported in Tables 4.17 and 4.18 for the ADL and Sports subscales respectively. To determine the significance of the difference between the areas under the curve, the critical z ratio was compared to 1.96. The results indicated that the ADL subscale was more responsive than either the physical function subscale or physical components summary score of the SF-36 and the Sports subscale was more responsive than the Sf-36 physical components summary score. However the Sports subscale was not more responsive than the physical function subscale of the SF-36.

4.2.7 Responsiveness to Change in Functional Status of FAAM ADL and Sports Subscales Compared to a Global Rating of Self Perceived Level of Function

The responsiveness of the FAAM ADL and Sports subscales were compared to a global rating of self-perceived level of function for activities of daily living and sports respectively. The global rating was scored from zero to 100 with zero being complete loss of function and 100 being the level of function before the onset of the patient's lower leg problem. Guyatt's responsiveness index and ROC Curves were used to compare the responsiveness of the FAAM and global rating of function. It was hypothesized that the FAAM ADL subscale would be more responsive than the global rating of function during activities of daily living and the FAAM Sports subscale would be more responsive than the global rating of function during sports.

Guyatt's responsiveness indices were calculated for the FAAM ADL and Sports subscales and activities of daily living (ADL) and Sports global ratings. The GRI values and their 95% confidence intervals are as follows: ADL subscale 2.75(2.02, 3.48), ADL global rating 1.94 (1.37, 2.50), Sports subscale 1.40 (0.93,1.86), and sports global rating 1.61 (2.06,1.16). Differences and 95% confidence interval for the differences between GRIs for the ADL and Sports subscales and ADL and Sports global ratings are as follows: 1) ADL subscale to ADL

Global rating 0.82 (.04,1.59), 2) Sports subscale to sports global rating 0.03 (-0.65,0.22). These results indicate that the ADL subscale was more responsive than the ADL global rating, however the Sports subscale was not found to be more responsive than the Sports global rating

ROC curves were constructed to compare the FAAM ADL and Sports subscales to the ADL and sports global rating. Two ROC curves were constructed to compare the area under the curved between the: ADL subscale and ADL global rating and Sports subscale and sports global rating (see Figures 4.39 and 4.40). The criterion measure to calculate sensitivity and specificity was whether the subject was in the group expected to change or the group expected to remain stable. The correlation between the areas and the critical ratio z was calculated as outlined by Haney and McNeil[34]. The results are reported in Table 4.19. To determine significance of the difference between the areas under the curve, critical z ratio was compared to1.96. The results indicated that the ADL was not more responsive than ADL global rating and the Sports subscale was not more responsive than sports global rating.

4.2.8 Convergent and Divergent Evidence to Support the Interpretation of the FAAM ADL and Sports Subscales

Convergent evidence was assessed by examining the relationship between the FAAM ADL and Sports subscales to concurrent measures of physical function. The concurrent measures of physical function included the physical function and the physical components summary scores of the SF-36 as well as the global rating of function. The global rating of function was provided separately for activities of daily and sports. The global ratings of function ranged from zero to 100 with zero being complete loss of function and 100 being the subject's prior level of function before the onset of his/her the lower leg problem. Divergent evidence was assessed by examining the relationship between the FAAM ADL and Sports subscales and

concurrent measures of emotional function. The concurrent measures of emotional function included the mental health and mental components summary scores of the SF-36.

It was hypothesized that there would be: 1) moderate to strong correlations (r>=0.6) between the FAAM ADL and Sports subscale scores and concurrent measures of physical function and 2) low correlations (r<=0.3) between the FAAM ADL and Sports subscales and concurrent measures of emotional function. Additionally it was hypothesized that the correlations between the FAAM ADL and Sports subscale scores and concurrent measures of physical function would be significantly greater than the correlations between the FAAM ADL and Sports subscales and concurrent measures of emotional function.

Testing for difference in correlation coefficients between the FAAM ADL and Sports subscales to concurrent measures of physical and emotional function was done based on the equation by Meng et al[64]. These calculated values were compared to a critical t value of 3.34 for alpha=.001 at 200 degrees of freedom. The initial scores from subjects were used in the analysis. The correlation coefficients between the FAAM ADL and Sports subscales and concurrent measures of physical and emotional function are presented in Table 4.20. All calculated values were significantly greater than the critical value at p<.001. These results indicated that the correlation of the FAAM ADL and Sports Subscales to the concurrent measures of function were significantly greater than their correlation to concurrent measures of emotional status.

The correlations between the FAAM ADL and Sports subscale scores and concurrent measures of physical function ranged from 0.84 to 0.78. The correlations between the FAAM ADL and Sports subscales to concurrent measures of emotional function ranged between 0.12 and -0.02. Additionally, as expected, the correlations between the FAAM ADL and Sports

subscale scores and concurrent measures of physical function were significantly greater than the correlations between the FAAM ADL and Sports subscales and concurrent measures of emotional function.

5 Summary and Conclusions

The overall purpose of this project was to develop a reliable and responsive self-reported health related quality of life instrument specific to those with foot and ankle disorders. The instrument consisted of two scales. One scale contained items related to activities of daily living, the other contained items related to sports activities. The development of the FAAM was accomplished in two stages. The purpose of stage one was to develop an instrument that contained items that had appropriate psychometric properties. Conclusions and summary regarding the results of this phase are discussed in section 5.1. The purpose of stage two was to assess the instrument's reliability, validity and responsiveness. Conclusions and summary regarding the results of this phase are discussed in section 5.2.

5.1 Summary of the Results to Produce the Final Version of the FAAM and ADL and Sports subscales

The fit of the graded response model was tested. This included assessing the assumptions for item response theory, assessing model fit, and assessing the property of parameter invariance. The assumptions of Item Response Theory were met with respect to guessing for a correct response is not a factor, the administration of the test is not under time constraints and unidimensionality. Also, the Graded Response Model fit the data. However, the property of parameter invariance was not met. This means that the results of this study can only be generalized to those patients that are similar to those patients included in this study. The subjects in phase one of this study had a mean age of 42.0years (SD 17.4, median 42.8, range 8 to 83).

There were 61.2% female and 38.1% male subjects. Mechanism of injury was related to activities of daily living for 19.8%; work for 9.5%; sports for 7.0%; post surgery 5.9%; and motor vehicle accident 1.5%. The duration of symptoms averaged 3.7 months (SD 8.6 months, median 1.5 months, range 1 day to 7.9 years). Diagnoses were as follows: 18% ankle joint pathology, 31.3% sprains and strains, 11.9% heel pathology, 14.7% fractures, 3.6% forefoot pathology, and 8.5% nonspecific leg pain.

5.1.1 Assessing the Assumptions of Item Response Theory

The initial step consisted of assessing whether or not the assumptions for use item response theory were met. For the graded response model, the assumptions include: 1) unidimensionality, 2) local dependence, 3) administration of the test is not under time constraints, and 4) guessing the correct answer is not a factor. There were no time constraints for completing the FAAM and since there is no correct answer per se, guessing the correct answer is not possible. Unidimensionality implies that the instrument measures a single domain. Local dependence pertains to the examinee's abilities, and implies that only one latent trait influences the examinee's response to the items. When the items on the instrument are represented by one factor, the assumptions of unidimensionality and local dependence will be met. Exploratory factor analyses were performed to determine dimensionality of the FAAM ADL and Sports scales. The results of the exploratory factor analysis for the preliminary version of the ADL subscale revealed two factors. Factor one accounted for 66.24 % of the variance with an eigenvalue of 17.22. Factor two accounted for 8.81% of the variance and had an eigenvalue of 2.29. The scree plot, which is a plot of the eigenvalues against the number of factors revealed two prominent factors. These results imply two factors underlie the 26-item ADL subscale. Items one through 22 had high factor loadings on the first factor while items 23 through 26 had high factor loadings on the second factor. The content of items 23 through 26 was related to pain

as follows: item 23 – general level of pain, item 24 – pain at rest, item 25- pain during normal activity and 26- pain first thing in the morning. Items 23 through 26 were thought to represent a unique factor and these items were omitted.

Factor analysis after omitting items 23 to 26 revealed a single factor that accounted for 74.09% of the variance and had an eigenvalue of 16.30. The second factor accounted for 3.88% of the variance and had an eigenvalue of 0.85. The scree plot for the 22 items also revealed a one-factor model fit the data. All 22 items had high factor loadings with the first factor.

