FOOT & ANKLE INTERNATIONAL

Evidence of Validity for the Foot and Ankle Ability Measure (FAAM)

RobRoy L. Martin, P.T., Ph.D., C.S.C.S.[1]; James J. Irrgang, Ph.D., P.T., A.T.C.[2]; Ray G. Burdett, Ph.D., P.T., C.Ped.[2]; Stephen F. Conti, M.D.[3]; Jessie M. Van Swearingen, Ph.D., P.T.[2]
Pittsburgh, PA

ABSTRACT

Background: **There is no universally accepted instrument that can be used to evaluate changes in self-reported physical function for individuals with leg, ankle, and foot musculoskeletal disorders. The objective of this study was to develop an instrument to meet this need: the Foot and Ankle Ability Measure (FAAM). Additionally, this study was designed to provide validity evidence for interpretation of FAAM scores.** ***Methods:*** **Final item reduction was completed using item response theory with 1027 subjects. Validity evidence was provided by 164 subjects that were expected to change and 79 subjects that were expected to remain stable. These subjects were given the FAAM and SF-36 to complete on two occasions 4 weeks apart.** ***Results:*** **The final version of the FAAM consists of the 21-item activities of daily living (ADL) and 8-item Sports subscales, which together produced information across the spectrum ability. Validity evidence was provided for test content, internal structure, score stability, and responsiveness. Test retest reliability was 0.89 and 0.87 for the ADL and Sports subscales, respectively. The minimal detectable change based on a 95% confidence interval 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 ($p < 0.05$). The minimal clinically important differences were 8 and 9 points for the ADL and Sports subscales, respectively. Guyatt responsive index and ROC analysis found the ADL subscale was more responsive than general measures of physical function while the Sports subscale was not. The ADL and Sport subscales demonstrated strong relationships with the SF-36 physical function subscale ($r = 0.84, 0.78$) and physical component summary score ($r = 0.78, 0.80$) and weak relationships with the SF-36 mental function subscale ($r = 0.18, 0.11$) and mental component summary score ($r = 0.05, -0.02$).** ***Conclusions:*** **The FAAM is a reliable, responsive, and valid measure of physical function for individuals with a broad range of musculoskeletal disorders of the lower leg, foot, and ankle.**

INTRODUCTION

Evaluative self-reported instruments use the response patterns of patients or subjects to measure changes in health status over time. If the instrument is created properly and evidence of validity is obtained, then the information collected can be used to interpret the effect of pathology and subsequent impairment on physical function. Information from this instrument also could be used to compare and assess the effectiveness of treatment interventions. There is no universally accepted self-reported evaluative instrument specific to those with leg, ankle, and foot musculoskeletal disorders.[10] The objective of this study was to develop an instrument to meet this need: the Foot and Ankle Ability Measure (FAAM). Additionally, this study was designed to provide validity evidence for interpretation of FAAM scores.

Four steps can be followed to develop a self-reported evaluative instrument: 1) generation of potential items, 2) initial item reduction, 3) final item reduction, and 4) acquisition of validity evidence to support interpretation of the score. A thorough list of possible items can be generated from a review of the literature, input from experts, and input from a sample of subjects for whom the instrument is intended. Items that do not represent the domain of interest or are repetitive, complex, too narrow in scope, or difficult for subjects or patients to

[1]Duquesne University, Pittsburgh, PA
[2]University of Pittsburgh, Pittsburgh, PA
[3]Human Motion Center of Allegheny Health System, Pittsburgh, PA

Corresponding Author:
RobRoy L. Martin, P.T., Ph.D., C.S.C.S.
Duquesne University
Physical Therapy
600 Forbes Avenue
111A RSHS
Pittsburgh, PA 15282
E-mail: martinr280@duq.edu
For information on prices and availability of reprints, call 410-494-4994 X226

interpret are removed during the initial item reduction. The objective of these first two steps is to prepare an interim instrument that can undergo psychometric analysis for final item reduction.

Psychometric procedures involving item response theory (IRT) can be used to complete final item reduction. The basic concept behind IRT is that the probability of choosing a response for each item is a function of the subject's or patient's ability and the difficulty level of each item.[3,4,7,9] The procedures for IRT involve constructing item characteristic curves which represent the probability of choosing a response for each item based on the subject's or patient's ability.[7] For example, an item with five potential responses, each response describing a level of proficiency for the activity in question, should have a characteristic curve consisting of five distinct and separate curves. Each of the five curves should have one peak and together the curves should span the spectrum of ability. Such an item should be responsive to changes in the status of the patient across the spectrum of ability. Individual items from the interim instrument can be selected for inclusion on the final instrument based on the appearance of their respective item characteristic curves.

In addition to item characteristic curves, IRT provides the amount of information that each item contributes at varying levels of ability.[7] Items that assess an individual's ability to perform activities that are easy and require a low level of ability should provide information in the low range of ability. Conversely, items that assess activities that are more challenging and require a high level of ability should provide information in the higher range of ability. Item information values can be summed to provide test information, which describes how much information the entire instrument provides across the spectrum of ability.[7] The more information an instrument provides, the more precise the instrument will be with less associated error of estimation.[7] The target test information function for an evaluative instrument should provide information across all ability ranges.[7] Therefore, an appropriate evaluative instrument should contain items that assess an individual's ability to perform activities that span from easy to more challenging. Items can be included or eliminated from the final instrument based on their ability to contribute information.

A number of assumptions need to be met for the results of the item characteristic curves and information functions to be valid. These assumptions include unidimensionality, local independence, administration of the test is not under time constraints, and guessing for a correct answer is not an option.[7] Unidimensionality implies that the instrument measures a single latent ability. Local independence implies that an individual's response to one question is independent of the responses to other questions after the individual's level of ability has been taken into account. This implies that only one latent ability accounts for the individual's response for each of the items contained on the instrument.[7] As such, factor analysis should demonstrate that the instrument is unidimensional. If it is determined that the items contained on the instrument represent one factor, then the assumptions of unidimensionality and local independence will likely be met.[18] Items from the interim instrument can be selected or eliminated based on their ability to fit into a one-factor model. Once IRT procedures are completed, items that have response characteristic curves and contribute information across the spectrum of ability should be included on the final instrument.

Validity evidence for this instrument needs to be obtained so that scores can be meaningfully interpreted. Interpreting the scores from an evaluative instrument requires evidence of the following: 1) items on the instrument represent the domain of interest (evidence of test content and internal structure), 2) scores remain stable when the underlying condition measured by the instrument remains stable (evidence of stability or test re-test reliability), 3) scores are related to other measures of the same or similar construct while not being unduly related to measures of different constructs (evidence of convergent and divergent validity), and 4) scores change with improvement or deterioration of the condition measured by the instrument (evidence of responsiveness).[12]

This study presents the methods and results associated with the four steps outlined for the development of the FAAM: 1) generation of potential items, 2) initial item reduction, 3) final item reduction, and 4) acquisition of evidence of validity. The goal of this project was to produce an instrument containing items that comprehensively assess physical performance of individuals with a broad range of leg, ankle, and foot musculoskeletal disorders. Additionally, this project was designed to provide evidence to support the use of the FAAM and allow meaningful interpretation of obtained scores.

