

Development of a Flexilevel Scale for use with computer-adaptive testing for assessing shoulder function

Karon F. Cook, PhD,^{a,b,c} Toni S. Roddey, PT, PhD, OCS, FAAOMPT,^d Kimberly J. O'Malley, PhD,^{a,e,f} and Gary M. Gartsman, MD,^g Houston and Austin, TX

In a 5-year study, a self-report measure of shoulder function—the Flexilevel Scale of Shoulder Function (FLEX-SF)—was developed by use of item response theory. A large pool of candidate items (N = 68) was developed. A questionnaire that included the 68 items, another scale of shoulder function, and clinical and demographic questions were administered to 400 persons with shoulder complaints. Patients' responses to the 68 items were calibrated by use of Andrich's rating scale model. Thirty-three items were selected from the pool and subdivided into three overlapping testlets targeting low, medium, and high shoulder function. A table translates raw scores on testlets to a common mathematical metric. The validity and reliability of the FLEX-SF was evaluated in a longitudinal study of 199 patients. The FLEX-SF scores were highly reliable and exhibited excellent validity (including responsiveness). We report on a simulation of a computer-adaptive test of shoulder function. This simulation is based on the developmental items we tested for use in the FLEX-SF. The results indicate that greater measurement efficiency can be achieved with a computer-adaptive test format. (J Shoulder Elbow Surg 2005;14:90S-94S.)

In recent years, interest in self-reported outcomes has increased substantially. The use of patients' subjective judgments to evaluate health outcomes implies confidence that the measures that elicit these self-reports are scientifically sound. Psychometrics is the science and mathematics that concerns itself with such issues.

From the ^aHouston Veterans Affairs Parkinson's Disease Research and Educational Center, ^bMeasurement Excellence and Training Resource Information Center (METRIC): A Veterans Affairs Health Services Research and Development Resource Center, ^cBaylor College of Medicine, ^dTexas Woman's University, ^eHealth Services Research and Development Center for Quality of Care and Utilization Studies, and ^gUniversity of Texas School of Medicine, Houston, and ^fPearson Educational Measurement, Austin.

Reprint requests: Karon F. Cook, PhD, PADRECC, Houston VAMC (127-PD), 2002 Holcombe Blvd, Houston, TX 77030 (E-mail: karonc@bcm.tmc.edu).

Copyright © 2005 by Journal of Shoulder and Elbow Surgery Board of Trustees.

1058-2746/2005/\$30.00

doi:10.1016/j.jse.2004.09.024

The development of an outcome measure by use of psychometric methods is a rigorous, expensive, and time-consuming project. The investment returns confidence that the scale's scores adequately and accurately portray the outcome of interest. Scientifically sound measurement is fundamental to excellence in research and clinical evaluations.

Our work in evaluating the psychometric properties of existing scales of self-reported shoulder outcome^{4,6,20} convinced us of the need to develop a new measure. This report details how, in a 5-year study, psychometric methods were used to develop a self-report measure of shoulder function—the Flexilevel Scale of Shoulder Function (FLEX-SF).⁵ We developed this scale by use of item response theory (IRT),⁸ a psychometric method that (1) accounts for differences in item difficulty and (2) supports Flexilevel Scales. A Flexilevel Scale is composed of 2 or more "testlets," or subsets of items, that target respondents with different levels of the trait being measured. The FLEX-SF measures self-reported shoulder function. Patients respond to an initial "routing item" that classifies them as having low, medium, or high shoulder function. They then respond only to the testlet that best targets their level of shoulder function.

In addition to describing the development of the FLEX-SF, we report on a simulation of a computer-adaptive test (CAT) of shoulder function. This simulation is based on the developmental items we tested for use in the FLEX-SF. CAT-based outcome measures are more efficient even than Flexilevel Scales. They hold substantial promise in the field of outcomes research.

