Clin Orthop Relat Res (2013) 471:3466–3474
DOI 10.1007/s11999-013-3097-1

Clinical Orthopaedics and Related Research®
A Publication of The Association of Bone and Joint Surgeons®

SYMPOSIUM: ABJS CARL T. BRIGHTON WORKSHOP ON OUTCOME MEASURES

Validation of PROMIS® Physical Function Computerized Adaptive Tests for Orthopaedic Foot and Ankle Outcome Research

**Man Hung PhD, MSTAT, MED, Judith F. Baumhauer MD, MPH,
L. Daniel Latt MD, PhD, Charles L. Saltzman MD, Nelson F. SooHoo MD,
Kenneth J. Hunt MD, and National Orthopaedic Foot & Ankle Outcomes Research Network**

Published online: 8 June 2013

Abstract

Background In 2012, the American Orthopaedic Foot & Ankle Society® established a national network for collecting and sharing data on treatment outcomes and improving patient care. One of the network's initiatives is to explore the use of computerized adaptive tests (CATs) for patient-level outcome reporting.

Questions/purposes We determined whether the CAT from the NIH Patient Reported Outcome Measurement Information System® (PROMIS®) Physical Function (PF) item bank provides efficient, reliable, valid, precise, and adequately covered point estimates of patients' physical function.

Methods After informed consent, 288 patients with a mean age of 51 years (range, 18–81 years) undergoing surgery for common foot and ankle problems completed a web-based questionnaire. Efficiency was determined by time for test administration. Reliability was assessed with person and item reliability estimates. Validity evaluation included content validity from expert review and construct

Each author certifies that he or she has no commercial associations (eg, consultancies, stock ownership, equity interest, patent/licensing arrangements, etc) that might pose a conflict of interest in connection with the submitted article.
All ICMJE Conflict of Interest Forms for authors and *Clinical Orthopaedics and Related Research* editors and board members are on file with the publication and can be viewed on request.
Each author certifies that his or her institution approved the human protocol for this investigation, that all investigations were conducted in conformity with ethical principles of research, and that informed consent for participation in the study was obtained.
Subjects were enrolled through a distributed consortium of foot and ankle clinics participating in the National Orthopaedic Foot & Ankle Outcomes Research Network, sponsored by the American Orthopaedic Foot & Ankle Society®. Institutions participating in this network include Baylor University (Houston, TX, USA), The University of Tennessee–Campbell Clinic (Memphis, TN, USA), Hospital for Special Surgery (New York, NY, USA), OrthoCarolina Research Institute (Charlotte, NC, USA), Stanford University (Redwood City, CA, USA), University of Arizona (Tucson, AZ, USA), University of California Los Angeles (Los Angeles, CA, USA), University of Iowa (Iowa City, IA, USA), University of Rochester (Rochester, NY, USA), and University of Utah (Salt Lake City, UT, USA).

Electronic supplementary material The online version of this article (doi:10.1007/s11999-013-3097-1) contains supplementary material, which is available to authorized users.

M. Hung (✉), C. L. Saltzman
Department of Orthopaedic Surgery Operations, University of Utah School of Medicine, 590 Wakara Way, Salt Lake City, UT 84108, USA
e-mail: Man.Hung@hsc.utah.edu

J. F. Baumhauer
Department of Orthopaedic Surgery, University of Rochester School of Medicine and Dentistry, Rochester, NY, USA

L. D. Latt
Department of Orthopaedic Surgery, University of Arizona, Tucson, AZ, USA

N. F. SooHoo
Department of Orthopaedic Surgery, University of California Los Angeles School of Medicine, Los Angeles, CA, USA

K. J. Hunt
Department of Orthopaedic Surgery, Stanford University, Redwood City, CA, USA

National Orthopaedic Foot & Ankle Outcomes Research Network
American Orthopaedic Foot and Ankle Society®, 6300 N River Road, Rosemont, IL, USA

validity measured against the PROMIS® Pain CAT and patient responses based on tradeoff perceptions. Precision was assessed by standard error of measurement (SEM) across patients' physical function levels. Instrument coverage was based on a person-item map.

Results Average time of test administration was 47 seconds. Reliability was 0.96 for person and 0.99 for item. Construct validity against the Pain CAT had an r value of -0.657 ($p < 0.001$). Precision had an SEM of less than 3.3 (equivalent to a Cronbach's alpha of ≥ 0.90) across a broad range of function. Concerning coverage, the ceiling effect was 0.32% and there was no floor effect.

Conclusions The PROMIS® PF CAT appears to be an excellent method for measuring outcomes for patients with foot and ankle surgery. Further validation of the PROMIS® item banks may ultimately provide a valid and reliable tool for measuring patient-reported outcomes after injuries and treatment.

Level of Evidence Level III, diagnostic study. See Instructions for Authors for a complete description of levels of evidence.

Introduction

Disorders of the foot and ankle can result in significant morbidity and limitations of physical function. The functional disability resulting from foot and ankle pathologies have an impact on physical abilities similar to those with other lower-extremity regions [13, 36]. Some conditions, such as ankle arthritis, may be equivalent or worse than that those reported for patients with end-stage kidney disease, congestive heart failure, or cervical spine pain and radiculopathy [13, 37]. As orthopaedic foot and ankle providers endeavor to optimize outcomes after treatment of these conditions, we recognize the need for consistent and valid tools to evaluate patients with foot and ankle disorders and their outcomes after treatment.

