Translating Shoulder Computerized Adaptive Testing Generated Outcome Measures into Clinical Practice

Ying-Chih Wang, OTR/L, PhD
Focus On Therapeutic Outcomes, Inc., Knoxville, Tennessee
Sensory Motor Performance Program, Rehabilitation Institute of Chicago, Chicago, Illinois
Department of Occupational Science & Technology, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin

Dennis L. Hart, PT, PhD
Focus On Therapeutic Outcomes, Inc., Knoxville, Tennessee

Karon F. Cook, PhD
Department of Rehabilitation Medicine, University of Washington, Seattle, Washington

Jerome E. Mioduski, MS
Focus On Therapeutic Outcomes, Inc., Knoxville, Tennessee

ABSTRACT:
Study Design: Prospective longitudinal cohort study.
Introduction: Increased use of computerized adaptive tests (CATs) to generate outcome measures during rehabilitation has stimulated questions concerning score interpretation.
Purpose of the Study: The purpose of the study was to describe meaningful interpretations of scores from patient self-report shoulder functional status (FS) outcome measures estimated using a shoulder CAT (score range 0–100).
Methods: We applied four approaches to the clinical interpretation of outcomes data from 30,987 patients with shoulder impairments receiving outpatient rehabilitation in 518 clinics in 30 states (United States) between August 2007 and July 2009. First, we used standard error of estimates to construct 95% confidence intervals for each CAT estimated score. Second, we estimated the percentile rank (PR) of FS scores. Third, we used two threshold approaches to define individual patient-level change: statistically reliable change (i.e., minimal detectable change or MDC) and clinically important change. Fourth, we developed and applied a functional staging model, the Shoulder Function Classification System (SFCS).
Results: Precision of a single score was estimated by FS score ±4. Based on score distribution, 25th, 50th, and 75th PRs corresponded to intake FS scores of 43, 52, and 59 and discharge FS scores of 59, 68, and 80, respectively. MDC calculations indicated that changes in FS scores of 11 or more units represented statistically reliable change. FS score increments of eight or more units were estimated to represent minimal clinically important improvement based on receiver operating characteristic. The five-level SFCS was judged to be clinically logical and provide insight for clinical interpretation of patient progress.
Conclusions: Results may improve clinical interpretation of CAT-generated outcome measures and assist clinicians using patient-reported outcomes during clinical practice.
Level of Evidence: Therapy level 2c.

Correspondence and reprint requests to Ying-Chih Wang, OTR/L, PhD, Rehabilitation Institute of Chicago, 345 East Superior, Suite 1406, Chicago, IL 60611; e-mail: <inga-wang@northwestern.edu>.

0894-1130/© 2010 Hanley & Belfus, an imprint of Elsevier Inc. All rights reserved.
doi:10.1016/j.jht.2010.06.001


Patient-reported outcomes (PROs) in health care have become critical outcomes in both clinical trials and long-term observational studies. Measures from PRO-generated outcomes supplement clinical evaluations and provide the bridge between a patient’s task performance in the clinic and actual function in real-world situations, and many such measures have been shown to be reliable, valid, and sensitive to change. As a result, the number of initiatives designed to promote quality or paying for performance is increasing and has been simulated in rehabilitation using PROs.

The present 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 outcomes of functional status (FS) estimated using a CAT for patients with shoulder impairments.
Development, simulation, validation, and use of the shoulder-specific CAT have been described. Briefly, the item bank for the shoulder CAT (37-item) was originally derived from the Flexilevel Scale of Shoulder Function, a flexilevel outcomes instrument for patients with shoulder impairments receiving rehabilitation. Previous results indicated that the 37-item shoulder CAT item bank met item response theory (IRT) assumptions of essential unidimensionality and local independence, and FS measures were precise, valid, sensitive to functional change, and practical.

Clinically meaningful interpretations of the shoulder CAT have not been studied. For example, if a patient with a rotator cuff sprain presented in therapy with an intake FS score of 32 and discharge score of 71 (0–100 scale), it is known that the patient has improved 39 points. However, additional information is needed to support clinical reasoning and care planning. The purpose of this study was to provide meaningful interpretations of FS outcome measures estimated using the shoulder CAT. Specifically, this work addressed questions of importance to clinicians including 1) How confident can I be in a reported score? 2) How does my patient compare to others? 3) How large a change represents a true change in shoulder function? 4) How much improvement represents a clinically important improvement to the patient? 5) What does a specific score mean? 6) Can I use the score to assist clinical practice?

