ELSEVIER

Journal of Clinical Epidemiology 58 (2005) 629–638

Journal of Clinical Epidemiology

Simulated computerized adaptive tests for measuring functional status were efficient with good discriminant validity in patients with hip, knee, or foot/ankle impairments

Dennis L. Hart[a,*], Jerome E. Mioduski[b], Paul W. Stratford[c]

[a]*Focus On Therapeutic Outcomes, Inc., 531 Yopps Cove Road, White Stone, VA 22578-2403, USA*
[b]*Focus On Therapeutic Outcomes, Inc., 6100 Lonas Dr, Knoxville, TN 37909 USA*
[c]*Department of Clinical Epidemiology and Biostatistics, McMaster University, 1400 Main St West, Hamilton, Ontario, Canada*

Accepted 7 December 2004

Abstract

Background and Objective: To develop computerized adaptive tests (CATs) designed to assess lower extremity functional status (FS) in people with lower extremity impairments using items from the Lower Extremity Functional Scale and compare discriminant validity of FS measures generated using all items analyzed with a rating scale Item Response Theory model (θ_{IRT}) and measures generated using the simulated CATs (θ_{CAT}).

Methods: Secondary analysis of retrospective intake rehabilitation data.

Results: Unidimensionality of items was strong, and local independence of items was adequate. Differential item functioning (DIF) affected item calibration related to body part, that is, hip, knee, or foot/ankle, but DIF did not affect item calibration for symptom acuity, gender, age, or surgical history. Therefore, patients were separated into three body part specific groups. The rating scale model fit all three data sets well. Three body part specific CATs were developed: each was 70% more efficient than using all LEFS items to estimate FS measures. θ_{IRT} and θ_{CAT} measures discriminated patients by symptom acuity, age, and surgical history in similar ways. θ_{CAT} measures were as precise as θ_{IRT} measures.

Conclusion: Body part-specific simulated CATs were efficient and produced precise measures of FS with good discriminant validity.

Keywords: Computerized adaptive testing; Item response theory; Lower Extremity Functional Scale; Rehabilitation

1. Introduction

Computerized adaptive testing (CAT) has transformed the process of estimating latent traits [1]. Latent traits or abilities cannot be directly observed, but can be estimated by analyzing a person's performance on a set of items [2]. For the purpose of this study of patients with lower extremity impairments, the latent trait of interest is lower extremity functional status (FS), which we operationally define as the patient's perception of their ability to perform functional tasks described in the FS items. FS is of interest because many people seek rehabilitation to improve functional deficits caused by lower extremity impairments [3].

CAT has its origins in mental [4], educational [5], and military [6] testing, but inexpensive, powerful computers have facilitated development of computerized adaptive tests (CATs) [1,6]. CATs have recently emerged in the medical [7,8] and rehabilitation [9,10] fields, and development of CAT measures of function in rehabilitation has been recommended [11–13].

CATs offer advantages compared to a computer administered or paper and pencil outcomes instruments. CATs (1) administer informative items, the difficulty of which are matched to the patient's level of ability reducing the number of inappropriate items administered; (2) administer fewer items, reducing respondent burden with little reduction in precision of patient ability estimates; 3) allow the level of measure precision to be established before testing improving control of measurement error during testing; and (4) simplify test revision by allowing adding and testing new items as needed [6,14]. CATs provide an efficient alternative to traditional paper-and-pencil or computer-administered tests, and allow outcomes data to be collected during the clinical encounter with reduced patient and scoring burden. Therefore,

* Corresponding author: Tel.: 804-436-9727; fax: 804-436-9328.
E-mail address: dsailhart@rivnet.net (D.L. Hart).

0895-4356/05/$ – see front matter © 2005 Elsevier Inc. All rights reserved.
doi: 10.1016/j.jclinepi.2004.12.004

CAT facilitates management of a central conflict in scale development: good measurement precision with low response burden [6,7] and is applicable to assessment of outcomes, that is, change in FS in patients receiving rehabilitation [9,10,15,16]. Recent symposia in health outcomes methodology and computer-based testing have emphasized the need to improve (1) outcomes assessment for advancing the science and practice of treatment-effectiveness evaluation [17], and (2) chart a path to development of better computer-based tests [18].

The foundation of CAT lies in Item Response Theory (IRT) methods [19–22]. Briefly, IRT comprises a set of mathematical models and associated statistical procedures that connect observed survey responses to a person's location on an unmeasured, underlying latent trait like FS. IRT models produce item and latent trait estimates that do not vary with population characteristics with respect to the underlying trait, standard errors conditional on trait level, and trait estimates linked to item content. IRT facilitates evaluation of whether items measure the trait of interest similarly in different subgroups of respondents, that is, differential item functioning (DIF) and assesses data fit to the model [23].

This article describes development of CATs using items from the Lower Extremity Functional Scale (LEFS), a common paper-and-pencil outcomes instrument for patients with lower extremity impairments receiving rehabilitation [3]. No articles have described IRT analyses or CAT applications of the LEFS. The overall purpose of this study was to develop CATs of the LEFS. Specific purposes were to (1) test unidimensionality and local independence of the LEFS items, (2) test LEFS item DIF, (3) develop CATs using LEFS items, and (4) compare the discriminant validity of FS measures generated using all LEFS items analyzed with an IRT rating scale model with measures generated from the simulated CATs.

2. Methods

2.1. Study design and setting

A secondary analysis of retrospective data collected from patients with lower extremity impairments prior to rehabilitation was conducted. Focus On Therapeutic Outcomes, Inc. (FOTO) Institutional Review Board approved the project.

2.2. Subjects

Patients ($n = 1772$, 48 ± 17 years, 14 to 89 years, 64% female) with lower extremity impairments were analyzed (Table 1). Patients, who represent a sample of convenience, received rehabilitation in 81 outpatient clinics in 20 states (United States) in the consecutive 24 months starting July 2002. All clinics were participating with FOTO (Knoxville, TN) a medical rehabilitation data management company.

