Clinical Interpretation of Outcome Measures Generated From a Lumbar Computerized Adaptive Test

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

Background. A computerized adaptive test (CAT) provides a way of efficiently estimating functional status in people with specific impairments.

Objective. The purpose of this study was to describe meaningful interpretations of functional status (FS) estimated using a lumbar CAT developed using items from the Back Pain Functional Scale (BPFS) and selected physical functioning items.

Design and Setting. This was a prospective longitudinal cohort study of 17,439 patients with lumbar spine impairments in 377 outpatient rehabilitation clinics in 30 states.

Outcome Measures. Patient self-reports of functional status were assessed using a lumbar CAT (0–100 scale).

Methods. Outcome data were interpreted using 4 methods. First, the standard error of the estimate was used to construct a 95% confidence interval for each CAT estimated score. Second, percentile ranks of FS scores were presented. Third, 2 threshold approaches were used to define individual patient–level change: minimal detectable change (MDC) and clinically important change. Fourth, a functional staging model, the Back Pain Function Classification System (BPFCS), was developed and applied.

Results. On average, precision of a single score was estimated by FS score±4. Based on score distribution, 25th, 50th and 75th percentile ranks corresponded to intake FS scores of 44, 51, and 59, and discharge FS scores of 54, 62, and 74, respectively. An MDC95 value of 8 or more represented statistically reliable change. Receiver operating characteristic analyses supported that changes in FS scores of 5 or more represented minimal clinically important improvement. The BPFCS appeared clinically logical and provided insight for clinical interpretation of patient progress.

Limitations. The BPFCS should be assessed for validity using prospective designs.

Conclusions. Results may improve clinical interpretation of CAT-generated outcome measures and assist clinicians using patient-reported outcomes during physical therapist practice.
Measures of health status that are based on patient-reported outcomes are evolving, with self-report questionnaires increasingly being used in both medical research and clinical practice. Measures based on patient-reported outcomes provide the patients’ perception of their health status and experience, and many measures have been shown to be reliable, valid, sensitive to change, responsive, and usable. As a result, several institutes are encouraging the medical research community to use patient-reported outcomes to support treatment effectiveness and longitudinally monitor patients’ progress. In the management of spinal disorders, outcome measures are recommended for following the patient’s self-reported function and supplementing traditional clinician examination data (e.g., physiologic measures such as range of motion and muscle strength [force-generating capacity]) and supporting clinical success.

The current study builds on previous work where we developed, simulated, applied, and validated body part–specific computerized adaptive testing (CAT) applications and developed functional staging models for patients 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 using a CAT for patients with lumbar impairments managed in outpatient rehabilitation clinics participating with Focus On Therapeutic Outcomes, Inc (FOTO), an international medical rehabilitation outcomes database management company.

The lumbar CAT was designed to efficiently evaluate each patient’s function by selecting items that provided the maximum information related to the patient’s functional status. In contrast to giving each patient fixed-length questionnaires, a CAT administration selects items from the item bank one at a time based on an administrative algorithm. The lumbar CAT starts by administering the most informative item at median-level difficulty (i.e., “Do you or would you have any difficulty at all with any of your usual work, housework, or school activities?”). Patients select answers to each item, and the CAT estimates the patient’s FS score with associated standard errors (SEs). The CAT continues to administer items until a stopping rule is satisfied. The final FS score represents a point estimate of FS for each patient on a linear scale of 0 to 100, with higher measures representing higher functioning.

Development, simulation, validation, and use of the 25-item lumbar spine–specific CAT have been described elsewhere. Briefly, the item bank for the lumbar CAT was developed using items from the Back Pain Functional Scale (BPFS), the physical functioning scale of the 36-Item Short-Form Health Survey questionnaire (SF-36), and selected physical functioning items from other scales. Previous results supported that the lumbar CAT met essential item response theory (IRT) assumptions of unidimensionality and local independence, and FS measures were precise, valid, sensitive to change and responsive, and practical, which supported lumbar CAT use in clinical and research applications. Differential item functioning, which examines whether items perform differently across defined groups, was negligible for levels of symptom acuity, sex, age, and surgical history in the lumbar CAT. The person reliability, analogous to Cronbach α measuring internal consistency, was .92. A responsiveness index of minimal detectable change (MDC) at the 95% confidence interval (CI) was 9. The lumbar CAT was 72% more efficient than using all 25 items to estimate FS measures (i.e., time to administer was shorter). On average, patients took less than 2 minutes (SD <1 minute) to answer 7 CAT items, which produced precise estimates of FS that adequately covered the content range with negligible floor and ceiling effects.

Clinically meaningful interpretations of the lumbar CAT have not been studied. For example, if a patient with low back seen for therapy has an intake FS score of 33 and a discharge FS score of 73 (0–100 scale), information other than that the patient has improved 40 points should provide additional information for clinical reasoning and care planning. Therefore, the purpose of this study was to describe meaningful interpretations of FS outcome measures estimated using the body part–specific lumbar CAT. Clinicians may gain more information by being able to answer the following questions: (1) How confident can I be in a reported score? (2) How does my patient compare with others? (3) How much change is likely to represent a true change? (4) How much improvement is likely to represent a clinically important improvement to the patient? (5) What does a specific score mean?