After examining the results of the factor analysis for the FAAM ADL subscale, it was concluded that a one-factor model was appropriate for the 22-item ADL subscale. Items 1 through 22 were primarily related to functional activities. Items 23 through 26, which were related to pain, were deleted from further consideration.

The results of the exploratory factor analysis for the Sports subscale were consistent with a one-factor model. This one factor account for 86.33% of the variance and had an eigenvalue of 6.91. The second factor accounted for 4.24% of the variance and had and eigenvalue of 0.34. The scree plot supported this conclusion. All eight items had high factor loadings on the first factor.

5.1.2 Assessment of Model Fit

There are two of models appropriate for an instrument when guessing is not an option and the possible responses can be ordered to represent varying degrees of ability. These two models are the partial credit and the graded response models. The partial credit model is a one-parameter model that is able to differentiate items based on item difficulty. The graded response model is two a parameter model, which includes both difficulty and discrimination parameters. The choice between the partial credit and graded response model is determined statistically by determining if the addition of the extra parameter in the graded response model added

significantly more information than the one- parameter partial credit model. If the addition of the extra parameters do not contribute to the fit of the model, then a one-parameter model is more appropriate. A likelihood ratio is obtained for each model. The difference between the likelihood ratios is tested statistically, using chi-squared values, to assess if the addition of extra parameters contributes to fit of the model. It was hypothesized that the two-parameter graded response model would fit the data better than the one-parameter partial credit model.

The results revealed that the two-parameter graded response model fit the data better than the one-parameter partial credit model for both the ADL and Sports subscales. The observed differences in the twice the negative log likelihood statistic between the one and two parameter model were 893.1 for the ADL subscale and 116.8 for the Sports subscale. Both of these values were greater than the critical value for the chi-squared test. In summary, the graded response model fit the data better than the partial credit model for both the ADL and Sports subscales. Therefore, the items on the FAAM ADL and Sports subscales can be differentiated based on how difficult the item is as well as how well the item differentiates between examinees based on the examinee's ability.

5.1.3 Assessment of Parameter Invariance

The property of invariance refers to the ability to reproduce the results with respect to model fit and parameter estimates[32]. Parameter invariance also implies that the difficulty and discrimination parameters for each item are unrelated to the characteristics of the examinee. Examinees with similar abilities, irrespective of their age, gender or diagnosis, will respond the same to each item.

Invariance was assessed by splitting the sample into two groups and comparing the results between the groups. The data were split into two groups randomly as well as by age (young versus old) and gender. Two methods were then used to assess the property of

invariance. The first method involved plotting the pairs of each item's difficulty and discrimination parameters calculated separately for each group. It was hypothesized that the plots of item difficulty and discrimination parameters for the subgroups should approximate a regression line with a slope of 1.0 and an intercept of zero[5]. The second method involved comparing negative likelihood statistics between a restricted and unrestricted model. An unrestricted model is a model where the item parameter estimates were determined separately for each subgroup. In the restricted model the item parameter estimates are constrained and set equal for each group. In the restricted model there is only one set of parameter estimates for each item for the entire sample. In the unrestricted model with two subgroups there will be two sets of parameter estimates for each item. It was hypothesized that the difference in negative twice log likelihood values between the restricted and unrestricted models would not be significantly different for the randomly selected groups as well as for the groups defined by age and gender.

The plots of item difficulty and discrimination parameters calculated separately for each subgroup should approximate a regression line with a slope of 1.0 and an intercept of zero if the property of invariance is upheld. Associated with this regression line would be a perfect correlation of 1.0 between the pairs of parameter estimates for each item. The three plots (random sample one versus two, young versus old and male versus female) for the ADL subscale grossly approximated the appropriate regression line. Some of the items deviated from the desired regression line. There was no pattern however, for items that did not consistently fit. The correlations between item discrimination parameters for three plots were .94 for the

randomly generated samples, .86 for young versus old and .85 for male versus females. For the item difficulties the correlations were .99, .98 and .96 respectively. There was no one item that consistently demonstrated invariance of the item parameters.

The plots of item difficulty and discrimination parameters for the Sports subscale had a larger deviation from the desired regression line then the ADL subscale. The correlations between item discrimination parameters were .92 for the randomly generated samples, .92 for young versus old and .82 for males versus females. The correlations between item difficulty parameters were .92, .93 and .99 respectively. There was no item that consistently demonstrated invariance of the item parameters.

The difference between the negative twice log likelihood values were greater than the critical value when comparing the restricted to unrestricted models in all three comparisons for the ADL and Sports subscale. This indicates that some of the items demonstrate differential item functioning.

Based on the item parameter plots and the comparison of the restricted and unrestricted models the property of invariance of item parameter estimates was not attained for either the ADL or Sports subscales. The discrimination parameters may have been more responsible than the difficulty parameters for the differential item functioning. Errors in the discrimination parameters can inflate the results of the item and test information function giving the impression that the test is more precise than it actually is. When the property of invariance in not upheld the test results cannot be generalized to all subpopulations of subjects. Also, item characteristics may be sample dependent, performance may be test dependent, ordered continuum of items based on their difficulty cannot be done, and results are test dependent not item dependent.

5.1.4 Assessing the Potential Responsiveness Across Ability Level for Each Item

An important feature of an evaluative index is responsiveness. This means that when the underlying condition that is being measured changes, the score on the instrument should change. In order for the FAAM to be responsive, the items must have a wide range of threshold difficulties and high levels of discrimination. Responsiveness of the individual items was assessed by examining the item characteristic curves. Item characteristic curves are a plot between the probability of choosing a particular response and the ability level of the examinee[17,23]. The item characteristic curve describes the relation between the trait an item assesses and the item response pattern across ability levels[17,23]. It was hypothesized that each item would have five distinct and separate curves, each with one peak, spanning the spectrum of ability.

Item characteristic curves were generated for each item on the 22-item on ADL subscale. All of the items had appropriate item characteristic curves except items 11-sleeping and 19-personal care. These items were not difficult and provided most information at the lower end of the ability spectrum. Item 11 functioned as an item with two potential responses as most subjects reported having either no difficulty or slight difficulty sleeping as a result of their foot or ankle problem. Even subjects on the low end of ability reported only slight difficulty sleeping as a result of their foot or ankle problem. Item 19 functioned more like an item with three responses as most subjects reported having moderate difficulty, slight difficulty, or no difficulty with personal care as a result of their foot or ankle problem. Items 11 and 19 did not function as expected across the spectrum of ability and were thus considered for deletion.

The slope of the item characteristic curve represents the item discrimination parameter[17,23]. A steeper slope implies the item has greater discrimination and the item is better able to separate examinees based on their ability. The discriminative parameter values for item

11 and 19 were 0.77 and 1.28 respectively. Discriminative parameters usually range from 0 to 2[17,23]. The other 20 items had discriminative parameters that averaged 2.13 (range 1.26-3.27) Item characteristic curves were also generated for each of the eight items on the Sports subscale. All eight items on the Sports subscale had item characteristic curves that spanned the range of function. Each of the eight items had five distinct and separate curves, each with one peak, spanning the spectrum of ability.

The underlying assumption of item response theory with respect to unidimensionality was met, therefore, item response theory could be used. Items 11 and 19 on the FAAM ADL subscale did not have appropriate item characteristic curves and might be candidates for elimination. All of the items on the FAAM Sports subscale had appropriate item characteristic curves.

5.1.5 Target Test Information Function

Item response theory can be used to generate an item information function that describes the amount of information that each item provides as a function of ability[33]. The item information function can be summed to provide a test information function. The more information the instrument provides the more precise the instrument. The more precise the instrument, the less associated error it has. A target test information function for an evaluative instrument should provide a maximum amount of information across all ability ranges[33]. It was hypothesized that the FAAM ADL and Sports subscale would produce a target test function that would be flat throughout the range of ability.

The test information function for the 22-item ADL scale provided more information in the lower ability ranges, while the Sports subscale produced test information function that

provided information in the higher ability ranges. An appropriate test information function is one that is flat throughout the range of ability. Combining the test information functions for the ADL and Sports subscale produced an appropriate test information function.