ITEM POOL DEVELOPMENT

A first step in the psychometric method is to develop a large pool of items that could, potentially, be included in the final measure. With a large pool of initial items, scale developers can be selective in choosing the best items for the measure. Scale developers can gather potential items in 3 major ways: (1) adapt published items from other physical function scales, (2) write items based on input from an expert panel, and (3) develop items based on patient interviews. We used each of these in developing the item pool for the FLEX-SF. Existing physical function scales

were examined, and items from those measures were adapted. An expert review committee consisting of a shoulder surgeon, 3 physical therapists, and a psychometrician met to write (and evaluate) items. Most importantly, we interviewed patients and asked about their shoulder problems and how shoulder dysfunction impacted their lives.

There were 68 candidate items in the initial pool. The research team developed a questionnaire that included these items, another scale of shoulder function,⁷ and clinical and demographic questions. The survey was administered to 400 persons with shoulder complaints. Data were collected and entered into an electronic database.

ITEM CALIBRATION

Previously published shoulder scales^{12,13,18,19} were developed by use of the classical psychometric model. For the FLEX-SF, we used a modern psychometric method called item response theory (IRT). In the classical approach, patients respond to items that query them about their ability to do specific tasks (eg, throw a ball or lift a 1-lb weight). Response options vary. In some scoring systems, there are only 2 options ("no = 0" or "yes = 1"). More often, respondents indicate on a Likert-type scale¹ the degree of difficulty they have performing specified functions. A shoulder scale might present patients with the following continuum of response options: "I can't do this = 0," "great difficulty = 1," "some difficulty = 2," "little difficulty = 3," and "no difficulty = 4." The number associated with the response chosen is the item score. To obtain a person's score for the whole scale and, in doing so, estimate his or her level of shoulder function, the item scores are manipulated mathematically. In most cases the mathematical manipulation is simply the addition or averaging of item scores.

This approach has limitations that concerned us in scaling shoulder function. First, with classically developed scales, no consideration is given to the relative difficulty of the tasks that constitute the scale's items. Two items on the Simple Shoulder Test¹³ ask respondents if they can lift an object to shoulder level without bending their elbow. One asks about lifting an object weighing 1 lb; the other asks about lifting an object weighing 8 lb. The classical method does not account for the different difficulties of the 2 tasks. A second limitation of the classical approach is that participants are asked to respond to all items, including ones that are not appropriate for them. For example, persons with a massive rotator cuff tear are asked to answer questions such as, "Can you throw a ball overhand 20 yards?"

IRT approaches overcome limitations of the classical psychometric model. After collecting patient responses to the 68 developmental items of the FLEX-SF,

we used a computer program³ to analyze (ie, calibrate) items and patients' responses with an IRT model—Andrich's rating scale model.² We calculated fit statistics for each of the 68 candidate items. Eight items failed to meet conventional standards for item fit and were dropped from further analyses.

STRUCTURING THE FLEX-SF

The IRT calibration ordered items from easiest to hardest. Figure 1 displays the relative difficulty of 5 sample items from the final form of the FLEX-SF. Thirty-three of the best items were subdivided into three overlapping testlets: an easy testlet, a middle-difficulty testlet, and a hard testlet. Each testlet was printed on a different color of paper. To begin the FLEX-SF, patients answered the question, "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?" Answers to this initial item were used to route respondents to the testlet that best targeted their function level. Respondents are directed to, for example, "Take only the items on the blue (or yellow or pink) piece of paper." A table translates scores on testlets to the corresponding FLEX-SF scores. FLEX-SF scores are calibrated to a common mathematical metric.

EVALUATING THE FLEX-SF

We conducted a longitudinal study of 199 patients to evaluate the final version of the FLEX-SF. Participants completed a packet of questionnaires at recruitment. A follow-up packet of questionnaires was mailed monthly for 3 months after recruitment. The packet included the American Shoulder and Elbow Surgeons (ASES) scale,¹⁸ the short form (SF)-12,²¹ and the FLEX-SF.⁵

Scale reliability

The reliability of an outcome's measure can be estimated in a number of ways. Test-retest reliability evaluates the stability of scores over time. Internal consistency estimates reliability based on the relationships among item scores and their consistency within the study population. To evaluate test-retest reliability, we calculated an intraclass correlation coefficient. The calculated value of the intraclass correlation coefficient was quite high, 0.90, with a 95% confidence interval of 0.84 to 0.94. To evaluate how internally consistent items were, we calculated Cronbach α values for each of the testlets individually. The values we obtained for this reliability analysis also were high. For the easy, medium-difficulty, and hard testlets, respectively, we calculated values of .96, .93, and .97.