To answer these questions, we utilized several approaches recommended by Jette et al., Hays et al., Schmitt and Di Fabio, and Stratford.
et al.\textsuperscript{32} and found to be positive in our own studies\textsuperscript{15–17} to derive more clinically meaningful interpretations of outcome measures. To do so, we: (1) constructed a 95% CI for each score point estimate, (2) established percentile ranks of FS scores, (3) assessed minimal detectable change (MDC), (4) assessed minimal clinically important improvement (MCII), and (5) developed and applied a functional staging approach. The first 4 methods provided statistical indexes, and the fifth method provided a graphical presentation to guide clinical interpretation of the patient’s improvement in interpreting estimated FS scores generated by the lumbar CAT.

**Method**

**Data Collection**

Data collection and the sample are described elsewhere.\textsuperscript{18} Patients seeking rehabilitation entered demographic data and completed self-report surveys using Patient Inquiry computer software developed by FOTO\textsuperscript{*} prior to initial evaluation and therapy. The CAT was administered at admission prior to initial evaluation and was administered again at discharge. Demographic data were entered by clinical staff.

Data were selected from the CAT database if patients: (1) were 18 years of age or older, (2) were managed for an orthopedic impairment of the lumbar spine, (3) received outpatient physical therapy, and (4) completed the CAT between July 2007 and August 2008.

**Setting and Participants**

A convenience sample of 17,439 patients with lumbar spine impairments receiving outpatient physical therapy in 377 outpatient clinics in 30 states (United States) was analyzed. The mean age of the patients was 51 years (SD=17, range=18–100). Sixty percent were female. Fifty-two percent of the patients reported their symptoms as chronic (onset more than 90 days earlier) versus subacute (22–90 days earlier) (25%) and acute (0–21 days earlier) (23%). Approximately 10% reported having no comorbid conditions, and 25% reported having 6 or more comorbid conditions. Patients received an average of 9 visits (SD=6). Identification of medical or surgical diagnoses was optional in the data collection, but of the patients with medical/surgical codes (57%), the most common diagnoses were associated with spinal pathology (ICD-9 codes 720–724) (29%); soft tissue disorders (ICD-9 codes 725–729) (18%); postsurgical conditions, including disectomy and fusion (5%); or sprains and strains, including sacroiliac region, lumbar spine, sacrum (ICD-9 codes 846–848) (4%).

**Patient Selection Bias**

To assess the potential for patient selection bias introduced by incomplete data, we used chi-square statistics with standardized deviates \((\text{observed minus expected})/\text{square root of expected})\) for categorical data and Student t tests for continuous variables to assess differences between patients who had both intake and discharge data compared with patients with just intake data.

**Approaches to Deriving Meaningful Interpretations of Measurements**

**Interpreting a single scale score:** How confident can I be in a reported score? Because each FS score represented a point estimate for each patient’s FS, the 95% CI band associated with the point estimate FS score was constructed to provide an estimate of precision of the measure (ie, FS score \(\pm 1.96 \times \text{CSEM}\)), where CSEM is the conditional standard error of measurement. We estimated 10 CSEMs, 1 for each of the 10 scale ranges (0–10, 11–20, \ldots, 91–100) by averaging the SEs in each FS scale range. Based on IRT measurement models, SE varies by level of FS. Therefore, different score ranges have different magnitudes of SEs. Extreme scores are likely to have larger SEs because less information is obtained from patients at the extremes (ie, patients with very low or very high functioning).

**Establishing the percentile rank of an FS score:** How does my patient’s functional score compare with other patients’ scores? The percentile rank (PR) of a point estimate FS score refers to the proportion of scores in a distribution to which a specific score is greater than or equal while comparing an individual’s score related to others in a defined population. To accommodate differences in FS scores at intake compared with discharge, we generated 2 PRs: one for patients at intake (PR\(i\)) and one for patients at discharge (PR\(d\)).

**Using 2 threshold approaches to define individual patient-level change.** To examine sensitivity to change, we used 2 threshold approaches to define individual patient-level change: (1) statistically reliable change and (2) clinically important change.\textsuperscript{31}

**How much change is likely to represent a true change?** Statistically reliable change, as described by Schmitt and Di Fabio,\textsuperscript{31} reflects the statistical significance of individual change. To assess statistically reliable change, we computed the MDC\textsuperscript{35} as: \(\text{MDC}_{95} = 1.96 \times \sqrt{2} \times \text{CSEM}\), where 1.96 represents the \(z\) value associated with a 95% CI. As above, we estimated 10 CSEMs, 1 for each of the 10 scale ranges (0–10, 11–20, \ldots, 91–100). For each CSEM, we multiplied the result by the square root of 2 to accommodate
the 2 measurements involved in measuring change: intake and discharge. As computed, MDC_{95} represents the smallest threshold for identifying statistically reliable change greater than random measurement error.35

How much improvement is likely to represent a clinically important improvement to the patient? To assess clinically important change, as recommended by Stratford and Riddle,34 we used an anchored-based longitudinal method35 with a 15-point Likert-type scale (−7 to +7)† to provide a global rating of change (GROC).36 In this study, we chose the upper limit of +3 or more (+3 = “somewhat better”) in defining the MCII because previous studies11-13 showed that this cut-score provided adequate assessment of important improvement. The threshold was determined using nonparametric receiver operating characteristic (ROC)37 curve analysis. Results have been described elsewhere.18 Here, we used the results to assist in clinical interpretation of the FS scores derived from the lumbar CAT.

