Development and Psychometric Evaluation of the Flexilevel Scale of Shoulder Function* (FLEX-SF)

Karon F. Cook, PhD[1]

Toni S. Roddey, PT, PhD, OCS, FAAOMPT[2]

Gary M. Gartsman, MD[3]

Sharon L. Olson, PT, PhD[2]

[1] VA Parkinson's Disease Research, Education and Clinical Center, Houston, TX and Dept. of Neurology, Baylor College of Medicine, Houston, TX

[2] School of Physical Therapy, Texas Woman's University, Houston, TX

[3] Department of Orthopedic Surgery, University of Texas Medical School and Texas Orthopedic Hospital, 7401 S. Main St., Houston, TX

Corresponding Author:
Karon F. Cook, PhD
Houston VA Parkinson's Disease Research, Education, and Clinical Center
VAMC (127-A) / 2002 Holcombe Blvd.
Houston, Texas 77030

Phone: 713 794-8936 Facsimile: 713 794-8888
e-Mail: karonc@bcm.tmc.edu

Number of Words: 4,350

*This research was supported by the Department of Veterans Affairs (HSR&D IIR 98-077) and Texas Orthopedic Hospital, Houston, Texas.

Complete Author Information

Corresponding author:
Karon F. Cook, Ph.D.
Houston VA Parkinson's Disease Research, Education, and Clinical Center
VAMC (127-A) / 2002 Holcombe Blvd.
Houston, Texas 77030

Phone: 713 794-8936 Facsimile: 713 794-8888
e-Mail: karonc@bcm.tmc.edu
Expertise: psychometrics, outcome assessment (Health Care), reliability, validity, item response theory

Co-Authors:
Toni S. Roddey, PT, PhD
Texas Woman's University, Department of Physical Therapy
1130 John Freeman Blvd.
Houston, TX 77030

Phone: 713 794-2067 Facsimile: 713 794-2071
e-Mail: hf_roddey@twu.edu
Expertise: physical therapy, orthopedics, shoulder rehabilitation, outcomes assessment

Gary Gartsman, MD
7401 South Main Street
Houston, TX 77030

Phone: 713 799-8600 Facsimile: 713 799-2932
e-Mail: gary@gartsman.com
Expertise: shoulder, orthopedic surgery, outcomes assessment, reliability, validity

Sharon L. Olson
Texas Woman's University, Department of Physical Therapy
1130 John Freeman Blvd.
Houston, TX 77030

Phone: 713 794-2090 Facsimile: 713 794-2071
e-Mail: hf_2olson@twu.edu
Expertise: orthopedics, geriatrics, shoulder, treatment, assessment, physical therapy, reliability

Development and Psychometric Evaluation of the Flexilevel Scale of Shoulder Function (FLEX-SF)

Abstract

BACKGROUND. Existing measures of self-reported shoulder function fail to measure effectively the full range of shoulder functioning. The classical approach for improving the reliability of a scale is adding items, but a scale with a substantial number of items imposes a large response burden on participants. A more efficient approach is to use modern psychometric methods to construct an “adaptive” scale in which patients only respond to items that are targeted at their level of shoulder function.

OBJECTIVE. We developed a flexilevel scale of shoulder function. This scale includes three “testlets” that target low, medium, and high shoulder function. Scores on the testlet were equated to a common mathematical metric.

DESIGN AND SUBJECTS. We developed an initial pool of 68 items. This pool was administered to 400 persons and responses were calibrated using a rating scale model. Subsets of items were identified for an easy, medium difficulty, and hard testlet. Properties of the scale were evaluated in a 3-month longitudinal study of 200 shoulder patients.

RESULTS. The FLEX-SF exhibited high reliability at both the scale level (ICC,3,1 = 0.90) and specific trait levels. The validity of the FLEX-SF was supported by its internal and external responsiveness (Guyatt responsiveness index = 1.12) and the pattern of its associations with other health status measures.

CONCLUSIONS. The FLEX-SF can be used as a primary endpoint in clinical trials even when there are relatively few people in each treatment group. The scale also has excellent properties for use in clinical settings tracking individual changes over time.

Key Words: shoulder, psychometrics, reliability, validity, outcomes assessment

Introduction

The competing goals of measurement precision and low response burden long have been recognized as a central conflict in scale development. However, the introduction of item response theory models[1-3] has fostered innovative scaling approaches such as adaptive scaling.[4,5] Adaptive scaling strategies have been employed in educational and psychological testing for almost 30 years[4-6]and in marketing for more than a decade.[7] In recent years, researchers have recognized the potential of adaptive scaling for the measurement of health outcomes.[8-10] Adaptive approaches increase scaling efficiency by requiring respondents to respond only to a subset of the pool of items—that subset that best targets their level of the trait being measured. Though computer-adaptive scales yield the largest increases in efficiency, adaptive scales can be created for the traditional paper and pencil format.[4,5]

This paper describes the development and evaluation of a paper and pencil adaptive scale of shoulder function, the Flexilevel Scale of Shoulder Function (FLEX-SF). The FLEX-SF consists of three subsets of items (testlets) that target different levels of shoulder functioning. Persons respond to a screening item which grossly classifies their shoulder function as low, medium, or high and are then routed to the testlet that matches their level of function.

Methods

Study Sample

A total of 612 participants were enrolled in the study. Seven participated in focused interviews, 400 completed the developmental form of the FLEX-SF, 5 participated in a pilot of the first draft of the FLEX-SF, and 200 were enrolled in a longitudinal study of the clinical validity of the scale. Half of those in the longitudinal study also participated in a test-retest evaluation of the scale. Institutional review boards at the Veteran's Administration Hospital, Baylor College of Medicine, and Harris County Hospital District reviewed and approved the study protocol. Persons were excluded from participating if they had undergone shoulder surgery less than 3 months prior to recruitment, were unable to read English, or were less than 18 years of age. Patients were recruited at 3 facilities, an orthopedic surgeon's office, a county Physical Therapy department, and the Houston Veteran's Affairs Medical Center Hospital.

Procedures

The study required several steps. We created an initial pool of 68 items and then modified them based on feedback from a pilot of the scale. Items were calibrated using Andrich's Rating Scale Model[11] and ordered according to difficulty. Redundant and misfitting items were dropped from the item pool. The remaining 33 items were assigned to one of three overlapping testlets (easy, medium difficulty, or hard). Scores for each testlet were equated to a common mathematical metric, and scoring rules were established. The final FLEX-SF and other self-reported outcome measures were administered to a sample of 200 participants in a longitudinal study of the reliability and validity (including responsiveness) of the scale.

Development of the FLEX-SF. We developed an initial pool of items by: (1) harvesting items from existing shoulder scales, (2) adapting items from other physical function scales, and (3) developing new items using patient interviews and input from an expert panel. Because we discovered in our previous work that the low range of functioning has not been well targeted by existing scales,[12] we paid particular attention to writing items that targeted low shoulder function.