METHODS

Generating Potential Items and Initial Item Reduction

A thorough list of potential items relating to symptoms, signs, and limitations in physical function associated with musculoskeletal disorders of the leg, ankle and foot was generated from a literature review and from input from physical therapists who treat individuals with foot and ankle related disorders, as well as from individuals with musculoskeletal leg, ankle, and foot pathology. Expert clinicians from the American Physical Therapy Association (APTA) Foot and Ankle Special

Interest Group participated in initial item reduction. These clinicians were mailed the list of potential items and 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 1 (important) or above were considered for inclusion on the interim instrument. The clinicians also were asked if they thought there should be two scales, one for activities of daily living and one for sporting activities.

Final Item Reduction

Subjects for Final Item Reduction

Potential subjects consisted of patients who were referred to physical therapy by a physician and were receiving treatment for a leg, ankle, or foot musculoskeletal disorder at one of the 45 Centers for Rehab Service's (CRS) outpatient orthopaedic clinics in southwestern Pennsylvania. Subjects who were receiving physical therapy treatment for coexisting musculoskeletal pathology in another body region were excluded from the study. There were 1027 subjects who received treatment from April, 1997, to December, 1999, and were included in this phase of the study. The average age of these subjects was 42.0 years (SD 17.39, median 42.81, range 8 to 83 years). Six hundred twenty-nine (61.2%) individuals were women, and 391 (38.1%) were men. Gender was not reported for 7 (0.7%) individuals. Duration of symptoms averaged 3.7 months (SD 8.55 months, median 1.45 months, range 1 day to 7.88 years). Using ICD-9 codes, diagnoses were categorized as follows: 193 (18%) ankle joint pathologies, 321 (31.3%) sprains/strains, 113 (11.9%) heel pathologies, 151 (14.7%) fractures, 37 (3.6%) forefoot pathologies and 87 (8.5%) nonspecific leg pain. Diagnosis could not be determined from analysis of information in the database for the remaining 125 (12%) subjects. The Institutional Review Board approved the use of subjects for final item reduction.

Procedure for Data Collection

Scores from the interim FAAM obtained on initial evaluation as well as demographic information could be extracted from the CRS clinical outcomes database. This information was collected as part of routine patient care.

Assumptions of Unidimensionality

Exploratory factor analysis was completed on the FAAM using PRELIS (Scientific Software International, Chicago, IL). Eigenvalues and factor loading patterns were used to identify and extract factors. The amount of variation explained by each factor is indicated by its eigenvalue. Items from the interim FAAM that did not fit a one-factor model were eliminated. Factor loading patterns were used to identify these items.

Model Fit

Multilog (Scientific Software Inc., Chicago IL) was used to calibrate items for the two-parameter graded response and one-parameter partial credit models. A likelihood ratio was obtained for each model. The fit of the one- and two-parameter models were compared using the difference in the negative twice the log likelihood statistics.[3] If the observed difference was greater than the critical value, the additional parameters estimated in the graded response model contributed significantly to model fit. At least 500 subjects were required to properly estimate model parameters and complete this analysis.[13]

Item Characteristic Curves

Item characteristic curves were constructed for each item using the item difficulty and discrimination parameters that were generated by Multilog. It was hypothesized that each item would have five distinct and separate response characteristic curves. Each curve would have one peak and together the five curves would span the spectrum of ability. Items from the interim instrument that did not have appropriate item characteristic curves were eliminated.

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. The item information values for each item at the nine ability levels were summed to produce the test information function.

Evidence of Validity

Subjects for Evidence of Validity

Two groups of subjects were used to provide validity evidence to support interpretation of the score, a group of subjects that were expected to change and a group that was expected to remain stable. The inclusion criterion for all subjects was that they received physical therapy treatment longer than 4 weeks for a leg, foot, or ankle musculoskeletal disorder. Subjects were excluded if they were receiving physical therapy treatment for coexisting musculoskeletal pathology in another body region. Subjects for the group expected to change consisted of 164 individuals receiving physical therapy treatment between July, 2002, and January, 2003 at one of the CRS clinics.

A list of potential subjects for the group expected to remain stable was obtained from the CRS database. This database was used to identify patients who were treated at least 1 year ago for a leg, foot, or ankle musculoskeletal disorder. One hundred and eighty potential subjects who were treated between January, 2000, and October, 2001, were identified for the group that was expected to remain stable. Using

mailing procedures outlined by Dillman,[3] information was obtained from 79 (42%) of the subjects in the group that was expected to remain stable. Age, gender, and duration of symptoms for the two groups are reported in Table 1. The physical therapy chart and ICD-9 codes were used to determine diagnoses for these subjects. The diagnosis profiles for the two groups also are reported in Table 1.

Fifty-one (21.2%) subjects, 38 individuals in the group expected to change and 13 individuals in the group expected to remain stable, reported undergoing foot and ankle-related surgery. These surgeries included open reduction and internal fixation (nine), ligament reconstruction (six), Achilles tendon repair (six), arthrodesis (five), bunionectomy (four), arthroscopic debridement (three), and total ankle replacement (one). Seven subjects reported having surgery for plantar fasciitis and 10 subjects reported nondescript foot, ankle, or toe surgery. The average time from surgery to the initial data collection in the group expected to change was 68 (SD 60, range 1 to 364) days. This information was not available for the group expected to remain stable. The Institutional Review Board approved the use of subjects to obtain this validity evidence.

Table 1: Demographic information and diagnoses profile for the two groups of subjects used to obtain evidence of validity

	Group Expected to Change	Group Expected to Remain Stable
Mean Age(years)	41.2	45.2
SD	16.3	15
Range	9–75	19–86
Gender		
Male	97	47
Female	67	32
Duration of symptoms (months)	4	4.7
SD	3.6	2.4
Range (days-years)	2–2.2	1–3.8
Diagnoses profile		
Joint/limb pain	55	39
Sprain/strain	47	24
Fractures	28	5
Plantar fasciitis	22	5
Bunion	3	1
Achilles rupture	2	0
Other	4	1
Missing	3	0

Procedure for Data Collection for Subjects Expected to Change

The FAAM and SF-36 were obtained on two separate occasions, approximately 4 weeks apart. These data were collected as part of routine patient care and could be extracted from the CRS clinical outcomes database. This included demographic information as well as initial and 4-week FAAM and SF-36 scores. Additionally, the response to the question, ''Over the past 4 weeks, how would you rate the overall change in your physical ability?'' was collected from the database. This question had seven potential responses: much worse, worse, slightly worse, no change, slightly improved, improved, and much improved.