At present, there is considerable variability among the available outcome instruments used in evaluating foot and ankle procedures and disorders. Many clinical instruments are used in the literature, and there is no broadly accepted consensus [10, 32, 41]. Computerized adaptive tests (CATs) using item response theory offer a possible solution to this important problem. A CAT is defined as a dynamically administered computer-based test in which responses to previous questions are used to select the most appropriate next question from an item bank, resulting in a measure that is both concise and precise. Item response theory holds promise that comparable measures can be obtained even if we ask different patients different questions; thus, it allows a test to be tailored to the ability or function of the patients, while still providing valid measurement. The integration of item response theory into CAT enables precise and efficient patient-reported outcome assessment and has been the focus of major NIH patient-reported outcome development efforts. Recently, CATs using item response theory have been successfully evaluated for the orthopaedic population using the Patient Reported Outcome Measurement Information System® (PROMIS®) [15]. The PROMIS® was created as part of the NIH in an effort to improve patient-reported outcome assessment [11]. The Assessment Center, as part of PROMIS®, was developed to enable administration of CATs in clinical research and contains a large repository of patient-reported outcome measures that can be scored in real time. These measures include physical functioning, symptoms, social behaviors, and experiences with treatment. PROMIS® is available to the public free of charge (see www.assessmentcenter.net and www.nihpromis.org). The main weakness of PROMIS® at present is its inability to integrate into existing electronic health record systems. Overcoming this weakness may facilitate its broader use in clinical research.

The PROMIS® Physical Function (PF) item bank v1.0 includes a total of 124 physical function items across five categories of physical functioning: upper extremity, lower extremity, axial, central, and instrumental activities of daily living. These items were selected from broadly accepted outcome instruments [34]. Subsequent evaluation using data from a normal population has demonstrated appropriate psychometric properties [15, 34]. While the PROMIS® PF CAT is not specifically tailored to any particular disease process, it has demonstrated ample validity and reliability evidence for some medical conditions [3] and in an orthopaedic population [15]. Yet, no study to date has evaluated the PROMIS® PF CAT for patients with disorders of the foot and ankle.

In 2012, the American Orthopaedic Foot & Ankle Society® (AOFAS) established the National Orthopaedic Foot & Ankle Outcomes Research (OFAR) Network, a national consortium for collecting and sharing data on treatment outcomes and improving patient care. The network currently includes 10 clinical sites, each with at least one fellowship-trained orthopaedic foot and ankle specialist, with a wide geographic distribution across the United States (Appendix 1). The breadth of the network allows access to a diverse population of patients.

We evaluated the measurement properties of the PROMIS® PF CAT for adult patients with common disorders of the foot and ankle using data collected through the OFAR Network in terms of efficiency, reliability, validity, precision, and coverage.

Patients and Methods

Patient Enrollment

The OFAR Network was developed by the AOFAS for use by the organization's membership. The goal of OFAR is, in part, to provide an infrastructure for multicenter prospective research and clinical trials. Ten geographically diverse sites were invited to participate in a proof of concept trial of the OFAR Network. Investigators at each of these 10 clinical sites volunteered to participate in the OFAR Network. In this prospective study, each site enrolled patients with any of six common foot and ankle disorders: ankle arthritis, ankle instability, adult acquired flatfoot deformity (pes planovalgus), hallux valgus (bunions), hallux rigidus, and hammer toe(s). These disorders were selected based on the typical case volume at the 10 sites and on the American Board of Orthopaedic Surgeons list of most frequent elective procedures for foot and ankle specialists. Institutional review board approval was obtained at each of the sites. All participating sites and investigators agreed to enroll a minimum number of 30 patients and to abide by the OFAR Network guidelines.

From March 2012 to June 2012, adult patients undergoing surgical treatment for one of the six specified foot and ankle disorders were invited by their treating provider to participate, provided that they were not younger than 18 years and that they had no active infection or ulceration at the area of the planned procedure.

Demographics

We recruited 323 patients for this study. Of the 323, 10 failed to meet study inclusion criteria and 25 had no data recorded for the PROMIS® PF CAT, resulting in a cohort of 288 patients. The cohort was 70% women, 87% white, and 5% Hispanic or Latino (Table 1). The patients' mean age was 51 years (SD, 15 years; range, 18–81 years). The distribution of the six foot and ankle disorders were 15% ankle instability, 13% ankle arthritis, 37% hallux valgus, 11% flatfoot deformity, 15% hallux rigidus, and 9% hammertoe.

Table 1. Patient demographic characteristics (n = 288)

Variable	Value
Age (years)*	51 ± 15 (18–31)
Sex (number of patients)†	
Male	74 (30%)
Female	170 (70%)
Race†	
White	208 (87%)
Black	11 (5%)
Asian	8 (3%)
White and Asian	1 (0.3%)
American Indian or Alaska Native	1 (0.3%)
Other	10 (3.5%)
Ethnicity†	
Not Hispanic or Latino	184 (95%)
Hispanic or Latino	10 (5%)
Diabetes†	
No	269 (95%)
Yes	14 (5%)
Rheumatoid arthritis†	
No	258 (91%)
Yes	25 (9%)

* Values are expressed as mean ± SD, with range in parentheses; †the subgroups that do not add up to the total sample size of 288 reflect missing data.