5.1.6 Conclusions for the Selection of Items for the Final FAAM ADL and Sports Subscales

The end result of this phase of the project was to reduce the number of items in the interim ADL subscale from 26 to 21 items. Four items related to pain eliminated because it appeared that they defined a separate factor. The fifth item that was eliminated was item 11-sleeping. Given that item 11 had the lowest inter-item correlation values, the lowest item to total score correlation, a item characteristic curve with only three distinct curves that spanned the spectrum of ability, a low discriminative ability compared to the other items and gave the most information at the lower end of the ability spectrum, item 11 was omitted from further consideration. Because item 19 had inter-item correlation values consistent with the other items, a more appropriate item characteristic curve, with four distinct curves that spanned the spectrum of ability, and had an ability to discriminate that was in order with the other items, it was retained to maintain the precision in the lower end of ability. The eight items on the Sports subscale were all retained.

5.2 Summary of the Evidence for Validity of the Final FAAM

The purpose of this phase was to evaluate the usefulness of the FAAM as a self-reported health related quality of life measure for those with foot and ankle impairments. This included an evaluation of reliability, responsiveness, and validity. Analyses included evaluation of the factorial structure, internal consistency, test-retest reliability, responsiveness and convergent and divergent evidence to support interpretation of the FAAM scores. Two groups were formed to

provide this evidence for interpretation of the FAAM scores. The group that was expected to change consisted of subjects currently receiving physical therapy for the treatment of a lower leg musculoskeletal disorder. Data were collected from this group at the initiation of care and again approximately four weeks later. The second group was expected to remain stable, which consisted of subjects who had been treated for a musculoskeletal disorder affecting the lower leg, foot and/or ankle at least one year ago. Data were collected from this group by mail twice approximately four weeks apart.

It was hypothesized that subjects currently receiving physical therapy would improve over the course of 4 weeks. We expected these subjects to perceive themselves to be somewhat or much better when comparing their final to initial status. One-hundred and seventeen (73.3%) subjects in the group expected to change described their perceived change in status as somewhat or much better. The average ADL and Sports subscale change scores subjects were 21.01 (SD 19.60 median 15.48 range -19 to 81) and 19.34 (SD 21.94 median 15.00 range -45 to 40) respectively.

Thirty-one (18.9%) subjects in the group that was expected to change described their change in functional status to be slightly better (24 or 14.6%), unchanged (5 or 3.0%), or slightly worse (21.2%). The average ADL and Sports subscale change scores for these 31 subject were 3.44 (SD 12.03 median -0.60 range -18 to 40) and 2.73 (SD 21.31 median 0.50 range -45 to 40) respectively.

It was hypothesized that subjects who had not received physical therapy for at least one year would perceive no change in their status between the initial and 4 week administration of the FAAM. Fifty-two (65.8%) subjects in the group expected to remain stable reported their status was unchanged. Thirteen (16.5%) subjects did not provide a perceived rating of change.

The remaining 14 (17.7%) subjects described their change in functional status to be much better (3 or 3.8%), somewhat better (3 or 3.8%), slightly better (2 or 2.5%), slightly worse (4 or 5.1%), somewhat worse (1 or 1.3% or much worse (1 or 1.3%). The change in ADL and Sports subscale scores and corresponding perceived change in status are reported in Table 5.1. The minimal detectable change for the ADL and Sports subscales was found 5.7 and 12.7 for the ADL and Sports subscale respectively. Therefore there were four subjects for the ADL and Sports subscale that had detectable change, two that improved and two that worsened.

5.2.1 Evidence of the Factorial Structure of the Final FAAM ADL and Sports Scales

Exploratory factor analyses were performed to examine the internal structure of the FAAM. The group expected to remain stable and the group expected change were analyzed separately. It was hypothesized that a one-factor model would fit the FAAM ADL and Sports subscales for both groups.

Factor analysis of the final 21-item ADL subscale for the group expected to change found that the items loaded on one factor that had an eigenvalue of 16.90 accounting for 80.46% of the variance and all items had high loadings on this factor. In the group expected to remain stable, the items loaded on two factors. Factor one had an eigenvalue of 16.46 that accounted for 78.37% of the variance and factor two had an eigenvalue of 2.58 that accounted for 12.28% of the variance. All items except item 18, personal care, had high loadings on the first factor. Item 18 was the only item that had a high factor loading on the second factor. In the stable group, all but one subject reported no difficulty with personal care. Therefore in the stable group it appears that this item was “too easy” for these subjects. This finding is consistent with the item characteristic curve and item information function for this item that was found in phase I of this

project (see section 4.1.7 and 4.1.8). In summary, the results indicated the factorial structure of the ADL subscale may be different in the groups defined by the subjects potential to change over a 4 week period.

The factor analysis of the final Sports subscale of the FAAM revealed that a one-factor model was appropriate for subjects in the group expected to change as well as in the group that was expected to remain stable. Because the factor structure for the two groups was the same, the groups could be combined for factor analysis. The findings were consistent with a one-factor model, which had an eigenvalue of 7.48 accounting for 93.48% of the variance and all of the items had high factor loadings on this factor. Thus it appears that the final Sports subscale represents one factor and this one factor model is similar across groups that are defined by the subjects' potential to change over a 4 week period of time.

5.2.2 Evidence for the Internal Consistency of the Final FAAM ADL and Sports Subscales

Internal consistency is a measure of consistency of the subjects' responses across items. Analysis of internal consistency for the ADL subscale in the group expected to remain stable also demonstrated that item 18 behaved differently from the remaining items. For item 18, the item to total score correlation was 0.27. The item to total score correlations for the other items ranged from0.66 to 0.87.

Because of the difference in factorial structure for the items in the ADL subscale between the group expected to change and the group expected to remain stable the groups could not be combined for the analysis on internal consistency. It was hypothesized that coefficient alpha would be greater than 0.90 for the subscales. The results of the analysis supported this hypothesis. For the ADL scale, coefficient alpha was 0.96 for the group expected to remain

stable and 0.98 for the group expected to change. The amount of error associated with a score at one point in time was calculated for the ADL subscale using coefficient alpha. The resulting SEM was 3.5 and 2.7 for the group expected to change and the group expected to remain stable respectively. The group expected to change and group expected to remain stable were combined for analysis of internal consistency of the Sports subscale. Coefficient alpha for the Sports scale was 0.98. The amount of error associated with a score at one point in time for the calculated for the Sports subscale was 5.12. Item to total score correlations were high for all items, ranging between 0.83 and 0.95.

Examining the internal consistency of other competing measures, coefficient alpha for the Foot Function Index was 0.96 in a sample of subjects with rheumatoid arthritis[13]. Coefficient alpha for the Lower Extremity Function Scale was 0.96 in a sample of individuals with a variety of lower extremity disorders[8]. Internal consistency of the physical function and role limitation because of physical health problems SF-36 scales were 0.93 and 0.84 respectively in subjects with general health problems[61].

5.2.3 Evidence for the Test Re-test Reliability of the Final FAAM ADL and Sports Subscales

Test re-test reliability assesses the stability of the scores over time. Repeated administration of the test should give the same score if there is no change in the construct being measured. It was hypothesized that the test re-test reliability would be good or excellent (> 0.9). The group expected to remain stable was used to estimate test re-test reliability. It was hypothesized that subjects who had not received physical therapy for at least one year would report no change in their status over the measurement period, however this was not the case for all subjects as previously described. Using all of the subjects in the group expected to remain

stable, the ICC test re-test reliability coefficients were .89 and .87 for the ADL and Sports subscales respectively. The SEM for the ADL subscale was 2.06 with a 95% confidence interval of plus or minus 4.03 points. Therefore, when evaluating change over a similar period of time, we can be 95% confident any change less than 4 points should be considered error. For the Sports subscale, the SEM was found to be 4.50 with a 95% confidence interval of plus or minus 8.82 points for the Sports subscale. Therefore, when evaluating change over a similar period of time, we can be 95% confident any change less than 9 points should be considered error.

The minimal detectable difference is obtained by using the test re-test reliability coefficient to constructing a confidence interval around the SEM**.** This CI takes into account the fact that 2 measurements at different points of time by including the square root of 2 in the calculation and therefore is appropriate to use to make decisions if a patient has changed with repeated measures over time. The 95% confidence interval for a minimal detectable was 5.07 for the ADL subscale. This means that we can be 95% confident that a change score of plus or minus 6 points is associated with a true change in score. The minimal detectable difference for the Sports subscale was 12.47. Therefore, we can be 95% confident that a change score of plus or minus 13 points is associated with a true change in score.