To answer these questions, we performed analyses found to be productive in our previous studies to derive clinically meaningful interpretations of outcomes measures, as well as analyses recommended by Jette et al., Hays et al., Schmitt and Di Fabio, and Stratford et al. Toward this end, we 1) constructed 95% confidence intervals (CIs) for each score point estimate, 2) established percentile ranks (PRs) for FS scores, 3) estimated statistically reliable change, 4) estimated minimal clinically important improvement (MCII), and 5) developed and applied a score-based functional staging system. The first four methods provided statistical indices and the fifth provided graphical presentation to assist in clinical interpretation of FS scores generated by the shoulder CAT.

METHODS

Data Collection

The data collection and sample have been described. Briefly, patients seeking outpatient rehabilitation entered demographic data and completed self-report surveys using Patient Inquiry computer software developed by FOTO (Knoxville, TN). Scores were input before initial evaluation and therapy, and at discharge. Under these two conditions, data were labeled intake and discharge, respectively.

For the purposes of this study, data were extracted from the FOTO CAT database if patients 1) were 18 years old or older, 2) were managed for an orthopedic impairment of the shoulder, 3) received outpatient physical therapy, and 4) completed the CAT between August 2007 and July 2009. This project was approved by the institutional review board for the Protection of Human Subjects from the Focus On Therapeutic Outcomes, Inc.

Setting and Participants

Data were from a convenience sample of 30,987 patients with shoulder impairments who met inclusion criteria and received outpatient physical therapy in one of the 518 outpatient clinics in 30 states (United States) served by FOTO. Patient mean (standard deviation [SD]) age was 53 (15) years (minimum = 18, maximum = 100). Fifty-one percent were females. Forty-seven percent reported that their symptoms were chronic (>90 days from date of onset) versus subacute (22–90 days) (30%) and acute (<22 days) (18%). Patients made an average of 11 (SD 11) clinic visits. Identification of medical or surgical diagnoses was optional in the data collection. Of the patients with medical/surgical codes (60%), the most common diagnoses were associated with soft tissue disorders (International Classification of Diseases, Ninth Revision [ICD-9] codes 725–729; [39%]). Eleven percent had postsurgical conditions such as tenotomy and repair of rotator cuff (CPT codes 23000 and 23929), and 4% had sprains and strains including those in the shoulder region, acromioclavicular joint, and rotator cuff (ICD-9 codes 840 including unspecified sprain or strain).

Shoulder CAT

The CAT administration was designed to efficiently evaluate each patient’s functional level by selecting items that provided the maximum information regarding a patient’s FS. In contrast to giving each patient fixed-length tests/questionnaires, items were selected from the 37-item bank one at a time based on the CAT administrative algorithm. The adaptive test started by administering the most informative item for persons with a median level of shoulder function, “How much difficulty do you have using your affected arm to reach a shelf that is shoulder height?” Patients were asked to rate their ability to perform this and other activities using a rating scale of five levels (i.e., 1–5) including 1) I can’t do this; 2) much difficulty; 3) some difficulty; 4) little difficulty; to 5) no difficulty. After each response, the patient’s estimated FS score and associated standard
errors (SEs) were updated. The computer algorithm continued to select the most informative items, given the current FS estimate until a stopping rule was satisfied. The final FS score represented a point estimate for each patient for their FS on a linear 0–100 scale with higher measures representing higher functioning. FS score change (functional status change [FSCH]) was defined by subtracting the FS score at intake from the FS score at discharge (FSCH = discharge FS – intake FS). FS, as assessed using the shoulder CAT, is operationally defined as the patient’s perception of their physical status and health condition as it relates to their abilities to perform daily functional tasks using their affected arm.

**Approaches to Deriving Meaningful Interpretations of Measurements**

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

The 95% CI band associated with the point estimate FS scores was constructed to provide estimates of measure precision (i.e., FS score ± 1.96 × SE).

*Establishing the PR of an FS Score: How Does My Patient’s Functional Score Compare to Other Patients’ Scores?*

The PR of an FS score refers to the percentage of scores in a given sample that are lower than it. To accommodate differences in FS scores at intake compared with discharge, we generated two PRs: a PR at intake (PRi) and one at discharge (PRd).

*Using Two Threshold Approaches to Define Individual Patient-Level Change*

To examine sensitivity to change, we used two threshold approaches to define individual patient-level change: statistically reliable change and clinically important change.

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

Statistically reliable change, as described by Schmitt and Di Fabio, reflects the statistical significance of individual change. To assess statistically reliable change, we computed the minimal detectable change (MDC) using the following formula: MDC = 1.96 × √2 × σ√(1 – reliability), where 1.96 represents the 95% CI, σ is the SD of the score, and “reliability” is a reliability coefficient (Cronbach’s a internal consistency). We multiplied the result by the square root of 2 to accommodate the two measurements involved in assessing change, intake, and discharge. As computed, MDC rep represents the smallest threshold for identifying statistically reliable change greater than random measurement error.