Data Collection

Each patient gave consent and was asked to complete a web-based questionnaire, which included demographic and comorbidity questions, the PROMIS® PF CAT, and the PROMIS® Pain Interference CAT, either on a computer or on an iPad during the preoperative period (no more than 30 days before the scheduled procedure). The questionnaire was created and completed using the PROMIS® Assessment Center, a secure database system housed at Northwestern University (Evanston, IL, USA). The PROMIS® Pain CAT has been evaluated with respect to precision, construct, and concurrent validity and results indicate that its 41-item bank is psychometrically sound [2]. Both the PROMIS® PF and Pain CATs were programmed to select an initial item with moderate difficulty from all items in their item banks to administer to the patients. Difficulty of an item in this context refers to whether an item measures higher functioning (more difficult) or lower functioning (less difficult) levels. The next items selected by the CAT algorithms were based on patients' responses to the previous items. The termination criteria for these tests were set at a maximum of 12 items and a maximum standard error of 0.33. The investigator or site personnel entered basic procedure-specific information for the surgical procedure performed.

Data Analysis

As the primary objective of this study was to assess and describe measurement properties of the PROMIS® PF CAT, this study was descriptive in nature, and power analysis to calculate the required sample size for the study was not applicable. We have, however, calculated the required sample size for analyzing the correlations. To achieve a 95% power to detect a difference of −0.6 between the null hypothesis correlation of 0 and the alternative hypothesis correlation of 0.6 using a two-sided hypothesis test with a significance level of 0.05, a minimum sample size of 30 was needed.

We applied a one-parameter item response theory model, the Rasch Partial Credit Model [8, 29, 30, 33], to examine the measurement properties (ie, efficiency, reliability, validity, precision, and coverage) of the PROMIS® PF CAT. The Rasch Partial Credit Model is a measurement model used to assess trait, ability, competence, or functional levels of individuals. It can be applied to questionnaires to assess instrument precision, dimensionality, scoring, and beyond. It also provides a mechanism for transforming data from ordinal scale to interval scale to meet the assumption of parametric statistical tests. The Rasch model allows both the item difficulty and the patients' physical function to be measured and placed on the same scale. It can provide robust support for instrument development, evaluation, and validation. However, before conducting Rasch analysis, we examined the data to assess the two major assumptions of the Rasch model (unidimensionality and fit). If unidimensionality and fit are adequate, the Rasch Partial Credit Model can be used to analyze the data.

Unidimensionality refers to whether an instrument measures predominantly one single construct or concept such as physical function. We evaluated unidimensionality by conducting principal component analysis of the residuals. If, after removing the first factor, the unexplained variance in the first contrast is less than 10% and the eigenvalue of the first contrast is less than 3, then unidimensionality for the PROMIS® PF CAT would be supported. Eigenvalue is a mathematical term to indicate variability in data and is normally used in factor analysis to assess dimensionality. Eigenvalues range from zero to the maximum number of items in an item bank. The higher the eigenvalue, the more variability there is in the data, which indicates departure from unidimensionality. We found, in our data analysis, after removing the first factor, the unexplained variance in the first contrast was 2.6% and the eigenvalue was 2.4. Hence, there was evidence to support unidimensionality.

Fit of the data to the Rasch model is also necessary for appropriate use of this model. Mean square (MNSQ) statistics are units of measurement to indicate whether the data fit the Rasch model. If the data fit the Rasch model, we can then apply the Rasch model to evaluate the data. Otherwise, we may utilize alternative methods or modification. The data fit the model if the infit MNSQ and the outfit MNSQ statistics are less than 2. Fit statistics close to 1 are considered the best [8, 14, 43]. Both of these statistics can range from zero to infinity. The outfit MNSQ is more sensitive to outliers and unexpected responses, whereas the infit MNSQ is more sensitive to discrepancies around the mean. We obtained a mean infit MNSQ of 0.92 and a mean outfit MNSQ of 0.96, indicating our data fit the model well. Thus, the Rasch Partial Credit Model was appropriate to evaluate the PROMIS® PF CAT data.

Measurement Properties

We examined the measurement properties of the PF CAT in the following five areas: efficiency, reliability, validity, precision, and coverage.

Efficiency

Efficiency is defined as the total amount of time to completing the instrument and whether the number of items that patients had to complete is reasonable. The PROMIS® Assessment Center's database automatically recorded the time that it took each patient to complete an item. Using that information, we computed the average time (in seconds) for patients to complete the entire PROMIS® PF CAT. We also examined the average number of items the PROMIS® PF CAT administered to the patients. (See Appendix 2 for a list of items administered by the PROMIS® PF CAT and Appendix 3 for a list of items contained in the entire PROMIS® PF item bank; supplemental materials are available with the online version of CORR®.)

Reliability

We assessed internal consistency reliabilities of the PROMIS® PF CAT in terms of person reliability and item reliability. Person reliability refers to the reproducibility of the ordering of patients' physical function measures (eg, Patient A has a higher physical function score than Patient B, Patient B has a higher physical function score than Patient C, etc) for the instrument.

Item reliability refers to the reproducibility of the ordering of item difficulty measures (eg, Item 1 is more difficult for people than Item 2, etc) for the PROMIS® PF CAT and this sample population. Reliability is an r value (correlation) and it ranges from −1 to 1 [12]. We considered reliability of 0.90 or greater as excellent and 0.80 or greater as good.

