Clinical Interpretation of a Lower-Extremity Functional Scale–Derived Computerized Adaptive Test

Ying-Chih Wang, Dennis L. Hart, Paul W. Stratford, Jerome E. Mioduski

Background. The increasing use of computerized adaptive tests (CATs) to generate outcome measures during rehabilitation has prompted questions concerning score interpretation.

Objective. The purpose of this study was to describe meaningful interpretations of functional status (FS) outcome measures estimated with a body part–specific CAT developed from the Lower-Extremity Functional Scale (LEFS).

Design. This investigation was a prospective cohort study of 8,714 people who had hip impairments and were receiving physical therapy in 257 outpatient clinics in 31 states (United States) between January 2005 and June 2007.

Methods. Four approaches were used to clinically interpret outcome data. First, the standard error of the estimate was used to construct the 90% confidence interval for each CAT-generated score estimate. Second, percentile ranks were applied to FS scores. Third, 2 threshold approaches were used to define individual subject–level change: statistically reliable change and clinically important change. The fourth approach was a functional staging method.

Results. The precision of a single score was estimated from the FS score $\pm 4$. On the basis of the score distribution, 25th, 50th, and 75th percentile ranks corresponded to intake FS scores of 40, 48, and 59 and discharge FS scores of 50, 61, and 75, respectively. The reliable change index supported the conclusion that changes in FS scores of 7 or more units represented statistically reliable change, and receiver operating characteristic analyses supported the conclusion that changes in FS scores of 6 or more units represented minimal clinically important improvement. Participants were classified into 5 hierarchical levels of FS using a functional staging method.

Limitations. Because this study was a secondary analysis of prospectively collected data via a proprietary database management company, generalizability of results may be limited to participating clinics.

Conclusions. The results demonstrated how outcome measures generated from the hip LEFS CAT can be interpreted to improve clinical meaning. This finding might facilitate the use of patient-reported outcomes by clinicians during rehabilitation services.
Measures of health status that are based on patient-reported outcomes (PROs) are routinely used in clinical trials, observational studies, medical research, and clinical practice. Measures from PROs represent patient-centered measures and are recommended by the Institute of Medicine and the US Food and Drug Administration because such measures provide patients’ perceptions of their health status without a third party’s interpretation and allow clinicians to quantify the effectiveness of their treatments and monitor patients’ progress. As a result, the number of initiatives designed to assess quality (ie, change in health status) or describe value-based purchasing (ie, payment based on quality or cost) is on the rise; one such initiative has been simulated for outpatient rehabilitation with PROs.

The use of PROs during rehabilitation is increasing, but one barrier to integrating PRO measures into clinical practice is that scores often do not have meaning for clinicians. Many PRO measures have strong reliability and validity and are sensitive to change, but relatively few have been shown to present clinically meaningful information. Therefore, one of the challenges facing PRO developers is describing the clinical meaning of PRO measures and any differences observed, that is, clinically meaningful or important changes that can be used to determine whether score changes indicate true and meaningful changes in follow-up or outcome studies. Another challenge is communicating scores derived from standardized outcome measures to clinicians directly involved with patient care. The lack of clinical interpretation of derived outcome measures impedes clinicians’ use of the outcome measures during treatment of patients.

The present study builds on previous work in which we developed, simulated, and applied body part–specific computerized adaptive tests (CATs) for people seeking rehabilitation for a variety of impairments in outpatient physical therapy clinics. Here, we examine the clinical interpretation of patient-reported measures of functional status (FS) estimated with a CAT for people with hip impairments managed in outpatient therapy clinics participating with Focus On Therapeutic Outcomes, Inc (an international medical rehabilitation outcome data-base management company). The use of CATs offers several advantages over standard computer-administered or paper-and-pencil outcome instruments. Instead of examiners being given the same fixed test, items are selected according to an examinee’s ability. After each response, the examinee’s ability estimate is updated, and a subsequent item that optimizes the psychometric properties of the new estimate is selected. Consequently, the number of items administered is minimized while measure precision is maintained. Functional status assessed with the hip CAT is operationally defined as a patient’s perception of his or her ability to perform functional tasks.