Using a functional staging approach: What does a specific score mean? Can I use the score to assist in care planning? Functional staging29 refers to a set of hierarchical outcome levels used to classify a patient into different stages describing the individual’s expected FS within each level. The concept of functional staging evolved from Bookmark38 and Keyform39 methods. The Bookmark method provides guidelines to set a standard following of a prescribed, rational system of procedures that result in the assignment of a number to differentiate among 2 or more degrees of performance.38 A Keyform is a unique product from IRT measurement models that provides a visual display of expected response patterns of an underlying measure.39 By combining features from both methods, functional staging provides a visual display of a clinically logical classification system based on IRT measurement methods.

No functional staging model exists for patients with lumbar impairments, so we: (1) built a conceptual model using FS items from the lumbar CAT item bank whose hierarchical order represents clinical reasoning and the underlying measurement construct, (2) determined the IRT-based cut-scores between stages based on the hierarchical order established in step 1, and (3) specified the expected performance within each level.

Building a conceptual model. We developed a functional classification system designed to be used in patients with lumbar spine impairments in both clinical and research settings that describes a functional hierarchy of physical activities: the Back Pain Function Classification System (BPPCS). The BPPCS is based on both the International Classification of Functioning, Disability and Health (ICF)40 framework of activities performed and a clinically logical hierarchical progression of functional stages paralleling the numeric FS scores from 0 (“low FS”) to 100 (“high FS”) in the lumbar CAT scale. For example, patients typically report improvements in functional ability as they progress during the treatment episode (ie, they perform more physically challenging activities, thereby moving to a higher functional stage). We conceptually based our functional stages on increasingly more difficult activities commonly reported to be difficult for patients with lumbar impairments.

Table 1 shows the general heading (activity level) and operational definition of each level of the BPPCS. We classified lumbar FS into 5 hierarchical functional levels: (1) is exceedingly limited in the ability to perform easy, routine functions; (2) exhibits extreme difficulty performing usual work or household activities; (3) exhibits moderate difficulty performing usual work or household activities; (4) exhibits little difficulty performing usual work or household activities; and (5) back to normal life performing rigorous daily activities. Higher levels represent better functioning (ie, higher FS scores).

Determining the cut-scores. Grouping activities into the 5 BPPCS functional staging levels required establishment of 4 cut-scores along the FS continuum. To determine the cut-scores, we performed 3 steps: (1) calibrated scores to the rating scale IRT model (RSM), (2) identified conceptual thresholds, and (3) associated conceptual thresholds with score thresholds.

To determine the cut-scores, we performed exploratory analysis based on our initial conceptualization of the hierarchical FS levels to identify item-category thresholds for specific items where patients at the current level were likely to accomplish the performance described in the BPPCS. To achieve this, we matched the response categories of lumbar CAT items to the operational descriptions of 2 adjacent levels.

As a result, the cut-score between level 1 (exceedingly limited functionally) and level 2 (extreme difficulty performing work or household activities) was determined by finding the threshold between the “1” and “2” responses for the item “BPFS—putting on shoes and socks” (see
footnote for original 6-point BPFSC). In a similar manner, we identified the other cut-scores. For example, the cut-score between level 2 and level 3 (moderate difficulty performing work or household activities) was determined by finding the threshold between the “2” and “3” responses for the item “BPFSClifting groceries from the floor.” The cut-score between level 3 and level 4 (little difficulty performing usual hobbies) was determined by finding the threshold between the “4” and “5” responses for the item “BPFSCperforming usual work or household activities.” Finally, the cut-score between level 4 and level 5 (get back to normal life; can perform rigorous daily activities) was determined by finding the threshold between the “2” and “3” responses for the item “BPFSCperforming vigorous activities.”

After conceptually identifying itemcategory thresholds, we performed statistical analyses to find the cut-scores between functional stages along the 0 to 100 FS continuous scale using Winsteps (version 3.68). To do so, we took advantage of the inherent feature of the IRT mathematics. We first analyzed the original lumbar CAT item bank data using the Andrich RSM. Within the RSM, each item was characterized by its category structure measure information (ie, category probability curves), which illustrates the probability of endorsing an item’s response at a given level of ability (eg, the probability of responding “limited a little” to the item “walking one block” given a patient’s estimated FS score). The cut-scores between functional stages were obtained by finding the step calibration (ie, Andrich threshold) at which adjacent categories are equally probable (ie, probability equals .5) based on what we have defined in the first step in determining the cut-scores. That is, we applied IRT statistical methods to find the threshold (a point on the FS scale) that represents the .5 probability of being in one category (eg, “unable”) or the next higher category (eg, “extreme difficulty”) in an item (eg, putting on your shoes or socks) as the cut-score between levels (in this case, levels 1 and 2).

In the final step in determining the cut-scores, the results of functional staging hierarchical levels were inspected. If the results were not logical (eg, clustered cut-scores), the same procedures were repeated until a clear cut-score was identified.