The patient interviews were co-lead by a physical therapist (Roddey) and a psychometrician (Cook). Interviews were audio-taped and transcribed, and transcripts were reviewed for recurrent themes and unique content. New items were added, and the entire item pool was reviewed by an expert panel consisting of 3 physical therapists, an orthopedic surgeon, and the project psychometrician. The panel evaluated the items for content coverage and suggested modifications in wording and format. The final developmental item pool consisted of 68 items. For each item, respondents indicated "how much difficulty" they had with the specified task. Response options and there corresponding scores were: "no difficulty" = 4, "little difficulty" = 3, "some difficulty" = 2, "much difficulty" = 1, "I can't do this"= 0, and " didn't do before shoulder problem" = N/A. The last category was included to give participants the option of indicating that the item was not appropriate for them because it referenced an activity they did not usually do regardless of the status of their shoulder. The number of persons who responded in this response category was valuable to us in deciding which items to retain for the final scale, our preference being for items for which most people chose response categories other than "didn't do before shoulder problem."

Calibration of developmental item pool. With Rasch models, estimates of persons' trait levels are based on probability functions that have two parameters, an item parameter, difficulty, (b_i) and a person parameter, theta, (θ_n).[13] In the simplest Rasch model (dichotomous model), there are only two item response categories (e.g., “yes/no”). The probability of a person answering “yes” to an item depends on the match between the item difficulty (b_i) and the person’s trait level (θ_n). In the present study, if person’s shoulder functioning is very high, they would be very likely to say “yes” to an item that asked if they could perform an easy task (e.g., Use your affected arm to pick up and drink out of a full water glass). When the value of theta is low and the item difficulty is high, the probability of a “yes” answer decreases. When trait level and item difficulty are equal, the probability of a “yes” response is equal to the median probability (0.50). In fact, the difficulty of an item (b_i) in the dichotomous Rasch model is defined as the point on the measurement continuum at which a person would have a 0.50 probability of endorsing the item. In the Rasch model, the distance between trait level and item difficulty (θ_n - b_i) is expressed as the exponent of base *e*. This constrains the estimated probabilities to a range of zero to one. In the simplest Rasch model (dichotomous model), there are only 2 possible response categories (e.g., “no/yes”, coded 0/1). The following equation expresses, for a given level of trait,θ_n, the probability of choosing a particular category response, *m*, for item *i*. As written, the expression defines the probability of endorsing (saying “yes”, coded as ‘1’).

$$\text{Probability } (m_i = 1|\theta_n) = e^{(\theta_n - b_i)} / \left(1 + e^{(\theta_n - b_i)}\right) \qquad (1)$$

The pool of items for the SF-FLEX was administered to 400 persons with shoulder complaints and then calibrated using Andrich’s rating scale model (RSM),[11] an extension of the dichotomous Rasch model for items with three or more response categories. With the RSM, all items have a fixed set of response categories, and the distances between adjacent response categories are equal across all items. A location parameter is estimated for each item, and *response threshold* values are estimated for the entire set of items. For example, for an item with 3 response categories (e.g., “never/sometimes/often”), the response threshold between an answer of “never” and “sometimes” is the distance from the item’s location parameter at which responding “sometimes” or “often” becomes more likely than responding “never”. The probability of responding in a given category, given trait level equal to theta (θ) is

$$P_{ix}(\theta_n) = \frac{\exp[\theta_n - (b_i + t_j)]}{\sum_{k=0}^{m_i} \exp[\sum_{j=0}^{k_i} \theta_n - (b_i + t_j)]} \qquad (2)$$

where

b_i = the item location parameter for item *i,*

t_j = response threshold parameters for the set of items.

Responses of “didn’t do before shoulder problem” were coded as “missing”.

BIGSTEPS,[14] the computer software used for the calibration, outputs statistics for evaluating the fit of the data to the rating scale model. “Outfit” and “infit” are modified chi-square statistics obtained using the computer program. These statistics summarize the degree to which individuals’ responses fit those predicted by the model.

A total of 8 items were found to be misfitting based on a criterion of infit and/or outfit statistics with values greater than 2.5.[3] These items were removed from further analyses. Based on the item parameter estimates obtained in the calibration, the remaining items were ordered according to difficulty along a continuum from very easy items to very difficult items. At some points along the continuum, there were as many as six items that had approximately the same difficulty level. Our expert panel decided which of the redundant items to retain. Care was taken to ensure that items adequately covered all shoulder movements (e.g., tasks were not limited to external rotation tasks). The choice of how many items to retain was somewhat arbitrary. With the adaptive scaling approach, we could construct a short test that measured low shoulder functioning as well as existing measures with much fewer items. With more items, we could construct a test that measured low shoulder function much more precisely than existing measures.. Because of our interest in research in shoulder rehabilitation in severely affected individuals, we chose the latter. The 33 retained items were divided into three 15-item testlets that targeted low, medium, and high shoulder function. The items of testlets overlapped with the testlet(s) adjacent to it. The range of the difficulty of items for each of the testlets was arbitrary. We chose these boundaries so that approximately equal portions of our study population were targeted by the low, medium, and high testlets. In Figure 1, the difficulties of several sample items are displayed.

Selection of Routing Item. The flexilevel scale required a way of classifying respondents as having low, medium, or high shoulder function. Based on their answers to all of the items that eventually comprised the testlets, we classified participants as low, medium, and high. This classification served as our proxy gold standard. Potential routing items were evaluated based on two criteria. The routing item chosen needed to have: (1) few missing answers and (2) no misclassifications across 2 categories (e.g., low functioning respondent classified as high functioning). The item that best met these criteria was chosen as the routing item, “How much difficulty do you have using your affected arm to place a can of soup (about 1 lb.) on a shelf at shoulder height?”

Pilot of the FLEX-SF. Each testlet was printed on a different color of paper. The routing item was printed on the first page of the FLEX-SF with a statement beside each response option instructing participants to, for example, “respond only to the items on the blue sheet” (See Appendix A). Once the FLEX-SF was in this form, we conducted interviews with a pilot sample of five persons who had shoulder problems. Participants completed the FLEX-SF and then discussed anything on the scale they found difficult or confusing. Based on this feedback, some items were dropped from testlets and replaced with items previously dropped because of redundancy.

Equating Testlet Scores. To distinguish between raw scores (obtained by summing item responses) and calibrated and linearly-transformed scores (from the Rasch calibration), we use the term “raw scores” to refer to the former and “FLEX-SF scores” to refer to the latter. We equated FLEX-SF scores from each testlet to a common mathematical metric. This was necessary so that a given FLEX-SF score, e.g., 30, meant the same whether it was obtained based on the easy, middle difficulty, or hard testlet. To accomplish this, we used a single sample, common item pool equating design.[15]

Longitudinal Study of Reliability and Validity Procedures. Two hundred persons participated in a longitudinal study in which we gathered data for the evaluation of the reliability and validity of the FLEX-SF. Participants completed a packet of questionnaires at recruitment. Follow-up packets were mailed monthly for 3 months after recruitment and included a posted, self-addressed envelope. Questionnaires included the American Shoulder and Elbow Surgeon’s Scale (ASES),[16] the SF-12,[17] and the FLEX-SF. In addition, the monthly follow-up questionnaires asked participants to rate their shoulder status (compared to the previous month) on a 7-category scale: “much worse”, “worse”, “no change”, “slight improvement”, “moderate improvement”, “large improvement”, and “very large improvement”. For convenience, we will refer to this scale as the Self-reported Change in Status scale.