**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, we used an anchored-based longitudinal method using a 15-point Likert-type scale (−7 to +7) to provide a global rating of change (GROC). Patients reported that their change was because of the effect of therapy at discharge: deterioration (scores −7 to −1), no change (score of 0), or improvement (scores +1 to +7). Using nonparametric receiver-operating-characteristic (ROC) curve analysis, the minimal FSCH that is required for a patient to report a GROC greater than or equal to a +3 was determined. There is debate regarding use of the GROC for determining this MCII. Cut-scores of +2 to +3, +1 to +3, or greater than or equal to +3 have been used to represent MCII change on a 15-point Likert scale. In this study, we chose the upper limit of +3 or more (+3 = “somewhat better”) in defining the MCII based on findings in previous studies that supported this level of change as an adequate estimate of important improvement. Results of ROC analysis of the shoulder CAT have been described elsewhere. Here, we used the results to assist in clinical interpretation of the FS scores derived from the shoulder CAT.

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

Functional staging refers to development of a set of hierarchical outcome levels for classifying patients into different stages that describe FS. No functional staging model exists for patients with shoulder impairments, so we 1) built a conceptual model based on clinical judgment, 2) determined the cut-scores between stages identified in step 1, and 3) specified expected performance at each level.

**Build a Conceptual Model**

We developed the Shoulder Function Classification System (SFCS) to reflect graded levels of increasingly
more difficult activities for patients with shoulder impairments. This hierarchical system is based on the International Disability and Health framework. Table 1 shows the general headings (activity levels) and operational definitions of each level of the SFCS (see Appendix 1 for the item labels and item descriptions of the shoulder CAT).

We identified five hierarchical functional levels pertaining to the use of the affected arm: 1) exceedingly limited shoulder—much difficulty performing easy routine activities, 2) poor shoulder—much difficulty performing routine daily activities, 3) fair shoulder—some difficulty performing routine daily activities, 4) good shoulder—little difficulty performing routine daily activities, and 5) excellent shoulder—no difficulty performing almost any rigorous activity.

To classify patients using the SFCS, two operationalizations are required. First, what constitutes the different levels of difficulty must be defined (i.e., extreme, moderate, little, and no difficulty). For this purpose, we were able to use the response categories of the items of the FS CAT item bank because, for each item, respondents indicated “how much difficulty” they had with the specified task using responses of “no difficulty,” “little difficulty,” “some difficulty,” “much difficulty,” and “I can’t do this.”

The second level of operationalization required identification of what constitutes easy (light) routine activities, routine (moderate) daily activities, and rigorous (heavy) activities. Easy routine daily activities were defined as using the affected arm for the following: taking off glasses or sunglasses, washing face, turning a faucet, and pulling on socks. Routine daily activities were defined as reaching an overhead shelf or moving a heavy skillet (e.g., cast iron skillet) from one stove burner to another. Rigorous activities included activities such as working overhead or moving heavy objects.

### Determine the Cut-scores

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

- **Step 2-a (Calibrated Scores to RSM)**

  We first performed a statistical analysis to obtain the category structure, which included measure information for each item using WINSTEPS data using the Andrich rating scale IRT model (RSM) and obtained category structure measure information (i.e., category probability curves) and item-category “threshold” (i.e., Andrich threshold) for each item.

- **Step 2-b (Identify Conceptual Thresholds)**

  We performed an exploratory analysis based on our initial conceptualization of the SFCS content and response scales of shoulder CAT items to identify logical thresholds between stages of the SFCS. For instance, the threshold between levels 1 and 2 was explored by finding a specific response category of an item that described the distinction between unable to perform light, routine daily tasks and to extreme difficulty with such tasks using the affected arm. Having identified this conceptual threshold, we then identified an item that operationalized this threshold. For example, we