Specifying the expected performance. Once the initial conceptual functional staging was developed and the cut-scores were determined by the structure calibration, we specified the expected performance in each stage based on the RSM model. Here, expected performance represents the patient’s predicted (ie, most likely) response.

Results
Patient Selection Bias

Over the testing time, 17,439 patients completed the lumbar CAT at intake, and of these, 6,607 completed the CAT at discharge (38% completion rate). Compared with patients with complete data (ie, intake and discharge), patients with just intake data (Tab. 2) tended to have lower FS intake scores (49.9

| Table 1. Back Pain Function Classification System (BPFCS) |
|---------------------------------|---------------------------------|---------------------------------|
| BPFS Level                      | Activity Level                  | Operational Definition          | FS Score Range* |
| Level 1                         | Exceedingly limited with routine functions | A person is unable to perform or is limited a lot in performing activities such as donning shoes or socks, bathing or dressing, getting into or out of a chair or bed, and walking around a room. | 0–23 |
| Level 2                         | Exhibits extreme difficulty performing usual work or household activities | A person has regained limited routine functioning but still has extreme difficulty performing activities such as work or household activities, driving a car for an hour, lifting groceries from the floor, and walking 1 block. | 24–38 |
| Level 3                         | Exhibits moderate difficulty performing usual work or household activities | A person has moderate difficulty performing daily activities around the home such as work or household activities, bending or stooping, climbing one flight of stairs, lifting overhead, and standing for an hour. | 39–57 |
| Level 4                         | Exhibits little difficulty performing usual work or household activities and hobbies | A person can perform with a little bit of difficulty indoors and outdoors activities such as usual hobbies, bending or stooping, standing for an hour, work or household tasks, and driving for an hour. | 58–77 |
| Level 5                         | Back to normal life performing vigorous daily activities | A person has no difficulty walking more than 1 mile or any limitation in performing activities requiring vigorous work, sports, or lifting tasks. | 78–100 |

*The functional scale (FS) score (0–100 scale) range was determined by the follow-up Bookmark and Keyform methods.
Comparisons of Patients Who Completed Intake and Discharge Computerized Adaptive Tests (CATs) Versus Patients Who Completed Intake CATs Only*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Intake CATs Only</th>
<th>Intake and Discharge CATs</th>
<th>P (^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>Mean (SD)</td>
<td>n</td>
</tr>
<tr>
<td>Intake FS</td>
<td>10,832</td>
<td>49.9</td>
<td>6,607</td>
</tr>
<tr>
<td>Age (y)</td>
<td>10,832</td>
<td>50.4</td>
<td>6,607</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>4,311</td>
<td>40</td>
<td>2,687</td>
</tr>
<tr>
<td>Female</td>
<td>6,521</td>
<td>60</td>
<td>3,919</td>
</tr>
<tr>
<td>Missing</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Acuity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acute</td>
<td>2,295</td>
<td>22</td>
<td>1,561</td>
</tr>
<tr>
<td>Subacute</td>
<td>2,724</td>
<td>25</td>
<td>1,713</td>
</tr>
<tr>
<td>Chronic</td>
<td>5,785</td>
<td>53</td>
<td>3,320</td>
</tr>
<tr>
<td>Missing</td>
<td>28</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>Surgery</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>8,854</td>
<td>82</td>
<td>5,461</td>
</tr>
<tr>
<td>1</td>
<td>1,277</td>
<td>12</td>
<td>783</td>
</tr>
<tr>
<td>2</td>
<td>384</td>
<td>4</td>
<td>224</td>
</tr>
<tr>
<td>3</td>
<td>158</td>
<td>1</td>
<td>71</td>
</tr>
<tr>
<td>≥4</td>
<td>132</td>
<td>1</td>
<td>65</td>
</tr>
<tr>
<td>Missing</td>
<td>27</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Exercise</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3×/wk or more</td>
<td>3,962</td>
<td>37</td>
<td>2,561</td>
</tr>
<tr>
<td>1–2×/wk</td>
<td>2,864</td>
<td>26</td>
<td>1,697</td>
</tr>
<tr>
<td>Seldom</td>
<td>4,006</td>
<td>37</td>
<td>2,349</td>
</tr>
<tr>
<td>Missing</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Payer</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fee-for-service</td>
<td>392</td>
<td>4</td>
<td>276</td>
</tr>
<tr>
<td>Litigation</td>
<td>24</td>
<td>0</td>
<td>28</td>
</tr>
<tr>
<td>Medicaid</td>
<td>560</td>
<td>5</td>
<td>184</td>
</tr>
<tr>
<td>Medicare A</td>
<td>435</td>
<td>4</td>
<td>254</td>
</tr>
<tr>
<td>Medicare B</td>
<td>1,536</td>
<td>14</td>
<td>1,176</td>
</tr>
<tr>
<td>Patient</td>
<td>146</td>
<td>1</td>
<td>54</td>
</tr>
<tr>
<td>HMO</td>
<td>1,135</td>
<td>10</td>
<td>577</td>
</tr>
<tr>
<td>PPO</td>
<td>4,203</td>
<td>39</td>
<td>2,421</td>
</tr>
<tr>
<td>Workers' compensation</td>
<td>960</td>
<td>9</td>
<td>744</td>
</tr>
<tr>
<td>Other</td>
<td>1,339</td>
<td>12</td>
<td>749</td>
</tr>
<tr>
<td>Missing</td>
<td>102</td>
<td>2</td>
<td>144</td>
</tr>
<tr>
<td>Comorbidities</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 or 1</td>
<td>2,690</td>
<td>25</td>
<td>1,641</td>
</tr>
<tr>
<td>2 or 3</td>
<td>3,072</td>
<td>28</td>
<td>1,967</td>
</tr>
<tr>
<td>4 or 5</td>
<td>2,292</td>
<td>22</td>
<td>1,480</td>
</tr>
<tr>
<td>≥6</td>
<td>2,752</td>
<td>25</td>
<td>1,508</td>
</tr>
<tr>
<td>Missing</td>
<td>26</td>
<td>0</td>
<td>11</td>
</tr>
</tbody>
</table>