The development, simulation, and use of the hip CAT were described previously. In brief, the item bank for the hip CAT was developed with items from the Lower-Extremity Functional Scale (LEFS), which has strong psychometrics and broad clinical and research acceptance. Previous results indicated that the LEFS CAT met the essential item response theory (IRT) assumptions of item unidimensionality and local independence. The LEFS items demonstrated differential item functioning for the lower extremity affected (hip, knee, or foot/ankle); that is, item difficulty parameter estimates varied across different body part impairment groups after a patient’s underlying ability was controlled for. For instance, people with knee or hip impairments perceived walking to be easier than people with foot/ankle impairments for walking between rooms, walking 2 blocks, walking 1.6 km (1 mile), and running. People with knee impairments perceived squatting to be more difficult compared with people with foot/ankle impairments. Therefore, the hip CAT was developed with items calibrated on the basis of data from people with hip impairments, making the hip CAT a body part-specific or condition-specific CAT.

When administered, on average, the hip CAT used 7 items to produce precise estimates of FS that adequately covered the content range with negligible floor and ceiling effects and that discriminated people in known clinical groups well. People who were older, had more chronic symptoms, had more surgeries, had more comorbidities, and did not exercise before receiving rehabilitation reported poorer (ie, lower) discharge FS than other people when intake FS was controlled for. Nonparametric receiver operating characteristic (ROC) analyses indicated that changes in FS scores of 6 or more units represented minimal clinically important improvement. When people were grouped by...
baseline FS measures and 4 separate ROC analyses were conducted (1 per quartile of intake FS measures), the results suggested that changes in FS scores of 11 or more, 6 or more, 4 or more, and 2 or more units represented clinically meaningful improvement for people in the first (intake FS, 0–40), second (intake FS, 40–48), third (intake FS, 48–59), and fourth (intake FS, 59–100) quartiles of intake FS measures, respectively. Clinically meaningful interpretations of the hip CAT have not been studied. For example, if a patient with a hip sprain presented in therapy with an intake FS score of 30, the following questions might be asked. How confident can I be about a reported score? How does my patient’s FS score compare with the scores of other patients? How much change is likely to represent a true change? How much improvement is likely to represent improvement that is clinically important to the patient? What does a specific score mean? To answer these questions, we constructed the 90% confidence interval (CI) for each score point estimate; established percentile ranks for FS scores; assessed statistically reliable change; assessed clinically important improvement; and described a functional staging approach. The first 4 methods provided statistical indexes, and the fifth method provided a graphical presentation to guide the clinical interpretation of a patient’s improvement in FS.

Method

Data Collection

The data collection process and the sample were described previously. In brief, data were collected by use of Patient Inquiry computer software. People seeking rehabilitation entered demographic data by using a computer in the clinic before their initial evaluation and therapy. Clinical staff could administer the CAT during physical therapy visits as a follow-up evaluation to monitor an individual’s status. Data were labeled “intake” when the CAT was completed before the initial evaluation (ie, at admission); data were labeled “discharge” when the CAT was completed at the time of discharge. The FS score change was defined by subtracting the FS score at intake from the FS score at discharge.

Data were selected from the CAT database if people were receiving treatment for a hip impairment (operationally defined as an orthopedic impairment of the pelvis, hip, or upper leg); if they were receiving outpatient physical therapy in clinics participating with Focus On Therapeutic Outcomes, Inc, in the United States between January 2005 and June 2007; and if they completed the LEFS CAT.

Setting and Participants

A convenience sample of 8,714 people who had hip impairments and were receiving outpatient physical therapy in 257 outpatient clinics in 31 states (United States) was analyzed. As described previously, the mean age of the participants was 56 years (SD=17, range=18–102). Most of the participants were women (65%). The participants were receiving treatment for hip (68%), upper-leg (25%), or pelvic (7%) impairments. Most of the participants reported that their symptoms were chronic (52%). The most common medical or surgical diagnoses were soft-tissue disorders of muscle, synovium, tendon, or bursa or enthesopathies (ICD-9 codes 725–729) (30%) and postsurgical conditions (CPT codes 20150–29999, including total hip replacement and open treatment of greater trochanteric fracture) (6%). More-detailed characteristics of the participants were described previously.

Hip LEFS CAT

During routine administration in the clinic, the CAT items were presented by asking the individual, “Today, do you or would you have any difficulty at all with:” followed by activities such as “performing heavy activities around your home.” The computer used the 5 original LEFS response categories. The LEFS rating scale categories are: (1) “extreme difficulty or unable to perform,” (2) “quite a bit of difficulty,” (3) “moderate difficulty,” (4) “a little bit of difficulty,” and (5) “no difficulty.” In addition, the participant could elect “not applicable” for any item; this response was recorded as missing data and was not used in the FS estimation.