Procedure for Data Collection for Subjects Expected to Remain Stable

The FAAM, SF-36, and global rating information was obtained on two separate occasions, approximately 4 weeks apart, using mailing procedures outlined by Dillman.[4] On the second mailing, responses to the question, ''Over the past 4 weeks, how would you rate the overall change in your physical ability?'' also were collected.

Evidence of Internal Structure

Exploratory factor analysis, using PRELIS, was completed separately for the group expected to remain stable and the group expected to change. Initial FAAM scores were used for this analysis.

An assessment of internal consistency was done using SPSS (Version 11.5, SPSS inc., Chicago, IL.) to calculate Cronbach's coefficient alpha. The initial FAAM scores were used for the assessment of internal consistency. The standard error of measure (SEM) to indicate the precision of measurement at a single point in time was calculated as $SEM = \sigma\sqrt{1 - r}$ where σ was the standard deviation of the scores and r was the coefficient alpha. A 95% confidence interval (CI) was then calculated to determine the measurement precision associated with a measure at a single point in time. If the factorial structure of the group expected to change and that of the group expected to remain stable were the same, then the groups could be combined for the analysis of internal consistency. If the factorial structure was different between these groups, then Cronbach's alpha would be calculated for both groups separately.

Evidence for Score Stability

Intra-class correlation coefficients (ICC 2,1)[15] using the initial and 4-week FAAM scores in the group that was expected to remain stable were calculated to provide evidence for score stability (test re-test reliability). To

estimate measurement precision associated with repeat measurements over an interval of approximately 4 weeks, we calculated the SEM using the ICC test retest reliability coefficient. The SEM was multiplied by $\sqrt{2}$ and a 95% CI was calculated to determine the minimal detectable change (MDC).[17]

Evidence of Responsiveness

Three analyses were done to assess responsiveness of the FAAM. Group level assessments of responsiveness included use of a two-way ANOVA with repeated measures and calculation of Guyatt's responsiveness index (GRI).[6] Assessment of responsiveness at the individual level was done by constructing of ROC curves.[1] A two-way ANOVA with repeated measures was calculated with SPSS, comparing initial and final FAAM scores of the groups that were expected to change and remain stable. It was hypothesized that the difference between initial and 4-week scores in the group expected to change would be greater than the difference over a similar period of time in the group that was expected to remain stable and therefore, a significant interaction would be found. The a priori alpha level for this analysis was set at 0.05.

The GRI was calculated by dividing the average change in score over the 4-week period in the group expected to change by the standard deviation of scores over the 4-week period in the group expected to remain stable.[5,19] It was hypothesized that the 95% CI for the GRI would not contain zero. In addition, to determine if the FAAM was more responsive than the SF-36 physical function and physical component scores, we calculated the difference and the associated 95% CI for the difference between the FAAM scores and the SF-36 physical function and physical component summary scores.

An assessment of responsiveness was done at the individual level to determine a change in score that could be interpreted as the minimal clinically important difference (MCID). The criterion used to determine the MCID was whether the patient perceived himself or herself to be improved or not improved after 4 weeks of physical therapy. For this analysis the group expected to change was dichotomized based on how subjects perceived their change in status between the initial and 4-week 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 improved'' (75 or 45.7%) or ''improved'' (42 or 25.6%) and were placed in the improved group. Thirty-one (18.9%) subjects described their change in functional status to be ''slightly improved'' (24 or 14.6%), ''unchanged'' (five or 3.0%) or ''slightly worse'' (2 or 1.2%) and were placed in the group that did not improve. SPSS calculated the sensitivity and specificity for each one point change in score. These values were then plotted to construct the ROC curves.

Evidence of Convergent and Divergent Validity

Convergent evidence was examined by assessing the associations between the FAAM and SF-36 physical function and physical component summary scores using Pearson correlation coefficients. Divergent evidence was examined by assessing the associations between the FAAM and SF-36 mental health and the mental health component summary scores.

Differences in the level of association between the variables that measure similar constructs and the variables that measure different constructs also were examined. Testing for differences in the correlation coefficients between the FAAM and concurrent measures of physical function and mental health was done based on the equation by Meng et al.[11] These calculated values were compared to a critical t value of 3.34 for alpha = 0.001 at 200 degrees of freedom. The a priori type I error rate was set at 0.001 to account for the multiple comparisons. The initial scores from subjects in both groups were used in the analysis of convergent and divergent validity.

Sample Size Estimation

Sample size estimation was done to ensure adequate power to assess for differences between two sample correlations. The sample of 1027 subjects used in the final item reduction was used to estimate the correlation between the FAAM and the SF-36. The correlation between the FAAM and the physical function and mental health subscale scores were 0.63 and 0.20 respectively. To account for multiple comparisons, the alpha level was set at 0.001. Based on this analysis, approximately, 220 subjects were required to detect a significant difference between the correlation values of 0.63 and 0.20, a power level of 80% using a one-tailed test.

RESULTS

Initial Item Reduction

Item generation produced 69 potential items, which were mailed to the members of the APTA Foot and Ankle Special Interest Group to be rated. Twenty-nine out of 43 (67.4%) surveys were returned. After reviewing all potential items, a decision was made to eliminate the items that did not assess physical performance. 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. Items relating to symptoms, the need for medication, cosmesis, ability to wear different shoes, and psychological limitations were eliminated. Of the

remaining items, 34 had a mean rating of one or greater and were believed to be important by the expert clinicians. Forty (94%) respondents thought there should be two separate scales, one for activities of daily living and one for sports. These 34 potential items were therefore divided into two subscales. The Activities of Daily Living (ADL) subscale contained 26 items pertaining to basic functional activities and pain. The Sports subscale contained eight items pertaining to higher level activities, such as those required in athletics.

Field-testing was done with the interim FAAM using 20 patients to ensure that the instrument was user-friendly for both clinicians and patients. An effort was made to ensure that the FAAM was easy to administer, complete, and score. This included making sure the directions were clear and easy to understand.

Final Item Reduction

Examining the Assumptions of Item Response Theory

PRELIS requires the use of complete data with no missing responses. Therefore, 659 (64.2%) of the 1027 subjects were used to evaluate both the ADL and Sports subscales. Factor analysis of the 26-item interim FAAM ADL subscale indicated that the items loaded on two factors. These two factors accounted for 75.0% of the variance and had Eigen values of 17.22 and 2.29. The factor loadings of each item on the first two principal components are reported in Table 2. Items 23 through 26 were items that differed from the other items as they pertained to pain. The items had high factor loadings on the second principal component and therefore were considered for elimination in an effort to allow the ADL subscale to conform to a one-factor model.

The factor analysis was repeated with items 23 through 26 omitted. The 22 items on the interim ADL subscale loaded on one factor which accounted for 74.09% of the variance and had an eigenvalue of 16.30. The factor loading of each item to the first principal component is found in Table 3. The resulting 22-item ADL subscale was then used for assessing model fit, the response characteristic curves, and test information function.