2.3. Data collection

The data collection process has been described [15,24]. Patients seeking rehabilitation entered demographic data and completed self-report surveys prior to initial evaluation, including a computer-administered LEFS survey, which was a fixed format survey in the exact format as the original paper-and-pencil LEFS [3]: it was not a computerized adaptive test. Because this is a secondary analysis of a retrospective data set, no conditions or restrictions were placed on selection of patients who received the LEFS. Collection of data was at the discretion of the treating therapist. Patients were selected for this study from the FOTO database if the patients completed a computer-administered LEFS survey at rehabilitation intake.

2.4. Outcome Instrument

Conceived by Binkley et al. [3], the LEFS is a 20-item patient self-report region-specific outcomes instrument. Items represent functional activities commonly affected in people with lower extremity impairments, like running on uneven ground or walking between rooms. LEFS items are scored on a five-point scale (0 to 4). LEFS scores vary from 0 (low) to 80 (more normal FS). The LEFS was designed to be applicable to patients with a spectrum of lower extremity problems including mild ankle sprains and total joint arthroplasty. Reports of LEFS psychometric properties are consistent with clinical practice and research applications and stable across patient problems and age groups [3,25–27]. FS, as assessed using LEFS items, represents the "activity" dimension of the World Health Organization's International Classification of Functioning, Disability, and Health [28].

2.5. Data analyses

2.5.1. Unidimensionality and local independence

Data were analyzed to determine how well unidimensionality and local independence IRT assumptions were met. Unidimensionality means items in a scale measure only one construct [19,23]. Local independence requires that any two items be uncorrelated when the latent trait is fixed [19]. Some have stated local independence follows automatically from unidimensionality [19,29], but the assumption of unidimensionality is violated to some degree in each practical situation. A set of items can be unidimensional and yet contain pairs of items that are correlated [30].

Many IRT models assume unidimensionality [19,22,23], but because unidimensionality and local independence cannot be strictly met [19,23], scale developers seek to derive sets of items that are "essentially unidimensional" [31] where one dimension is dominant, possibly in the presence of one or more minor dimensions.

We used factor analytic methods using weighted least-square methods for factor analysis of categorical data [32] to investigate the assumption of unidimensionality and local

Table 1
Patient characteristics at rehabilitation intake

	Body part affected		
Characteristic	Hip (n = 444)	Knee (n = 949)	Foot/ankle (n = 379)
Diagnoses (%)			
Osteoarthritis	7	11	0
Internal derangement of joint	38	53	4
Effusion of joint	0	0	3
Enthesopathies[a]	20	0	9
Soft tissue disorders[b]	17	2	30
Flat foot	NA	NA	1
Congenital disorders	0	1	1
Abnormality of gait	0	6	0
Dislocation of joint	0	14	0
Fractures	3	0	10
Strains and sprains	14	13	25
Contusions	1	0	1
Other	0	0	16
Age (mean ± SD, min, max in years)	54 ± 17, 15, 88	47 ± 17, 14, 89	45 ± 15, 14, 88
Age 14 to <45 (%)	29	44	52
Age 45 to 65 (%)	45	41	39
Age >65 (%)	26	15	8
Gender (% female)	70	61	65
Acuity of symptoms (%)			
Acute	13	16	10
Subacute	27	29	35
Chronic	60	54	55
Surgical history (%)			
None	93	78	87
One or more	7	22	13
Exercise History (%)			
At least 3×/week	34	37	41
1–2×/week	24	25	24
Seldom or never	41	38	35
Medication use (%)	57	64	56
Type of Referring Physician (%)			
Primary care	60	32	41
Orthopedic surgeon	15	47	34
Physiatrist	14	1	3
Neurologist	2	0	1
Occupational medicine	2	2	1
Podiatrist	0	0	9
Other	7	18	11
Payer Source (%)			
Indemnity	11	9	14
Medicaid	2	2	5
Medicare	19	15	7
Patient private pay	0	1	2
HMO	43	34	43
PPO	15	16	12
Workers' compensation	6	14	14
Other	4	9	3

Abbreviations: SD, standard deviation; min, minimum; max, maximum.

[a] Enthesopathies are disorders of peripheral ligamentous or muscular attachments.

[b] Soft tissue disorders include synovium, tendon, bursa, muscle, fascia.

independence because traditional factor analysis may overestimate the number of factors and underestimate the factor loadings when analyzing skewed categorical data [33]. Presence of a dominant factor in the LEFS items was assessed with exploratory factor analyses (EFA) of latent trait variables followed by confirmatory factor analyses (CFA) [34]. Eigenvalue analyses were conducted, and results were evaluated with scree plots. Model fit was evaluated using comparative fit index (CFI) [35], the Tucker-Lewis index (TLI) [36], and the root-mean-square error of approximation (RMSEA) [37]. The CFI and TLI measure fit of a model relative to the null model, the CFI incorporates a correction for model complexity, and the TLI accounts for degrees of freedom [38]. The RMSEA was used because it is a fit statistic that accounts for model parsimony. The TLI and CFI range from 0 (poor fit) to 1 (good fit). Values of CFI and TLI greater than 0.90 are indicative of good model fit; RMSEA values less than 0.1 suggest adequate fit [37]. We also evaluated item loadings and residual correlation between items. Analyses were conducted using Mplus software (Muthén & Muthén, Los Angeles, CA) [32].

2.5.2. *IRT model selection*

The Andrich [39] rating scale IRT model (RSM) was selected because it is a latent structure model for polytomous responses to a set of test items, which is the format of the LEFS. The RSM is an extension of the Rasch model for dichotomous responses [40] where response categories are scored such that the total score for all items constitutes a latent ability estimate (θ) for respondents, in the present case, an estimate of the lower extremity FS. Rasch models, now being studied in health care [41], specify one-parameter logistic functions, which allow items to vary in their difficulty level (β) but assume items are equally discriminating [23,40,42]. The RSM extends the dichotomous model by assuming response categories are ordered and equidistant across items [39]. With the RSM, response categories are assigned thresholds for the entire set of items in the rating scale [43]. Thresholds are measures of θ where the probability of endorsing two adjacent responses is equal [39]. Item difficulty (β) is estimated by a single-scale location parameter, which represents the average difficulty (β) for each item relative to category thresholds [23,44]. The RSM permits estimates of FS ability (θ), category thresholds, and item difficulty (β) to be placed on the same metric, so the association between a patient's underlying level of the latent FS trait and the probability of a particular item response can be plotted using a nonlinear monotonic function, that is, item–characteristic curve [23].