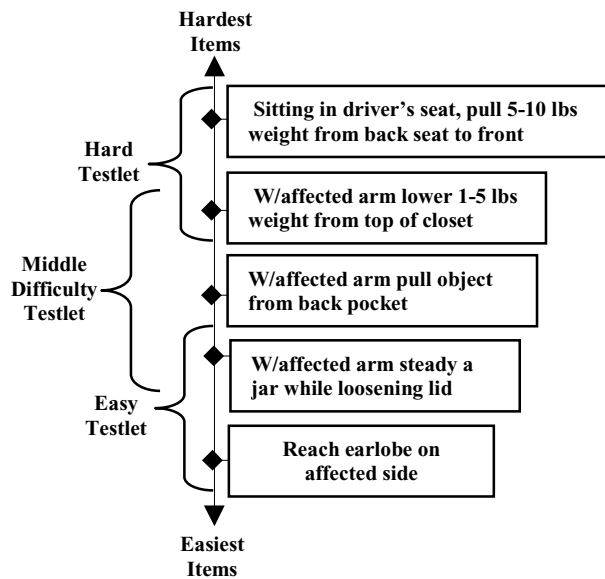


Figure 1 Example items from the 3 testlets of the FLEX-SF.

Score-level reliability

With classically developed shoulder scales, only a scale-level estimate of reliability can be calculated. This estimate averages the reliability of the scale in measuring low through high scores and obscures the fact that scales typically are more precise in measuring some levels of shoulder function than in measuring others. A unique feature of scales developed by use of IRT models is that the reliability of every possible score on the outcome measure is estimated. We compared the score-level reliability of the FLEX-SF with that of the ASES scale. The details of these analyses and our results are reported in detail elsewhere.⁵ The FLEX-SF scores proved more reliable than the ASES scores across all levels of the scale. In some cases the differences were quite substantial.

Validity

Validity calculations estimate the degree to which a scale measures the outcome it is purported to measure. Inappropriately, many researchers will describe a scale as "valid" or "not valid." Validity, however, is not a dichotomous characteristic. The purpose of psychometric analyses is to gather and evaluate evidence regarding the validity of using a scale's scores in a specified population for a specified purpose. There are many ways to collect evidence to support (or reject) an outcome measure's validity. To ensure that items of the FLEX-SF represented self-reported shoulder function well, we interviewed patients and convened an expert panel to review potential scale items.

Once the FLEX-SF was developed, we used addi-

tional methods to evaluate its validity. For example, a set of a priori hypotheses formalized our expectations regarding the associations among FLEX-SF scores and scores on other outcome measures. These expectations were largely upheld. FLEX-SF scores proved also to be quite responsive; that is, changes in FLEX-SF scores corresponded well with changes in external indicators such as scores on other outcome measures.¹⁰ To estimate the minimally clinically important difference for FLEX-SF scores, we identified the subset of patients who reported a noticeable change in status. The averaged score change in this subset was 3.02. This value supports responsiveness of FLEX-SF scores to changes in clinical status. Details of additional validity assessments have been reported elsewhere.⁵

COMPUTER-ADAPTIVE SHOULDER TESTING

On the basis of our psychometric analyses of the FLEX-SF, we concluded that FLEX-SF scores were reliable and valid in monitoring the status of patients with shoulder problems. In addition, we found that using the Flexilevel testing strategy appreciably reduced response burden without sacrificing the scientific soundness of the measure.