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 factor loadings of each item to the first principal component are found in Table 4.

Table 2: Factor loadings for each item to the first two principal components for the 26-Item ADL subscale

Item content	PC 1	PC 2
1) Standing	0.83	0.10
2) Walking on even ground	0.89	−0.02
3) Walking on uneven ground without shoes	0.86	−0.02
4) Walking up hills	0.93	−0.11
5) Walking down hills	0.92	−0.12
6) Going up stairs	0.90	−0.10
7) Going down stairs	0.88	−0.09
8) Walking on uneven ground	0.90	−0.12
9) Stepping up and down curbs	0.88	−0.09
10) Squatting	0.80	−0.19
11) Sleeping	0.65	0.33
12) Coming up on your toes	0.75	−0.22
13) Walking initially	0.82	−0.01
14) Walking 5 minutes or less	0.90	−0.13
15) Walking approximately 10 minutes	0.90	−0.16
16) Walking 15 minutes or greater	0.89	−0.16
17) Home responsibilities	0.89	−0.07
18) Activities of daily living	0.86	−0.01
19) Personal care	0.80	−0.01
20) Light to moderate work (standing and walking)	0.92	−0.03
21) Heavy work (push/pulling, climbing, carrying)	0.87	−0.16
22) Recreational activities	0.76	−0.17
23) General level of pain	0.50	0.73
24) Pain at rest	0.52	0.71
25) Pain during your normal activity	0.65	0.59
26) Pain first thing in the morning	0.40	0.71

PC = principal component.

Model Fit

The modified 22-item ADL and eight-item Sports subscales were calibrated separately. 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. 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 was greater than the

Table 3: Factor loadings of the items to the first principal component for the 22-Item ADL subscale

Item content	PC 1
1) Standing	0.82
2) Walking on even ground	0.89
3) Walking on uneven ground without shoes	0.87
4) Walking up hills	0.93
5) Walking down hills	0.92
6) Going up stairs	0.91
7) Going down stairs	0.89
8) Walking on uneven ground	0.91
9) Stepping up and down curbs	0.89
10) Squatting	0.81
11) Sleeping	0.66
12) Coming up on your toes	0.77
13) Walking initially	0.82
14) Walking 5 minutes or less	0.91
15) Walking approximately 10 minutes	0.91
16) Walking 15 minutes or more	0.89
17) Home responsibilities	0.89
18) Activities of daily living	0.86
19) Personal care	0.79
20) Light to moderate work (standing and walking)	0.92
21) Heavy work (push/pulling, climbing, carrying)	0.88
22) Recreational activities	0.77

PC = principal component.

critical value of 32.67 for $p = 0.05$ and 21 degrees of freedom. The negative twice log likelihood statistics for the Sports subscale were −977.6 and −860.8 for the one- and two-parameter models respectively. The observed difference of 116.8 was greater than the critical value of 14.07 for $p = 0.05$ and 7 degrees of freedom.

Item Characteristic Curves

The item parameters generated by Multilog were entered into an Excel spread sheet to create the item characteristic curves for each of the items on the ADL and Sports subscales. Inspection of the item characteristic curves for the ADL subscale revealed all items, except items 11 (sleeping) and 19 (personal care), had response curves that spanned the spectrum of ability. The item characteristic curve for item 6 (going up stairs) is an example of an item that had response curves that spanned the spectrum of ability and is presented in Figure 1. The item characteristic curve for item 11 is presented in Figure 2. Item characteristic curves also were plotted for eight items on the interim Sports subscale. All eight had response curves that were similar to that displayed in Figure 1.

Table 4: Factor loadings for each item to the first principal component for the Sports subscale

Item content	PC 1
1) Running	0.94
2) Jumping	0.96
3) Landing	0.95
4) Starting and stopping quickly	0.91
5) Cutting/lateral movements	0.92
6) Low impact activities	0.93
7) Ability to perform activity with your normal technique	0.92
8) Ability to participate in your desired sport as long as you would like	0.90

PC = principal component.

Target Test Information Function

The test information function for the modified 22-item ADL subscale is shown in Figure 3. Most information was supplied at the lower end of ability for the ADL subscale. Items 11 (sleeping) and 19 (personal care) were noted to give the most information at the lower end of ability and because of this, a decision was made to consider eliminating items 11 and 19. Item 11 was deleted first because it had the lowest factor loading to the first principle component (Table 3). The test information function, after deleting item 11, did not substantially change. Item 19 was then deleted. The test information function was recalculated, and a decrease in information was noted throughout the range of ability. Therefore, item 19 was retained to maximize the instrument's precision of measurement across the range of ability. The test information function for the eight-item Sports subscale is shown in Figure 4. As expected, the test information function provided most information in the higher ability ranges.

The final version of the 21-item ADL and eight-item Sports subscale can be found in appendix 1. In an effort to prevent potential problems associated with missing data, scores for the FAAM ADL and Sports subscales were generated only when subjects completed 90% or more of the items (19 of 21 for the ADL and seven of eight for the sports subscales). The ADL and Sports subscales

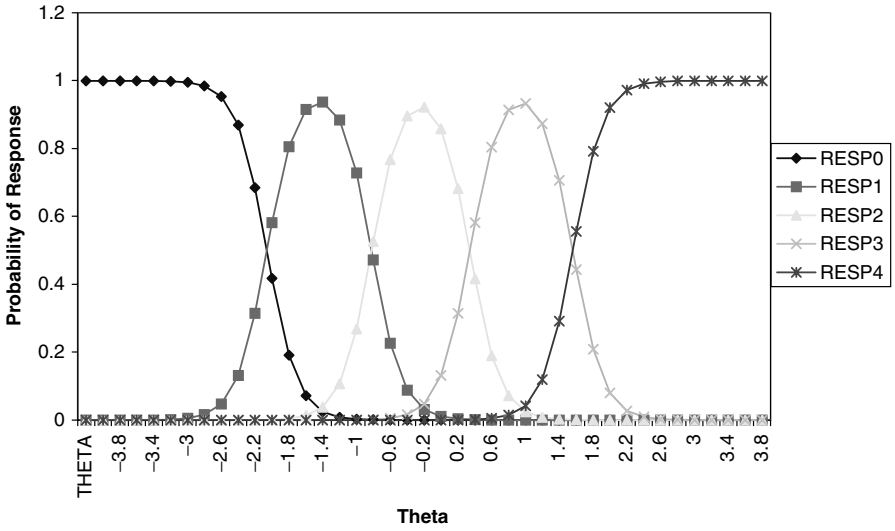

Fig. 1: Item characteristic curve for item 6 (walking up stairs).