#### TABLE 1. Shoulder Function Classification System

<table>
<thead>
<tr>
<th>Activity Level</th>
<th>Operational Definition</th>
<th>FS Score Range*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1 Exceedingly limited shoulder</td>
<td>Unable to perform or limited a lot performing light routine activities using the affected arm like taking off glasses, flushing the toilet, turning a faucet or washing the face.</td>
<td>0–24</td>
</tr>
<tr>
<td>Level 2 Poor shoulder</td>
<td>Regained limited functioning and can perform light daily activities with some difficulty but still has much difficulty performing moderate upper extremity activities such as picking up and drink out of a full water glass, pulling a chair out from a table, and tightening a jar lid.</td>
<td>25–43</td>
</tr>
<tr>
<td>Level 3 Fair shoulder</td>
<td>Can perform light daily activities with little difficulty and moderate daily activities with some difficulty but still has much difficulty performing heavy upper extremity activities such as lowering a lightweight object from the top shelf of a closet and reaching an overhead shelf.</td>
<td>44–59</td>
</tr>
<tr>
<td>Level 4 Good shoulder</td>
<td>Can perform routine daily activities using the affected arm with no difficulty but still has a little difficulty performing heavy activities that require a combination of good range of motion, muscle strength, and endurance such as reaching an overhead shelf, working overhead for more than 2 min, or touching an object on the back seat while sitting in the front seat of a car.</td>
<td>60–80</td>
</tr>
<tr>
<td>Level 5 Excellent shoulder</td>
<td>No difficulty using the affected arm to perform almost any rigorous activity.</td>
<td>81–100</td>
</tr>
</tbody>
</table>

**FS** = functional status.

*The FS score (0–100 scale) range was determined by the follow-up Bookmark and Keyform methods.
hypothesized that the item “taking off glasses or sunglasses using your affected arm” would allow the distinction between SFCS levels 1 and 2. We then identified the shoulder FS level at which people had a 50/50 probability of endorsing category “1” (I can’t do this) versus endorsing category “2” (much difficulty) or higher.

In a similar manner, we identified cut-scores between the other SFCS functional stages. For example, the cut-score between level 2 and level 3 was defined as the threshold between responses “2” and “3” for the item “reach across to the middle of a table with your affected arm to get a salt shaker.” The cut-score between level 3 and level 4 was defined as the threshold between responses “3” and “4” for the item “move a heavy skillet (e.g., cast iron skillet) from one stove burner to another.” Finally, the cut-score between level 4 and level 5 was defined as the threshold between responses “4” and “5” for the item “work overhead for more than 2 minutes.”

- Step 2-c (Associate Conceptual Thresholds with Score Thresholds)

After categorizing structure measure information for each item and conceptual thresholds, we superimposed the conceptual thresholds on the category structure statistical framework and inspected the functional staging hierarchy. If the results were not clinically logical (such as clustered cut-scores or lack of meaningful clinical interpretation), the procedures were repeated. We progressed until satisfied the initial functional staging model appeared clinically and statistically logical.

Specify the Expected Performance

Once the initial conceptual functional staging with cut-scores was developed, we specified the expected performance in each stage based on the Andrich47 rating scale (RSM) measurement model. Here, the expected performance represents the predicted (i.e., most likely) response categories that the patient would report.

RESULTS

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

Because SEs at discharge were similar to SEs at intake with average of 0.09 score unit differences (range 0.00–0.24), we report only SEs associated with intake FS score estimations for defining CIs. The mean SE value across all levels of FS was 2.3 yielding a 95% CI of ±5 (i.e., 1.96 × 2.3). However, for the 94% of patients with FS intake scores between 20 and 80, the SE was 1.8, and the 95% CI was FS point estimate ±4 (i.e., 1.96 × 1.8). For patients with <20 or >80, the 95% CI was FS point estimate ±7 (i.e., 1.96 × 3.6).

Establishing the PR of an FS Score: How Does My Patient Compare to Others?

The mean FS scores at intake and discharge were 51 and 69, respectively, with SDs of 15 at both time points. Based on the score distribution in this sample, 25th, 50th, and 75th PRs corresponded to intake FS scores of 43, 52, and 59 and discharge FS scores of 59, 68, and 80, respectively. Table 2 provides PRs for additional FS intake (PRi) and discharge scores (PRd).

Using Threshold Approaches to Define Meaningful Change

How Much Change is Likely to Represent a True Change?

Statistically Reliable Change

Based on our previous data set21 of 400 patients with shoulder impairments, the internal consistency reliability coefficient was 0.97. The reliability estimates were 0.93 when we used the imputation approach based on the IRT model using the current shoulder CAT data (N = 30,987). To be conservative, 0.93 was used. With the SD at intake equal to 14.84,
the minimal score change (discharge FS – intake FS) required for MDC95 was 10.88 FS score units (i.e., $1.96 \times \sqrt{2 \times 14.84 \times \sqrt{(1 - 0.93)}}$). Therefore, MDC analyses-supported 11 or more FS change units represented statistically reliable change.