Reliability. To evaluate test-retest reliability, a sub-sample comprised of the first 100 participants enrolled in the longitudinal study completed initial questionnaires in the physician’s office and were given an envelope containing a second copy of the FLEX-SF. Participants were instructed to take the retest packet home, complete it between 24-48 hours after recruitment, and return the completed questionnaire. The retest questionnaire included a version of the Self-reported Change in Status scale that asked participants to rate their shoulder status in comparison to their status at recruitment (24-48 hours previously).

Validity. The inclusion of patient interviews and review of items by an expert panel were efforts to ensure that items of the FLEX-SF well represented the construct being measured (construct validity). Also, we developed hypotheses about associations among scores on the FLEX-SF and other outcome measures. We hypothesized that there would be a:

1. low correlation ($r<0.5$) with the SF-12 mental health subscale,
2. moderate correlation ($r>0.6$) with the SF-12 physical function subscale, and
3. moderately high correlation ($r>0.7$) with the ASES.

The choice of values of 0.5, 0.6, and 0.7 to represent low, moderate, and moderately high correlations is arbitrary but consistent with cutoff values used in the validation of other scales of physical function.[18]

Data Analysis

Using the longitudinal sample of 200, data were analyzed to evaluate the reliability of FLEX-SF scores and to assess their construct validity. Our assessment of construct validity included calculations of minimal clinically important difference, responsiveness, receiver operating curves, and testing of hypothesized relationships between the scores of the FLEX-SF and scores of other measures.

Reliability Assessments

Calculation of Scale-Level Reliability. An intraclass correlation coefficient (ICC) was calculated to estimate the test-retest reliability of the FLEX-SF. ICCs are a family of ANOVA-based estimates of reliability that compare true and total variance in a sample of scores.[19] The ICC used for this study was derived from a two-way mixed model in which raters rate all targets (notated as ICC (3,1)). The ICC and its confidence intervals were obtained using SPSS.[20] Cronbach's alpha values were calculated for each of the FLEX-SF testlets to evaluate their inter-item consistency.

Calculation of Trait-Level Reliability. Because of our previous work,[12] we were particularly interested in comparing the trait-specific reliability of the FLEX-SF to another shoulder outcome measure (ASES). To accomplish this, we modified the methodology used in the previous study, details of which are explained elsewhere.[12] Briefly, we calibrated both the ASES and the FLEX-SF scores using Andrich's rating scale model,[11] and then linearly transformed the scores of all scales to have a range of 0-

100. The study sample was divided into 3 subgroups based on which testlet was taken. Based on the standard errors of measurement, we calculated the 95% confidence interval for several score levels. This allowed us to compare, within each subgroup, the precision of individual scores on the ASES and on the FLEX-SF. Because the transformation to a 0-100 metric was made in each of the 3 subgroups, the comparisons are relative and appropriate *within subgroups only*, not across subgroups.

Validity Assessments.

Minimally Clinically Important Difference. Different methods have been recommended for estimating the minimally clinically important difference (MCID).[21,22] We defined the MCID as the average amount of within-person FLEX-SF score change in patients identifying themselves as just improved or just worse. “Just improved” was defined as marking one response category higher than the response, “no change” on the Self-reported Change in Status scale, “just worse”, as marking one response category below. The magnitude of change necessary for patients to perceive themselves as “worse”, however, should not be assumed a priori to be the same as that for a perception of “better”. We evaluated the comparability of the worse and improved groups’ change scores by calculating a t-test of their difference scores. The differences were not significant ($p = 0.48$), so we combined the groups to obtain estimates of the MCID at three occasions (one month compared to baseline, two months compared to one month, and three months compared to two months). The median MCID was calculated for the three occasions.

Responsiveness. A measure has *internal responsiveness* to the degree that its scores demonstrate the capacity to change over a specified period of time, and *external responsiveness* to the degree that changes correspond to changes in some external standard such as clinical status or an accepted health status score.[23]

We evaluated the internal responsiveness of the FLEX-SF using Guyatt's responsiveness index (RI).[24] The RI is an effect size statistic that expresses the magnitude of change on a measure with respect to some estimate of variation. The RI is expressed as

$$RI=\Delta_x /\mathrm{sqrt}(2*MSE_x) \qquad (2)$$

where Δ_x is the minimally clinically important change and MSE_x is the mean square error. We estimated MSE using a repeated measures analysis of variance (ANOVA) of the FLEX-SF scores from the test and retest administrations. This approach to estimating responsiveness defines clinically important change as that which exceeds the amount expected in a clinically stable population. We limited our analysis to the scores of persons who reported experiencing "no change" in the interval between test and retest administrations.

We evaluated the external responsiveness of the FLEX-SF using the linear regression method proposed by Husted and colleagues.[23] We regressed patients' Self-reported Change in Status scores from the one-month follow-up questionnaire upon the differences in patients' baseline and one-month FLEX-SF scores. The significance of the regression coefficients served as a test of the external responsiveness for SSF- FLEX scores.

Receiver Operating Curves. We calculated receiver operating characteristic (ROC) curves[25] for the FLEX-SF scores. As Deyo and Cantor[25] point out, this approach is analogous to evaluating the performance of a diagnostic test. In the current case, the condition diagnosed is whether a clinically important change in shoulder function has occurred. For the three monthly follow-up comparisons we divided our sample into "improved" and "not improved" based on responses to the Self-reported Change in Status scores. Those reporting that they had not changed or were worse were classified as not improved; all others were classified as improved. We calculated the area under the ROC curve where sensitivity (proportion correctly classified as undergoing change) is plotted against one minus the FLEX-SF specificity (proportion correctly classified as not undergoing change).[20] When a scale perfectly predicts classification on the dichotomous outcome variable (e.g., "improved" or "not improved"), the area under the ROC is 1.0; when a scale predicts no better than chance, the area under the ROC is 0.50. Though there are no recognized standards for what is a "good" value, the magnitude of the area for any given scale can be evaluated in comparison to other scales. For the current study, we calculated ROC areas for the ASES, the SF-12 physical, and the SF-12 mental scores.

Information. We plotted the test *information* curve for the entire set of FLEX-SF items. In IRT, information is defined as the reciprocal of the square of the standard error of measurement.[2] A plot of test information against the trait being measured, θ, is a graphic display of the relative amount of measurement precision of a scale at different trait levels.

Results

Descriptive Statistics

The 400 persons whose responses were used to calibrate the FLEX-SF participated anonymously, and minimal demographic and clinical information was collected for them. The sample was 64% male. Of the 393 participants who reported surgical status, 134 (34%) were post-surgical and 259 (66%) were non-surgical.