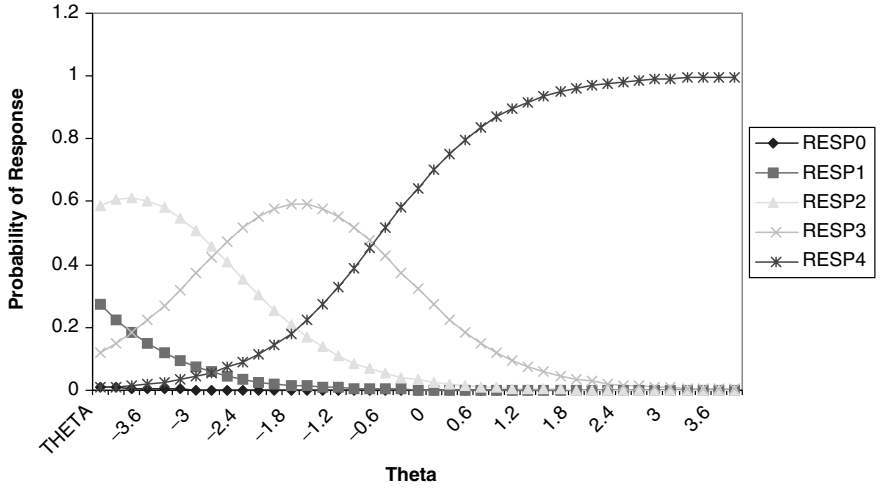

Fig. 2: Item characteristic curve for item 11 (sleeping).

were scored separately. The response to each item on the ADL subscale was scored from 4 to 0, with 4 being ''no difficulty'' and 0 being ''unable to do.'' Responses marked as not applicable were not counted. The scores on each of the items were added together to get the item score total. The total number of items with a response was multiplied by 4 to get the highest potential score. If all 21 items were answered, the highest potential score was 84. If one item was unanswered the highest score was 80, if two were unanswered the total highest score was 76. The total item score was divided by the highest potential score and then multiplied by 100 to produce the ADL score that ranged from 0 to 100. The Sports subscale was scored in a similar manner. If all eight items were answered the highest potential score was 32. If one item was unanswered the highest potential score was 28. As with the ADL subscale, the item score total was divided by the highest potential score and multiplied by 100. A higher score represents a higher level of physical function for both the ADL and Sports subscales.

Results for Evidence of Validity

Subjects Expected to Change

Subjects were included in the analysis of the ADL subscale if they responded to at least 19 of the 21 items. Out of the 164 subjects, 151 (87%) met this criteria. For analysis of the Sports subscale, subjects were included if they responded to at least seven of the eight items. Of the 164 subjects, 130 (79.2%) met this

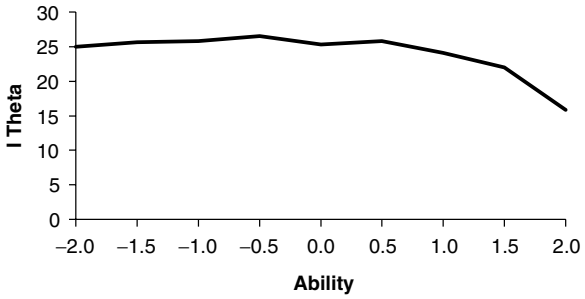

Fig. 3: Test information function for the 22-Item ADL subscale.

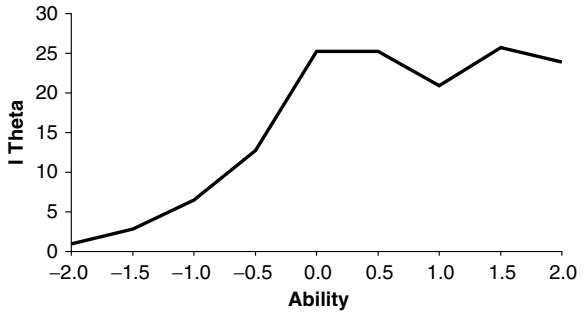

Fig. 4: Test information function for the Sports subscale.

criteria. The average score for the initial ADL and Sports subscales were 58.0 (SD 24.8, median 59.5, range 5 to 98) and 25.2 (SD 26.7, median 15.0, range 0 to 94), respectively. The average time between completing the initial and final surveys was 32.3 days (SD 12.1, median 28, range 23 to 106 days). The average 4-week ADL and Sports subscale scores were 74.9 (SD 20.0 median 77.5 range 13–100) and 43.9 (SD 30.0, median 45.1, range 0 to 100), respectively.

Subjects Expected to Remain Stable

Of 79 subjects from whom initial baseline information was obtained, 74 (93.6%) answered 19 of the 21 items on the ADL subscale and 70 (88.6%) answered seven of the eight items on the Sports subscale. The average score for initial ADL and Sports subscale scores were 91.5 (SD 13.6, median 97.5, range 37 to 100) and 78.6 (SD 23.8, median 88.4, range 13 to 100), respectively. The time period between the initial and followup surveys averaged 65.6 (SD 19.8, median 67.0, range 31 to 101) days. The average score for final ADL and Sports subscale scores were 92.6 (SD 13.2, median 97.6, range 27 to 100) and 81.9 (SD 23.3, median 93.8, range 13 to 100), respectively.

Evidence of Internal Structure

PRELIS requires use of complete data with no missing responses. Therefore, in the group expected to change, 112 (68.3%) of the 164 subjects and 106 (81.5%) of the 130 subjects were used to evaluate the ADL and Sports subscales respectively. In the group expected to remain stable, 61 (77.2%) of the 79 subjects were used to evaluate both the ADL and Sports subscales.

Exploratory factor analysis found the items on the ADL subscale loaded on one factor in the group expected to change. This single factor accounted for 80.46% of the variance and had an eigenvalue of 16.90. In the group expected to remain stable the items loaded on two factors. The first factor 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 factor loadings for the group expected to change and the group expected to remain stable are shown in Table 5. No attempt was made to perform a factor analysis on the combined sample because the factorial structure was different between the two groups.

The principal component analysis of the Sport subscale in the group expected to change revealed all items loaded on one factor that accounted for 86.7% of the variance and had an eigenvalue of 6.94. In the group expected to remain stable, all items also loaded on one factor that accounted for 86.4% of the variance and had an eigenvalue of 6.91. Factor analysis was performed combining data from both groups with an effective sample size of 167 (83.5%) of the total 200 subjects. The resulting principal component analysis found the items loaded on one factor that accounted for 93.48% of the variance and had an eigenvalue of 7.48. The factor loadings for each of these analyses are presented in Table 6.

The assessment of internal consistency for the ADL subscale was done separately for the groups that were expected to change and for those expected to remain stable. In the group that was expected to change, coefficient alpha was 0.98 with a SEM of 3.5 and a 95% CI of $+/-6.9$ points. In the group expected to remain stable, coefficient alpha was 0.96 with a SEM of 2.7 and a 95% CI of $+/-5.3$. The assessment of internal consistency for the Sports subscale was done using data from the combined samples yielding coefficient alpha of 0.98 with a SEM of 5.1 with a 95% CI of $+/-10.0$.