CAT Algorithm

As described previously, the hip CAT used an item bank developed from the LEFS. In brief, the adaptive test started with the administration of the most informative item at a median level of difficulty (eg, going up or down 10 stairs, about one flight of stairs). A participant’s ability (FS) and associated standard errors (SE) were estimated. There were 2 stopping rules: (1) the SE for the provisional ability was less than 4 of 100 FS units (ability estimates were scaled from 0 to 100, with higher measures representing higher functioning), and (2) each change in provisional ability estimates for the last 3 administered items was less than 1 of 100. If a stopping rule was not met, the computer selected the most informative item given the current FS estimate. The computer continued to administer items until a stopping rule was satisfied, at which time a final estimate of ability and its SE were provided. The final FS score represented a point estimate of the lower-extremity FS of each participant on a scale from 0 to 100, with higher measures representing...
higher functioning. When this CAT algorithm is used, an individual can receive different set of items when taking CATs at different times, depending on changes in the person’s FS. Figure 1 summarizes the CAT algorithm.

Approaches to Deriving Meaningful Interpretations of Measurements

Interpreting a single scale score: How confident can I be about a reported score? Because the final FS score represented a point estimate of the FS of each participant, the 90% CI associated with the FS score point estimate was constructed to provide an estimate of precision. The width of the CI band represented an estimate of the precision of the measure. If a participant had a discharge FS score overlapping the 90% CI range of the intake FS score, the score change might not represent improvement but instead might be attributable to measurement error.

Establishing the percentile rank of an FS score: How does my patient compare with others? The percentile rank of a score represents the percentage of scores in a distribution to which a specific score is greater than or equal to in a defined population. The percentile rank provides additional information regarding the location of a patient’s condition along the functional continuum relative to the condition of a group of people with similar characteristics. Although the percentile rank could be based on a healthy population, we believed that people in such a population (a more general norm) would probably respond “no difficulty” to LEFS items or would obtain nearly maximum scores on the hip CAT (ie, FS scores close to 100). Therefore, we used the percentile rank to provide additional information regarding the location of a patient’s condition along the functional continuum relative to the condition of a group of people with similar hip impairments. To accommodate differences in FS scores at intake and at discharge, we generated 2 percentile ranks: a percentile rank for people at intake and a percentile rank for people at discharge.

Using 2 threshold approaches to define individual patient-level change. To examine the sensitivity of change, we used 2 threshold approaches to define individual patient-level change: statistically reliable change and clinically important change.34

![Diagram](https://example.com/diagram.png)
How much change is likely to represent a true change? Statistically reliable change reflects the statistical significance of individual change. We computed the reliable change index (RCI), \(^{17,35,36}\) a z test of a longitudinal change between intake and discharge, by using the following mathematical formula: 
\[
\text{RCI} = (X_2 - X_1)/\left(\sigma\sqrt{2(1-r)}\right)
\]
where \(X_2\) is the score at discharge, \(X_1\) is the score at intake, \(\sigma\) is the standard deviation of the score at intake, and \(r\) is the reliability coefficient.

When the RCI is 1.65 or larger, change is considered statistically reliable or significant at \(P<0.10.\) As computed, the minimal score change (ie, FS score change) required for a statistically reliable change \((X_2 - X_1 = 1.65\times\sigma\sqrt{2(1-r)})\) is equivalent to the upper limit of the 90% CI of the minimal detectable change (MDC\(_{90}\)). The RCI and minimal detectable change statistics can be interpreted, as 90% of truly stable people will display random fluctuations when assessed on 2 occasions within the range of MDC\(_{90}.\) Therefore, the MDC\(_{90}\) represents the smallest threshold for identifying statistically reliable change that is greater than random measurement error. A 90% CI was selected so that the results could be compared directly with the results of initial LEFS testing.\(^{14}\)

How much improvement is likely to represent an improvement that is clinically important to a patient? To assess clinically important change,\(^{17,39–41}\) as recommended by Stratford and Riddle,\(^{42}\) we used an anchor-based longitudinal method\(^{43,44}\) and a 15-point Likert-type scale (–7 to +7)\(^1\) to provide a global rating of change.\(^{40}\)

**Minimal clinically important improvement** was defined as a cut score of +3 or greater (+3 = somewhat better). The threshold was determined with non-parametric ROC\(^3^1\) analyses. The results were reported previously.\(^{18}\)

Here, we used the results to assist in the clinical interpretation of the FS derived from the hip CAT.