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

Clinically Important Change

There were 5,132 patients who completed both the FS and reported their global impressions of change with the GROC item. Of these, 566 (11.0%) reported little or no change (i.e., GROC scores ≤ 3), and 4,567 (89.0%) reported improvement (i.e., GROC scores ≥ 3). ROC analyses supported eight or more FS change units represented clinically meaningful improvement (i.e., MCII). Because previous studies showed that change is dependent of baseline FS measures, we performed ROC analyses using patients grouped by quartile of baseline FS scores. Results suggested that 23 or more, 10 or more, five or more, and two or more FS change scores represented clinically meaningful improvement for patients in the first (intake FS 0–43), second (intake FS 43–52), third (intake FS 52–60), and fourth (intake FS >60–100) 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 SFCS. This figure shows the expected response (the horizontal bars) to a given item as a function of the underlying upper extremity function (i.e., FS) estimated by the shoulder CAT. In this figure, the shoulder CAT items are listed in order of difficulty in the left column: more challenging items are listed on the top (i.e., BULB—using your affected arm to work overhead for more than 2 minutes and BCKSEATRE—while sitting in the front seat of a car, use your affected arm to touch an object on the back seat). Appendix 1 lists the item labels and item descriptions in the same order as shown in Figure 1. Beneath the figure is the FS score continuum ranging from 0 to 100 (higher values represent more functioning). Levels of functional staging from level 1 (left or lower functioning) to level 5 (right or higher functioning) are identified in the figure. Our association of conceptual and score thresholds resulted in cutscores of 24 between functional levels 1 and 2, 43

![Functional staging using the Shoulder Function Classification System. FS = functional status.](image-url)
between functional levels 2 and 3, and so on. Expected responses of each shoulder CAT item at each functional level can be obtained by drawing a vertical line over an FS score (x-axis) in Figure 1. By doing so, clinicians can see all the expected responses of all items even when the patient did not answer all the items during the shoulder CAT.

Table 3 presents the frequency distribution of the functional staging classification. Patients were classified into functional staging levels based on their intake (rows) and discharge (column) FS scores. The results represent an initial validation of the functional staging process.

### A Clinical Example

To illustrate how to use these strategies to enhance clinically meaningful interpretation, we will use a hypothetical example to answer the questions asked in the introduction. We selected an actual example out of the CAT database. Mr Smith (male, age 73) came to the clinic due to subacute rotator cuff sprain. His initial FS score was 32 at admission and 71 at discharge (39 FS score change) after receiving 10 outpatient therapy visits. To visualize his responses to our shoulder CAT, we plotted his responses in Figure 2: yellow circles identify the responses at intake, and purple circles identify responses at discharge. Mr Smith answered six items at intake and nine items at discharge using the shoulder CAT.

The 95% CI estimate of his intake FS score location was 28 to 36. Compared with other patients with a variety of shoulder impairments, Mr Smith’s PR at intake (PRi) was nine. The SFCS functional staging classified Mr Smith as exhibiting much difficulty performing easy routine activities (level 2). During the intake shoulder CAT administration, Mr Smith reported having much difficulty using his affected arm taking off glasses or sunglasses, flushing the toilet, putting on underpants, washing the side of his face, turning a faucet, and reaching a shelf at shoulder height.

At discharge, the 95% CI range for Mr Smith’s score was 67 to 75. Compared with other patients at discharge, Mr Smith’s PR at discharge (PRd) was 59. The functional staging classification suggested that Mr Smith improved to level 4 (exhibiting little difficulty performing routine daily activities). With an improvement of 39 FS score units (FSCH = 39), Mr Smith’s improvement was considered to be statistically reliable (FSCH > MDC95 cut-score of 11) and clinically meaningful (FS score change by quartile > MCII cut-score of 23) as supported by Mr Smith’s perspective that his condition was “a good deal better.” At the time of discharge, Mr Smith reported having no difficulty on two functional tasks including lowering a lightweight object (1–5 lbs) from the top shelf of a closet and pulling a medium-weight object (5–10 lbs) from under a bed. He reported having little difficulty reaching to shoulder height or overhead like reaching a shelf that is shoulder height, touching an object on the back seat, reaching an overhead shelf, placing a can of soup (1 lb) on a shelf overhead, reaching and pulling a string that controls a light or fan, and moving a heavy object like a heavy skillet from one stove burner to another.