Validity

Both content and construct validities were investigated. Content validity suggests the items in the PROMIS® PF

CAT do in fact measure patients' physical functioning and the items are appropriate and comprehensive relative to the foot and ankle patient population. This was determined by two methods. First, we formed a panel of six fellowship-trained foot and ankle surgeons and asked each to carefully examine the content of the PROMIS® PF CAT and provide face validity input. Second, we reviewed PROMIS® literature documenting the PROMIS® PF CAT item bank development process [34].

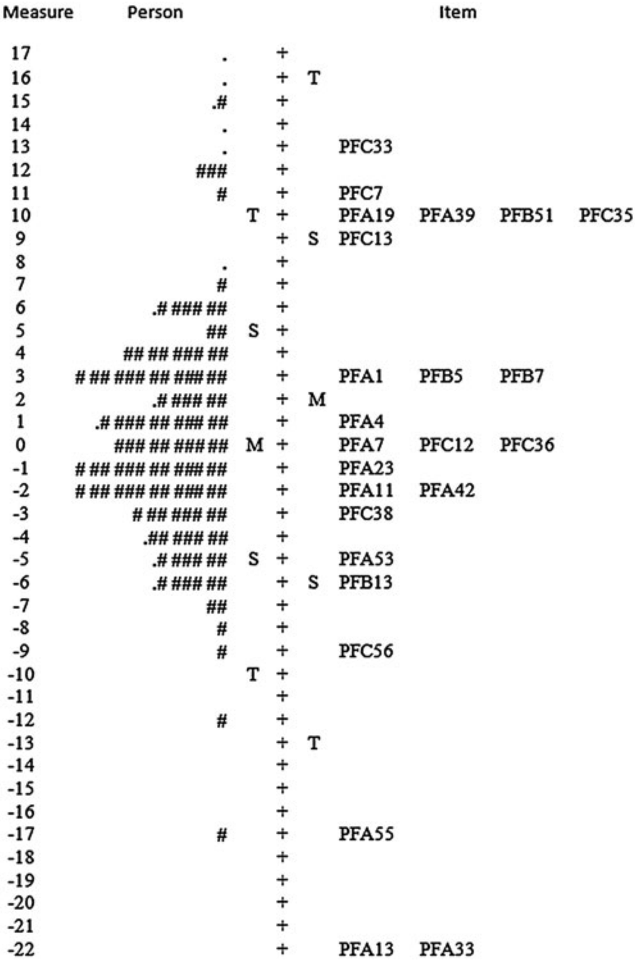
Construct validity describes evidence that the instrument has a relationship with related measures, or it is able to differentiate between known groups of patients. We used Pearson correlation analysis to determine whether the PROMIS® PF CAT had any relationship with the PROMIS® Pain CAT measures using standard criteria to indicate the strength of correlation: low correlation: $r < 0.3$; moderate correlation: $r = 0.3$ to 0.5; and high correlation, $r > 0.5$. Additionally, we evaluated for evidence that the PROMIS® PF CAT was able to differentiate between patients with greater morbidity to those with less morbidity. Patients with comorbidities were defined as those who had history of at least one ailment (hypertension, rheumatoid arthritis, osteoporosis, high blood pressure, asthma, anxiety, diabetes, breast cancer, thyroid disease, hay fever, hepatitis, gout, sleep apnea, etc) or were willing to trade at least 1 year in current state of life for perfect foot and ankle function. Patients without comorbidities were defined as those who did not have any ailment. We used the independent-samples t-test to determine whether the person measures between these two groups of patients differed; statistical significance was set at $p < 0.05$.

Precision

Standard error of measurement (SEM) is an index that quantifies the degree the measurement is free of error and is commonly used to indicate measurement precision [17, 18]. Smaller SEM suggests greater precision. An SEM of 3.3 or less (equivalent to a Cronbach's alpha of ≥ 0.90) is generally accepted as excellent precision and 5.0 or less (equivalent to a Cronbach's alpha of ≥ 0.80) as good precision [8].

Coverage

Instrument coverage determines whether the range of items can cover the entire spectrum of the patient population's physical functioning levels [17]. To examine coverage, we computed the persons' measures (eg, patients' PROMIS® PF scores) and the PROMIS® PF CAT items' difficulty measures and plotted these two distributions of measures side by side vertically to form a person-item or Wright map (Fig. 1). This person-item map is essentially a vertical ruler, separating patients on the left and the items on the right. The top of the ruler corresponds to high physical function levels and the bottom corresponds to low physical function levels. When a patient is aligned with an item on the same location of the ruler, that patient's physical functioning level is said to be well targeted by this item. When all the patients are well targeted by the entire set of items in the instrument, the instrument is said to have adequate coverage (ie, lack of ceiling and floor effects).



```
Measure      Person                         Item

 17                         .      +
 16                         .      +  T
 15                        .#      +
 14                         .      +
 13                         .      +     PFC33
 12                       ###      +
 11                         #      +     PFC7
 10                             T  +     PFA19   PFA39   PFB51   PFC35
  9                                +  S  PFC13
  8                         .      +
  7                         #      +
  6                .# ### ##       +
  5                      ##     S  +
  4            ## ## ### ##        +
  3      # ## ### ## ### ##        +     PFA1    PFB5    PFB7
  2                .# ### ##       +  M
  1         .# ### ## ### ##       +     PFA4
  0          ### ## ### ##      M  +     PFA7    PFC12   PFC36
 -1      # ## ### ## ### ##        +     PFA23
 -2      # ## ### ## ### ##        +     PFA11   PFA42
 -3            # ## ### ##         +     PFC38
 -4              .## ### ##        +
 -5                .# ### ##    S  +     PFA53
 -6                .# ### ##       +  S  PFB13
 -7                      ##        +
 -8                       #        +
 -9                       #        +     PFC56
-10                             T  +
-11                                +
-12                       #        +
-13                                +  T
-14                                +
-15                                +
-16                                +
-17                       #        +     PFA55
-18                                +
-19                                +
-20                                +
-21                                +
-22                                +     PFA13   PFA33
```