Of the 200 participants in the longitudinal study, one withdrew days after recruitment requesting that her data not be included. The remaining sample was 53% male, with a mean age of 52 (SD=16). Return rates for the mailed packets were consistently high throughout the longitudinal study. For the test-retest, 1 month, 2 month, and 3 month packets, rates were 77% (77/100), 63% (126/199), 62% (123/199), and 61% (121/199), respectively.

Reliability

Scale-Level Reliability. The FLEX-SF exhibited high reliability. The test-retest ICC (3,1) was 0.90 with a 95% confidence interval of 0.84 to 0.94. Cronbach’s alpha values for the easy, medium difficulty, and hard testlets were, respectively, 0.96, 0.93, and 0.97.

Trait-Level Reliability. We evaluated the trait-level reliability of the FLEX-SF by comparing it to the ASES in three study subgroups categorized by which testlet they completed. For the subgroup of participants who took the easy testlet, we computed the width of the 95% confidence intervals for scores of 25, 35, 45, 55, and 65. These are presented in Figure 2. The results for the medium and hard testlet are presented in Figures 3 and 4.

As can be seen in Figures 2-4, the 95% confidence intervals for FLEX-SF scores were substantially smaller than those for the ASES in all three subsamples. Some of these differences were quite large. In the subsample of participants who took the middle difficulty testlets, the FLEX-SF 95% confidence intervals for scores of 25 and 55 were less than half the size of those for the ASES. Increased precision was evident in the subsamples who took the easy and hard testlets as well.

Validity

Hypotheses Regarding Associations with FLEX-SF. The associations between the FLEX-SF and the other outcome measures are presented in Table 1. Our expectations regarding the magnitude of these associations were largely upheld. The only association that was not as predicted was the Baseline SF-12 Physical subscale correlation (0.53) that was not above the arbitrary cutoff we set for a moderate correlation (r>0.6).

Minimally Clinically Important Difference. We combined the subset of patients who reported being just worse with those who reported being just better. The average change in FLEX-SF score for this combined group was 3.02. Therefore, our MCID was estimated at 3.02.

Responsiveness

Internal Responsiveness. The repeated measures ANOVA of the clinically stable participants in our test-retest sample yielded a mean squared error (MSE) of 3.63. Based on this value and an MCID of 3.02, the internal responsiveness index for the FLEX-SF was calculated to be 1.12. The magnitude of this statistic is comparable to values obtained for other outcome measures. For example, the internal responsiveness of the Sickness Impact Profile and the Functional Independence Measure have been calculated as 1.15[26] and 1.29,[27] respectively.

External Responsiveness. The results of the linear regression supported the responsiveness of the FLEX-SF (Table 2). Across all time intervals, the mean regression coefficients (beta *(b)*) was 0.116. This value estimates that a one-unit change in FLEX-SF scores (scores have a 50-point range) represented a 0.116 change in Self-reported Change in Status scale (scores have 7-point range). These results should be interpreted with caution, however, since change in status was measured on an ordinal scale.

The median R^2 value for the models was 0.191 ($p<0.001$) indicating that the model explained a significant, though small, portion of the variance. As seen in Table 3, ASES scores also exhibited external responsiveness. The regression models for the SF-12 subscales were not statistically significant, and therefore, the responsiveness of these more generalized scales is not supported.

Receiver Operating Curves. ROC areas are reported in Table 3. As expected the FLEX-SF (mean ROC = 0.75) and ASES (mean ROC = 0.74) performed better than the SF-12 physical (mean ROC = 0.62) and mental health subscales (mean ROC = 0.54).

Information. The test information function for the full item pool of the FLEX-SF is displayed as Figure 5. As expected, the largest portion of the test information is at the middle of the θ continuum. This occurs because information is summative and the magnitude of information in the middle is a function not only of the medium difficulty testlet, but also of the fact that both the easiest and hardest testlets overlap with the middle testlet. Also, the information of polytomous items is the sum of the category information. Unlike dichotomous items whose information functions have a single peak, the information for a polytomous item will have as many peaks as it has categories. The overlapping bell-shaped curves within and across item cause the test information to be greatest at the middle values of θ. The net result is that, in order to have adequate information in measuring the lower and higher levels of shoulder function, a scale will have much more information than may be needed for measuring middle levels of shoulder function.

The amount of information needed at the ends of the measurement continuum depends upon the purpose for which the measure is used and the trait-level of the study population. One way of evaluating the precision of a scale with respect to the sample, however, is to define its “effective measurement range.”[28] In a Rasch model, this range extends from the response threshold (point of median probability) with the lowest value to the one with the highest value. For the FLEX-SF this score range was 6-47.

Conclusions

We were successful in our efforts to develop a paper and pencil, adaptive scale that offers precision without imposing a large response burden on patients. The analysis of the longitudinal study demonstrated that the FLEX-SF has excellent scale-level reliability (test-retest ICC of 0.90, Cronbach alpha values of 0.93 to 0.97). However, we have noted that traditional reliability estimates are an *average* of the scale reliability across all levels of the variable being measured.[3,29] The FLEX-SF scale distinguished itself from existing measures with regard to trait-level reliability. At all levels of shoulder function, the 95% confidence intervals for the FLEX-SF scores were consistently smaller than those for the ASES.

The implications of these results extend beyond the measurement of shoulder outcome. Using an adaptive scaling approach, we were able to develop an outcome measure that was reliable, responsive, and minimized the burden imposed on respondents. Because the measure is adaptive, patients and research participants are not asked to answer questions that are clearly mismatched to their level of the trait being measured. This advantage is obtained without sacrificing reliability and responsiveness. The FLEX-SF had a responsiveness index of 1.2. Based on Guyatt's[24] power calculations, a responsiveness index of 1.0 would require, for an independent groups t-test, a sample size of 19 per group (assuming alpha=0.05, 1-tail test, B=0.10). And, for related group comparisons under the same assumptions, a sample size of 11 per group would be required. The practical importance of this for research purposes is that the

FLEX-SF can be used as a primary endpoint in clinical trials that have relatively few people in each treatment group.

The results of this study indicate that substantial measurement efficiency can be gained using an adaptive testing strategy. Future research should examine the additional efficiency gained in computer-adaptive testing where branching occurs after response to each item.

Acknowledgements

*This research was supported by the Department of Veterans Affairs (HSR&D IIR 98-077) and Texas Orthopedic Hospital, Houston, Texas.

References

1. Embretson SE, Reise SP. Item response theory for Psychologists. Mahway, NJ: Lawrence Erlbaum Associates, Publishers; 2000.

2. Hambleton R, Swaminathan H. Item Response Theory: Principles and Applications. Norwell: Kluwer Academic Publisher; 1985.

3. Wright BD, Masters GN. Rating Scale Analysis. Chicago: Mesa Press;1982.

4. Lord FM. The self-scoring flexilevel test. Journal of Educational Measurement 1971;8:147-151.

5. Lord FM. A theoretical study of the measurement effectiveness of flexilevel tests. Journal of Educational Measurement 1971;31:805-813.