### DISCUSSION

During the past two decades, valid and reliable questionnaires to measure the patient’s experience of disease and illness such as FS have proliferated, but this information has done little to change clinician behavior or improve patient health. Results of the present study followed recommended approaches to derive clinically meaningful interpretations of FS estimates generated from a shoulder CAT to facilitate use of the measures by clinicians in their routine clinical practice and advance clinical science of rehabilitation of the upper extremity. Instead of using a single method, we integrated several different methods including traditional score distribution statistics (i.e., SE, percentile), sensitivity to change and responsiveness indices (MDC, MCII), and functional staging to enhance clinical interpretation of patient-reported measures estimated using a body part-specific CAT.

### TABLE 3. Frequency Distribution of the Functional Staging Classification: Patients Were Classified into Functional Staging Levels Based on Their Intake (Rows) and Discharge (Column) FS Scores

<table>
<thead>
<tr>
<th>Intake FS Staging</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Total</th>
<th>Percent (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
<td>83</td>
<td>209</td>
<td>248</td>
<td>124</td>
<td>673</td>
<td>4.9</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>299</td>
<td>1,095</td>
<td>1,253</td>
<td>505</td>
<td>3,159</td>
<td>22.9</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>150</td>
<td>1,824</td>
<td>3,236</td>
<td>1,428</td>
<td>6,649</td>
<td>48.2</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>10</td>
<td>279</td>
<td>1,510</td>
<td>1,189</td>
<td>2,991</td>
<td>21.7</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>60</td>
<td>257</td>
<td>321</td>
<td>2.3</td>
</tr>
<tr>
<td>Total</td>
<td>30</td>
<td>543</td>
<td>3,410</td>
<td>6,307</td>
<td>3,503</td>
<td>13,793</td>
<td>100.0</td>
</tr>
<tr>
<td>Percent (%)</td>
<td>0.2</td>
<td>3.9</td>
<td>24.7</td>
<td>45.7</td>
<td>25.4</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

FS = functional status.
Discussion of the pros and cons of interpretation using the SE of measurement has been reported.22–24 Briefly, because SEs vary by level of FS under IRT models, it allows assessment of score precision (i.e., point estimates and MDC) along the FS continuum. However, the same methods demonstrate that extreme scores are likely to have larger SEs, which tend to make conclusions related to low or high scores difficult. As a result, we set that the 95% CI was FS point estimate for patients with 20 or 80.

We initially computed the sensitivity to change MDC index based on a single internal consistency reliability (Cronbach's alpha) estimate of 0.97 derived from our previous data set of 400 patients who answered our full-length shoulder CAT survey. However, we felt that the high reliability coefficient may lead to inflated results of the estimated MDC index because the CAT test—retest reliability coefficient, which we do not have, is expected to be lower than internal consistency reliability coefficient, which should lead to a larger MDC index. To test this hypothesis, we used two approaches to estimate the reliability: 1) imputing the missing values in our shoulder CAT data, and 2) using reliability estimates recommended by Nicewander and Thomasson.53 The reliability estimates were 0.93 when we used the imputation approach based on the IRT model. The reliability estimates were 0.71 and 0.59 when we used estimates based on Jensen's inequality (equations 19 and 20).53 For references, studies applying CAT methods have reported different degree of test—retest reliability using CAT: 0.75–0.90,54 0.79,55 0.70–0.95.56 Because we currently do not have test—retest reliability estimate of the shoulder CAT based on a real sample, the true impact of the sensitivity to change index was unknown and warrants further investigation.

One of the strengths of this study is the functional staging performed to enhance clinical interpretation of the shoulder CAT-generated outcome measures. The use of visual displays and functional staging approaches in health care to assist clinical practice has been described.32,57–61 Briefly, by using the visual

FIGURE 2. Clinical example. The patient’s (Mr Smith) responses are circled in the figure: yellow circles identify the responses at intake, and purple circles identify responses at discharge. Notation: (1–5 scale): 1 = I can’t do this; 2 = much difficulty; 3 = some difficulty; 4 = little difficulty; 5 = no difficulty; the “:” is the threshold cut-score between contiguous responses per item. FSCH = FS discharge – FS admission. FSCH = functional status change; GROC = global rating of change.
display of functional staging (Figure 1), clinicians can see the observed and predict missing responses by drawing a vertical line over an FS measure (x-axis) or circling the responses. Patients’ unexpected responses (e.g., patients at a higher functional level reported to have much difficulty performing some specific easy tasks) are easy to identify, and these may be useful to help clinicians manage patients who may consider whether there is a logical reason why the client had an unexpected response. Clinicians can set goals by identifying items that refer to functional activities appropriate as short-term goals and those appropriate as long-term goals. If a clinician has the patient’s FS scores reported at two or more points in time, the clinician can track the patient’s progress longitudinally by drawing vertical lines and inspecting overall FS change. In addition, clinicians can use the functional staging to estimate what a patient’s response would be to a specific question asked at intake by the CAT but not asked by the CAT at discharge. Results from the functional staging can supplement a patient’s specific report when the patient is assessed using a CAT administrative platform. The report may print both the actual (i.e., answered by the patient) and predicted responses (i.e., those items that were not administered) for all functional items within a CAT. Because items in the functional staging figure were listed in descending order of difficulty, the functional staging report may be helpful in describing how the items selected in the CAT corresponded to the order of difficulty of all functional activities. The functional staging classifications may prove valuable for interpreting numerical FS scores.