Fig. 1 This person-item map shows the spread of items and patients along a standardized linear scale labeled "Measures." A vertical dash ruler separates persons (ie, patients) on the left and items on the right. The top of the ruler represents high physical function levels, whereas the bottom represents low physical function levels. The map shows a normal distribution for the PROMIS® PF CAT, with approximately ½ of the subjects' physical functioning above the mean and ½ below, indicating efficient utilization of items. Instrument coverage was found to be excellent, with minimal ceiling effect (0.32%) and no floor effect. # = two patients; . = one patient; M = mean; S = 1 SD from the mean; T = 2 SDs from the mean. PFA19, PFC33, etc, are item identification numbers (see Appendix 2; supplemental materials are available with the online version of CORR®).

The extent to which an instrument lacks items to cover the lower end of a population's physical function is called the floor effect; the extent to which the instrument lacks items covering the upper end of a population's physical function is called the ceiling effect.

Results

Efficiency

The mean number of items administered by the PROMIS® PF CAT was 4 (range, 2–12; mode, 4; median, 4), requiring an average of only 47 seconds (SD, 49 seconds) to complete the instrument. A graph of item measures (Fig. 1) showed a normal distribution, with approximately ½ of the subjects' physical functioning above the mean and ½ below, indicating efficient utilization of items.

Reliability

The internal consistency reliability of the instrument was excellent, with a person reliability of 0.96. This suggests similar ordering of persons would occur with repeated studies. The item reliability was also high at 0.99, suggesting the ordering of item difficulty would remain the same regardless of altered patient populations or foot and ankle conditions (Fig. 1).

Validity

Construct validity was high as evaluated by all methods. Patients with comorbidities had lower PROMIS® PF scores (mean, 38; SD, 5) than those without comorbidities (mean, 44; SD, 9) ($t = 5.103$; $p < 0.001$). Also, there was a strong relationship between the PROMIS® PF and Pain CATs ($r = -0.657$; $p < 0.001$), suggesting physical function decreases as pain increases.

Precision

Measurement precision, represented by the SEM, was less than 3.3 across a broad range of physical function, suggesting excellent instrument precision (Fig. 2).

Coverage

A person-item map shows that the items covered the lower levels of patients' physical function completely (Fig. 1). The majority of patients with higher physical functioning were also covered by the items; only a small number of these patients were not covered by the items, reflecting a slight ceiling effect. Overall, instrument coverage was found to be excellent, with a minimal ceiling effect (0.32%) and no floor effect.

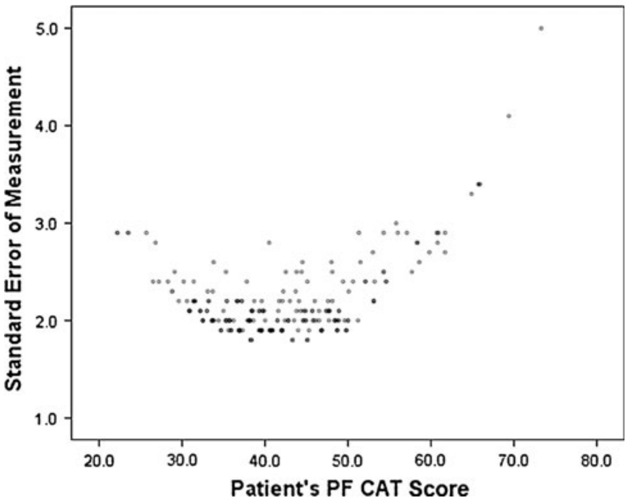


Fig. 2 Measurement precision of the PROMIS® PF CAT, represented by the SEM, was primarily less than 3.3 across a broad range of physical function, suggesting excellent instrument precision.

Discussion

There are many instruments currently available for assessing patient-reported outcomes in foot and ankle surgery. There has been considerable uncertainty as to which of these outcome tools is best for reporting the results of treatment for patients with foot and ankle disorders [10, 40]. The currently available patient-reported outcome instruments include generic instruments such as the SF-36 [31], designed for broad use in a variety of medical conditions, and more specialized questionnaires such as the Foot Function Index (FFI) [9] and the AOFAS Clinical Rating Systems [20]. Most of these instruments were developed based on classical test theory and thus suffer from a number of problems that limit their utility in a clinical setting. Two of the most important limitations are that they are time consuming for patients and that they cover only a narrow range of clinical conditions and disease severities. Item response theory-based outcome instruments, such as the PROMIS® PF CAT, should effectively overcome these limitations because a few relevant items targeted to each individual can be automatically selected from a large item bank. The PROMIS® PF CAT instrument is an appealing alternative to the existing tools as it is more efficient to administer, minimizes responder burden, is available in the public domain, and allows comparison across a broad spectrum of medical and surgical conditions. Our goal was to evaluate the efficiency, reliability, validity, precision, and coverage of the PROMIS® PF CAT for assessing physical function in patients with common foot and ankle disorders by using a sample of patients drawn from the multicenter AOFAS-sponsored OFAR Network.