* FS = functional status, HMO = health maintenance organization, PPO = preferred provider organization.

** P values for chi-square statistic.

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

Because the SEs at discharge were similar to the SEs at intake, with an average of 0.15 score unit differences (minimum=0.00, maximum=0.33), we report only SEs associated with intake FS score estimations for brevity. Table 3 shows the CSEMs, one for each of the 10 intake scale ranges (0–10, 11–20, ..., 91–100). The CSEM was smallest (ie, 2.2) over the scale range of 41 to 60 and increased at both ends of the scale range. Few patients (1.1%) had intake FS scores less than 20, and 1.7% of patients had intake FS scores greater than 80, so we calculated 2 FS point estimates: one for the majority of patients with intake FS scores between 20 and 80 (ie, FS estimate=6.9) and one for patients with <20 or >80 FS intake scores (FS estimate=7.3).

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

The mean (SD) FS scores at intake and discharge were 50 (13) and 65 (16), respectively. On average, patients improved by 14 FS score units. Based on score distribution, 25th, 50th, and 75th percentile ranks corresponded to intake FS scores of 44, 51, and 59 and discharge FS scores of...
54, 62, and 74, respectively. Table 4 lists detailed PRs based on FS intake scores (PRi) and FS discharge scores (PRd).

**How Much Change Is Likely to Represent a True Change?**

On average, the mean of 95% CI upper limits of MDC95 values for all patients was 13.9, but the mean MDC95 value for the 97% of patients with FS intake scores between 20 and 80 was 7.8 (Tab. 3). For patients with FS scores <20 or >80, the mean MDC95 value was 15.

**How Much Improvement Is Likely to Represent a Clinically Important Improvement to the Patient?**

There were 2,612 patients with both GROC and FS change data. Of these, 449 (17.2%) reported no change (ie, GROC scores <3), and 2,163 (82.8%) reported improvement (ie, GROC scores ≥3). Receiver operating characteristic analyses supported 5 or more for all patients: 9, 5, 3, and 5 or more FS change units per quartile of intake score represented clinically meaningful improvement (ie, MCII) (Tab. 3). We conducted the ROC analyses in the following ways: (1) based on the entire sample using all patients and (2) based on score quartile. For all patients, the results supported that 5 or more represented the cutoff score of the MCII. The results suggested that 9 or more, 5 or more, 3 or more, and 5 or more FS change scores represented clinically meaningful improvement for patients in the first, second, third, and fourth quartiles of FS intake measures, respectively.

**Using a Functional Staging Approach: What Does a Specific Score Mean? Can I Use the Score to Assist Clinical Practice?**

Figure 1 displays the functional staging of our BPFCS. The figure shows the expected response (the horizontal bars) to a given item as a function of the underlying lower-extremity ability (ie, FS) estimated by the lumbar CAT. In Figure 1, the lumbar CAT items are listed in descending order of difficulty in the left column: more challenging items are listed on the top (ie, “Vigorous activities” and “Participating in a recreational sport”), and the easier items are listed at the bottom (ie, “Walking around a room” and “Bathing or dressing yourself”). Beneath the figure is the FS score continuum, ranging from 0 to 100 (higher values represent more functioning toward the right), separated by different levels of functional staging from level 1 (left or lower functioning) to level 5 (right or higher functioning). Lower levels (eg, levels 1–2) describe the patient’s limited functional ability, and higher stages (eg, levels 4–5)