Evidence of Score Stability

The ICC(2,1) for the ADL subscale was ICC(2,1). 89 with a SEM of 2.1 and the MDC based upon the 95% CI was $+/-5.7$ points. The ICC(2,1) for the Sports subscale was. 87 with a SEM of 4.5 and the MDC based on the 95% CI was $+/-12.3$ points.

Responsiveness to Change in Functional Status

The average difference between the initial and 4-week ADL subscale scores in the group that was expected

Table 5: Factor loadings for the items to the principal components for the ADL subscale

	Group expected to change	Group expected to remain stable	
Item Content	**PC1**	**PC1**	**PC2**
1) Standing	0.81	0.85	−0.14
2) Walking on even ground	0.90	0.87	0.05
3) Walking on uneven ground without shoes	0.84	0.70	−0.16
4) Walking up hills	0.93	0.88	−0.18
5) Walking down hills	0.93	0.87	−0.18
6) Going up stairs	0.90	0.87	−0.16
7) Going down stairs	0.92	0.89	−0.15
8) Walking on uneven ground	0.90	0.79	0.03
9) Stepping up and down curbs	0.92	0.73	−0.09
10) Squatting	0.84	0.68	0.23
11) Coming up on your toes	0.78	0.81	0.12
12) Walking initially	0.83	0.78	−0.11
13) Walking 5 minutes or less	0.90	0.79	−0.34
14) Walking approximately 10 minutes	0.91	0.80	−0.28
15) Walking 15 minutes or more	0.89	0.76	−0.12
16) Home responsibilities	0.90	0.79	0.35
17) Activities of daily living	0.82	0.78	0.04
18) Personal care	0.71	0.28	0.88
19) Light to moderate work (standing and walking)	0.86	0.88	0.19
20) Heavy work (push/pulling, climbing, carrying)	0.87	0.82	0.38
21) Recreational activities	0.67	0.71	0.29

PC = principal component.

Table 6: Factor loadings for each item to the principal components for the Sports subscale

Item Content	**Group expected to change PC 1**	**Group expected to remain stable PC 1**	**Groups combined PC 1**
1) Running	0.89	0.92	0.95
2) Jumping	0.92	0.93	0.96
3) Landing	0.92	0.90	0.96
4) Starting and stopping quickly	0.91	0.83	0.94
5) Cutting/lateral movements	0.91	0.86	0.94
6) Low impact activities	0.87	0.81	0.91
7) Ability to perform activity with your normal technique	0.87	0.90	0.93
8) Ability to participate in your desired sport as long as you would like	0.83	0.88	0.92

PC = principal component.

Table 7: Analysis of variance summary table comparing the initial and final ADL subscale scores

Source	Type III sum of squares	df	Mean square	F-value	Significance
Time	6012.455	1	6012.455	41.074	$p < .001$
Time* Group	6230.304	1	6230.304	42.562	$p < .001$
Error	29568.930	202	146.381		

df = degrees of freedom.

Table 8: Analysis of variance summary table comparing the initial and final Sports subscale scores

Source	Type III sum of squares	df	Mean square	F-value	Significance
Time	5289.264	1	5289.264	22.377	$p < .001$
Time * Group	5310.420	1	5310.420	22.466	$p < .001$
Error	39001.365	165	236.372		

df = degrees of freedom.

to change was 17.1 (SD 19.88, median 12.59, range −25 to 81). The average difference between the initial and 4-week ADL subscale scores in the group that was expected to remain stable was −0.2 (SD 6.21, median 0.00, range −19 to 24). The analysis of variance summary table is present in Table 7. The group by time interaction was significant ($F(1, 202) = 42.562$ $p < 0.001$). The average change on the sports subscale in the group that was expected to change was 17.2 (SD 24.8, median 11.6, range −34 to 97). The average change in the group that was expected to remain stable was 0.0 (SD 12.3, median 0.00, range −28 to 33). The analysis of variance summary table is presented in Table 8. The group by time interaction was significant ($F(1, 165) = 22.466$ $p < 0.001$).

The GRI for the ADL subscale was 2.75 with a 95% CI ranging from 2.02 to 3.48. The GRI for the Sports subscale was 1.40 with a 95% CI ranging from 0.93 to 1.86. Differences and 95% confidence intervals for the differences between GRIs were as follows: 1) ADL subscale and SF-36 physical function subscale 0.98 (0.27,1.69), 2) ADL subscale and SF-36 physical component summary score 1.63 (1.02,2,24), 3) Sports subscale and SF-36 physical function subscale −0.38 (−1.00,0.25), and 4) Sports subscale and SF-36 physical component summary score 0.27 (−.022, 0.77).

The ROC analyses using the subjects in the group expected to change and the dichotomy (i.e. improved vs. not improved after 4-weeks of physical therapy) as the criterion measure are presented in Figures 5 and 6

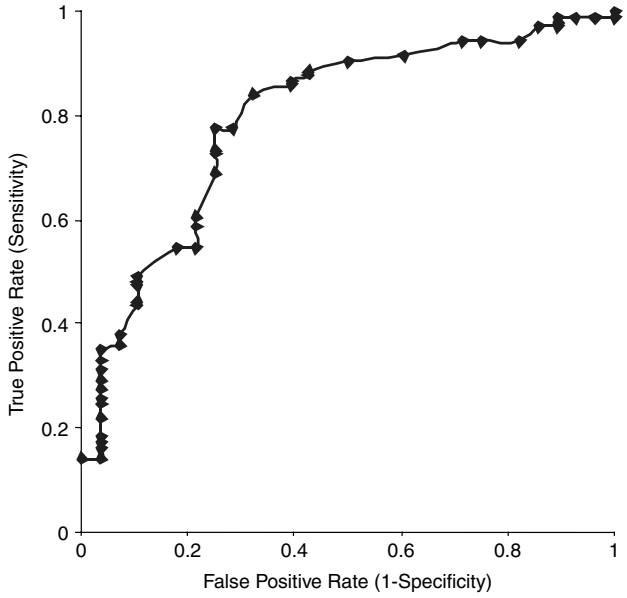

Fig. 5: The ROC curve of the ADL subscale using the dichotomized group expected to change.

for the ADL and Sport subscales, respectively. The area under the ROC curve for the ADL subscale was 0.80 with a 95% CI ranging from 0.89 to 0.71. The change score that best distinguished between a patient that perceived himself or herself to be improved from a patient that did not perceive himself or herself to be

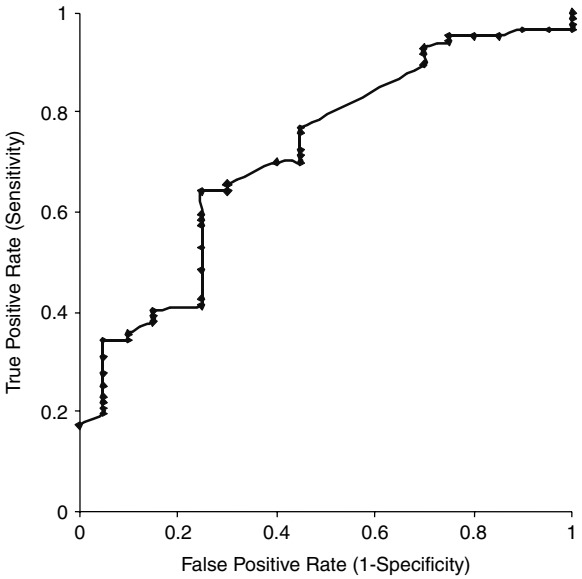

Fig. 6: ROC curve of the Sports subscale using the dichotomized group expected to change.

improved was eight 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 eight-point change in the ADL subscale score.