A primary limitation of this study is the small sample size within the limited number of diseases studied. The patients recruited included those with six common foot and ankle conditions. The overall sample size is 288, but the sample sizes for each of the six conditions are relatively small, restricting detailed subgroup analyses on conditions. A larger sample that includes a wide range of foot and ankle conditions and a larger sample size within each condition would help to establish stronger validity evidence and to generalize to a larger patient population with foot and ankle diseases. Additionally, there is currently no gold standard instrument against which to assess criterion validity.

The criteria for selecting among patient-reported outcome instruments are validity, reliability, and responsiveness in evaluating the health of the targeted population [10, 21, 25]. The validity and reliability of the SF-36, FFI, and AOFAS systems have been addressed in previous studies [4, 6, 9, 10, 20, 22, 23, 25–27, 31, 35, 38–40, 42]. The validity of the AOFAS Clinical Rating Systems has been questioned while the FFI has validity evidence primarily in patients with rheumatoid arthritis [39]. The AOFAS and FFI questionnaires remain among the most commonly used tools for foot and ankle conditions despite the limited evidence supporting their validity for the broad range of conditions for which they are used [10, 40]. In contrast, the SF-36 has been extensively validated and tested for reliability but has low levels of responsiveness relative to region-specific tools [1, 5, 7, 19, 24, 27, 28]. The OFAR Network is in the process of evaluating responsiveness of the PROMIS® PF CAT relative to legacy instruments. Demonstration of high levels of responsiveness of the PROMIS® PF CAT in comparison to existing legacy tools would further support its use for patients with foot and ankle problems.

Our study provides strong validity evidence for the PROMIS® PF CAT in patients with common foot and ankle conditions. After foot and ankle experts reviewed the items in the PROMIS® PF CAT instrument, pilot testing in a sample cohort of patients demonstrated construct validity with high levels of correlation between the PROMIS® PF and Pain CAT scores. In addition, the PROMIS® PF CAT was precise and efficient to administer, with minimal ceiling and floor effects across disease severity. Our data suggest lower PROMIS® PF CAT scores correspond to patients with more severe comorbidity.

In conclusion, assessment of patient-reported outcomes has become increasingly important in evaluating the efficacy of medical and surgical treatments. Our study provides validity support for the PROMIS® PF CAT for patients with foot and ankle conditions. Further study of the responsiveness of this instrument relative to legacy scales, such as SF-36 and FFI, may determine whether the PROMIS® PF CAT will ultimately become a preferred alternative to legacy patient-reported outcome measures in patients with foot and ankle disorders. Such a paradigm shift in outcome instruments may increase compliance and accuracy, reduce respondent burden, and allow direct comparison to other health conditions. Further work is underway to optimize the PROMIS® CATs and the PROMIS® Lower Extremity Physical Function CAT [16, 17] for broad application to measure outcomes after injuries and treatment.

Acknowledgments The authors thank Ian Alexander MD (Cleveland Clinic–Medina Hospital, Medina, OH, USA) for his extensive contributions to the AOFAS OFAR Network and this project. We acknowledge the contributing investigators of the OFAR Network, including Timothy R. Daniels MD (University of Toronto, Toronto, ON, Canada), James Brodsky MD (Baylor University, Houston, TX, USA), Jon Deland MD (Hospital for Special Surgery, New York, NY, USA), Scott Ellis MD (Hospital for Special Surgery, New York, NY, USA), Phinit Phisitkul MD (University of Iowa, Iowa City, IA, USA), W. Hodges Davis MD (OrthoCarolina Research Institute, Charlotte, NC, USA), Susan Ishikawa MD (The University of Tennessee–Campbell Clinic, Memphis, TN, USA), Loretta B. Chou MD (Stanford University, Redwood City, CA, USA), John Ketz MD (University of Rochester, Rochester, NY, USA), Margaret Chilvers MD (University of Arizona, Tucson, AZ, USA), Sam Flemister MD (University of Rochester, Rochester, NY, USA), Florian Nickisch MD (University of Utah, Salt Lake City, UT, USA), and Timothy Beals MD (University of Utah, Salt Lake City, UT, USA). We are thankful for the contributions of each of the 10 clinical testing sites in the OFAR Network. We also thank Lousanne Lofgren BS (AOFAS, Rosemont, IL, USA) and Rose Olea BS (AOFAS, Rosemont, IL, USA) for their tireless work on this project. Special thanks go to Christine Cheng BS (University of Utah, Salt Lake City, UT, USA) and Shirley D. Hon BS (University of Utah, Salt Lake City, UT, USA) for their tremendous support in data collection, analysis, manuscript preparation, and review. Finally, the study design, statistical and psychometric support, and computing resources provided by the Center for Outcomes Research and Assessment at the University of Utah Department of Orthopaedics are gratefully acknowledged.