2.5.3. *Development of the LEFS IRT hierarchical structure and item fit*

Hierarchical structure of a scale pertains to the ordering or calibration of items by level of difficulty as evidenced by patients' responses to the items [15,45]. Item difficulty calibrations or β parameters were estimated in logits using RSM. Scores were linearly transformed to a range of 0 (low) to 100 (more normal function) representing a measure of physical functioning, which we operationally define as an FS scale.

IRT models are "falsifiable" models, that is, a given model may or may not be appropriate for a particular set of data [22]. Assessing item structure and data fit investigates the success of the selected model in predicting or explaining the data. Infit and outfit mean square statistics for the sample were examined as an assessment of whether the data fit the RSM. Mean square fit statistics measure adherence of items to Rasch model restrictions [42]. The RSM requires an item to have a greater probability of producing higher ratings for persons with more ability compared to persons with less ability. Patients with a certain functional ability should have a higher probability of scoring higher on easier items than more difficult items. Mean square fit statistics are centered at 1.0, and their values increment higher with increasing violations of expected results.

IRT models also allow assessment of item fit [15,16,42,45,46]. The extent to which each item fits the FS construct was assessed by item goodness-of-fit statistics [42]. Goodness-of-fit is an assessment of how well item calibrations (estimated for the entire sample) fit the data with respect to sample individuals [45]. Item infit provides information about responses given to items near patient ability. Item outfit is an outlier-sensitive statistic that assesses items that are far from patient ability levels. Poor item fit was operationally defined as infit or outfit $<.6$ or >1.4 [47]. Items with poor fit were identified for possible deletion [15,42]. To test interitem consistency reliability, Cronbach's alpha values were calculated using raw LEFS responses.

2.5.4. *DIF*

An item demonstrates DIF if patients having the same ability but coming from different groups do not have the same probability of selecting a given item response [19,22,48]. Items were assessed for DIF by selecting groups of patients by body part impaired (hip, knee, foot/ankle), age group (young = 14 to <45, middle = 45 to 65, and older >65 years), gender (male, female), and acuity of symptoms (number of calendar days between date of onset of symptoms and date of initial evaluation, that is, acute = 21 days or less, subacute = 22 to <90 days, chronic = 90 days or more). There is evidence that patients with different affected body parts respond differently to physical functioning items [49], and age [24,50,51] and acuity [24,50,51] have been shown to affect patient self-report of FS.

Each item was assessed for DIF using the RSM in WINSTEPS software [52] by comparing pairs of item calibrations estimated for each group of patients by body part, age, gender, and acuity using independent *t*-test statistics. For DIF analyses, we anchored each person's ability from the overall sample to keep ability levels constant within each subgroup contrast [53]. Because of the number of repeated tests within each variable and sample size was large, the significance level of .01 was adjusted by a Bonferroni correction. Number of significant ($P < .01$) differences in item calibrations and the magnitude of the differences provide an assessment of DIF [52].

2.5.5. *CAT development*

We developed CATs following the logic of Thissen and Mislevy [54] using software developed specifically for this project (CAT Development and Testing Software, version 2.1.0, FOTO, Inc., Knoxville, TN) [55]. The basic components of the adaptive test included: selecting the starting item;

estimating theta; assessing stopping rules; and selecting subsequent items.

The adaptive test started by administering the most informative item for the scale [19], which provides a response from which a good initial provisional estimate of FS can be generated [54]. The computer estimated patient ability using maximum likelihood estimation employing a Newton-Raphson estimation technique [56]. The estimation process makes no assumptions about the distribution of the interviewees, and is Bayesian in the sense that the item difficulties come from a source outside of the current interviewee's data. After each response, a provisional estimate of ability and its standard error (SE) were calculated.

Stopping rules were assessed following each ability estimate. There were two stopping rules: (1) SE for the provisional ability was less than 4 out of 100 FS units, and (2) change in provisional abilities for the last three items was each less than 1 out of 100. The patient's level of FS from the adaptive test, that is, θ_{CAT}, was estimated using responses from all items administered.

If, after each theta estimate, no stopping rule was satisfied, the computer selected a subsequent item to administer that was most informative given the patient's current provisional ability estimate [54]. In this way, the adaptive test was designed to maximize the amount of information per patient ability given the subset of items administered and minimize test length. After administering another item, new estimates of FS ability and SE were generated, and stopping rules were reassessed.

2.5.6. *Simulated CAT*

The computerized adaptive test was used to estimate a measure of FS, that is, θ_{CAT}, with an SE for each patient in the original data set where each patient had answered all items using the computer-administered survey. The adaptive test recorded which and how many items were used before a stopping rule was satisfied.

2.5.7. *Relative precision*

Relative precision (RP) estimates [57] were used to examine how much more or less discriminating the CATs FS measures, that is, θ_{CAT}, were compared to FS measures generated using all LEFS items, that is, θ_{IRT}. We operationally define precision as a measure's relative success in discriminating differences in FS across levels of selected independent variables. Calculated in this way, precision depends upon the degree to which measures of FS differentiate groups of patients being compared (between-group variance) and error (within-group variance) [57–60].

Estimates of RP were calculated for each pair of measures, that is, θ_{IRT} and θ_{CAT}, per independent variable, that is, body part, age, gender, and acuity, by computing ratios of ANOVA F statistics: θ_{CAT} F divided by θ_{IRT} F. Measures are more efficient relative to one another if they classify patients with greater accuracy (less error) [57–63]. For this study, we standardized comparisons by keeping the subject sample constant across measures and within independent variables.