6. Wainer H. Computerized Adaptive Testing: A Primer. Hillsdale: Lawrence Erlbaum Associates;1990.

7. Singh J, Howell RD, Rhoads GK. Adaptive designs for Likert-type data: an approach for implementing marketing surveys. Journal of Marketing Research 1990;28:304-321.

8. McHorney CA. Generic health measurement: past accomplishments and a measurement paradigm for the 21st century. Ann Intern Med 1997;127:743-750.

9. Revicki DA, Cella DF. Health status assessment for the twenty-first century: item response theory, item banking and computer adaptive testing. Qual Life Res 1997;6:595-600.

10. Ware JE, Jr., Bjorner JB, Kosinski M. Practical implications of item response theory and computerized adaptive testing: a brief summary of ongoing studies of widely used headache impact scales. Med Care 2000;38(9 Suppl):II73-II82.

11. Andrich DA. A rating formulation for ordered response categories. Psychometrika 1978;43:561-573.

12. Cook KF, Gartsman GM, Roddey TS, Olson SL. The measurement level and trait-specific reliability of 4 scales of shoulder functioning: an empiric investigation. Arch Phys Med Rehabil 2001;82(11):1558-1565.

13. Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Danmarks Paedogogiske Institut. Copenhagen: Danmarks Paedogogiske Institut;1960.

14. BIGSTEPS: [Rasch-model computer program]. Chicago: MESA Press;1997.

15. Angoff WH. Scales, Norms, and Equivalent Scores. Princeton: Educational Testing Service;1984.

16. Richards RR, Bigliani LU, Gartsman GM, Iannotti JP, Zuckerman JD. A Standardized Method for the Assessment of Shoulder Function. J Shoulder Elbow Surg 1994;3:347-352.

17. Ware J, Jr., Kosinski M, Keller SD. A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. Med Care 1996;34(3):220-233.

18. Binkley JM, Stratford PW, Lott SA, Riddle DL. The Lower Extremity Functional Scale (LEFS): scale development, measurement properties, and clinical application. North American Orthopaedic Rehabilitation Research Network. Phys Ther 1999;79(4):371-383.

19. Shrout P, Fleiss J. Intraclass Correlations: Uses in Assessing Rater Reliability. American Psychological Association 1979;420-428.

20. SPSS for Windows. Chicago: SPSS Inc.;1999.

21. Guyatt GH, Kirshner B, Jaeschke R. Measuring health status: what are the necessary measurement properties? J Clin Epidemiol 1992;45(12):1341-1345.

22. Redelmeier DA, Guyatt GH, Goldstein RS. Assessing the minimal important difference in symptoms: a comparison of two techniques [see comments]. J Clin Epidemiol 1996;49(11):1215-1219.

23. Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. J Clin Epidemiol 2000;53(5):459-468.

24. Guyatt GH, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chronic Dis 1987;40(2):171-178.

25. Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. J Chronic Dis 1986;39(11):897-906.

26. Stucki G, Liang MH, Fossel AH, Katz JN. Relative responsiveness of condition-specific and generic health status measures in degenerative lumbar spinal stenosis. J Clin Epidemiol 1995;48(11):1369-1378.

27. Wallace D, Duncan PW, Lai SM. Comparison of the responsiveness of the Barthel Index and the motor component of the Functional Independence Measure in stroke the impact of using different methods for measuring responsiveness. Journal of Clinical Epidemiology 2002;55(9):922-928.

28. Linacre JM, Wright BD. A User's Guide to BIGSTEPS: Rasch-model Computer Program, version 2.7. Chigago: Mesa Press;1997.

29. Cook K, Dodd B, Fitzpatrick S. A comparison of polytomous item response models in the context of testlet scoring. Journal of Outcome Measurement 1999;3(1).

Figure Legend

Figure 1: Sample items of the FLEX-SF and how they are grouped by testlet.

Figure 2: Width of 95% Confidence intervals of ASES and Flexilevel scores for respondents routed to the easiest testlet. (Scores of participants have been rescaled so that the lowest score is 0 and the highest is 100. Widths of intervals are comparable within but not across figures.)

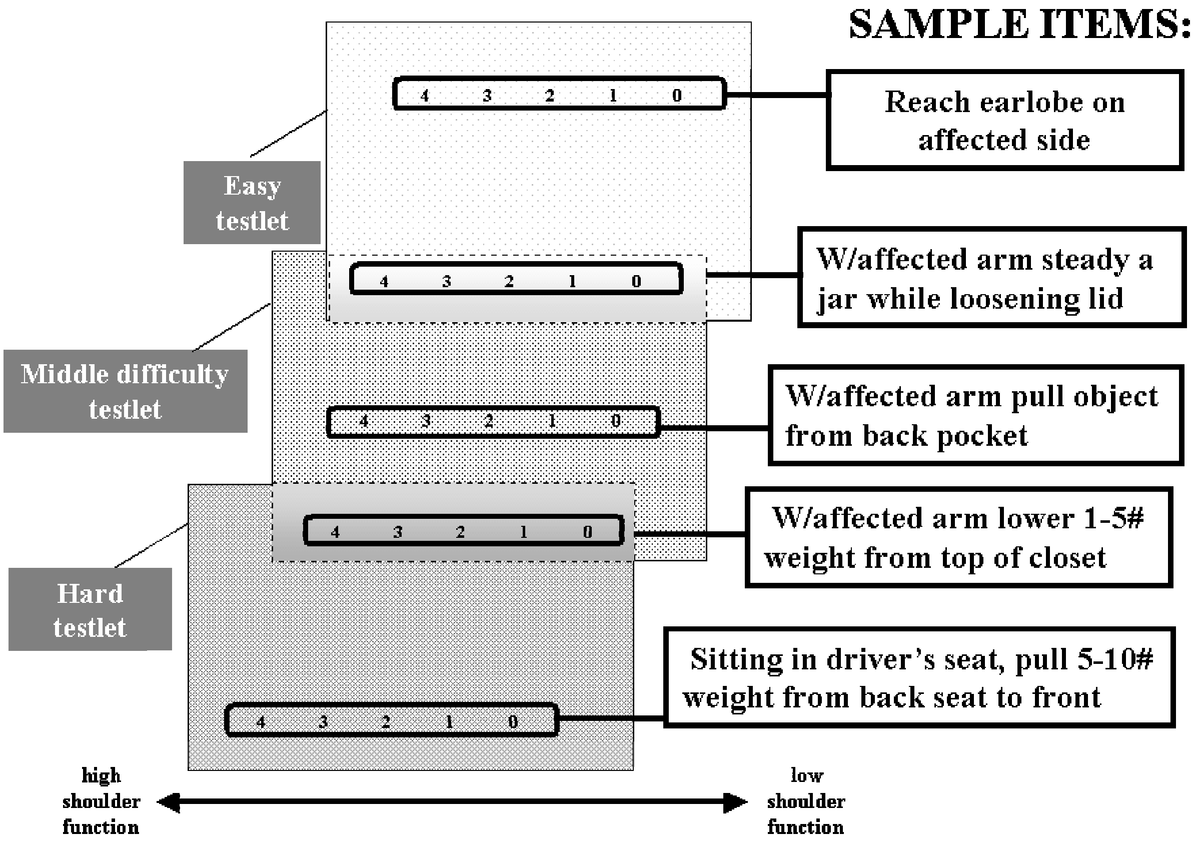
Figure 3: 95% Confidence intervals of ASES and Flexilevel scores for respondents routed to the medium difficulty testlet. (Scores of participants have been rescaled so that the lowest score is 0 and the highest is 100. Widths of intervals are comparable within but not across figures.)