Appendix 1

National Orthopaedic Foot & Ankle Outcomes Research Network Sites and Principal Investigators

Site	Principal investigator
The University of Tennessee–Campbell Clinic (Memphis, TN, USA)	Sue Ishikawa MD, G. Andrew Murphy MD, David Richardson MD
Stanford University (Redwood City, CA, USA)	Kenneth J. Hunt MD
University of Arizona (Tucson, AZ, USA)	L. Daniel Latt MD
University of Utah (Salt Lake City, UT, USA)	Charles L. Saltzman MD, Man Hung PhD
University of Iowa (Iowa City, IA, USA)	Phinit Phisitkul MD
University of Rochester (Rochester, NY, USA)	Judith F. Baumhauer MD
Hospital for Special Surgery (New York, NY, USA)	Jonathan T. Deland MD, Scott Ellis MD
University of California Los Angeles (Los Angeles, CA, USA)	Nelson Soohoo MD
OrthoCarolina Research Institute (Charlotte, NC, USA)	W. Hodges Davis MD
Baylor University (Houston, TX, USA)	James W. Brodsky MD

References

1. Amadio PC, Silverstein MD, Ilstrup MD, Schleck CK, Jensen LM. Outcome assessment for carpal tunnel surgery: the relative responsiveness of generic, arthritis-specific, disease-specific, and physical exam measures. *J Hand Surg Am.* 1996;21:338–346.
2. Amtmann D, Cook KF, Jensen MP, Chen WH, Choi S, Revicki D, Cella D, Rothrock N, Keefe F, Callahan L. Development of a PROMIS item bank to measure pain interference. *Pain.* 2010; 150:173–182.
3. Bajaj JS, Thacker LR, Wade JB, Sanyal AJ, Heuman DM, Sterling RK, Gibson DP, Stravitz RT, Puri P, Fuchs M, Luketic V, Noble N, White M, Bell D, Revicki DA. PROMIS computerised adaptive tests are dynamic instruments to measure health-related quality of life in patients with cirrhosis. *Aliment Pharmacol Ther.* 2011;34:1123–1132.
4. Beaton DE, Hogg-Johnson S, Bombardier C. Evaluating changes in health status: reliability and responsiveness of five generic health status measures in workers with musculoskeletal disorders. *J Clin Epidemiol.* 1997;50:79–93.
5. Beaton DE, Richards RR. Assessing the reliability and responsiveness of 5 shoulder questionnaires. *J Shoulder Elbow Surg.* 1998;7:565–572.
6. Beskin J, Brage M, Guyton G, Saltzman C, Sands A, SooHoo NF, Kadel N, Stroud C, Thordarson D, Sangeorzan B. Reproducibility of the Foot Function Index: a report of the AOFAS outcomes committee. *Foot Ankle Int.* 2005;26:962–967.
7. Bessette L, Sangha O, Kuntz KM, Keller RB, Lew RA, Fossel AH, Katz JN. Comparative responsiveness of generic vs. disease-specific and weighted vs. unweighted health status measures in carpal tunnel syndrome. *Med Care.* 1998;36:491–502.
8. Bond TG, Fox CM. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences.* Mahwah, NJ: Lawrence Erlbaum; 2001.
9. 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.
10. Button G, Pinney S. A meta-analysis of outcome rating scales in foot and ankle surgery: is there a valid, reliable, and responsive system? *Foot Ankle Int.* 2004;25:521–525.
11. Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, Ader D, Fries JF, Bruce B, Rose M; PROMIS Cooperative Group. The Patient Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. *Med Care.* 2007;45:S3–S11.
12. Davidshofer KR, Murphy CO. *Psychological Testing: Principles and Applications.* 6th ed. Upper Saddle River, NJ: Pearson/Prentice Hall; 2005.
13. Glazebrook M, Daniels T, Younger A, Foote CJ, Penner M, Wing K, Lau J, Leighton R, Dunbar M. Comparison of health-related quality of life between patients with end-stage ankle and hip arthrosis. *J Bone Joint Surg Am.* 2008;90:499–505.
14. Hung M, Carter M, Hayden C, Dzierzon R, Morales J, Snow L, Butler J, Bateman K, Samore M. Psychometric assessment of the patient activation measure short form (PAM-13) in rural settings. *Qual Life Res.* 2013;22:521–529.
15. Hung M, Clegg DO, Greene T, Saltzman CL. Evaluation of the PROMIS physical function item bank in orthopaedic patients. *J Orthop Res.* 2011;29:947–953.
16. Hung M, Clegg DO, Greene T, Weir C, Saltzman CL. A lower extremity physical function computerized adaptive testing instrument. *Foot Ankle Int.* 2012;33:326–335.
17. Hung M, Nickisch F, Beals T, Greene T, Clegg DO, Saltzman CL. A new paradigm for patient-reported outcomes assessment in foot and ankle research: computerized adaptive testing. *Foot Ankle Int.* 2012;33:621–626.
18. Jette AM, Haley SM, Ni P, Olarsch S, Moed R. Creating a computer adaptive test version of the late-life function and disability instrument. *J Gerontol A Biol Sci Med Sci.* 2008;63: 1246–1256.
19. Katz JN, Gelberman RH, Wright EA, Lew RA, Liang MH. Responsiveness of self-reported and objective measures of disease severity in carpal tunnel syndrome. *Med Care.* 1994;32: 1127–1132.
20. 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.
21. Kocher MS, Horan MP, Briggs KK, Richardson TR, O'Holleran J, Hawkins RJ. Reliability, validity, and responsiveness of the American Shoulder and Elbow Surgeons Subjective Shoulder Scale in patients with shoulder instability, rotator cuff disease, and glenohumeral arthritis. *J Bone Joint Surg Am.* 2005;87: 2006–2011.
22. Kuyvenhoven MM, Gorter KJ, Zuithoff P, Budiman-Mak E, Conrad KJ, Post MW. The Foot Function Index with Verbal Rating Scales (FFI-5pt): a clinimetric evaluation and comparison with the original FFI. *J Rheumatol.* 2002;29:1023–1028.
23. Liang MH, Fossel AH, Larson MG. Comparison of five health status instruments for orthopedic evaluation. *Med Care.* 1990;28: 632–642.
24. Lingard EA, Katz JN, Wright RJ, Wright EA, Sledge CB. Validity and responsiveness of the Knee Society Clinical Rating