The magnitude of the F-value from the ANOVA represents a measure of precision. If the RP ratio is equal to 1, both methods of estimating function are equally discriminatory. If the RP >1 the measurement method in the numerator is superior in differentiating function compared to method in the denominator. The greater the F-value, the greater the amount of systematic variance a measurement method accounts for and, therefore, the greater its ability to discriminate groups of patients. When the subject sample is held constant, the greater F-value represents the most discriminating measurement method [57,58]. The ability of a measurement method to discriminate differences (validity) and be sensitive to changes (precision) in clinical outcome measures is clinically important and relevant for future outcomes research. Estimates of RP provide estimates of discriminant validity. Confidence intervals for the RP statistics were obtained using a bootstrap algorithm [64]. A total of 1,000 bootstrap samples with replacement were generated from each independent variable comparison, and F statistics and RP values were calculated for each resampling, which provided an estimate of the distribution for each RP. The 25th and 975th values of the RP distribution identified the 95% confidence interval [59,64].

3. Results

3.1. *Unidimensionality and local independence*

EFA of the 1,772 patients with complete scores on all 20 LEFS items produced a scree plot analysis that supported one dominant factor (first three eigenvalues = 13.1, 1.7, 0.7) with the first three factors explaining 66, 9, and 4% of data variance. In CFA, a three-factor model fit better than a one-factor model, but the correlations between the three factors were high (>0.62) suggesting one dominant factor. Fit statistics from the one- to three-factor models were CFI = 0.93, 0.94, 0.94, TLI = 0.98, 0.98, 0.99, and RMSEA = 0.22, 0.18, 0.16, for one-, two-, and three-factor models, respectively.

These statistics represent mixed results regarding fit for the model. Although the percentage of item variance accounted for was high, the magnitude of coefficients was strong, and the CFI and TLI indices were acceptable, the RMSEA was not. Plus, the number of residuals greater than absolute value of .10 was higher than desired suggesting possible local dependence between items related to item pairs (i.e., items assessing similar tasks). Assessment of factor loadings and residual correlations suggested the item rolling in bed did not load well on the dominant factor and negatively correlated with several other items and was deleted.

The 19-item set was reanalyzed producing similar (Table 2) results (first three eigenvalues = 12.7, 1.5, 0.70, with the first three factors explaining 67, 8, and 4% of data variance, and fit statistics from a one- to three-factor models were CFI = 0.94, 0.95, 0.96, TLI = 0.98, 0.99, 0.99, and

RMSEA = 0.21, 0.17, 0.14 for one-, two-, and three-factor models, respectively). Assessment of residual correlations suggested the item running on uneven ground negatively correlated with other items, that is, running on even ground. However, deleting the item running on uneven ground provided a negligible change in CFI, TLI, and RMSEA statistics. Because the percentage of item variance accounted for was high, the magnitude of coefficients was strong, and the CFI and TLI indices were acceptable, we believe the 19-item set represents a unidimensional pool with acceptable local independence.

3.2. LEFS IRT hierarchical structure and data fit

Findings supported the 19-item LEFS data fit the RSM well (mean square infit .99 and outfit 1.02 statistics, person reliability .95 and separation 4.25, root-mean-square error 2.90 and Cronbach's α .96). However, one item (sitting for 1 hr) had infit and outfit statistics >2 and was deleted. Subsequent RSM analyses supported the 18-item LEFS data fit the RSM as well as the 19-item LEFS (mean square infit .98 and outfit 1.01 statistics, person reliability .95 and separation 4.34, root-mean-square error 2.97, and Cronbach's α .96). The 18-item LEFS scale was used for subsequent analyses because the items represented a unidimensional scale, all items had adequate fit, and person separation was good.

3.3. DIF

Because the sample size was large, many DIF *t*-test statistics were significant ($P < .01$), but differences in item calibrations were small, that is, <2 out of 100 FS units. Therefore, we decided to consider pairs of item calibrations with differences in difficulty levels $\geqslant 5$ out of 100 FS units with significant ($P < .01$) Bonferonni adjusted *t*-test statistics as clinically important DIF. Of the 18 LEFS items, 4, 1, 0, 0, and 1 items displayed clinically important DIF in at least one pair of item calibration comparisons per level of body part, age, gender, and symptom acuity, respectively. We believe these results represent clinically important DIF by body part, but negligible DIF by age, gender, or acuity.

Table 2
Factor loadings for 19-item set[a]

Item name	Three-factor solution			One-factor solution
	Factor 1	Factor 2	Factor 3	
Working	0.614			0.823
Bathing	0.769			0.799
Walking between rooms	0.641			0.802
Shoes	0.803			0.722
Squatting	0.518			0.757
Lifting	0.837			0.815
Light activities	0.831			0.864
Heavy activities	0.689			0.857
Car	0.868			0.783
Stairs	0.434			0.815
Standing	0.364			0.811
Sitting	0.687			0.518
Hobby		0.511		0.719
Run even		0.861		0.939
Run uneven		0.908		0.956
Turning		0.945		0.962
Hopping		0.855		0.899
Walking blocks			0.717	0.938
Walking a mile			0.662	0.930
Factor correlations				
Factor 2	0.651			
Factor 3	0.619	0.588		

[a] Factor loadings from confirmatory factor analyses.

Results of DIF by body part appear clinically logical. For example, patients with knee or hip impairments perceived walking to be easier than patients with foot or ankle impairments for walking between rooms, walking two blocks, walking a mile, and running on uneven ground. The item by body part with the largest DIF was squatting. Patients with knee impairments perceived squatting as more difficult compared to patients with foot or ankle impairments (squatting item difficulty 55.3[0.3] for patients with knee impairments compared to 45.7[0.5], β[SE], for patients with foot/ankle impairments [$t = 9.6$, $df = 809$, $P < .01$]). Because of the DIF by body part results, patients were grouped body part, items were recalibrated separately by body part affected, three separate CATs were developed, and RP was tested separately by body part.

3.4. CAT development and simulation

The FS scales for patients with hip, knee, or foot/ankle impairments are displayed in Table 3. CATs were generated for each scale. The CATs used on average 6 (SD = 2.3, median = 5, minimum = 4, maximum = 17), 6 (SD = 1.4, median = 5, minimum = 4, maximum = 15), or 6 (SD = 1.4, median = 5, minimum = 4, maximum = 15) items before a stopping rule was satisfied (hip, knee, foot/ankle CATs, respectively). FS measures, that is, θ_{IRT} and θ_{CAT}, correlated well for all CATs ($r = 0.968$, 0.965, 0.968 for hip, knee, or foot/ankle CATs, respectively). Frequency counts of items used in the CATs varied across body part specific CATs (frequency reports available upon request) reflecting differing items and information per abilities tables per body part, but each of the CATs used all items.