**Using a functional staging approach: what does a specific score mean?** Functional staging involves the classification of a patient into hierarchical outcome stages for the purpose of describing the patient’s expected FS in each stage.\(^3\) The development of a functional staging approach involves 3 steps: select a conceptual model or classification system that has hierarchical order and is in accordance with the underlying measurement construct, determine cut scores on the basis of the staging definitions established in the first step, and specify the expected performance in each stage.

**Select a conceptual model.** To classify lower-extremity function, we used the classification of walking described by Perry et al\(^{45}\) as the framework to construct the stages. As defined by Perry et al,\(^{45}\) lower-extremity function can be classified into several hierarchical functional levels: physiological ambulator, limited household ambulator, independent household ambulator, limited community ambulator, and independent community ambulator. Because the original target population in the classification of Perry et al was people who had sustained a stroke, we established the functional staging approach for our hip CAT by combining 2 “household ambulator” levels (2 low functional levels) and then added a higher level, “active community ambulator” (a more challenging functional level), to further distinguish the functional performance of subjects at the high end of functioning.

By definition,\(^{45}\) a “physiological ambulator” is an individual who can walk only for the purpose of exercise (stage 1). A “household ambulator” is an individual who can walk continuously for distances that are considered reasonable for walking inside the home but limited for walking in the community because of endurance, strength, or safety concerns (stage 2). A “limited community ambulator” is an individual who can walk regularly in the home and occasionally in the community (stage 3). An “independent community ambulator” is an individual who can walk for distances of at least 400 m (0.25 mile) independently in the community without physical or safety concerns (stage 4). Finally, an “active community ambulator” was defined as a individual who not only can walk 1.6 km (1 mile) with no difficulty but also can run on even ground with little difficulty (stage 5).

**Determine the cut scores.** Because our functional staging approach was based on the framework of the classification of walking ability, we selected specific walking items as our primary interest: walking between rooms, walking 2 blocks, walking 1 mile, and running on even ground. Four cut scores were needed to establish functional stages 1 to 5.

To determine the cut scores for functional stages along the FS continuous scale from 0 to 100, we took advantage of the inherent feature of IRT mathematics.\(^{32}\) We first analyzed the original hip CAT item bank\(^{20}\) by using the Andrich\(^{46}\) rating scale IRT model. This measurement model was selected because it is a latent structure model for polytomous responses to a set of test items and because it was the model used to develop the hip CAT. Within the
Andrich rating scale IRT model, each item was characterized by its category structure measure information (i.e., category probability curve), which illustrates the probability of endorsing the response to an item at a given level of ability. The cut scores between functional stages were determined by finding the “category boundaries” on the ability continuum at which the probability of endorsing one of the pair of contiguous responses was .5. Here, we use the term “category boundaries” to refer to the structure calibration (i.e., Andrich threshold) or step calibration at which adjacent categories were equally probable.

We performed an exploratory analysis on the basis of our initial defined hierarchical stages to identify item category thresholds for specific items for which patients at a given stage were very likely (i.e., a probability of at least 50%) to achieve the performance described. As a result, the cut scores between a physiological ambulator and a household ambulator and between a household ambulator and a limited community ambulator were determined by finding the thresholds between the “1” and “2” responses and between the “3” and “4” responses for the item “walking between rooms” (see original LEFS rating scale categories in earlier description of hip LEFS CAT). The cut score between a limited community ambulator and an independent community ambulator was determined by finding the threshold between the “2” and “3” responses for the item “walking 1.6 km (1 mile).” Finally, the cut score between an independent community ambulator and an active community ambulator was determined by finding the lower threshold between the “3” and “4” responses for the item “running on even ground.”

### Results

**Interpreting a Single Scale Score: How Confident Can I Be About a Reported Score?**

The SEs at discharge were similar to the SEs at intake, with differences of 0 to 0.25 score unit. Here, for brevity, we present only SEs associated with intake FS score estimates. On average, the SE for all levels of FS was 2.9, and the SE for 96% of subjects with FS intake scores between 20 and 80 was 2.4. For all levels of FS, the 90% CI was 95% CI for participants with FS scores between 20 and 80 was FS point estimate ±4.

**Establishing the Percentile Rank of an FS Score: How Does My Patient Compare With Others?**

The mean FS scores at intake and discharge were 49 (SD = 15) and 63 (SD = 18), respectively. On average, subjects improved by 13 FS score units. The 25th, 50th, and 75th percentile ranks corresponded to intake FS scores of 20, 25, and 30 and discharge FS scores of 50, 50, and 50, respectively. Table 1 lists detailed percentile ranks based on FS intake scores and FS discharge scores.