System in comparison with the SF-36 and WOMAC. *J Bone Joint Surg Am.* 2001;83:1856–1864.

25. Lohr KN, Aaronson NK, Alonso J, Burnam MA, Patrick DL, Perrin EB, Roberts JS. Evaluating quality-of-life and health status instruments: development of scientific review criteria. *Clin Ther.* 1996;18:979–992.
26. Mangione CM, Goldman L, Orav EJ, Marcantonio ER, Pedan A, Ludwig LE, Donaldson MC, Sugarbaker DJ, Poss R, Lee TH. Health-related quality of life after elective surgery: measurement of longitudinal changes. *J Gen Intern Med.* 1997;12:686–697.
27. Martin DP, Engelberg R, Agel J, Swiontkowski MF. Comparison of the Musculoskeletal Function Assessment questionnaire with the Short Form-36, the Western Ontario and McMaster Universities Osteoarthritis Index, and the Sickness Impact Profile health-status measures. *J Bone Joint Surg Am.* 1997;79:1323–1335.
28. Marx RG, Jones EC, Allen AA, Altchek DW, O'Brien SJ, Roeda SA, Williams RJ, Warren RF, Wickiewicz TZ. Reliability, validity, and responsiveness of four knee outcome scales for athletic patients. *J Bone Joint Surg Am.* 2001;83:1459–1469.
29. Masters GN. A Rasch model for partial credit scoring. *Psychometrika.* 1982;47:149–174.
30. Masters GN. The analysis of partial credit scoring. *Applied Measurement in Education.* 1988;1:279–297.
31. McHorney CA, Ware JE, 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.
32. Naal FD, Impellizzeri FM, Rippstein PF. Which are the most frequently used outcome instruments in studies on total ankle arthroplasty? *Clin Orthop Relat Res.* 2010;468:815–826.
33. Rasch G. *Probabilistic Models for Some Intelligence and Attainment Tests.* Chicago, IL: The University of Chicago Press; 1980.
34. Rose M, Bjorner JB, Becker J, Fries JF, Ware JE. Evaluation of a preliminary physical function item bank supported the expected advantages of the Patient-Reported Outcomes Measurement Information System (PROMIS). *J Clin Epidemiol.* 2008;61: 17–33.
35. Saag KG, Saltzman CL, Brown K, Budiman-Mak E. The Foot Function Index for measuring rheumatoid arthritis pain: evaluating side-to-side reliability. *Foot Ankle Int.* 1996;17:506–510.
36. Salaffi F, Carotti M, Grassi W. Health-related quality of life in patients with hip or knee osteoarthritis: comparison of generic and disease-specific instruments. *Clin Rheumatol.* 2005;24: 29–37.
37. Saltzman CL, Zimmerman MB, O'Rourke M, Brown TD, Buckwalter JA, Johnston R. Impact of comorbidities on the measurement of health in patients with ankle osteoarthritis. *J Bone Joint Surg Am.* 2006;88:2366–2372.
38. Soderman P, Malchau H. Validity and reliability of the Swedish WOMAC osteoarthritis index: a self-administered disease-specific questionnaire (WOMAC) versus generic instruments (SF-36 and NHP). *Acta Orthop Scand.* 2000;71:39–46.
39. SooHoo NF, Samimi D, Vyas R, Botzler T. Evaluation of the validity of the Foot Function Index in measuring outcomes in patients with foot and ankle disorders. *Foot Ankle Int.* 2006;27: 38–42.
40. SooHoo NF, Shuler MS, Fleming LL. Evaluation of the validity of the AOFAS Clinical Rating Systems by correlation to the SF-36. *Foot Ankle Int.* 2003;24:50–55.
41. Suk M, Hanson B, Norvell D, Helfet DL. *AO Handbook of Musculoskeletal Outcomes Measures and Instruments.* New York, NY: Thieme; 2009.
42. Tuttleman M, Pillemer SR, Tilley BC, Fowler SE, Buckley LM, Alaracon GS, Trentham DE, Neuner R, Clegg DO, Leisen JC, Heyse SP. A cross sectional assessment of health status instruments in patients with rheumatoid arthritis participating in a clinical trial. *J Rheumatol.* 1997;24:1910–1915.
43. Wright BD, Masters GN. *Rating Scale Analysis: Rasch Measurement.* Chicago, IL: MESA Press; 1982.