3.5. RP

Neither θ_{CAT} nor θ_{IRT} measures of FS discriminated patients by gender regardless of body part affected, and neither θ_{CAT} nor θ_{IRT} measures of FS discriminated patients by age for patients with foot/ankle impairments (Table 4). All other θ_{CAT} and θ_{IRT} measures discriminated groups of patients in clinically logical and similar ways for the other independent variables. RP 95% CIs supported similar discriminating abilities of θ_{CAT} and θ_{IRT} measures.

4. Discussion

Results (1) support body part specific, that is, hip, knee, foot/ankle, CATs can be generated from LEFS items; (2) measures

Table 3
Item characteristics of the three body part specific scales

	Hip (n = 444)		Knee (n = 949)		Foot/ankle (n = 379)	
Item	Calibration (SE)	Infit/outfit	Calibration (SE)	Infit/outfit	Calibration (SE)	Infit/outfit
Turning	60.5(0.5)	0.99/0.85	64.0(0.4)	0.89/0.63	64.6(0.6)	0.94/0.69
Run uneven	60.8(0.5)	0.84/0.74	63.6(0.4)	0.83/0.65	66.0(0.6)	0.72/0.56
Run even	58.9(0.5)	0.89/0.83	61.2(0.4)	0.90/0.77	62.3(0.6)	0.93/0.81
Hopping	57.2(0.5)	1.02/0.90	60.9(0.4)	1.06/1.20	62.2(0.6)	1.11/1.45
Squatting	49.5(0.5)	1.09/1.10	55.3(0.4)	1.42/1.39	45.8(0.6)	1.32/1.31
Walking a mile	52.1(0.5)	0.96/0.90	52.4(0.3)	0.91/0.80	56.7(0.5)	0.86/0.79
Hobby	51.8(0.5)	1.20/1.40	51.6(0.4)	1.34/1.46	53.5(0.6)	1.27/1.30
Standing	50.6(0.5)	1.05/1.18	50.2(0.3)	0.99/0.95	52.9(0.5)	1.07/1.12
Heavy activities	52.2(0.5)	0.77/0.74	50.1(0.3)	0.80/0.78	47.8(0.5)	0.80/0.76
Stairs	46.6(0.5)	0.88/0.87	48.7(0.3)	1.01/1.05	46.9(0.5)	0.92/0.99
Working	47.3(0.5)	0.72/0.83	45.5(0.3)	0.83/0.88	47.4(0.5)	0.85/0.89
Walking blocks	43.5(0.5)	0.97/0.88	44.3(0.3)	0.90/0.83	48.5(0.5)	0.87/0.83
Lifting	44.5(0.5)	1.00/1.00	38.8(0.4)	1.23/1.16	36.3(0.6)	1.25/1.09
Car	41.1(0.5)	1.11/1.41	37.7(0.3)	0.87/1.02	34.8(0.6)	0.80/0.95
Bathing	39.4(0.6)	1.18/1.18	36.7(0.4)	1.17/1.15	35.2(0.6)	0.95/1.42
Light activities	39.6(0.5)	0.68/0.65	36.3(0.3)	0.66/0.69	37.0(0.6)	0.67/0.73
Shoes	38.7(0.6)	1.32/1.40	34.0(0.4)	1.16/1.39	34.4(0.7)	1.37/1.43
Walking between rooms	32.2(0.6)	1.02/1.23	30.4(0.4)	0.81/0.82	36.4(0.6)	1.02/1.34
Scale						
Mean(SD)	51(13)		45(13)		50(13)	
Infit	0.99		0.98		0.98	
Outfit	1.01		0.98		1.07	
Reliability	0.94		0.95		0.95	
Separation	4.14		4.59		4.25	
RMSE	3.18		2.84		2.94	
95% MDC	8.81		7.87		8.15	
Cronbach	0.96		0.96		0.96	

Item calibrations scaled 0 to 100 with higher values representing better lower extremity function. Items sorted by item calibrations for patients with knee impairments.

Abbreviations: SE, standard error; reliability, person reliability; separation, person separation; RMSE, root-mean-square error; mean(SD), average person measures (standard deviation); 95% MDC, 95% minimal detectable change; Cronbach, Cronbach's alpha.

of lower extremity FS generated using these CATs can discriminate known groups of patients in clinically logical ways; (3) θ_{CAT} measures were similar to θ_{IRT} measures in their discriminating abilities, but (4) because θ_{CAT} measures were estimated using on average six LEFS items, the CATs were 67% more efficient compared to using 18 unidimensional LEFS items and 70% more efficient compared to using all 20 original LEFS items. This represents a clear superiority for the body part-specific LEFS CATs when respondent burden is of concern.

Although most factor analytic results supported the IRT assumption of unidimensionality for 19 LEFS items, RMSEA values were disappointing, and there were more negative residuals than desired. These results can be interpreted to mean the LEFS items represent an essentially unidimensional scale [31] with adequate local independence of items, but the potential exists for some dependence of items. The practical significance of these results may be to underestimate the SE of θ_{CAT} and end the CATs too soon [30]. Our study design did not permit testing these possibilities. However, RP estimates were as expected, that is, θ_{CAT} measures were similar to θ_{IRT} measures in their discriminating ability, which may imply negligible underestimated θ_{CAT} SE.

The finding that FS items displayed DIF by body part was not unexpected psychometrically [49] or clinically. Our findings and others [49,65–67] suggest DIF will commonly affect calibration of FS or activities of daily living items. In most instances, it is reasonable to expect DIF in FS items, particularly by body part affected [49], because it is clinically appropriate for patients with different impairments to perceive the level of difficulty differently per item during the performance of a functional task depending on the body part affected. For example, in our analyses, people with hip impairments considered lifting an object like a bag of groceries from the floor to be more difficult compared to people with foot or ankle impairments ($t = -11.1$, $df = 809$, $P < .01$). It could be hypothesized that people with hip impairments could not shift their body weight away from their hip during the lift as well as people with foot or ankle impairments. Future studies should examine the relation between clinical correlates and psychometric DIF results as well as the practical implications of DIF when present.