There are several limitations of this study. Because this study was a secondary analysis of prospectively collected data via a proprietary database management company, Focus On Therapeutic Outcomes, Inc. (FOTO), the researchers were not in control of the data collection procedure. Generalizability of results may be limited because there was the potential for patient selection bias related to which patients were asked to take the CAT. There may be differences between participating clinics compared with clinics that do not collect data using FOTO. Bias may occur on the patients asked to take the CAT and on self-selection (i.e., those patients who completed the survey when asked). We reported the clinical interpretation of the shoulder CAT based on a sample of patients with various shoulder disorders. Similar procedures should be validated in patients with specific shoulder disorders. Meanwhile, because there is no well-established shoulder functional classification system, we developed the SFCS based on our initial conceptualization of the hierarchical functional staging levels. Although we used clinical experience and logic to develop the functional staging classification, future research should evaluate the relevance of these stages using methods such as those proposed by Kane. Future research should investigate whether classification systems should be tailored to distinct diagnostic groups of patients, for example, rotator cuff syndrome, shoulder joint pain, shoulder sprains, and strains.

CONCLUSION

Results may improve clinical interpretation of CAT-generated outcome measures and assist clinicians using PROs during clinical practice.

REFERENCES

14. Hart DL, Mioduski JE, Stratford PW. Simulated computerized adaptive tests for assessing functional status were efficient with good discriminant validity in patients with hip,
APPENDIX 1

SHOULDER CAT (37-ITEM)