The area under the ROC curve for the Sport subscale was 0.72 with a 95% CI ranging from 0.78 to 0.66. The change score that best distinguished an improved from an unimproved patient after 4 weeks of physical therapy was nine 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 nine-point change score.

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

The correlation coefficients between the ADL and Sports subscales and concurrent measures of physical and mental health are presented in Table 9. The calculated t-values to assess the difference in the correlation coefficients between the ADL and Sports subscales to measures of physical and mental functioning are presented in Tables 10 and 11, respectively.

Table 9: Correlation coefficients between the ADL and sports subscales to concurrent measures of physical and emotional function

	ADL subscale	Sports subscale
Physical Function Subscale	0.84	0.78
Physical Component Summary Score	0.84	0.80
Mental Health Subscale	0.18	0.11
Mental Component Summary Score	0.05	−0.02

DISCUSSION

The FAAM was developed to meet the need for a self-reported evaluative instrument that comprehensively assesses physical function of individuals with musculoskeletal disorders of the leg, foot, and ankle. These results indicate that the FAAM is a reliable, valid, and responsive measure of self-reported physical function for individuals participating in physical therapy, with or without operative intervention, for a broad range of musculoskeletal disorders of the leg, foot, and ankle. Specifically this study provided evidence that the FAAM

Table 10: Comparison of relationship between ADL subscale and concurrent measures of physical and mental functioning

Relationship of ADL subscale with:	Global rating of function	Physical function subscale	Physical component summary score
Mental Health Subscale	14.3*	14.4*	15.5*
Mental Health Component Summary Score	14.9*	16.1*	18.2*

Note. Values in the cells are the t-values to compare the relationship of the ADL subscale with measures of physical and mental health.
*$p < 0.001$.

Table 11: Comparison of relationship between Sports subscale and concurrent measures of physical and mental functioning

Relationship of Sports subscale with:	Global rating of function	Physical function subscale	Physical component summary score
Mental Health Subscale	19.3*	10.5*	12.1*
Mental Health Component Summary Score	22.3*	11.8*	11.4*

Note. Values in the cells are the t-values to compare the relationship of the Sports subscale with measures of physical and mental health.
*$p < 0.001$.

ADL and Sports subscales contain items that represent the domain of interest, scores remain stable when the underlying condition remains stable, scores relate to other measures of the same or similar construct and did not relate to measures of a different construct, and scores change as the individual's status changes.

The procedures used to develop the FAAM incorporated IRT. The methods for IRT differ from classic methods used to develop evaluative instruments. Classic test development assesses the instrument as a unit where IRT allows for individual item assessment.[7] Typically, when instruments are constructed using classic methods, items are selected based on what the developers of the instrument deem should be on the instrument. Using classic methods, it can be difficult to assess the quality of information each item supplies as the patient's or subject's characteristics cannot be differentiated from the characteristics of the test.[7] Also, classic test theory offers no means to select or remove items based on their contribution of information to the instrument.[7] Using IRT, the response characteristic curves and item information were analyzed individually for each item. These kinds of analyses should improve the quality and precision of the information obtained by the instrument. The results of IRT and assessment of individual item importance by expert clinicians provide validity evidence for the content and internal structure of the FAAM. The results of IRT analysis also support concurrent use of the ADL and Sports subscales to collect information throughout the range of ability.

In addition to evidence based on test content and internal structure, evidence to support the interpretation of scores also were obtained by estimating the error associated with measurement at a single point in time as well as the MDC and MCID. These values were 6.9, 5.7, and 8 points and 10, 12.3, and 9 points, respectively, for the ADL and Sports subscales. How a clinician or researcher implements the instrument will determine which of these values are appropriate to use. With a patient who on initial evaluation scores a 60 on the ADL subscale, given that the error associated with a score at a single point in time is $+/-6.9$ for the ADL subscale, the clinician can be 95% confident that the true score for this patient is between 66.9 and 53.1. We also can be 95% confident that other individuals who score between 66.9 and 53.1 have the same true score as the patient with an observed score of 60.

The MDC and MCID need to be taken into account when setting goals for treatment and when trying to determine if a patient's score changed over a period of time. A change in the ADL score would be considered greater than measurement error associated with repeated measurements if the change score exceeded the MDC. Thus, for a patient with an initial score of 60, changes in subsequent scores would be considered greater than measurement error if they exceeded 65.7. Given the MCID for the ADL scale, if a patient's change score exceeds 8, there is a high likelihood that the patient would considered himself or herself to be improved. Conversely, if the change score is less than 8, there is a high likelihood the patient would consider himself or herself not to have improved. In interpreting the meaning of ''improved,'' it is important to remember that we operationally defined this as the patient's perceived change in status of ''much improved'' or ''improved'' after approximately 4 weeks of physical therapy. The error associated with measurement at a single point in time, MDC, and MCID for the sports scale should be interpreted in a similar manner.

Sensitivity and specificity need to be considered when evaluating the MCID value. Sensitivity is the proportion of subjects who had a meaningful clinical change and also had a change in score equal to or greater than the MCID value. Specificity is the proportion of subjects who did not undergo a meaningful clinical change and had a change in score below the MCID value. A perfect instrument would have a MCID score that had a sensitivity and specificity of one and therefore a 100% accuracy of determining if an individual had a clinically meaningful change. In actuality there are no perfect instruments and for the purpose of identifying

clinically meaningful change, the value for the MCID is chosen such that the sensitivity and specificity values are maximized.

Our results provide convergent and divergent evidence for validity of the FAAM ADL and Sports Subscales. As expected, the FAAM was found to have relatively high correlations with concurrent measures of physical function and relatively low correlations with concurrent measures of mental health. The relationship between the FAAM and concurrent measures of physical function were significantly different than the associations between the FAAM and concurrent measures of mental health. This provides evidence that the FAAM is a measure of physical function as opposed to mental function.