Our DIF analysis results and those of others [49,65–67] were different than those reported by Haley et al. [53]. In their Physical & Movement domain of their Activity Measure for Postacute Care, only one item, "reaching overhead while standing," demonstrated DIF across diagnostic groups.

Table 4
Relative precision results

									F-statistics		
Body part	Variable	θ_{IRT}			θ_{CAT}			df [a]	θ_{IRT}	θ_{CAT}	RP (95% CI)
Hip (n = 444)	Acuity	Acute	Subacute	Chronic	Acute	Subacute	Chronic				
		44(2)	49(1)	53(1)	45(2)	49(1)	53(1)	2/441	8.02*	8.77*	1.1(0.8,1.5)
	Age	Young	Middle	Older	Young	Middle	Older				
		56(1)	51(1)	44(1)	56(1)	51(1)	45(1)	2/438	22.70*	19.36*	0.9(0.7,1.0)
	Gender	Male	Female		Male	Female					
		49(1)	51(1)		50(1)	52(1)		1/439	1.50	0.96	NA
Knee (n = 949)	Acuity	Acute	Subacute	Chronic	Acute	Subacute	Chronic				
		43(1)	47(1)	46(1)	42(1)	46(1)	45(1)	2/946	4.29*	4.44*	1.0(0.6,1.6)
	Age	Young	Middle	Older	Young	Middle	Older				
		47(1)	45(1)	42(1)	46(1)	45(1)	42(1)	2/942	6.75*	6.09*	0.9(0.7,1.2)
	Gender	Male	Female		Male	Female					
		46(1)	45(1)		46(1)	44(1)		1/942	2.47	2.41	NA
Foot/ankle (n = 379)	Acuity	Acute	Subacute	Chronic	Acute	Subacute	Chronic				
		47(2)	49(1)	53(1)	48(2)	49(1)	53(1)	2/376	4.11*	3.44*	0.8(0.5,1.2)
	Age	Young	Middle	Older	Young	Middle	Older				
		51(1)	50(1)	51(3)	52(1)	50(1)	51(3)	2/372	0.09	0.24	NA
	Gender	Male	Female		Male	Female					
		52(1)	50(1)		52(1)	51(1)		1/374	1.97	0.73	NA

Abbreviations: θ_{IRT}, person measures estimated using Rating Scale model and all items; θ_{CAT}, person measures estimated using computerized adaptive tests; RP, relative precision; 95% CI, 95% confidence interval; NA, not applicable because at least one *F*-statistics was not significant.

[a] Degrees of freedom for ANOVA for main effect/error.

* *F*-statistics is significant ($P < .05$).

Both studies used similar DIF analytic techniques, assessed physical activities, and used five response categories per item. However, the current sample and Haley et al.'s sample were quite different, which may explain the differences in DIF results. In our study, patients were younger, had specific orthopedic conditions, and all were able to be treated in outpatient clinics, while the patients in the Haley et al sample were older, many did not have specific orthopedic conditions, and were treated in a variety of postacute care settings, including inpatient rehabilitation facilities, transitional care units, ambulatory care facilities, and at home. Further testing is needed to describe possible biomechanical and clinical factors as well as psychometric techniques that may affect item DIF.

Identification and management of between groups DIF is evolving with no clear generally accepted method. In this study we used a simple comparison of item calibrations because the RSM assumes item discrimination is the same across all items [39]. However, an important psychometric issue is how much difference in item calibrations needs to be present to provide a practical impact on the measure of FS. We made a decision to consider five or more FS units out of 100 to be clinically important, but there are no standards governing this decision. Future analyses should examine this cut point.

The management of between groups DIF is also evolving. Some statistically control for the effects of DIF using structural equation modeling [65] or item calibration adjustment [67,68], while deletion or revision of items with DIF [66] or separating items into different unidimensional scales [30] have also been recommended. Here, we separated patients into three separate samples and generated CATs separately for each group. Future studies need to assess the practical impact of various methods of DIF management.

Improved efficiency, that is, 70% reduction in test length, of the LEFS CATs compared to the original 20 item LEFS is meaningful for patients because of reduced respondent burden for data entry. When improved efficiency of CATs is associated with equal or minimal reduction in measure precision, as in this sample, CATs are recommended, particularly if patient fatigue, age, or medical comorbidities may negatively influence validity of data collection if data collection is burdensome. However, it is common for θ_{CAT} measures to have the same or more error than θ_{IRT} measures, which is evident in the RP lower 95% CI bounds. Those interested in using CATs have to balance reduced patient burden with the possibility of increased measure error before deciding to use a computerized adaptive test.

4.1. Limitations

The Rasch model selected assumes all items are equally discriminating [39]. According to other studies [69,70], items used to assess function commonly differ in their discriminating ability [19,22,23]. Future studies should evaluate whether other IRT models fit LEFS data better, and if they do, assessment of practical implications of differences in model fit for CAT measure RP and discriminant validity are recommended.

Eighteen LEFS items represent a small item bank for testing functional abilities of patients with lower extremity impairments, and 18 items are fewer than the item bank of

25 polytomous items previously found to be adequate for a rating scale model CAT [43]. However, the LEFS CATs used four to six items to estimate a patient's FS before a stopping rule was satisfied for 87% of the patients with foot/ankle or knee impairments and 73% of patients with hip impairments. Only 6, 3, and 2% of patients with hip, knee and foot/ankle impairments, respectively, required more than 10 items before a stopping rule was satisfied. Therefore, 18 LEFS items appear to be enough to estimate a precise, that is, limited SE, measure of FS in these patients given the stopping rules of a SE <4 out of 100 or a θ_{CAT} that was stable (i.e., change over the last three items was <1 out of 100). In the current medical/business environment, less burden means more efficiency and less cost, including outcomes data collection [41,71].

There were a limited number of comparisons for between group DIF testing. Becuase results of DIF testing suggested DIF by body part, testing of DIF across more patient demographic variables is warranted. Further, there are a number of approaches to the analysis of DIF [67,68,72], and future research should compare the results obtained here with those based on other DIF statistics.