Three subjects in the group expected to remain stable reported a change in status that matched a change in global rating of function. When these three subjects were eliminated from this analysis, the ICC test re-test reliability coefficients increased to .91 for the ADL. The SEM decreased to 1.5 with a minimal detectable change at 95% confidence of 4.1. The ICC for the Sports Subscale remained at .87. The SEM decreases to 3.4 with a minimal detectable change at 95% confidence of 9.5.

Test re-test reliability of the FAAM was better than that of the Foot Function Index, Lower Extremity Function Scale and physical function and role limitation because of physical health problems scales of the SF-36. Test re-test reliability for the Foot Function Index and Lower Extremity Function Score are 0.87 and 0.86 respectively[8,13]. Test re-test reliability for the physical function and role limitation because of physical health problems scales of the SF-36 are 0.71 and 0.57 in subjects with sciatica[70]. The Ankle Arthritis Scale had higher levels of test re-test reliability (ICC of 0.97)[24]. The Lower Extremity Function Scale had minimal detectable change of plus or minus 5.3 points at a 90% confidence level[7]. The minimal detectable change for the FAAM ADL and Sports subscale was re-calculated using the 90% confidence interval and was found to be plus or minus 4.78 and 10.47 respectively..

5.2.4 Evidence to Support the Responsiveness of the Final FAAM ADL and Sports Subscales

Responsiveness is the ability to detect change when a change has occurred. A multiple group design was used in our analysis to demonstrate responsiveness. To accomplish this, groups that were expected to undergo differential rates of change over a 4-week period were compared. Responsiveness was assessed using a two-way repeated measures ANOVA, Guyatt's responsiveness index and ROC curve analyses.

In the first analysis, it was hypothesized that the group expected to change would have a greater change in the FAAM score over a four-week period than the group expected to remain stable. The two-way repeated measures ANOVA found a significant interaction indicating that the group expected to change had a greater change score over the four-week period compared to those that were expected to remain stable. For the second analysis of responsiveness, Guyatt's responsiveness index was calculated by dividing the change score of the group that was expected

to change by the standard deviation of the stable group. It was hypothesized that the 95% confidence interval for the GRI would not contain zero, indicating that the GRI was significantly different from zero. Guyatt's responsiveness index was significantly different than zero for both the FAAM ADL and Sports subscale.

In the final analysis of responsiveness, a ROC curve was used to evaluate responsiveness. Sensitivity of change is the ability to detect change when a change has occurred. Specificity of change is the ability to correctly detect when an individual has not changed. A perfect instrument would have sensitivity, specificity and area under the ROC curve of 1.00. The area under the ROC curve represents ability to discriminate those subjects who improved from those that did not improve based on a criterion measure of change. For the ADL subscale the area under the curve was 0.80. The optimal change score to identify a changed from an unchanged individual is that point on the ROC curve that lies closest to the upper left hand corner of the graph. This change score represents the minimal clinically important change. The minimal clinically important change for the ADL subscale was 4 points, which had a sensitivity and specificity of 0.76 and 0.88 respectively. Therefore, 76% of subjects who were in the group expected to change had a change score greater than 4 points. Conversely 88% of subjects who were in the group expected to remain stable had a change score of less than 4 points.

The minimal clinically important change for the Sports subscale was 5 points, which had a sensitivity and specificity of 0.63 and 0.81 respectively. Therefore, 63% of subjects who were in the group expected to change had change score of 5 points or more. Conversely 81% of subjects who were in the group expected to remain stable had change score of less than 5 points.

A second ROC analysis was done with the group expected to change separated into two groups. One group consisted of the 117 subjects that reported being somewhat or much better.

The second group consisted of the 31 subjects that reported being slightly better, unchanged or slightly worse. The minimal clinically important change for the ADL subscale was 8 points, which had a sensitivity and specificity of 0.77 and 0.75 respectively. Therefore, in the group expected to change 77% of subjects who reported that they were improved had a change score greater than 8 points. Conversely 75% of subjects in the group that was expected to change who reported that they were not improved had a change score less than 8 points.

With this second ROC analysis, the minimal clinically important change for the Sports subscale was 9 points, which had a sensitivity and specificity of 0.64 and 0.75 respectively. Therefore, in the group expected to change 64% of subjects who reported that they were improved had a change score greater than 9 points. Conversely 75% of subjects in the group that was expected to change who reported that they were not improved had a change score less than 9 points.

The error associated with a measurement at a single point in time, the minimal detectable change and the minimal clinically important difference for the ADL scale were 6.9, 5.7 and 8 points and for the sports scale they were 10, 12.3 and 9 points respectively for the Sports subscale. The small difference between these three values for the ADL subscale can be attributed to measurement error. The difference associated with these three values for Sports subscale may be more than one could attribute to measurement error. This discrepancy may be due to the fact that the values were calculated using difference techniques**.** The errors associated with measurement at a single point in time were calculated using coefficient alpha to determine the SEM and are therefore related to content sampling. The errors associated with minimal detectable difference were calculated using ICC to determine the SEM and are related the stability of the measure over time. The minimal clinically important difference was determined

in the ROC analysis and the change in score that represented the highest values for sensitivity and specificity. Also, the error associated with measurement at a single point in time were calculated using the group that was expected to remain stable combined with the group that was to change, the minimal detectable difference was calculated using the group that was expected to remain stable, and the minimal clinically important difference was calculated using the group expected to change. When using the Sports subscale, it may be appropriate to use a change score of 12 points or greater to identify those who not only significantly changed but also those who had a clinically important change.

The authors of the LEFS investigated responsiveness of the LEFS. The minimal clinically important change for the Lower Extremity Function Scale was found to be 9 points using an ROC analysis. The subject's prognosis for improvement, as determined by the clinician on the initial evaluation, was used as the criterion measure of change. The change of nine points was associated with a sensitivity of 0.81 and specificity of 0.70[8].

Beaton offered a method to interpret and compare responsiveness studies based on three criteria: subjects involved in the study, criteria of how change is quantified and the methods used to measure responsiveness[4]. There are two major differences when comparing responsiveness between this study and the LEFS[8]. The first difference is related to the subjects that were used to investigate responsiveness. The authors investigating responsiveness of the LEFS used subjects with any lower extremity musculoskeletal disorder. This included 80 (74.8%) subjects with knee, thigh or leg disorders and two (2.9%) subjects with hip disorders. Twenty-two (20.5%) subjects the subjects had foot or ankle disorders. The second difference was with respect to the criteria of how change was quantified. To study responsiveness of the LEFS, the criterion for changes was based on the clinician's prognostic rating of change for the patient following the initial

examination. In our study we used two methods to quantify change. In the first method, the construct for change was a comparison of those currently receiving treatment with those that have not received treatment in over a year. The second method to quantify change was based on the subject's perceived change in status. This was done in the group receiving treatment. The analysis of both studies included presenting the results at the individual level, using the ROC analysis[8]. Our analysis also included presenting the results at the group level using Guyatt's responsiveness index.

5.2.5 Evidence to Support that the FAAM ADL and Sports Subscales are More Responsive to Change than a General Measure of Functional Status

It was hypothesized that the FAAM ADL and Sports subscales would be more responsive than a general measure of functional status. The physical function and physical components summary of the SF-36 were used as the general measure of functional status and were compared to the FAAM ADL and Sports subscales.

The most optimal change score needed to differentiates an individual in the group expect to change from an individual in the group expected to remain stable was found to be 4 points for the FAAM ADL subscale, 6 points physical function scale of the SF-36 and 5 points for the physical components summary score. The ADL subscale not only had a smaller change score to differentiate subjects, but the most optimal change score also had higher sensitivity and specificity values with a significantly larger area under the curve than the general measures of function. The ROC and GRI analysis found the ADL subscale to be more responsive than the both the physical function subscale and the physical component summary score.

The most optimal change score needed to differential an individual in the group expect to change from an individual in the group expected to remain stable was found to be 5 points for

the Sports subscale of the FAAM compared to1 point physical function scale and 3 points for the physical component summary score of the SF-36. The results of the ROC curve analyses demonstrated the FAAM Sports subscale was more responsive than the SF-36 physical component summary score at the individual level. However, at the group level GRI analysis the Sports subscale was nor more responsive than the physical component summary score. Both the ROC and GRI analysis found the Sports subscale was not more responsive than the physical function subscale.