---

### Table 3.
CSEM, MDC95, and MCII per FS Range of the Lumbar CAT*

<table>
<thead>
<tr>
<th>FS Range at Intake</th>
<th>No. of Patients (%)</th>
<th>CSEM</th>
<th>95% CSEM</th>
<th>MDC95</th>
<th>MCII</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–10</td>
<td>0.5</td>
<td>11.4</td>
<td>22.2</td>
<td>31.5</td>
<td>9</td>
</tr>
<tr>
<td>11–20</td>
<td>0.6</td>
<td>5.3</td>
<td>10.4</td>
<td>14.7</td>
<td>9</td>
</tr>
<tr>
<td>21–30</td>
<td>3.3</td>
<td>3.5</td>
<td>6.9</td>
<td>9.8</td>
<td>9</td>
</tr>
<tr>
<td>31–40</td>
<td>15.9</td>
<td>2.6</td>
<td>5.1</td>
<td>7.2</td>
<td>9</td>
</tr>
<tr>
<td>41–50</td>
<td>27.2</td>
<td>2.2</td>
<td>4.2</td>
<td>6.0</td>
<td>9</td>
</tr>
<tr>
<td>51–60</td>
<td>30.8</td>
<td>2.2</td>
<td>4.3</td>
<td>6.1</td>
<td>3</td>
</tr>
<tr>
<td>61–70</td>
<td>15.6</td>
<td>2.7</td>
<td>5.3</td>
<td>7.5</td>
<td>5</td>
</tr>
<tr>
<td>71–80</td>
<td>4.4</td>
<td>3.7</td>
<td>7.3</td>
<td>10.3</td>
<td>5</td>
</tr>
<tr>
<td>81–90</td>
<td>0.8</td>
<td>5.4</td>
<td>10.6</td>
<td>15.0</td>
<td>5</td>
</tr>
<tr>
<td>91–100</td>
<td>0.9</td>
<td>11.4</td>
<td>22.2</td>
<td>31.5</td>
<td>5</td>
</tr>
</tbody>
</table>

*FS= functional status, CSEM=conditional standard error of measurement at intake, 95% CSEM=1.96 × CSEM, MDC95=minimal detectable change (95% confidence interval) given intake FS, MCII=minimal clinically important improvement per quartile of FS intake scores, CAT=computerized adaptive test.

### Table 4.
Lumbar Computerized Adaptive Test (CAT) Percentile Ranks Based on Intake and Discharge Functional Status Scores (0–100 Scale)*

<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</td>
<td>0</td>
<td>46</td>
<td>35</td>
<td>12</td>
<td>66</td>
<td>90</td>
<td>56</td>
</tr>
<tr>
<td>25</td>
<td>2</td>
<td>1</td>
<td>48</td>
<td>39</td>
<td>14</td>
<td>68</td>
<td>92</td>
<td>59</td>
</tr>
<tr>
<td>30</td>
<td>4</td>
<td>1</td>
<td>50</td>
<td>44</td>
<td>16</td>
<td>70</td>
<td>94</td>
<td>64</td>
</tr>
<tr>
<td>32</td>
<td>5</td>
<td>1</td>
<td>52</td>
<td>52</td>
<td>20</td>
<td>72</td>
<td>96</td>
<td>71</td>
</tr>
<tr>
<td>34</td>
<td>6</td>
<td>2</td>
<td>54</td>
<td>57</td>
<td>23</td>
<td>76</td>
<td>97</td>
<td>77</td>
</tr>
<tr>
<td>36</td>
<td>8</td>
<td>2</td>
<td>56</td>
<td>68</td>
<td>30</td>
<td>85</td>
<td>99</td>
<td>88</td>
</tr>
<tr>
<td>38</td>
<td>14</td>
<td>4</td>
<td>58</td>
<td>73</td>
<td>34</td>
<td>90</td>
<td>99</td>
<td>88</td>
</tr>
<tr>
<td>40</td>
<td>17</td>
<td>5</td>
<td>60</td>
<td>77</td>
<td>38</td>
<td>95</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>42</td>
<td>21</td>
<td>6</td>
<td>62</td>
<td>86</td>
<td>49</td>
<td>97</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>44</td>
<td>24</td>
<td>8</td>
<td>64</td>
<td>88</td>
<td>52</td>
<td>99</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

*Based on score distribution, 25th, 50th, and 75th percentile ranks corresponded to intake FS scores of 44, 51, and 59 and discharge FS scores of 54, 62, and 74, respectively. Score=FS score estimated by the lumbar CAT at either intake or discharge, PRi=percentile rank at intake, PRd=percentile rank at discharge.
indicate the patient can perform more rigorous activities and have little difficulty performing work tasks. The threshold probability results identified 23 as the cut-score between functional levels 1 and 2, 38 as the cut-score between functional levels 2 and 3, and so forth. Items are displayed using either 3 response categories for physical functioning items or 6 levels for BPFS items.

Using the functional staging method, we can compare the patient’s FS score with the functional stages to better interpret the patient’s FS score. The expected responses of each lumbar CAT item at each functional level can be obtained by drawing a vertical line over an FS measure (x-axis) on Figure 1. By doing so, clinicians can see all of the expected responses to all items even when the patient did not answer the items during administration of the lumbar CAT. If a clinician has a patient’s FS scores reported at 2 points in time, the clinician can track the patient’s progress by drawing 2 vertical lines and inspecting the overall FS change. The information also is useful when the clinician would like to understand the patient’s answer to a specific question asked by the CAT at intake but not at discharge.

In this sample of patients with lumbar impairments receiving outpatient physical therapy, the percentages of patients in each functional staging level at intake and discharge are presented in Table 5. Patients were classified into functional staging levels based on their intake (rows) and discharge (column) FS scores.