Acknowledgments

The authors thank John M. Linacre, PhD for his comments regarding statistical analyses, and Karon F. Cook, PhD, for her insightful comments regarding statistical analyses, results, and manuscript edits.

References

[1] Wainer H. Introduction and history. In: Wainer H, editor. Computerized adaptive testing. A primer. 2nd ed Mahway, NJ: Lawrence Erlbaum Associates; 2000. p 1–21.

[2] Hambleton RK. Emergence of item response modeling in instrument development and data analysis. Med Care 2000;38(Suppl II):II-60–5.

[3] Binkley JM, Stratford PW, Lott SA, Riddle DL. The lower extremity functional scale (LEFS): scale development, measurement properties, and clinical application. Phys Ther 1999;79:371–83.

[4] Lord FM, Novick MR. Statistical theories of mental test scores. Reading, MA: Addison-Wesley; 1968.

[5] Lord FM. Some test theory for tailored testing. In Holtzman WH, editor. Computer-assisted instruction, testing, and guidance. New York: Harper and Row; 1970. p 139–83.

[6] Sands WA, Waters BK, McBride JR, editors. Computerized adaptive testing. From inquiry to operation. Washington, DC: American Psychological Association; 1997.

[7] Ware JE, Bjorner JB, Kosinski M. Practical implications of Item Response Theory and computerized adaptive testing. A brief summary of ongoing studies of widely used headache impact scales. Med Care 2000;38(9, Suppl II):II-73–82.

[8] Ware JE, Kosinski M, Bjorner JB, et al. Applications of computerized adaptive testing (CAT) to the assessment of headache impact. Qual Life Res 2003;12:935–52.

[9] Dijkers MP. A computer adaptive testing simulation applied to the FIM instrument motor component. Arch Phys Med Rehabil 2003;84: 384–93.

[10] Haley SM, Coster WJ, Andres PL, Kosinski M, Ni P. Score comparability of short forms and computerized adaptive testing: simulation study with the activity measure for post-acute care. Arch Phys Med Rehabil 2004;85:661–6.

[11] Haley S, Jette A. Extending the frontier of rehabilitation outcome measurement and research. J Rehabil Outcome Meas 2000;4(4):31–41.

[12] Ware JE. Conceptualization and measurement of health-related quality of life: comments on an evolving field. Arch Phys Med Rehabil 2003;84(Suppl 2):S43–51.

[13] Velozo CA, Kielfofner G, Lai J-S. The use of Rasch analysis to produce scale-free measurement of functional activity. Am J Occup Ther 1999;53:83–90.

[14] Dodd BG, De Ayala RJ, Koch WR. Computerized adaptive testing with polytomous items. Appl Psychol Meas 1995;19(1):5–22.

[15] Hart DL, Wright BD. Development of an index of physical functional health status in rehabilitation. Arch Phys Med Rehabil 2002;83: 655–65.

[16] Cook KF, Roddey RS, Gartsman GM, Olson SL. Development and psychometric evaluation of the Flexilevel Scale of Shoulder Function. Med Care 2003;41(7):823–35.

[17] Patrick DL, Chiang Y. Convening health outcomes methodologists. Med Care 2000;38(9, Suppl II):II-3–6.

[18] Mills CN, Potenza MT, Fremer JJ, Ward WC, Eds. Computer-based testing. Building the foundation for future assessments. Mahwah, NJ: Lawrence Erlbaum; 2002.

[19] Lord FM. Applications of Item Response Theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum; 1980.

[20] van der Linden W, Hambleton RK. Handbook of modern Item Response Theory. New York: Springer-Verlang; 1997.

[21] Embretson SE, Reise SP. Item Response Theory for psychologists. Mahwah, NJ: Lawrence Erlbaum Associates; 2000.

[22] Hambleton RK, Swaminathan H, Rogers HJ. Fundamentals of Item Response Theory. Newbury Park, CA: Sage Publications; 1991.

[23] Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Med Care 2000;38(9, Suppl II):II-28–42.

[24] Resnik L, Hart DL. Using clinical outcomes to identify expert physical therapists. Phys Ther 2003;83(11):990–1002.

[25] Alcock GK, Stratford PW. Validation of the Lower Extremity Functional Scale on athletic subjects with ankle sprains. Physiother Can 2002;54:233–40.

[26] Stratford PW, Binkley JM, Watson J, Heath-Jones T. Validation of the LEFS on patients with total joint arthroplasty. Physiother Can 2000;52:97–105,110.

[27] Stratford PW. Getting more from the literature: estimating the standard error of measurement from reliability studies. Physiother Can 2004;56: 27–30.

[28] World Health Organization. International classification of functioning, disability and health. Geneva: World Health Organization; 2001.

[29] Lazarsfeld PF, Henry NW. Latent structure analysis. Boston, MA: Houghton-Mifflin; 1968.

[30] Wainer H, Mislevy RJ. Item response theory, item calibration, and proficiency estimation. In: Wainer H, editor. Computerized adaptive testing. A primer. 2nd ed. Mahway, NJ: Lawrence Erlbaum Associates. 2000; 61–100.

[31] Stout WF. A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. Psychometrika 1990;55(2):293–325.

[32] Muthén LK, Muthén BO. Mplus user's guide. Los Angeles, CA: Muthén & Muthén Publishers; 2001.

[33] Bjorner JB, Kosinski M, Ware JE. The feasibility of applying item response theory to measures of migraine impact: a re-analysis of three clinical studies. Qual Life Res 2003;12:887–902.

[34] Fabrigar LR, Wegener DT, MacCallum RC, Stahan EJ. Evaluating the use of exploratory factor analysis in psychological research. Psychol Methods 1999;4(3):272–99.

[35] Bentler P. Comparative fit indices in structural models. Psychol Bull 1990;107:238–46.

[36] Tucker L, Lewis C. A reliability coefficient for maximum likelihood factor analysis. Psychometrica 1973;38:1–10.

[37] Loehlin JC. Latent variable models: factor, path, and structural analysis. Mahway, NJ: Lawrence Erlbaum; 1998.