5.2.6 Evidence to Support the FAAM is More Responsive than a Global Rating of Self Perceived Level of Function

It was hypothesized that the FAAM ADL and Sports subscales would be more responsive to changes in function than a self-perceived global rating of function. The results of the GRI and ROC curve analysis did not support this hypothesis. Based on the ROC analysis, to determine a clinically meaningful change in status a change of 4 and 7 points were needed for the ADL subscale and the global rating of activities of daily living were needed respectively. The areas under the ROC curves were not significantly different. To determine a clinically meaningful change in status a four point change was needed for the Sports subscale and a one point change was need for the global rating for sports. The GRI analysis found the ADL subscale to be more responsive than ADL global rating. The Sports subscale was not more responsive than the global rating of sports.

The findings do not necessarily support the use of the global rating instead of the FAAM, but rather, they should be used to supplement one another. Although the global rating seems to be as responsive of the FAAM, it does not supply as much detail about what specific activities

have significantly improved. Also, information about limitations and changes with specific activities are required for evaluation, treatment planning and goal setting.

5.2.7 Convergent and Divergent Evidence to Support the Interpretation of the FAAM

Convergent evidence came from assessing the association of the FAAM to concurrent measures of physical function including the physical function and physical component summary scores of the SF-36 and the global rating of function. The hypothesis was there would be a strong to moderate correlation ($r>0.6$) between the FAAM and the concurrent measures of function. The results supported this hypothesis. The correlations between the FAAM ADL subscale and the physical function subscale and physical components summary score of the SF-36 and global rating of function during ADL were 0.84, 0.84 and 0.83 respectively. The correlations between the FAAM Sports subscale and the physical function scale and physical components summary score of the SF-36 and global rating of function during sports were 0.78, 0.80 and 0.89 respectively.

A similar analysis was done with the Lower Extremity Function Score. The correlation between the Lower Extremity Function Scale and the SF-36 physical function subscale and SF-36 physical component summary score were 0.80 and 0.64 respectively[8].

Divergent evidence came from assessing the association between the FAAM and concurrent measures of emotional function including the mental health scale and mental components summary score of the SF-36. It was hypothesized that there would be weak correlations ($r<.3$) between the FAAM and concurrent measures of emotional function. The results supported this hypothesis. The correlations between the FAAM ADL subscale and the mental health scale and mental components summary score of the SF-36 were 0.18 and 0.05 respectively. The correlations between the FAAM subscale and the mental health scale and mental components summary score of the SF-36 were 0.11 and -0.02 respectively.

A similar analysis was done with the Lower Extremity Function Score. The correlation between the Lower Extremity Function Scale and the mental health subscale and mental component summary score of the SF-36 were 0.23 and 0.30 respectively[8].

The hypotheses that correlations between the FAAM ADL and Sports subscale scores to concurrent measures of physical function would be significantly greater than the correlations of the FAAM ADL and Sports subscales with concurrent measures of emotional function were also met as calculated t value was grater than the critical value for all analyses.

5.3 Conclusion

The purpose of the first stage of this project was to develop an instrument that contained items demonstrating appropriate psychometric properties. This included selecting items based on the pattern of factor loadings, inter-item correlations, item to total score correlations, and coefficient alpha. The initial 26-item interim version of the ADL subscale did not meet the assumption of unidimensionality that is required for item response theory. Four items related to pain were omitted which allowed the 22-item subscale meet the assumptions of unidimensionality. The fit of the graded response model was appropriate for the 22-item ADL, however the property of item parameter invariance was not achieved. This was evidenced by differential item functioning for both item discrimination and difficulty parameters. All of the items had appropriate item characteristic curves with exception of item 11-sleeping and item 19-personal care. Item 11 was not able to discriminate examinees as well as item 19. The test information function for the 22 item ADL subscale supplied information throughout the spectrum of ability, however, it was noted that there was more information at the lower end of the ability spectrum. Item 11 was subsequently omitted because it had: 1) the lowest inter-item correlation values, 2) the lowest item to total score correlation, 3) an item characteristic curve

with only three distinct curves that spanned the spectrum of ability, and 4) a low discrimination compared to the other items. Because item 19 had inter-item correlation values consistent with the other items, had a more appropriate item characteristic curve, and had an ability to discriminate that was in order with the other items, it was retained to improve precision of measurement in the lower end of ability.

The eight-item Sports subscale met the assumption of unidimensionality and the fit of the graded response model was determined to be appropriate. The property of item parameter invariance was not achieved as evidenced by differential item functioning for both the discrimination and difficulty parameters. The item characteristic curves were appropriate and the test information function provided the majority of information at the higher end of the ability spectrum.

The final ADL subscale consisted of 21 items all related to functional activities during daily living, and the final FAAM Sports subscale consisted of eight items related to sports activities. The properties of the items included on the ADL and sports scales can however, can only be generalized to a population similar to the one used to estimate the item parameters because the property of parameter invariance was not achieved. The test information function indicated that the ADL subscale supplied most information at the lower end of the ability spectrum while the Sports subscale supplied more information at the higher end of ability. Therefore, to cover the full range of ability it is appropriate to give both scales.

The purpose of the second part of this project was to assess reliability, responsiveness and validity of the final version of the FAAM. The results of this part of the study can be generalized to populations that have similar characteristics as the subjects in this study. The subjects in this part of the study had an average age of 42.5 (range 9-86) with a duration of

symptoms averaging 4.5 months (range 1day- 2.2 years). The diagnosis profile of subjects consisted of the following: joint and limb pain (n=94), sprains and strains (n=71), fractures (n=33), plantar fasciitis (n=27), bunion (n=5), and Achilles rupture (n=2). These subjects were not receiving physical therapy for a coexisting pathology, however, they may have had other disorders such as diabetes and cardiovascular disease, representative of those with similar demographic characteristics. The average baseline scores of FAAM ADL and Sports scores were 60.8 and 39.0 respectively.

The results of the factor analysis revealed a one factor model was appropriate for the ADL subscale in the group expected to change but not for the group expected to remain stable. This did not significantly alter the internal consistency as coefficient alpha was greater than 0.90 for both groups. The error associated with a measurement at a single point in time for the group receiving physical therapy was +/- 6.9 points at a 95 % confidence level. The Sports scale fit a one-factor mode and internal consistency was greater than 0.90 when the groups were combined. The error associated with measurement at a single point of time for the Sports subscale was +/- 10 points at a 95% confidence level. The test re-test reliability was 0.89 and 0.87 for the ADL and Sports subscales respectively. The minimal detectable difference at a 95% confidence level was 5.7 and 12.3 points for the ADL and Sports subscales respectively.

The ADL subscale was also more responsive to changes in functional status than general measures of physical function. At a group level of analysis, Guyatt's responsiveness index indicated that the ADL scale was more responsive than a global rating of functional status, however the sports subscale was not more responsive than the physical function subscale of the SF-36 or a global rating of function. At an individual level of analysis, a ROC analysis indicated that the Sports subscale was more responsive than the physical component summary score.

The values associated with the measurement error at a single point of time, the minimal detectable difference and the minimal clinically important change may vary depending on the baseline level of function of the subjects. We have uniformly assigned values to these measures across all of our subjects, irrespective of their baseline functional level. Future research should include an assessment of reliability and responsiveness across different functional levels.

Evidence of validity of the FAAM ADL and sports scales was demonstrated by relatively strong correlations with concurrent measures of physical function and relatively low correlations to measures of emotional status.

The FAAM and LEFS are two evaluative measures of functional limitations. The FAAM was specifically developed for those with musculoskeletal disorders of the lower leg. The LEFS was developed to encompass the entire lower extremity containing items relevant to those with hip and knee disorders as well lower leg disorders. Future research should compare the two instruments in patients with lower leg, foot/ankle disorders.

When using the FAAM to describe the outcome for treatment intervention it should be used with a general measure of health status that will measure other components of health such as emotional status. It should also be complimented by measures of pain and impairment such as strength, range of motion and joint effusion.

In conclusion, this study provides evidence of reliability, responsiveness and validity for the FAAM ADL and Sports subscales. The FAAM is appropriate to measure the effects of treatment for individuals with lower leg musculoskeletal disorders.

APPENDIX A
Initial Items Selection:

Procedures to produce the final version of the FAAM are included in stage I. This stage consists of five steps: 1) define the purpose of the instrument, 2) item generation, 3) initial item reduction, 4) instrument construction, and 5) final item reduction. Steps one through four were completed as preliminary work. The goal of this preliminary work was to construct an interim FAAI that would be appropriate for psychometric testing of individual items.