**A Clinical Example**

To illustrate how to use these strategies to enhance clinically meaningful interpretation, we will answer the questions posed. Our patient, “Mr Jones” (male, age 25 years), came to the clinic due to acute non-specific or mechanical low back pain. His FS score at admission was 33, and his discharge FS score after receiving 4 outpatient therapy visits was 73 (FS score change of 40). To visualize his responses to our lumbar CAT, we plotted all of his responses in Figure 2: yellow circles identify his responses at intake, and purple
circles identify his responses at discharge. Mr Jones received 8 items at intake and 14 items at discharge using the lumbar CAT. The increase in the number of items administered may be due to Mr Jones’ higher functional abilities reported at discharge. Specifically, more questions are required for higher (as well as lower) functional scores during the CAT algorithm to accurately estimate the patient’s function and satisfy the CAT stopping rules.

The 95% CI estimate of his intake FS score location was 28 to 38. Compared with other patients with a variety of lumbar impairments, Mr Jones’ PR at intake (PRi) was 5, indicating that his FS (estimated by FS score) exceeded that of 5% of the patients at intake who also had lumbar impairments. The functional staging algorithm classified Mr Jones as having “extreme difficulty performing work or household activities” (level 2). During the lumbar CAT administration, Mr Jones reported having a lot of limitation walking one block and performing his usual work. He reported having a little limitation walking around a room and bathing or dressing.

At discharge, Mr Jones’ 95% CI estimate of his score location was 66 to 80. Compared with other patients at discharge, Mr Jones’ PR at discharge (PRd) was 72. The functional staging classification suggested Mr Jones improved to level 4 (“little difficulty performing usual hobbies”). With an improvement of FS score change of 40, Mr Jones’ improvement was considered to be greater than measurement error (FS score change was >7, which was >MDC95) and clinically meaningful (FS score change by quartile > MCII cut-score of 9), as supported by Mr Jones’ perspective that his condition was “a very great deal better.” At the time of discharge, Mr Jones reported having no difficulty performing many functional tasks, including participating in recreational sports, walking more than a mile, and climbing several flights of stairs. He reported having little difficulty performing heavy activities and standing for an hour and a little limitation in little lifting or carrying groceries and lifting overhead to a cabinet.

### Discussion

The results of the current study followed approaches recommended by Jette et al., Hays et al., Schmitt and Di Fabio, and Stratford et al. to derive more clinically meaningful interpretations of outcomes measures, which might facilitate use of these measures by clinicians in their routine clinical practice. Current and previous and our preliminary results support the validity of the model (Tab. 5).

#### Table 5.
Frequency Distribution of the Functional Staging Classification: Patients Were Classified Into Functional Staging Levels Based on Their Intake (Rows) and Discharge (Columns) Functional Status (FS) Scores

<table>
<thead>
<tr>
<th>Discharge FS staging</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Total</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intake FS staging</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>9</td>
<td>22</td>
<td>36</td>
<td>30</td>
<td>15</td>
<td>112</td>
<td>1.7</td>
</tr>
<tr>
<td>2</td>
<td>17</td>
<td>118</td>
<td>354</td>
<td>244</td>
<td>81</td>
<td>814</td>
<td>12.4</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>83</td>
<td>1,268</td>
<td>1,743</td>
<td>560</td>
<td>3,665</td>
<td>55.9</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>3</td>
<td>1,142</td>
<td>1,007</td>
<td>656</td>
<td>1,810</td>
<td>27.6</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>34</td>
<td>116</td>
<td>153</td>
<td>2.4</td>
</tr>
<tr>
<td>Total</td>
<td>40</td>
<td>226</td>
<td>1,802</td>
<td>3,058</td>
<td>1,428</td>
<td>6,554</td>
<td>100</td>
</tr>
<tr>
<td>%</td>
<td>0.6</td>
<td>3.4</td>
<td>27.5</td>
<td>46.7</td>
<td>21.8</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

The primary strength of this study is that functional staging allows clinical interpretation of FS scores generated by the lumbar CAT through a visual display of functional stage levels using IRT methods. Because there is no well-established back pain functional classification system based on clinical measurement that is commonly used in medical research and clinical practice, we developed the BPFCS based on IRT methods and our initial conceptualization of the hierarchical functional staging levels and previous studies. The functional staging model seems clinically relevant, and our preliminary results support the validity of the model (Tab. 5). However, validation of our functional staging model is needed.

Numerous methods have been proposed to enhance clinical interpretation of outcomes measures, but the majority have focused on sensitivity to change and responsiveness using...
Ostelo et al. developed practical guidance on frequently used measures of pain and FS for patients with low back pain. They proposed the following minimally important change values:

- Visual analog scale (0–100): 15 (15%)
- Numerical rating scale (0–10): 2 (20%)
- Roland Disability Questionnaire (0–24): 5 (21%)
- Oswestry Disability Index (0–100): 10 (10%)
- Quebec Back Pain Disability Questionnaire (0–100): 20 (20%)

None of these scales is linear. In the current CAT study using a 0–100 linear scale, clinically meaningful improvement was indicated by changes in FS scores of at least 9 (9%), 5 (5%), 3 (3%), and 5 (5%) units by quartile of intake scores. The apparent improved responsiveness of the lumbar CAT measures compared with the measures Ostelo et al. studied may occur because IRT methods improved linearity of categorically scored scales.
Clinical Interpretation of a Lumbar CAT