[38] March HW, Balla JR, Hau K. An evaluation of increment fit indices: a clarification of mathematical and empirical properties. In: Marcoulides GA, Schumacker RE, editors. Advanced structural equation modeling: issues and techniques. Mahwah, NJ: Lawrence Erlbaum; 1996. p 315–53.

[39] Andrich D. A rating formulation for ordered response categories. Psychometrika 1978;43:561–73.

[40] Rasch G. Probabilistic models for some intelligence and attainment tests. Chicago, IL: MESA Press; 1980.

[41] McHorney CA, Monahan PO. Postscript. Applications of Rasch analysis in health care. Med Care 2004;42(1, Suppl):I-73–8.

[42] Wright BD, Masters GN. Rating scale analysis. Chicago, IL: MESA Press; 1982.

[43] Dodd BG. The effect of item selection procedure and stepsize on computerized adaptive attitude measurement using the rating scale model. Appl Psychol Meas 1990;14(4):355–66.

[44] Andersen EB. The rating scale model. In: van der Linden W, Hambleton RK, editors. Handbook of Modern Item Response Theory. New York: Springer-Verlag; 1997. p 67–84.

[45] Haley SM, McHorney CA, Ware JE. Evaluation of the MOS SF-36 physical functioning scale (PF-10): I. Unidimensionality and reproducibility of the Rasch item scale. J Clin Epidemiol 1994;47(6): 671–84.

[46] White L, Velozo C. Assessing unidimensionality of the Oswestry Low Back Pain Disability Questionnaire. Arch Phys Med Rehabil 2002;83:822–31.

[47] Bond TG, Fox CM. Applying the Rasch model: fundamental measurement in the human sciences. Mahwah, NJ: Lawrence Erlbaum Associates; 2001.

[48] Steinberg L, Thissen D, Wainer H. Validity. In: Wainer H, editor. Computerized adaptive testing. A primer. 2nd ed. Mahway, NJ: Lawrence Erlbaum Associates; 2000. p 188–229.

[49] Hart DL. Assessment of unidimensionality of physical functioning in patients receiving therapy in acute, orthopedic outpatient centers. J Outcome Meas 2000;4(1):413–30.

[50] Jette DU, Jette AM. Physical therapy and health outcomes in patients with knee impairments. Phys Ther 1996;76(11):1178–87.

[51] Hart DL. The power of outcomes: FOTO Industrial Outcomes Tool—initial assessment. Work 2001;16:39–51.

[52] Linacre JM. A user's guide to WINSTEPS. Chicago, IL: MESA Press; 2004.

[53] Haley SM, Coster WJ, Andres PL, Ludlow LH, Ni P, Bond TLY, Sinclair SJ, Jette AM. Activity outcome measurement for postacute care. Med Care 2004;42(1 Suppl):I-49–61.

[54] Thissen D, Mislevy RJ. Testing algorithms. In: Wainer H, editor. Computerized adaptive testing. A primer. 2nd ed. Mahway, NJ: Lawrence Erlbaum Associates; 2000. p 101–34.

[55] Hart DL, Mioduski JE. CAT development and testing software user's guide. Knoxville, TN: FOTO, Inc.; 2004.

[56] Linacre JM. Estimating measures with known polytomous item difficulties. Rasch Meas Trans 1998;12(2):638.

[57] McHorney CA, Ware JE, Rogers W, Raczek AE, Rachel Lu JF. The validity and relative precision of MOS short and long form health status scales and Dartmouth COOP Charts. Results from the Medical Outcomes Study. Med Care 1992;30:MS253–65.

[58] Werneke M, Hart DL. Discriminant validity and relative precision for classifying patients with nonspecific neck and back pain by anatomic pain patterns. Spine 2003;28:161–6.

[59] Norquist JM, Fitzpatrick R, Dawson J, Jenkinson C. Comparing alternative Rasch-based methods vs raw scores in measuring change in health. Med Care 2004;42(1 Suppl):I-25–36.

[60] Fitzpatrick R, Norquist JM, Dawson J, Jenkinson C. Rasch scoring of outcomes of total hip replacement. J Clin Epidemiol 2003;56(1): 68–74.

[61] Liang MH, Larson MG, Cullen KE, Schwartz JA. Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. Arthritis Rheum 1985;28:542–7.

[62] 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;31:247–63.

[63] Ware JE, Kosinski M, Bayliss MS, McHorney CA, Rogers WH, Raczek A. Comparisons of methods for the scoring and statistical analysis of SF-36 health profiles and summary measures: results from the Medical Outcomes Study. Med Care 1995;33(4):AS264–79.

[64] McHorney CA, Haley SM, Ware JE. Evaluation of the MOS SF-36 physical functioning scale (PF-10): II. Comparison of relative precision using Likert and Rasch scoring methods. J Clin Epidemiol 1997;50(4):451–61.

[65] Fleishman JA, Spector WD, Altman BM. Impact of differential item functioning on age and gender differences in functional disability. J Gerontol Soc Sci 2002;57B(5):S275–82.

[66] Bjorner JB, Kreiner S, Ware JE, Damsgaard MT, Bech P. Differential item functioning in the Danish translation of the SF-36. J Clin Epidemiol 1998;51(11):1189–202.

[67] Tennant A, Penta M, Tesio L, Grimby G, Thonnard JL, Slade A, et al. Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model. Med Care 2004;42 (1, Suppl):I-37–48.

[68] Crane PK, van Belle G, Larson EB. Test bias in a cognitive test: differential item functioning in the CASI. Stat Med 2004;23:241–56.

[69] McHorney CA, Cohen AS. Equating health status measures with Item Response Theory. Illustrations with functional status items. Med Care 2000;38(9, Suppl II):II-43–59.

[70] McHorney CA. Use of item response theory to link 3 modules of functional status from the Asset and Health Dynamics Among the Oldest Old Study. Arch Phys Med Rehabil 2002;83:383–94.

[71] Lohr KN. Health outcomes methodology symposium. Summary and recommendations. Med Care 2000;38(9, Suppl II):II-194–208.

[72] Lai JS, Cella D, Chang CH, Bode RK, Heinemann AW. Item banking to improve, shorten and computerize self-reported fatigue: an illustration of steps to create a core item bank from the FACIT-Fatigue Scale. Qual Life Res 2003;12:485–501.