The first step is to define the purpose of the instrument. This is an evaluative instrument because its primary objective is to measure changes in an individual's functional status over time as a result of treatment. Of specific interest is how treatment of foot and ankle related disorders impacts functional status.

The goal of step two and three was to establish evidence based on test content. This was accomplished in two phases. The first phase was to generate an exhaustive list of all possible items that may assess functional status and disability related to impairments of the foot and ankle. The second phase involved initial item reduction. The objective of this second phase was to remove items that were considered to be unimportant; in other words items that were repetitive, complex, too narrow in scope and/or difficult for the subject to interpret.

Generation of an exhaustive list of all possible items was done through literature review, input from experts and input from a sample of subjects for which the instrument was intended. Instruments that have been developed specifically for the foot and ankle as well as those developed for other regions of the lower extremity were reviewed to generate an initial list of items. The instrument s reviewed included the OFAS Clinical Rating System[48], AOS[24], FFI[13], LEFS[8], KOS[38], and the DASH[37]. Expert clinicians that evaluate and manage individuals with impairments of the foot and ankle interest were asked to contribute ideas, topics and items that

they felt needed to be included on the instrument. Individuals with foot and ankle related disorders were also asked to contribute ideas as well as critique possible items to make sure that the FAAM addressed areas they felt should be included on the instrument.

Item generation produced 69 items. Experts to assist in selecting items were obtained from a list of members from the American Physical Therapy Association (APTA) Foot and Ankle Special Interest Group. Surveys were mailed to members of the APTA Foot and Ankle Special Interest Group and they were asked to give their opinion on the relative importance of the 69 potential items. The experts were asked to rate each item on a scale ranging from -2 (not important) to +2 (very important). Items that attained a mean score of one or above were included on the interim FAAM. The experts were also asked if they felt there should be two scales, one for lower level every day activities and one for higher level sporting activities.

Twenty-nine out of 43(67.4%) surveys were returned. Ninety four percent of the respondents felt there should be two separate scales, one for activities of daily living and one for sports. The average rating for each potential items in seven categories are displayed in Tables A1 through A7. The seven categories included a miscellaneous functional limitation category (28 items) (Table A1). Four categories were more specific and assessed the ability to walk (11items) (Table A2) negotiate stairs (six items) (Table A3), participate in sports (five items) (Table A4) and do work (three items) (Table A5). The other two categories dealt with symptoms (13 items) (Table A6) and nonspecific possible item of interest (four items) (Table A7)

The end result of the analysis was an interim FAAM that included two scales; an activity of daily living scale that contained 26 items and a sports scale that contained eight items (see appendix B). The symptom categories and psychological aspect were eliminated because the purpose of the instrument was to measure functional ability. The only exception to this was the

inclusion of pain. Pain was included on the interim FAAM because pain may be a patient's major complaint and most limiting factor. Walking distance could be measured by time or distance. It was felt that time represented a more universally accepted way to assess walking distance. The items related to the use of an assistive device, need for medication, cosmesis, and ability to wear different shoes were all eliminated because of their inability to fit into the question format ("How does your foot and ankle limit your ability to perform the following activity?") and the response pattern (severe, moderate, mild or no difficulty).

Construction of the FAAM included choosing the most suitable wording as well as organizing the directions, the items, and the possible responses. The wording of the directions was done not limit the subjects to either a performance or capacity bias.

Responses ranged from "no difficulty at all" to "unable to perform". Responses to questions were set up in a Likert format with five potential responses. Scoring of the FAAM was set up as an ability scale with a higher score representing a higher degree of functioning.

Finally, field-testing was done with 20 patients to ensure the index was user friendly, for both clinicians and patients. An effort was made to ensure the FAAM was easy to administer, complete and score. This included making sure the directions were clear and easily read.

APPENDIX B

Interim Foot and Ankle Ability Measaure (FAAM)

Please answer **every question** with one response that most closely describes to your condition within the past week.
If the activity in question is limited by something other than your foot or ankle mark not applicable (N/A).

	No difficulty	Slight difficulty	Moderate difficulty	Extreme difficulty	Unable to do	N/A
Standing	□	□	□	□	□	□
Walking on even ground	□	□	□	□	□	□
Walking on even ground without shoes	□	□	□	□	□	□
Walking up hills	□	□	□	□	□	□
Walking down hills	□	□	□	□	□	□
Going up stairs	□	□	□	□	□	□
Going down stairs	□	□	□	□	□	□
Walking on uneven ground	□	□	□	□	□	□
Stepping up and down curbs	□	□	□	□	□	□
Squatting	□	□	□	□	□	□
Sleeping	□	□	□	□	□	□
Coming up on your toes	□	□	□	□	□	□
Walking initially	□	□	□	□	□	□
Walking 5 minutes or less						
Walking approximately 10 minutes	□	□	□	□	□	□
Walking 15 minutes or greater	□	□	□	□	□	□

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

	No difficulty at all	Slight difficulty	Moderate difficulty	Extreme difficulty	Unable to do	N/A
Home Responsibilities	□	□	□	□	□	□
Activities of daily living	□	□	□	□	□	□
Personal care	□	□	□	□	□	□
Light to moderate work (standing, walking)	□	□	□	□	□	□
Heavy work (push/pulling, climbing, carrying)	□	□	□	□	□	□
Recreational activities	□	□	□	□	□	□

Please rate your pain level as it relates to your **foot and ankle**:

	None	Mild	Moderate	Severe	Unbearable
General level of pain	□	□	□	□	□
At rest	□	□	□	□	□
During your normal activity	□	□	□	□	□
First thing in the morning	□	□	□	□	□

FAAM Sports Scale

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

	No difficulty at all	Slight difficulty	Moderate difficulty	Extreme difficulty	Unable to do	N/A
Running	□	□	□	□	□	□
Jumping	□	□	□	□	□	□
Landing	□	□	□	□	□	□
Starting and stopping quickly	□	□	□	□	□	□
Cutting/lateral movements	□	□	□	□	□	□
Low impact activities	□	□	□	□	□	□
Ability to perform activity with your normal technique	□	□	□	□	□	□
Ability to participate in your desired sport as long as you would like	□	□	□	□	□	□

Appendix C

Item Characteristic Curves for the 22 Item ADL Subscale **(Figure C)**

Appendix D

Item Characteristic Curves for the Sports Subscale **(Figure D)**

APPENDIX E

Final Foot and Ankle Ability Measaure (FAAM)

Please answer **every question** with one response that most closely describes to your condition within the past week.
If the activity in question is limited by something other than your foot or ankle mark not applicable (N/A).

	No difficulty	Slight difficulty	Moderate difficulty	Extreme difficulty	Unable to do	N/A
Standing	□	□	□	□	□	□
Walking on even ground	□	□	□	□	□	□
Walking on even ground without shoes	□	□	□	□	□	□
Walking up hills	□	□	□	□	□	□
Walking down hills	□	□	□	□	□	□
Going up stairs	□	□	□	□	□	□
Going down stairs	□	□	□	□	□	□
Walking on uneven ground	□	□	□	□	□	□
Stepping up and down curbs	□	□	□	□	□	□
Squatting	□	□	□	□	□	□
Coming up on your toes	□	□	□	□	□	□
Walking initially	□	□	□	□	□	□
Walking 5 minutes or less						
Walking approximately 10 minutes	□	□	□	□	□	□
Walking 15 minutes or greater	□	□	□	□	□	□

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

	No difficulty at all	Slight difficulty	Moderate difficulty	Extreme difficulty	Unable to do	N/A
Home Responsibilities	□	□	□	□	□	□
Activities of daily living	□	□	□	□	□	□
Personal care	□	□	□	□	□	□
Light to moderate work (standing, walking)	□	□	□	□	□	□
Heavy work (push/pulling, climbing, carrying)	□	□	□	□	□	□
Recreational activities	□	□	□	□	□	□

FAAM Sports Scale

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

	No difficulty at all	Slight difficulty	Moderate difficulty	Extreme difficulty	Unable to do	N/A
Running	□	□	□	□	□	□
Jumping	□	□	□	□	□	□
Landing	□	□	□	□	□	□
Starting and stopping quickly	□	□	□	□	□	□
Cutting/lateral movements	□	□	□	□	□	□
Low impact activities	□	□	□	□	□	□
Ability to perform activity with your normal technique	□	□	□	□	□	□
Ability to participate in your desired sport as long as you would like	□	□	□	□	□	□

BIBLIOGRAPHY

1. Airaksinen O, Kolari PJ, Miettinen H. Elastic bandages and intermittent pneumatic compression for treatment of acute ankle sprains. Arch.Phys.Med.Rehabil. 1990; 71:380-383.