The measurement model that allows Flexilevel testing makes possible an even more efficient mode of outcomes measurement—computer-adaptive testing (CAT). (Note: A scale administered by computer is not computer-adaptive. In the former, an exact equivalent of the scale in its paper-and-pencil form is administered to the patient. The delivery mode only is changed.) With CAT measures, after the first item, the presentation of each successive item is determined by the person's responses to previous items. In a CAT-based outcome measure, a computer algorithm estimates a patient's level of the outcome being measured (eg, shoulder function), updates the estimate each time the patient answers another item, and selects the next item to present based on its match with the computer's updated estimate of the respondent's trait level.

There currently are no CAT-based shoulder outcome scales. To demonstrate how such a scale would work, however, we simulated a CAT of shoulder function based on data we collected in developing the FLEX-SF. This simulation demonstrates how CATs operate to maximize the efficiency of adaptive scaling.

SIMULATION METHODS

For our simulation of computer-adaptive testing of shoulder function, we required (1) a software program for adaptive testing, (2) a pool of items calibrated to an IRT model, (3) a data set of patients'

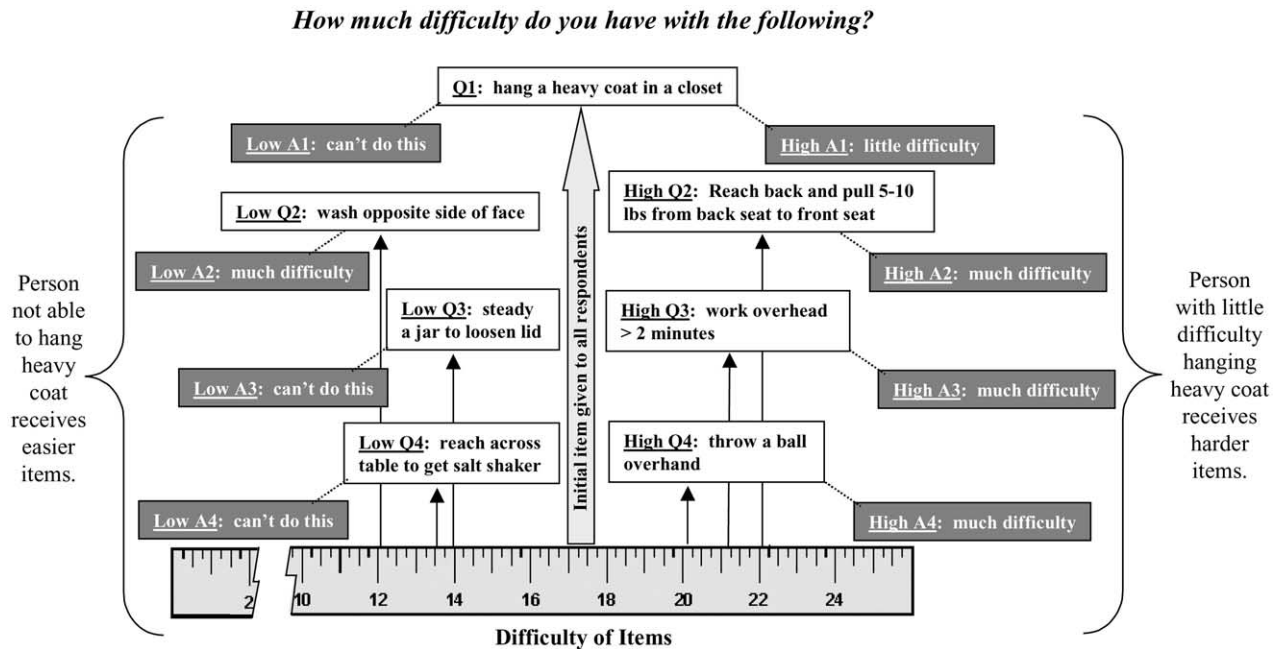


Figure 2 Questions (Q) and answers (A) for 2 patients (one with high shoulder function and one with low shoulder function) in a simulated computer-adaptive administration of shoulder function questions.

responses to this pool of items, (4) a stopping rule, and (5) a flashy acronym. For the simulation, we used a software program developed at the University of Texas at Austin. At the time of this study, we had access only to a version of this program that simulated CAT by use of an IRT model called the partial credit model (PCM).¹⁵ As stated above, our pool of shoulder function items was calibrated with a different IRT model.² Therefore, we recalibrated 60 of the developmental pool of items using the PCM. Scores were calibrated to range from 1 to 25.