### Table 1.

Hip Computerized Adaptive Test (CAT) Percentile Ranks Based on Intake and Discharge Functional Status (FS) Scores

<table>
<thead>
<tr>
<th>Score</th>
<th>PRi (%)</th>
<th>PRd (%)</th>
<th>Score</th>
<th>PRi (%)</th>
<th>PRd (%)</th>
<th>Score</th>
<th>PRi (%)</th>
<th>PRd (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>1.9</td>
<td>0.3</td>
<td>46</td>
<td>45.8</td>
<td>17.7</td>
<td>66</td>
<td>89.0</td>
<td>61.6</td>
</tr>
<tr>
<td>25</td>
<td>3.3</td>
<td>0.5</td>
<td>48</td>
<td>51.4</td>
<td>21.8</td>
<td>68</td>
<td>91.1</td>
<td>65.4</td>
</tr>
<tr>
<td>30</td>
<td>7.2</td>
<td>1.1</td>
<td>50</td>
<td>56.1</td>
<td>26.5</td>
<td>70</td>
<td>92.6</td>
<td>68.5</td>
</tr>
<tr>
<td>32</td>
<td>9.7</td>
<td>1.7</td>
<td>52</td>
<td>60.8</td>
<td>31.0</td>
<td>72</td>
<td>93.9</td>
<td>71.6</td>
</tr>
<tr>
<td>34</td>
<td>12.9</td>
<td>2.8</td>
<td>54</td>
<td>66.4</td>
<td>36.3</td>
<td>75</td>
<td>95.4</td>
<td>76.1</td>
</tr>
<tr>
<td>36</td>
<td>17.3</td>
<td>4.0</td>
<td>56</td>
<td>70.0</td>
<td>39.7</td>
<td>85</td>
<td>97.7</td>
<td>84.1</td>
</tr>
<tr>
<td>38</td>
<td>22.2</td>
<td>5.7</td>
<td>58</td>
<td>74.7</td>
<td>44.9</td>
<td>90</td>
<td>98.6</td>
<td>89.8</td>
</tr>
<tr>
<td>40</td>
<td>27.7</td>
<td>8.2</td>
<td>60</td>
<td>78.9</td>
<td>48.5</td>
<td>95</td>
<td>98.9</td>
<td>91.6</td>
</tr>
<tr>
<td>42</td>
<td>32.6</td>
<td>10.7</td>
<td>62</td>
<td>82.0</td>
<td>52.7</td>
<td>97</td>
<td>99.0</td>
<td>93.0</td>
</tr>
<tr>
<td>44</td>
<td>40.6</td>
<td>14.6</td>
<td>64</td>
<td>85.6</td>
<td>56.5</td>
<td>98</td>
<td>100.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

The scale ranged from 0 to 100. Scores were estimated with the hip CAT at intake or discharge. The 25th, 50th, and 75th percentile ranks corresponded to intake FS scores of 40, 48, and 59 and discharge FS scores of 50, 61, and 75, respectively. PRi = percentile rank at intake, PRd = percentile rank at discharge.

### Specify the expected performance.

Once the initial conceptual functional staging approach was developed and the cut scores were determined by the structure calibration, we specified the expected performance in each stage on the basis of the Andrich rating scale IRT measurement model. Here, the expected performance represents the response categories that a subject most likely would report.
Using Threshold Approaches to Define Meaningful Change

How much change is likely to represent a true change? On the basis of our previous full-length LEFS data set of 444 subjects with hip impairments, the internal consistency reliability coefficient was .96.20 With a standard deviation at intake of 15, the minimal score change (discharge FS – intake FS) needed for the RCI to be 1.65 or larger was 7 FS score units. Therefore, RCI analyses indicated that changes in FS scores of 7 or more units represented statistically reliable change.

How much improvement is likely to represent an improvement that is clinically important to the patient? As described previously,18 ROC analyses indicated that changes in FS scores of 11 or more, 6 or more, 4 or more, and 2 or more units represented clinically meaningful improvement for subjects in the first (intake FS, 0–40), second (intake FS, 0–60), third (intake FS, 0–70), fourth (intake FS, 0–70), and fifth (intake FS, 0–100) stages of functional impairment.