Figure 4: 95% Confidence intervals of ASES and Flexilevel scores for respondents routed to the most difficulty testlet. (Scores of participants have been rescaled so that the lowest score is 0 and the highest is 100. Widths of intervals are comparable within but not across figures.)

Figure 5: Test Information Function for the FLEX-SF and Effective Measurement Range

SAMPLE ITEMS:

Figure 1. Sample items of the FLEX-SF and how they are grouped by testlet.


SAMPLE ITEMS:
4 3 2 1 0
Reach earlobe on affected side
Easy testlet
W/affected arm steady a jar while loosening lid
Middle difficulty testlet
W/affected arm pull object from back pocket
W/affected arm lower 1-5# weight from top of closet
Hard testlet
Sitting in driver's seat, pull 5-10# weight from back seat to front
high shoulder function
low shoulder function

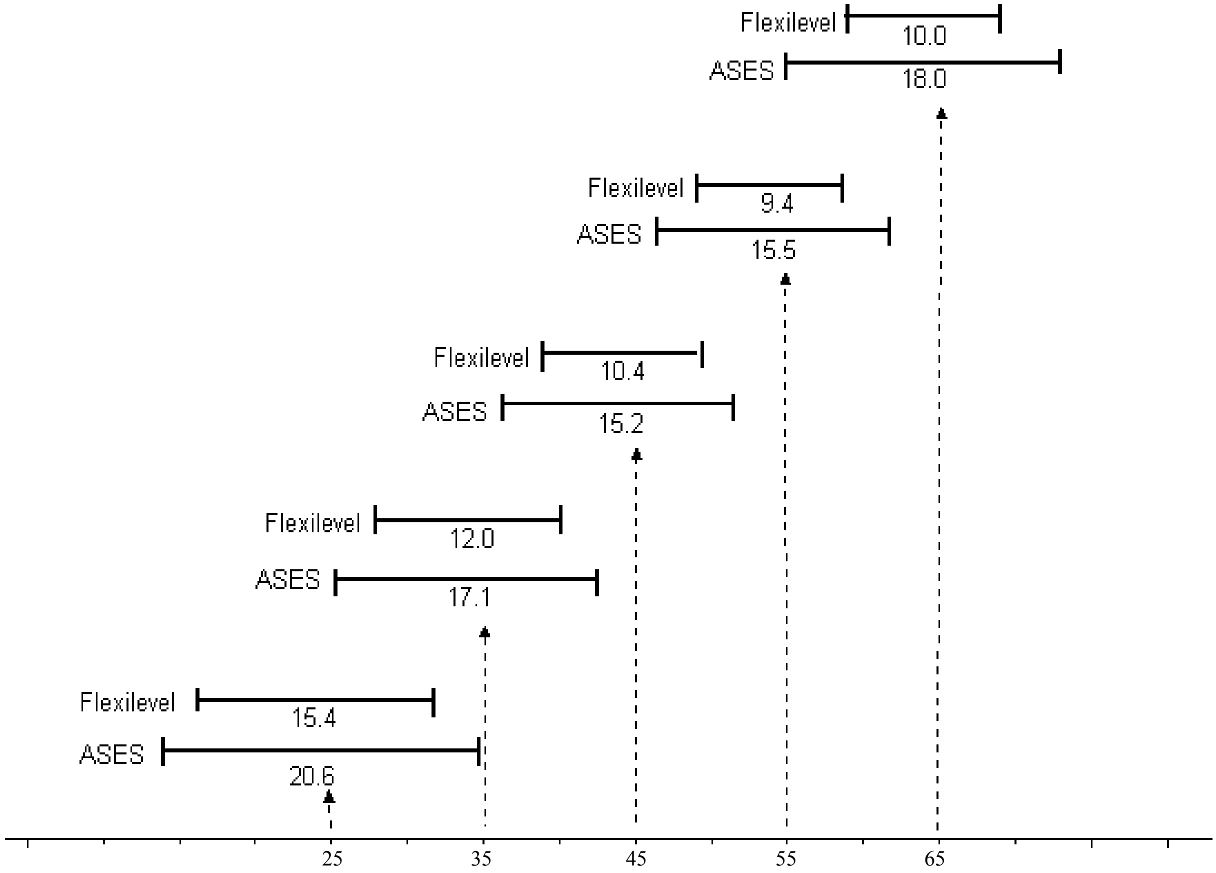


Figure 2. Width of 95% Confidence intervals of ASES and Flexilevel scores for respondents routed to the easiest testlet. (Scores of participants have been rescaled so that the lowest score is 0 and the highest is 100. Widths of intervals are comparable within but not across figures.)