2. Andresen EM, Patrick DL, Carter WB, Malmgren JA. Comparing the performance of health status measures for healthy older adults. J.Am.Geriatr.Soc 1995; 43:1030-1034.

3. Angus PD, Cowell HR. Triple arthrodesis. A critical long-term review. J.Bone Joint Surg.Br 1986; 68:260-265.

4. Beaton DE. Understanding the Relevance of Measured Change Through Studies of Responsiveness. Spine. 2000; 25:3192-3199.

5. Bejar II. A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. Journal of Education Measurement 1980; 17:283-296.

6. Bellacosa RA, Pollak RA. Patient expectations of elective foot surgery. J.Foot.Ankle.Surg. 1993; 32:580-583.

7. Berger M.B., Bobbit R.A., Carter W.B., Gilson B.S. The SIP; Development and final revision of a health status measure. Med.Care 1981; 19:787-805.

8. Binkley JM, Stratford PW, Lott SA, Riddle DL. The Lower Extremity Functional Scale (LEFS): scale development, measurement properties, and clinical application. North American Orthopaedic Rehabilitation Research Network. Phys.Ther. 1999; 79:371-383.

9. Bork CE, Francis JB. Developing effective questionnaires. Phys.Ther. 1985; 65:907-911.

10. Boyden EM, Kitaoka HB, Cahalan TD, An KN. Late versus early repair of Achilles tendon rupture. Clinical and biomechanical evaluation. Clin.Orthop 1995; 150-158.

11. Brazier JE, Harper R, Jones NM, O'Cathain A, Thomas KJ, Usherwood T, et al. Validating the SF-36 health survey questionnaire: new outcome measure for primary care. BMJ. 1992; 305:160-164.

12. Brazier JE, Walters SJ, Nicholl JP, Kohler B. Using the SF-36 and Euroqol on an elderly population. Qual.Life Res 1996; 5:195-204.

13. Budiman-Mak E, Conrad KJ, Roach KE. The Foot Function Index: a measure of foot pain and disability. J.Clin.Epidemiol. 1991; 44:561-570.

14. Cooper DM, Frederick RD. Treatment of idiopathic clubfoot. J.Bone Joint Surg.Am. 1995; 77a:1477-1489.

15. Daly PJ, Kitaoka HB, Chao EY. Plantar fasciotomy for intractable plantar fasciitis: clinical results and biomechanical evaluation. Foot.Ankle. 1992; 13:188-195.

16. Dawson J, Fitzpatrick R, Murray D, Carr A. Comparison of measures to assess outcomes in total hip replacement surgery. Qual.Health Care 1996; 5:81-88.

17. DeAyala RJ. An introduction to polytomous item response theory models. Measurement and Evaluation in Counseling and Development. 1993; 25:172-189.

18. Demottaz JD, Mazur JM, Thomas WH, Sledge CB, Simon SR. Clinical study of total ankle replacement with gait analysis. A preliminary report. J.Bone Joint Surg.Am. 1979; 61:976-988.

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

20. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. Control.Clin.Trials. 1991; 12:142S-158S.

21. Di Fabio RP, Boissonnault W. Physical therapy and health-related outcomes for patients with common orthopaedic diagnoses. J.Orthop Sports Phys.Ther. 1998; 27:219-230.

22. Dillman DA. Mail and Telephone Surveys: The Total Design Method. New York, N.Y.: Wiley-Intersience, 1978.

23. Dodd BG, DeAyala RJ, Koch WR. Computerizedadaptive testing with polytomous items. Applied Psyhcological Measurement. 1995; 19:5-22.

24. Domsic RT, Saltzman CL. Ankle osteoarthritis scale. Foot.Ankle.Int. 1998; 19:466-471.

25. Eiff MP, Smith AT, Smith GE. Early mobilization versus immobilization in the treatment of lateral ankle sprains. Am.J.Sports Med. 1994; 22:83-88.

26. Ellwood PM. Shattuck lecture-outcome management: A technology of patient experience. New England Journal of Medicine. 1998; 318:1549-1556.

27. Evanski PH, Waugh TR. Management of arthritis of the ankle. An alternative of arthrodesis. Clin.Orthop 1977; 110-115.

28. Garratt AM, Ruta DA, Russell I, Macleod K, Brunt P, McKinlay A, et al. Developing a condition-specific measure of health for patients with dyspepsia and ulcer-related symptoms. J.Clin.Epidemiol. 1996; 49:565-571.

29. Geigle R, Jones SB. Outcomes measurement: a report from the front. Inquiry. 1990; 27:7-13.

30. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J.Chronic.Dis. 1987; 40:171-178.

31. Haley SM, McHorney CA, Ware JEJ. Evaluation of the MOS SF-36 physical functioning scale (PF-10): I. Unidimensionality and reproducibility of the Rasch item scale. J.Clin.Epidemiol. 1994; 47:671-684.

32. Hambleton RK, Jones RW. Comparison of classical test theory and item response theory and their applications to test development. Education Measurement: Issues and Practices. 1993; 12:38-47.

33. Hambleton RK, Swaminathan H. Fundamentals of Item Response Theory. New Park, C.A.: Sage Publications, 1991.

34. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983; 148:839-843.

35. Hays RD, Hadorn D. Responsiveness to change: an aspect of validity, not a separate dimension. Qual.Life Res 1992; 1:73-75.

36. Herberts P, Goldie IF, Korner L, Larsson U, Lindborg G, Zachrisson BE. Endoprosthetic arthroplasty of the ankle joint. A clinical and radiological follow-up. Acta Orthop Scand. 1982; 53:687-696.

37. Hudak PL, Amadio PC, Bombardier C. Development of an upper extremity outcome measure: The DASH (disabilities of the arm, shoulder, and hand). American Journal of Industrial Medicine. 1996; 29:602-608.

38. Irrgang JJ, Snyder-Mackler L, Wainner RS, Fu FH, Harner CD. Development of a patient-reported measure of function of the knee. J.Bone Joint Surg.Am. 1998; 80:1132-1145.

39. Jeager RM. Statistics: A Spectator Sport. second ed. Newbury Park, C.A.: Sage Publications, 1990.

40. Jette AM. Physical disablement concepts for physical therapy research and practice. Phys.Ther. 1994; 74:380-386.

41. Jette DU, Downing J. The relationship of cardiovascular and psychological impairments to the health status of patients enrolled in cardiac rehabilitation programs. Phys.Ther. 1996; 76:130-139.

42. Jette DU, Jette AM. Physical therapy and health outcomes in patients with knee impairments. Phys.Ther. 1996; 76:1178-1187.

43. Kaikkonen A, Kannus P, Jarvinen M. A performance test protocol and scoring scale for the evaluation of ankle injuries. Am.J.Sports Med. 1994; 22:462-469.

44. Kaplan RM, Anderson JP, Wu AW, Mathews WC, Kozin F, Orenstein D. The Quality of Well-being Scale. Applications in AIDS, cystic fibrosis, and arthritis. Med.Care 1989; 27:S27-S43

45. Katz JN, Harris TM, Larson MG, Krushell RJ, Brown CH, Fossel AH, et al. Predictors of functional outcomes after arthroscopic partial meniscectomy. J.Rheumatol. 1992; 19:1938-1942.

46. Kessler RC, Mroczek DK. Measuring the effects of medical interventions. Med.Care 1995; 33:AS109-AS119

47. Kirshner B, Guyatt G. A methodological framework for assessing health indices. J.Chronic.Dis. 1985; 38:27-36.

48. Kitaoka HB, Alexander IJ, Adelaar RS, Nunley JA, Myerson MS, Sanders M. Clinical rating systems for the ankle-hindfoot, midfoot, hallux, and lesser toes. Foot.Ankle.Int. 1994; 15:349-353.

49. Kitaoka HB, Anderson PJ, Morrey BF. Revision of ankle arthrodesis with external fixation for non-union. J.Bone Joint Surg.Am. 1992; 74:1191-1200.