Comparison of the FAAM to general measures of physical function in this sample of subjects found that the ADL subscale was more responsive to changes in functional status than the physical function and physical components summary scores of the SF-36, while the Sports subscale was not. The Sports subscale may not have been responsive to changes in status because subjects in the group expected to change were functioning at a relatively low-level ability. The test information curve for the sports scale indicated that the sports scale provided the greatest information at the higher levels of ability and relatively little information at lower levels of ability. Thus, it is likely that the sports would be more responsive in a sample of subjects who had higher levels of ability. Future research to compare responsiveness of the sports scale to responsiveness of the physical function and physical components summary scale of the SF-36 should make use of subjects who are functioning at higher levels of ability.

In addition to the SF-36, the American Orthopaedic Foot and Ankle Society (AOFAS) Clinical Rating Systems[8] and Foot Function Index (FFI)[2] are instruments commonly reported in the literature. The AFOAS Clinical Rating Systems were developed to be used with individuals with a broad range of foot and ankle musculoskeletal disorders, including disorders at the ankle-hindfoot, midfoot, hallux, and lesser toes.[8] The FFI was developed to be used for individuals with rheumatoid arthritis.[2] Without direct statistical comparison using the same subjects, comparing the FAAM to these instruments is difficult. However, if the instruments are to be compared important issues to consider are item content and evidence to support the instruments' usefulness. In terms of item content, the AOFAS Clinical Rating Systems combine the scores from items relating to pain, range of motion, alignment, and calluses, as well as the scores from items related to physical performance.[8] The FFI also combines the scores from items related to pain and physical performance.[2] There is little evidence to support combining scores from items that assess symptoms and the scores from items that assess physical performance. Also, related to item content the AOFAS Clinical Rating Systems and FFI may not have adequate representation assessing activities that require a high level of ability (i.e. sports activities).[2,8] This could potentially cause a ceiling effect and inadequate sensitivity to change when individuals are only limited at the high end of ability.

There is limited evidence to support the usefulness of the AOFAS Clinical Rating Systems and the FFI. The AOFAS Clinical Rating Systems had poor relationships to measures of physical function and therefore its ability to measure health status has been questioned.[16] The evidence of reliability and validity to support the use of the FFI can only be generalized to individuals with rheumatoid arthritis.[2,14] The evidence to support the usefulness of the FAAM in this study can be generalized to individuals receiving outpatient physical therapy over a 4-week period for a musculoskeletal disorder of the leg, ankle, or foot. This includes individuals who are undergoing conservative as well as surgical intervention.

Limitations of this study are associated with values defining the measurement error at a single point of time, MDC and MCID. Additional validity evidence to support interpretation of the FAAM score will be required for applications in other settings or over a different time frame. These values also may vary depending on the baseline level of function of the subjects. We have uniformly assigned values to these indices across all of our subjects, regardless of their baseline functional level. Future research should include an assessment of reliability and responsiveness across different functional levels.

Appendix 1 **Foot and Ankle Ability Measure (FAAM)**
Activities of Daily Living subscale

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	□	□	□	□	□	□

How would you rate your current level of function during your usual activities of daily living from 0 to 100 with 100 being your level of function prior to your foot or ankle problem and 0 being the inability to perform any of your usual daily activities?

□□□.0 %

Foot and Ankle Ability Measure (FAAM)
Sports subscale

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	☐	☐	☐	☐	☐	☐

How would you rate your current level of function during your sports related activities from 0 to 100 with 100 being your level of function prior to your foot or ankle problem and 0 being the inability to perform any of your usual daily activities?

☐☐☐.0 %

Overall, how would you rate your current level of function?

☐ Normal ☐ Nearly normal ☐ Abnormal ☐ Severely abnormal

REFERENCES

1. **Beaton, DE:** Understanding the relevance of measured change through studies of responsiveness. Spine **25**:3192–3199, 2000.
2. **Budiman-Mak, E; Conrad, KJ; Roach, KE:** The Foot Function Index: a measure of foot pain and disability. J. Clin. Epidemiol. **44**:561–570, 1991.
3. **DeAyala, RJ:** An introduction to polytomous item response theory models. Measurement and Evaluation in Counseling and Development **25**:172–189, 1993.
4. **Dodd, BG; DeAyala, RJ; Koch, WR:** Computerized adaptive testing with polytomous items. Applied Psychological Measurement **19**:5–22, 1995.
5. **Dillman, DA:** *Mail and Telephone Surveys: The Total Design Method.* New York, Wiley-Intersience, 1978.
6. **Guyatt, G; Walter, S; Norman, G:** Measuring change over time: assessing the usefulness of evaluative instruments. J. Chronic Dis. **40**:171–178, 1987.
7. **Hambleton, RK; Jones, RW:** Comparison of classical test theory and item response theory and their applications to test development. Education Measurement: Issues and Practices **12**:38–47, 1993.
8. **Kitaoka, HB; Alexander, IJ; Adelaar, RS; et al.:** Clinical rating systems for the ankle-hindfoot, midfoot, hallux, and lesser toes. Foot Ankle Int. **15**:349–353, 1994.
9. **Lord, FM:** The relation of test score to the trait underlying the test. Educ. Psychol. Meas. **13**:517–548, 1953.
10. **Martin, RL; Irrgang, JJ; Burdett, RG; Conti, SF; VanSwearingen, JM:** A systematic review of self-reported evaluative measures of function and disability for the foot and ankle. J. Orthop. Sports Phys. Ther. submitted for publication, 2004.
11. **Meng, X; Roenthal, R; Rubin DB:** Comparing correlation coefficients. Psychol. Bull. **111**:172–175, 1992.
12. **Messick, S:** Meaning and values in test validation: The science and ethics of assessment. Educational Researcher. **18**:5–11, 1989.
13. **Reise, SP; Yu, J:** Parameter recovery in the graded response model using MULTILOG. Journal of Education Measurement. **27**:133–144, 1990.
14. **Saag, KG; Saltzman, CL; Brown, CK, Budiman-Mak, E:** The Foot Function Index for measuring rheumatoid arthritis pain: Evaluating side-to–side reliability. Foot Ankle Int. **17**:506–510, 1996.
15. **Shrout, PE; Fleiss, JL:** Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. **86**:420–428, 1979.
16. **SooHoo, NF; Shuler, M; Fleming, LL:** Evaluation of the validity of the AOFAS Clinical Rating Systems by correlation to the SF-36. Foot Ankle Int. **24**:50–55, 2003.
17. **Strafford, PW; Binkley, J; Soloman, P; et al.:** Defining the minimal level of detectable change for the Roland-Morris questionnaire. Phys Ther. **76**:359–365, 1996.
18. **Teresi, JA; Goldman, RR; Cross, P; et al.:** Item bias in cognitive screening measures: Comparisons of elderly white, Afro-American, Hispanic and high and low education subgroups. J. Clin. Epidemiol. **48**:473–483,1995.
19. **Tuley, MR; Mulrow, CD; McMahan, CA:** Estimating and testing an index of responsiveness and the relationship of the index to power. J. Clin. Epidemiol. **44**:417–421,1991.