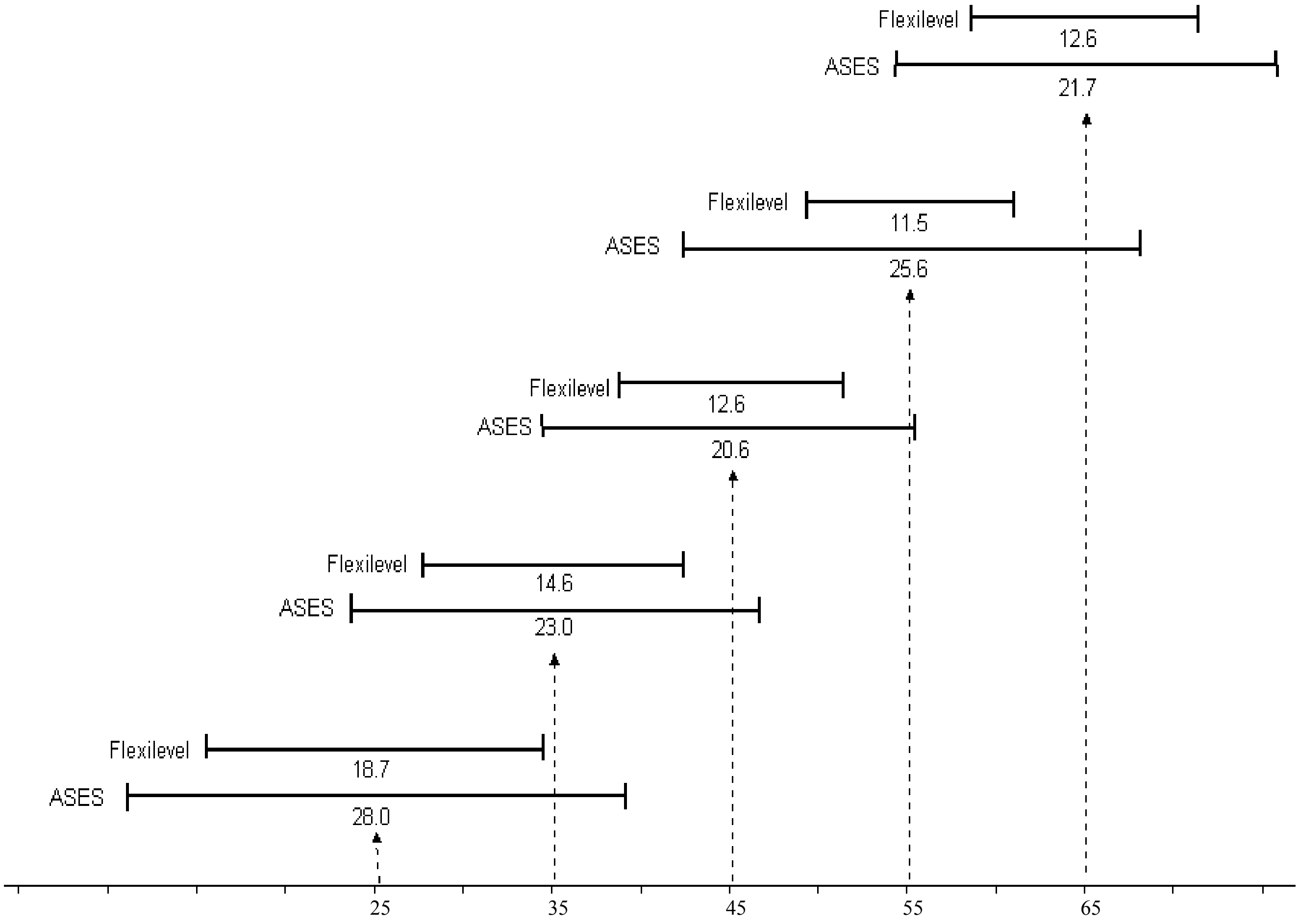


Figure 3. 95% Confidence intervals of ASES and Flexilevel scores for respondents routed to the medium difficulty testlet. (Scores of participants have been rescaled so that the lowest score is 0 and the highest is 100. Widths of intervals are comparable within but not across figures.)

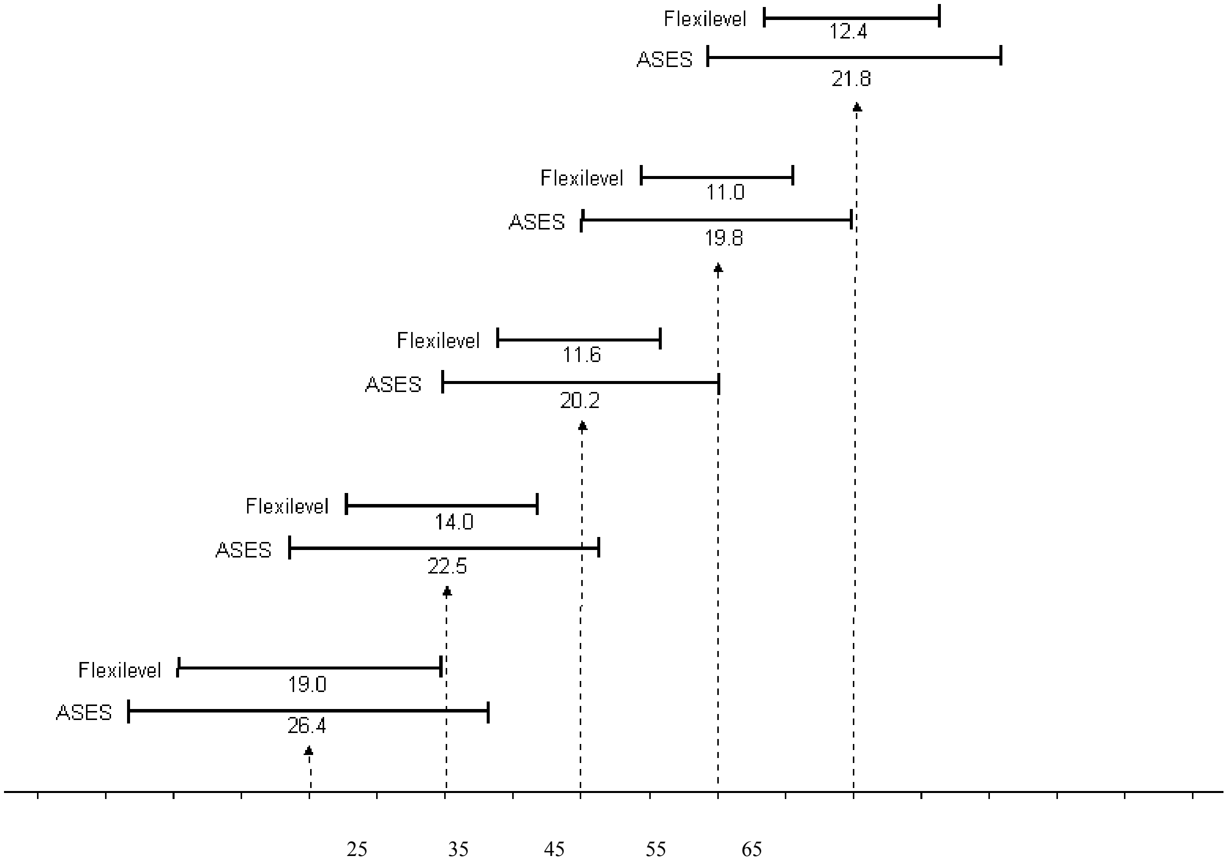


Figure 4. 95% Confidence intervals of ASES and Flexilevel scores for respondents routed to the most difficulty testlet. (Scores of participants have been rescaled so that the lowest score is 0 and the highest is 100. Widths of intervals are comparable within but not across figures.)