<table>
<thead>
<tr>
<th>Item number</th>
<th>Item Label</th>
<th>Item Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BULB</td>
<td>Use your affected arm to work overhead for more than 2 min</td>
</tr>
<tr>
<td>2</td>
<td>BCKSEATRE</td>
<td>While sitting in the front seat of a car, use your affected arm to touch an object on the back seat</td>
</tr>
<tr>
<td>3</td>
<td>OVRHEADSHE</td>
<td>Use your affected arm to reach an overhead shelf</td>
</tr>
<tr>
<td>4</td>
<td>MEDBOX</td>
<td>Use your affected arm to lower a lightweight object (1–5 lbs) from the top shelf of a closet</td>
</tr>
<tr>
<td>5</td>
<td>BED</td>
<td>With your affected arm, pull a medium-weight object (5–10 lbs) from under a bed</td>
</tr>
<tr>
<td>6</td>
<td>OVRHEADI</td>
<td>Use your affected arm to place a can of soup (1 lb) on a shelf overhead</td>
</tr>
<tr>
<td>7</td>
<td>SKILLET</td>
<td>Use your affected arm to move a heavy skillet (e.g., cast iron skillet) from one stove burner to another</td>
</tr>
<tr>
<td>8</td>
<td>STRING</td>
<td>Use your affected arm to reach and pull a string that controls a light or fan</td>
</tr>
<tr>
<td>9</td>
<td>COLLAR</td>
<td>Adjust the back of your collar with your affected hand</td>
</tr>
<tr>
<td>10</td>
<td>BALLUND</td>
<td>Use your affected arm to throw a ball underhand</td>
</tr>
<tr>
<td>11</td>
<td>POTATOES</td>
<td>Use your affected arm to stir a large bowl of thick food such as mashed potatoes</td>
</tr>
<tr>
<td>12</td>
<td>SLIDE</td>
<td>Use your affected arm to slide hanging clothes in a closet from one end of the rod to the other</td>
</tr>
<tr>
<td>13</td>
<td>COMB</td>
<td>Comb or brush your hair using your affected arm</td>
</tr>
<tr>
<td>14</td>
<td>POCKET</td>
<td>Use your affected arm to pull something out of your back pocket</td>
</tr>
<tr>
<td>15</td>
<td>SHELF</td>
<td>Use your affected arm to reach a shelf that is shoulder height</td>
</tr>
<tr>
<td>16</td>
<td>SAFESTRP</td>
<td>Use your affected arm to reach across body to get a car’s shoulder strap (safety belt)</td>
</tr>
<tr>
<td>17</td>
<td>PULLBOX</td>
<td>With your affected arm, slide a medium weight (5–10 lb) box across a table by pulling it completely to you</td>
</tr>
<tr>
<td>18</td>
<td>WHEELSA</td>
<td>Use your affected arm, turn a steering wheel in the same direction as your affected arm (e.g., turn right if it is your right shoulder that is affected)</td>
</tr>
<tr>
<td>19</td>
<td>SHOULDER1</td>
<td>Use your affected arm to place a can of soup (1 lb) on a shelf at shoulder height</td>
</tr>
<tr>
<td>20</td>
<td>CROOK</td>
<td>With your affected arm, carry something of medium weight in the crook of your arm (where your elbow bends)</td>
</tr>
<tr>
<td>21</td>
<td>JAR</td>
<td>With your affected arm, tighten a jar lid</td>
</tr>
<tr>
<td>22</td>
<td>PULLCHAIR</td>
<td>Pull a chair out from a table, using your affected arm</td>
</tr>
<tr>
<td>23</td>
<td>TIE</td>
<td>Get a scarf or necktie over your head and around your neck, using both hands</td>
</tr>
<tr>
<td>24</td>
<td>PUSHCHAIR</td>
<td>Push yourself out of a chair, using both arms</td>
</tr>
<tr>
<td>25</td>
<td>CHEST</td>
<td>With your affected arm, lift the lid of a chest that sits on the floor</td>
</tr>
<tr>
<td>26</td>
<td>JARSTEADY</td>
<td>With your affected arm, steady a jar while you loosen the jar lid</td>
</tr>
<tr>
<td>27</td>
<td>SALT</td>
<td>While sitting, reach across to the middle of a table with your affected arm, to get a salt shaker</td>
</tr>
<tr>
<td>28</td>
<td>DEODORANT</td>
<td>Put deodorant under the arm opposite your affected shoulder</td>
</tr>
<tr>
<td>29</td>
<td>WATER</td>
<td>Use your affected arm to pick up and drink out of a full water glass</td>
</tr>
<tr>
<td>30</td>
<td>TABLE</td>
<td>While sitting, lift your affected hand and put it on a table in front of you</td>
</tr>
<tr>
<td>31</td>
<td>UNDERPANTS</td>
<td>Put on underpants (panties, briefs, or boxers) using both hands</td>
</tr>
<tr>
<td>32</td>
<td>SOCKSON</td>
<td>Pull on your socks using both hands</td>
</tr>
<tr>
<td>33</td>
<td>EAROPP</td>
<td>Use your affected arm to reach the earlobe on the side opposite your affected shoulder</td>
</tr>
<tr>
<td>34</td>
<td>FAESA</td>
<td>With your affected arm, wash the side of your face on the same side as your affected shoulder</td>
</tr>
<tr>
<td>35</td>
<td>FAUCETSA</td>
<td>Using your affected arm, turn a faucet in the same direction as your affected arm (e.g., turn right if it is your right shoulder that is affected)</td>
</tr>
<tr>
<td>36</td>
<td>FLUSHING</td>
<td>Flush the toilet using your affected arm</td>
</tr>
<tr>
<td>37</td>
<td>GLASSES</td>
<td>Take off glasses or sunglasses using your affected arm</td>
</tr>
</tbody>
</table>

Rating scale (1–5 scale): 1 = I can’t do this; 2 = much difficulty; 3 = some difficulty; 4 = little difficulty; 5 = no difficulty.
JHT Read for Credit
Quiz: Article #169

Record your answers on the Return Answer Form found on the tear-out coupon at the back of this issue or to complete online and use a credit card, go to JHTReadforCredit.com. There is only one best answer for each question.

#1. Using the MDC the authors suggest that 8 unit increments on the
a. SFCS was statistically significant
b. SFCS was clinically important
c. FS was statistically significant
d. FS was clinically important

#2. The authors suggest that the
a. CAT scores are unhelpful clinically
b. CAT scores are the gold standard in shoulder outcome measures
c. five-level on the SFCS is clinically meaningful
d. twenty five-level on the SFCS is statistically meaningful

#3. The FS scoring was on a _______ scale
   a. 1-100
   b. 1-10
   c. visual analog
   d. therapist rated

#4. The critical score was
   a. the difference in CAT at intake and discharge
   b. the difference in FS at intake and discharge
   c. the CAT + the FS at discharge
   d. the CAT – the FS at discharge

#5. This study is intended as a useful clinical tool
   a. false
   b. true

When submitting to the HTCC for re-certification, please batch your JHT RFC certificates in groups of 3 or more to get full credit.