Hypotheses await future study. A single outcome score provides only a general sense of the patient’s status, which is insufficient when a clinician wants to interpret an outcome measure. A total score does not reveal how a patient performs on specific items, and a CAT may not ask the same items over time if the patient improved functionally. Using functional staging enhances clinical interpretation of CAT-generated outcome measures because a clinician can identify expected response patterns related to underlying FS measures from the visual display of the functional levels. Other authors have described how functional staging models in health care can assist in interpretation of clinical outcomes.26,32 Our results support these previous descriptions by showing how clinicians can use the visual display of functional staging (Figs. 1 and 2) to see observed responses and predict missing responses, as well as how they can apply clinically pertinent descriptions to the patient’s current or anticipated FS. Besides observing fitted items (ie, observed response = expected response), ratings that deviate from expected patterns appear useful to therapists. Clinicians should consider whether there is a logical reason why the client received an unexpected pattern of items or responses. Clinicians can inspect items where the patient reported having little difficulty and those where the patient reported having extreme difficulty. Clinicians can set clinical goals by using the items that the patient is likely to accomplish as short-term goals and challenging items as long-term goals. If a clinician has the patient’s FS scores reported at 2 points in time, the clinician can track the patient’s progress by drawing 2 vertical lines and inspecting the overall FS change. Clinicians can use functional staging to track a specific question asked by the CAT at intake but not at discharge to monitor a patient’s status.

There are several limitations of this study. Because this study was a secondary analysis of prospectively collected data via a proprietary database management company (FOTO), the researchers were not in control of the data collection procedure. Generalizability of results may be limited to patients entering the data and participating clinics because there was the potential for selection biases related to differences in characteristics of patients taking the CAT and other variables, as well as simply being treated in participating clinics compared with clinics not participating. Missing values in demographic information are inevitable. We reported the clinical interpretation of the lumbar CAT based on a sample of patients with various lumbar spine disorders. Similar procedures should be validated across different specific diagnostic groups in patients with lumbar spine disorders.

In addition, the validity of utilizing a retrospective GROC approach has been criticized because of the potential for recall bias, and the rating results might be influenced by status at discharge since the GROC relies on retrospective judgments experienced over weeks or months. The retrospective rating is valid to the extent that clinicians routinely ask patients whether they have noticed a change and incorporate the response into the decision-making process. Moreover, the GROC approach does not consider measurement precision and thus has unknown reliability.35 Nonetheless, many researchers have proposed the retrospective GROC approach as one external anchor to capture a patient’s perception of important improvement, along with ROC analysis as a valid sensitivity-to-change method.11–15,36

There might be confusion related to the association between our estimates of measurement error, MDC and MCII, which represent different concepts and distributions and require different estimation methods. Estimates of SE are a direct result of IRT methods and demonstrated measurement error varied over the FS continuum. Our estimates of 95% CI upper limits of MDC were based on the SEs from IRT methods, so SEs and MDCs differed over the FS continuum. Estimates of MCII were based on classical test theory methods that assume measurement error is constant over the FS continuum (although we adjusted the MCII estimates by quartile of intake FS measures), and we did not report 95% CIs of MCII, which is psychometrically inappropriate. Therefore, although it is common to see values of SE smaller than those of MDC, which are smaller than those of MCII, using classical test theory methods, one should approach comparisons of our values with caution because values of SE and MDC were developed using IRT methods and values of MCII were developed using classical test theory methods. Further study is warranted to improve understanding of these values developed from different psychometric philosophies.

The results of this study advance the clinical interpretation of the FS scores from the lumbar CAT. The methods are identical to those previously published for CATs that estimate FS measures for patients with hip,15 knee,16 and foot and ankle17 impairments. Although all of the methods are identical, the current project applied those methods to a sample of patients with lumbar impairments. Using IRT methods with
Clinical Interpretation of a Lumbar CAT

different items assessing FS in different types of patients produces different thresholds used to identify the functional stages, which means the functional stages will be different for each type of patient. We used the same sample in a validation study, which addressed the measurement properties of the lumbar CAT. Given that the “validation” and “clinical interpretation” used different analytical methods and the article would have become too lengthy if combining 2 parts, we feel that this article is necessary.

Conclusion
With clinical interpretation and immediate test scoring and reporting using CAT administration, clinicians, researchers, and policy makers will benefit as more constructs can be assessed, displayed, reported, and interpreted.

Dr Wang, Dr Hart, and Mr Stratford provided data collection and participants. Dr Wang and Dr Hart provided data analysis. Dr Wang provided project management. Dr Hart, Dr Mioduski provided consultation (including review of manuscript before submission).

Dr Wang, Dr Hart, and Dr Mioduski acknowledge that they are employees of Focus On Therapeutic Outcomes, Inc (FOTO), the database management company that manages the data analyzed in the study. Dr Mioduski programmed the software used to collect the data and managed the data collected from the CAT. Mr Werneke is independent of FOTO but works for CentraState Medical Center, which uses FOTO software to manage patients receiving outpatient rehabilitation. Mr Werneke was involved in the conceptual development of the functional staging classification system of the project. Mr Stratford is independent of FOTO but is one of the developers of the Back Pain Functional Scale, on which part of the lumbar CAT was developed.

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

This article was submitted November 10, 2009, and was accepted April 27, 2010.


References