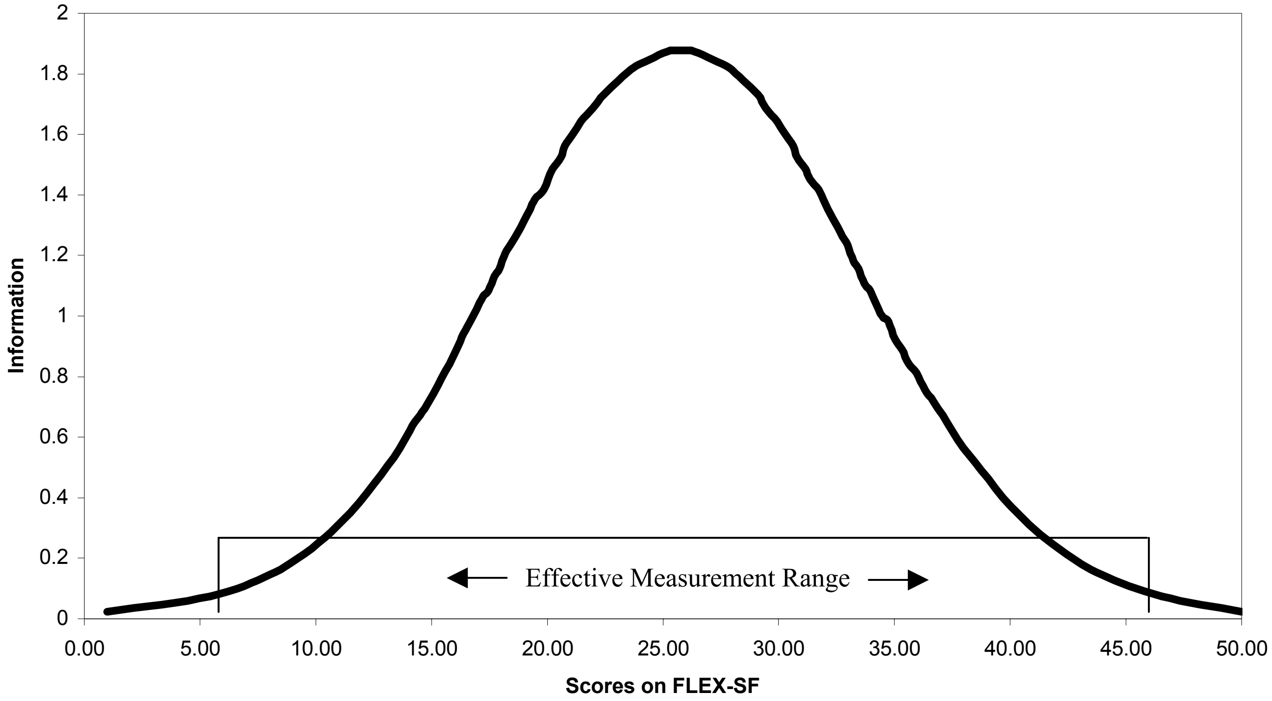


Figure 5. Test Information Function for the FLEX-SF and Effective Measurement Range

Table 1. Spearman Rho Correlations between FLEX-SF Scores and ASES and SF-12 Physical and Mental Health Subscales at Baseline, 1 Month, 2 Month, and 3 Month

FLEX-SF	**Baseline**	**1 Month**	**2 Month**	**3 Month**
ASES	.75	.83	.87	.83
SF-12 Physical	.53	.64	.71	.65
SF-12 Mental Health	.15	.29	.16	.23

Table 2. Summary information for regression of Self-reported Change in Status scores from the one-month follow-up upon differences in patients' baseline and one-month FLEX-SF scores.

	FLEX-SF	ASES	SF-12 Mental	SF-12 Physical
R^2	0.191	0.176	0.005	0.034
beta	0.116	0.107	0.009	0.036
Standard Error	0.022	0.022	0.012	0.019
Probability	<0.0005	<0.0005	0.465	0.060

Table 3. Areas Under the ROC Curves of the FLEX-SF, ASES, and SF-12 Physical and Mental Health Subscales.

	Baseline - 1 Month				1 Month - 2 Month				2 Month - 3 Month				Mean
		+	- / 0	.		+	- / 0	.		+	- / 0	.	
FLEX-SF	.730	73	46	80	.801	62	38	99	.714	57	36	106	.75
ASES	.689	67	43	89	.801	62	38	99	.724	57	36	106	.74
SF-12 Physical	.571	66	38	95	.644	56	33	110	.642	51	33	115	.62
SF-12 Mental	.481	66	38	95	.452	56	33	110	.678	51	33	115	.54

+ = number improved

- / 0 = number not improved

. = number missing (including those who did not complete survey, or did not complete the referent portion of the survey)

Appendix A

Branching Item: How much difficulty do you have using your affected arm
to place a can of soup (about 1 lb.) on a shelf at shoulder height?

I CAN'T DO THIS	Easy testlet
GREAT DIFFICULTY	Easy testlet
SOME DIFFICULTY	Middle difficulty testlet
LITTLE DIFFICULTY	Middle difficulty testlet
NO DIFFICULTY	Hard Testlet

Items	Testlet	Item Difficulty
Use your affected arm to reach the earlobe on the same side as your affected shoulder.	E	19.1
Using your affected arm, turn a faucet in the opposite direction as your affected arm (e.g. turn left if it is your right shoulder that is affected).	E	20.3
Use your affected arm to reach the earlobe on the opposite side of your affected shoulder.	E	20.8
Put on underpants (panties, briefs, or boxers) using both hands.	E	21.0
While sitting, lift your affected hand and put it on a table in front of you.	E	21.1
Use your affected arm to pick up and drink out of a full water glass.	E	21.4
With your affected arm, wash the side of your face opposite your affected shoulder.	E	21.5
Put deodorant under the arm opposite your affected shoulder.	E	22.3
While sitting, reach across to the middle of a table with your affected arm, to get a salt shaker.	M,E	23.7
With your affected arm, steady a jar while you loosen the jar lid.	M,E	23.7
Push yourself out of a chair, using both arms.	M,E	23.9
Pull a chair out from a table, using your affected arm.	M,E	24.5
With your affected arm, tighten a jar lid.	M,E	24.9
With your affected arm, carry something of medium weight in the crook of your arm (where your elbow bends).	M,E	24.9
Using 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).	M,E	25.2
Using your affected arm, turn a steering wheel in the opposite direction as your affected arm (e.g. turn left if it is your right shoulder that is affected).	M	25.4
With your affected arm, slide a medium weight (5-10 lbs.) box across a table by pulling it completely to you.	M	25.8
With your affected arm, slide a medium weight (5-10 lbs.) box across a table by pushing it away from you.	H,M	25.9
Use your affected arm to pull something out of your back pocket.	M	25.9
Use your affected arm to slide hanging clothes in a closet from one end of the rod to the other.	H,M	26.1
Use your affected arm to reach across body to get a car's shoulder strap (safety belt).	H,M	26.4

Items	Testlet	Item Difficulty
Use your affected arm to reach and pull the string that controls a light or fan.	H,M	26.5
Use your affected arm to place a can of soup (1 lb) on a shelf overhead.	H	27.1
With your affected arm, pull a medium weight object (5-10 lbs.) from under a bed.	H	27.3
Use your affected arm to hang a heavy coat in a closet.	H,M	28.3
Use your affected arm to lower a lightweight object (1-5 lbs.) from the top shelf of a closet.	H	28.5
Use your affected arm to reach an overhead shelf.	H	29.1
Use your affected arm to reach the small of your back with your thumb.	H	29.5
Use your affected arm to reach the middle of your back.	H	31.7
Use your affected arm to place a gallon of milk (8-10 lbs.) on a shelf overhead.	H	31.7
While sitting in the front seat of a car, use your affected arm to touch an object on the back seat.	H	32.0
Using your affected arm, work overhead for more than 2 minutes.	H	32.6
While sitting in the front seat of a car, pull a medium weight object (5-10 lbs.) from the bask seat to the front seat of the car, using your affected arm.	H	34.0

E = Easy Testlet
M = Medium Testlet
H = Hard Testlet