Table 2. Performance Expected in Each Functional Stage

<table>
<thead>
<tr>
<th>Stage</th>
<th>Functional Status Score Range</th>
<th>Classification</th>
<th>Running on Even Ground</th>
<th>Walking 1.6 km (1 Mile)</th>
<th>Walking 2 Blocks</th>
<th>Walking Between Rooms</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0–20</td>
<td>Physiological ambulator</td>
<td>Unable</td>
<td>Unable</td>
<td>Unable</td>
<td>Unable</td>
</tr>
<tr>
<td>2</td>
<td>21–35</td>
<td>Household ambulator</td>
<td>Unable</td>
<td>Unable</td>
<td>A lot of difficulty</td>
<td>Moderate difficulty</td>
</tr>
<tr>
<td>3</td>
<td>36–48</td>
<td>Limited community ambulator</td>
<td>Unable</td>
<td>A lot of difficulty</td>
<td>Moderate difficulty</td>
<td>A little difficulty</td>
</tr>
<tr>
<td>4</td>
<td>49–61</td>
<td>Independent community ambulator</td>
<td>Moderate difficulty</td>
<td>A little difficulty</td>
<td>A little difficulty</td>
<td>No difficulty</td>
</tr>
<tr>
<td>5</td>
<td>62–100</td>
<td>Active community ambulator</td>
<td>A little difficulty</td>
<td>No difficulty</td>
<td>No difficulty</td>
<td>No difficulty</td>
</tr>
</tbody>
</table>
Using a Functional Staging Approach: What Does a Specific Score Mean?

Using the functional staging method, we could compare a participant’s FS score with the functional stages to better interpret the participant’s FS score. For example, participants classified in functional stage 3 (scores 36–48) reported being limited community ambulators, having moderate difficulty walking 2 blocks and having quite a bit of difficulty to extreme difficulty walking 1.6 km (1 mile). Participants classified in functional stage 4 (scores 49–61) reported being independent community ambulators, having a little bit of difficulty to moderate difficulty walking 1.6 km (1 mile) and having moderate difficulty to quite a bit of difficulty running on even ground. A guideline for interpreting FS stages is provided in Table 2, in which the expected performance of walking items in each stage is specified and simplified for easy use. We also simplified the original LEFS rating scale categories and replaced them with “unable” (if the expected response was “extreme difficulty”), “a lot of difficulty” (for “quite a bit of difficulty”), “moderate difficulty,” “a little difficulty” (for “a little bit of difficulty”), and “no difficulty.”

We then classified participants with hip impairments by using our functional staging method. For our sample of 8,714 participants receiving therapy for hip impairments, both intake and discharge data were available for 4,032 participants (i.e., 46% completion rate). Of these 4,032 participants, 2,204 (55%) improved to the next (i.e., higher) functional stage from intake to discharge, and 220 (5%) dropped to a lower functional stage.

Table 3 shows the frequency distribution of the functional stage classification on the basis of intake and discharge FS scores. The percentages of participants in each functional stage at intake were 1.3% (stage 1), 12.2% (stage 2), 36.9% (stage 3), 50.1% (stage 4), and 19.5% (stage 5). At discharge, the percentages of participants in each functional stage at intake were 0.3% (stage 1), 3.2% (stage 2), 18.2% (stage 3), 28.4% (stage 4), and 49.8% (stage 5). For each functional stage, the percentages of participants who

### Table 3.

<table>
<thead>
<tr>
<th>Intake Functional Stage</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Total no. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>8</td>
<td>21</td>
<td>8</td>
<td>9</td>
<td>54 (1.3)</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>81</td>
<td>179</td>
<td>142</td>
<td>85</td>
<td>491 (12.2)</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>13</td>
<td>435</td>
<td>540</td>
<td>480</td>
<td>1,489 (36.9)</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>7</td>
<td>90</td>
<td>383</td>
<td>732</td>
<td>1,212 (30.1)</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>2</td>
<td>9</td>
<td>74</td>
<td>701</td>
<td>786 (19.5)</td>
</tr>
<tr>
<td>Total no. (%)</td>
<td>13 (0.3)</td>
<td>131 (3.3)</td>
<td>734 (18.2)</td>
<td>1,147 (28.4)</td>
<td>2,007 (49.8)</td>
<td>4,032 (100)</td>
</tr>
</tbody>
</table>

*Participants were classified into functional stages on the basis of their intake and discharge functional status scores. Data are reported as number (percentage) of participants.*
improved to a higher functional stage from intake to discharge were 85% (stage 1), 82% (stage 2), 68% (stage 3), and 60% (stage 4).