50. Kopec JA, Esdaile JM, Abrahamowicz M, Abenhaim L, Wood-Dauphinee S, Lamping DL, et al. The Quebec Back Pain Disability Scale: conceptualization and development. J.Clin.Epidemiol. 1996; 49:151-161.

51. Lemon B, Pupp GR. Long-term efficacy of total SILASTIC implants: a subjective analysis. J.Foot.Ankle.Surg. 1997; 36:341-346.

52. Lord FM. Thr relation of test score to the trait underlying the test. Psychological Measurement. 1953; 13:517-548.

53. Lysholm J, Gillquist J. Evaluation of knee ligament surgery results with special emphasis on use of a scoring scale. Am.J.Sports Med. 1982; 10:150-154.

54. MacKenzie CR, Charlson ME, DiGioia D, Kelley K. A patient-specific measure of change in maximal function. Arch.Intern.Med. 1986; 146:1325-1329.

55. MacKenzie EJ, Burgess AR, McAndrew MP, Swiontkowski MF, Cushing BM, deLateur BJ, et al. Patient-oriented functional outcome after unilateral lower extremity fracture. J.Orthop Trauma. 1993; 7:393-401.

56. Maffulli N, Testa V, Capasso G, Bifulco G, Binfield PM. Results of percutaneous longitudinal tenotomy for Achilles tendonopathy in middle- and long-distance runners. Am.J.Sports Med. 1997; 25:835-840.

57. Manoli A, Prasad P, Levine RS. Foot and ankle severity scale (FASS). Foot.Ankle.Int. 1997; 18:598-602.

58. Mazur JM, Schwartz E, Simon SR. Ankle arthrodesis. Long-term follow-up with gait analysis. J.Bone Joint Surg.Am. 1979; 61:964-975.

59. McDowell I, Newell C. Measuring Health: A Guide to Rating Scales and Questionnaires. New York, N.Y.: Oxford University Press, 1987.

60. McHorney CA, Haley SM, Ware JEJ. Evaluation of the MOS SF-36 Physical Functioning Scale (PF-10): II. Comparison of relative precision using Likert and Rasch scoring methods. J.Clin.Epidemiol. 1997; 50:451-461.

61. McHorney CA, Ware JEJ, Raczek AE. The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. Med.Care 1993; 31:247-263.

62. McHorney CA, Ware JEJ, Rogers W, Raczek AE, Lu JF. The validity and relative precision of MOS short- and long-form health status scales and Dartmouth COOP charts. Results from the Medical Outcomes Study. Med.Care 1992; 30:MS253-MS265

63. Meenan RF, Gertman PM, Mason JH. Measuring health status in arthritis. Arthritis and Rheumatism. 1980; 23:146-152.

64. Meng X, Roenthal R, Sax G. Compaing correlation coefficients. Psychological Bulletin. 1957; 111:172-175.

65 Messick S. Meaning and values in test validation: The science and ethics of assessment. Educational Researcher 1989; 18:5-11.

66. Nagi S. Some Conceptual Issues in Disability and Rehabilitation. In: Sussman M, editor. Sociology and Rehabilitation. Washington, DC: American Sociology Association, 1965:100-113.

67. Nagi S. Disability Concept revisited: Implication for Prevention. In: Pope AM, Tarlov AR, editors. Disability in America: Toward a National Agenda for Prevention. Washington, D.C.: National Academy Press, 1991:309-327.

68. Olerud C, Molander H. A scoring scale for symptom evaluation after ankle fracture. Arch.Orthop Trauma.Surg. 1984; 103:190-194.

69. Patrick DL, Deyo RA. Generic and disease-specific measures in assessing health status and quality of life. Med.Care 1989; 27:S217-S232

70. Patrick DL, Deyo RA, Atlas SJ, Singer DE, Chapin A, Keller RB. Assessing health-related quality of life in patients with sciatica. Spine 1995; 20:1899-1908.

71. Puno RM, Grossfeld SL, Henry SL, Seligson D, Harkess J, Tsai TM. Functional outcome of patients with salvageable limbs with grades III-B and III-C open fractures of the tibia. Microsurgery. 1996; 17:167-173.

72. Reise SP, Yu J. Parameter recovery in the graded response model using MULTILOG. Journal of Education Measurement 1990; 27:133-144.

73. Rosenbaum D, Becker HP, Sterk J, Gerngross H, Claes L. Functional evaluation of the 10-year outcome after modified Evans repair for chronic ankle instability. Foot.Ankle.Int. 1997; 18:765-771.

74. Sanders R, Fortin P, DiPasquale T, Walling A. Operative treatment in 120 displaced intraarticular calcaneal fractures. Results using a prognostic computed tomography scan classification. Clin.Orthop 1993; 87-95.

75. Sharma S. Applied Multivariate Techniques. New York, NY: John Wiley and Sons, Inc., 1996.

76. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979;86:420-228.

77. Silverstein B, Fisher WP, Kilgore KM, Harley JP, Harvey RF. Applying psychometric criteria to functional assessment in medical rehabilitation: II. Defining interval measures. Arch.Phys.Med.Rehabil. 1992; 73:507-518.

78. Stewart AL, Hays RD, Ware JEJ. The MOS short-form general health survey. Reliability and validity in a patient population. Med.Care 1988; 26:724-735.

79. SAS program for fit of the IRT model. Stone CA. 1998;

80. Anonymous editor. Classification error and goodness-of-fit in IRT models. 1994;

81. Stratford PW, Binkley FM, Riddle DL. Health status measures: strategies and analytic methods for assessing change scores. Phys.Ther. 1996; 76:1109-1123.

82. Takakura Y, Tanaka Y, Sugimoto K, Tamai S, Masuhara K. Ankle arthroplasty. A comparative study of cemented metal and uncemented ceramic prostheses. Clin.Orthop 1990; 209-216.

83. Tarlov AR, Ware JEJ, Greenfield S, Nelson EC, Perrin E, Zubkoff M. The Medical Outcomes Study. An application of methods for monitoring the results of medical care. JAMA 1989; 262:925-930.

84. Teresi JA, Goldman RR, Cross P, Gurland B, Kleinman M, Wilder D. Item bias in cognitive screening measures: Comparisons of elderly white, afro-american, hispanic and high and low education subgroups. Journal of Clinical Epidemiology. 1995; 48:473-483.

85. Tugwell P, Bombardier C, Buchanan W, Goldsmith C, Grace E, Hanna B. The MACTAR patient preference disability questionnaire: An individualized functional priority approach for assessing improvement in physical disability in clinical trials in rheumatoid arthritis. J.Rheumatol. 1987; 14:446-451.

86. Tuley MR, Mulrow CD, McMahan CA. Estimating and testing an index of responsiveness and the relationship of the index to power. J.Clin.Epidemiol. 1991; 44:417-421.

87. Unger AS, Inglis AE, Mow CS, Figgie HE. Total ankle arthroplasty in rheumatoid arthritis: a long-term follow-up study. Foot.Ankle. 1988; 8:173-179.

88. Ware JEJ, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med.Care 1992; 30:473-483.

89. Weinberger M, Oddone EZ, Samsa GP, Landsman PB. Are health-related quality-of-life measures affected by the mode of administration? J.Clin.Epidemiol. 1996; 49:135-140.

90. Weinberger M, Samsa GP, Tierney WM, Belyea MJ, Hiner SL. Generic versus disease specific health status measures: comparing the sickness impact profile and the arthritis impact measurement scales. J.Rheumatol. 1992; 19:543-546.

91. Westaway MD, Stratford PW, Binkley JM. The patient-specific functional scale: validation of its use in persons with neck dysfunction. J.Orthop Sports Phys.Ther. 1998; 27:331-338.

92. Whyte J. Toward a methodology for rehabilitation research. Am.J.Phys.Med.Rehabil. 1994; 73:428-435.

93. Williams JI, Naylor CD. How should health status measures be assessed? Cautionary notes on procrustean frameworks. J.Clin.Epidemiol. 1992; 45:1347-1351.

94. World Health Organization. International Classification of Impairments. Geneva, Switzerland: World Health Organization, 1980.

95. Wright JG, Feinstein AR. A comparative contrast of clinimetric and psychometric methods for constructing indexes and rating scales. J.Clin.Epidemiol. 1992; 45:1201-1218.

96. Wright JG, Young NL. A comparison of different indices of responsiveness. J.Clin.Epidemiol. 1997; 50:239-246.