**A Clinical Example**

To illustrate how to use these strategies to enhance clinically meaningful interpretations, we will answer the questions posed earlier. A patient (referred to here as “Jack”) sustained an unspecified sprain of hip and thigh (ICD-9 code 843.90). Jack, an 18-year-old man, sought treatment at a clinic because of limited function related to his sprain. His intake FS score was 30, and his discharge FS score, after 7 outpatient therapy visits, was 77 (FS score change = 47). He considered his overall condition to be “a great deal better” and reported a global rating of change of +6.

To visualize Jack’s responses to our hip CAT, we plotted all of his responses in Figure 3. The structure of Figure 3 is equivalent to that of Figure 2, except that Jack’s responses are circled: light gray circles identify the responses at intake, and dark gray circles identify the responses at discharge.

The 90% CI estimate of Jack’s intake FS score location was 26 to 34. Compared with the percentile ranks for other patients with a variety of hip impairments, Jack’s percentile rank at intake was 7, indicating that his lower-extremity function (estimated by FS) exceeded that of 7% of the patients who also had hip impairments at intake. The functional staging algorithm classified Jack as a household ambulator (stage 2).

At discharge, the 90% CI estimate of Jack’s FS score location was 73 to 81. Compared with the percentile ranks for other patients, Jack’s percentile rank at discharge was 76. The functional staging classification suggested that Jack improved to being classified as an active community ambulator (stage 5). Jack’s improvement, indicated by an FS score change of 47, was considered to be statistically reliable (RCI = 11.08 [1.65]; FS score change, >7) and clinically meaningful (FS score change by quartile, >11); in addition, the improvement was supported by Jack’s perspective that his condition was “a great deal better.”

**Discussion**

Computerized adaptive tests are commonly used as standardized tests for licensure, certification, and admission but have only recently begun to be used to collect routine clinical data in busy outpatient rehabilitation clinics. In the present study, we followed the ap-
Interpretation of LEFS CAT

proaches recommended by Jette et al,5 Stratford et al,15 Schmitt and Di Fabio,34 and Hays et al35 to derive more clinically meaningful interpretations of outcome measures, which might facilitate the use of these measures by clinicians.

Several methods have been proposed to enhance clinical interpretation; they have focused on responsiveness or have used a single numeric index to define the minimal detectable change.52,53 In the present study, clinical interpretation was provided by several different approaches, including traditional score distribution (eg, standard error, percentile), responsiveness indexes, and functional staging. Given that these approaches offer different perspectives, we used them to generate a more comprehensive and clinically relevant interpretation of the FS scores estimated with the hip CAT, so that clinicians may select an approach depending on need.

As described by Liang,16 providing information on function to clinicians has not led to changes in the process of care or to improved patient health. We believe that it is the responsibility of instrument developers to identify meaningful intraindividual change and to report the clinical interpretation of the measures. If these steps are taken, clinicians, managers of outpatient rehabilitation clinics, and researchers can use the measures in their clinical practices for the purpose of improving outcomes.

The item bank for the hip CAT was developed from items in the LEFS.14 On the basis of the original LEFS development and validation results, Binkley et al14 suggested that the potential error associated with a score on the LEFS (original LEFS score range of 0–100) at a given point in time was 5.3 (or 6.6%) scale points (90% CI), the MDC90 was 9 (or 11.3%) scale points, and the minimal clinically important difference was 9 (or 11.3%) scale points. In the present study, a 90% CI was selected so that the results could be compared directly with the results of the initial LEFS testing by Binkley et al.14 With the hip CAT (hip CAT score range = 0–100), the 90% CI was FS point estimate ± 4 (or 4%), the minimal change in FS scores required for a statistically reliable change was 7 units (or 7%, equivalent to the MDC90), and minimal clinically important improvement was indicated by changes in FS scores of at least 6 units (or 6%) overall and 11, 6, 4, and 2 units by quartile. The overall better performance of the hip CAT scores may have been attributable to the use of IRT methods, which improve the linearity of categorically scored scales,54 or the CAT administrative algorithm, which uses the most informative items given a patient’s current FS estimate. The use of items whose difficulty may not be matched to a patient’s functional ability may introduce error in FS measures.55 Additionally, we focused only on hip impairments and did not include other lower-extremity musculoskeletal impairments, and our sample size was relatively large (N = 8,714) compared with that in the study of Binkley et al14 (N = 107).

One of the strengths of the present study was the functional staging approach, which allowed clinical interpretation through a visual display of scales with strong, clinically logical, hierarchically structured item banks developed by use of IRT methods. Without such a mathematical structure, functional staging would be dependent on clinical intuition. The display transforms IRT results into a format that can be applied directly by clinicians without a background in IRT methods.

The basic idea for functional staging was derived from bookmark56 and key form57 methods. The bookmark method provides guidelines for setting a standard by following a prescribed, rational system of procedures that result in the assignment of a number to differentiate between 2 or more degrees of performance. A key form is a unique product of IRT methods that provides a visual display of the expected response patterns for the underlying measures. By combining features from both methods, functional staging provides a classification system and a visual display measurement tool. With functional staging, a rating instrument can be presented as a simple classification form.58–61 Clinicians can use the visual display to inspect the expected performance of a patient on the basis of the FS estimates derived from the hip CAT. Additionally, clinicians can derive an approximation of an individual FS score or examine unexpected ratings without a computerized analysis, as described by Linacre57 and Kielhofner et al.61 In this regard, improving the interpretation of FS measures may assist in a clinician’s understanding of a patient’s abilities.58

One concern about using functional staging is whether such a specific classification system should be tailored to groups of people with distinct diagnoses, such as people with total hip replacement and people with arthritis, because the discharge functional goals of these groups may be different. If a particular patient’s responses do not clearly fit into a particular stage, then differential item functioning analysis should be performed to inspect the item hierarchical order of that patient’s specific diagnostic group. The other concern is the need to validate the functional staging classification system because the thresholds were determined arbitrarily. Further studies should be done to validate the functional staging classification and to determine whether different functional staging systems should be estab-
lished for groups of people with distinct diagnoses.

There are several limitations of the present study. First, to estimate the RCI, we used single internal consistency reliability. Using internal consistency reliability instead of a test-retest reliability coefficient may inflate the interpretation of the precision of a measure. Test-retest reliability has not yet been determined for CAT estimates of FS; thus, the impact magnitude is unknown. Investigations of the estimation of the RCI with a test-retest reliability coefficient are warranted.

Second, we used score distributions from people with similar diagnoses to establish our percentile ranks based on the assumption that a more general norm from a healthy population would produce nearly maximum scores on the hip CAT, whereas scores from people with similar diagnoses would be more representative of people treated for hip impairments in outpatient rehabilitation. Nonetheless, the true score distribution of general or healthy populations is unknown, and establishing norms based on general or healthy populations would be valuable.

Finally, 4,032 patients completed the hip CAT at rehabilitation discharge. These data represented a completion rate (ie, percentage of patients with both intake data and discharge data) of 46%. The missing values led to a concern about potentially biased estimates. Compared with patients with just intake data in this sample, patients with both intake and discharge data in this sample were older ($t=7.8; df=8,197; P<.001$), exercised more ($\chi^2=7.1; df=2; P=.029$), were more likely to have had a surgical procedure on a hip ($\chi^2=11.1; df=4; P=.025$), and had lower FS intake scores ($t=2.6; df=8,957; P=.009$). The groups did not differ in sex ($\chi^2=0.2; df=1$; $P=.686$), acuity of symptoms ($\chi^2=3.5; df=2; P=.172$), number of comorbidities ($\chi^2=2.0; df=3; P=.579$), or medication use for their conditions ($\chi^2=2.2; df=1; P=.136$) at intake. However, based on the fact that the hip CAT was administered in 257 outpatient physical therapy clinics in 31 states (United States) and that the data set was large, we believe that the potential impact of subject selection bias was reduced, although the true impact is unknown.

Conclusion

We demonstrated clinically meaningful interpretations of patient-reported FS outcome measures estimated with a body-part-specific CAT for use in routine clinical practice. The hip CAT currently is used routinely in many outpatient rehabilitation clinics across the United States and is integrated with an electronic health records system in Israel, attesting to its efficiency and utility. The results of the present study should improve the clinical interpretation of outcomes and encourage future studies.

Dr Wang, Dr Hart, and Mr Stratford provided concept/idea/research design. Dr Wang and Dr Hart provided writing and data analysis. Dr Hart and Mr Mioduski provided data collection and project management. Mr Mioduski provided facilities/equipment and clerical support. Mr Stratford provided consultation (including review of manuscript before submission).

This project was approved by the Institutional Review Board for the Protection of Human Subjects of Focus On Therapeutic Outcomes, Inc.

This article was received November 8, 2008, and was accepted May 6, 2009.


References


34 Schmitz JS, Di Fabio RP. Reliable change and minimum important difference (MID) proportions facilitated group responsiveness comparisons using individual threshold criteria. *J Clin Epidemiol.* 2004;57:1008–1018.