The responses of persons who answered all 60 items (N = 213) were input into the simulation program. This data served as a proxy for patients' responses to the items as presented by the computer. The simulation rests on the assumption that patients would have answered the items of the CAT in the same way that they answered the items in the paper-and-pencil version.

With traditional measures, assessment stops when the patient completes all of the items of the scale. With CAT measures, patients respond only to a subset of the available pool of items, so a stopping rule must be identified. Stopping rules may specify that the assessment ends (1) after a given number of items are administered, (2) after a specified level of precision is reached, or (3) after some combination of 1 and 2. For our study, we chose to specify the level of precision. Typically, a measurement's precision improves as patients respond to more items. This improvement is evidenced in a decrease in the magnitude of the standard

error of measurement (SEM). In a shoulder outcomes CAT, a patient's estimated shoulder function is updated with every response, as is the SEM statistic. For the current demonstration, we programmed the computer algorithm to cease administering items once the SEM fell to or below 1.5 (indicating 95% probability that the patients' actual shoulder function lies within ± 3 points of the calculated score).

To distinguish the FLEX-SF from the simulated, CAT-based shoulder outcome measure, we will refer to the latter as the Shoulder Health Outcomes Computer Adaptive Test (SHO-CAT). (A publicly available version of the SHO-CAT was planned for release in 2004. Information regarding obtaining the software will appear on the Web site of the Measurement Excellence and Training Resource Information Center [METRIC] at <http://www.measurementexperts.org>.)

SIMULATION RESULTS

The process CAT uses to select items and estimate patients' shoulder function is iterative and analogous to a children's game of "I spy." In this game, after every guess, the guesser is told, "You're getting warmer" or "You're getting colder." On the basis of this continuing feedback, the guesser gradually "hones in" on the target object. In much the same way, a computer-adaptive shoulder measure hones in on a good estimate of the patient's shoulder function. This estimate improves each time a patient responds to an item. The progress of the estimation procedure

can be evaluated by tracking changes in the value of the SEM. As the estimate becomes more accurate ("warmer"), the SEM decreases. For the current simulation, we programmed the computer algorithm to stop administering items once the SEM dropped below 1.5.

For the majority of the patients in our simulated SHO-CAT, the SEM-based stopping rule was reached after only a few items. Of the patients, 61% (129/213) were administered 5 items or fewer and 73% (155/213) were administered 10 items or fewer. For 12 patients (6%), the CAT reached the specified SEM after only 3 items had been administered.

For demonstration purposes, we selected 2 examinees to demonstrate how the SHO-CAT tailors the administration of items to the shoulder function of the patient (Figure 2). The same initial item was administered to each patient: "How much difficulty do you have hanging a heavy coat in a closet?" Figure 2 displays the assessment path for 2 patients, one with relatively low shoulder function (left side of figure) and another with relatively high shoulder function (right side of figure). For both of these patients, the stopping rule was reached after 4 questions had been answered.

To the initial item regarding hanging a heavy coat in a closet, the lower-functioning patient (P-LOW) responded "can't do this" and the higher functioning patient (P-HIGH) responded "no difficulty." From here, the assessment paths for the 2 diverge. P-LOW gets the easier question, "How much difficulty do you have washing the side of your face opposite your affected shoulder?" In contrast, the second item administered to P-HIGH asks about the difficulty of "while sitting in the driver's seat of a car, reaching into the backseat and pulling a 5-10 lb weight to the front seat." As displayed in the figure, this individualizing of the items administered continues until the SEM drops to 1.5 or below.

CONCLUSIONS

We concur with the many health outcome researchers who have trumpeted the promising future of adaptive testing for the measurement of patient-reported outcomes.^{9,11,14,16,17,22-24} Our evaluation of the FLEX-SF and our simulation of the SHO-CAT convince us that the adaptive scaling methods have great potential in improving the measurement of shoulder function.

REFERENCES

- Anastasi A. Psychological testing. New York: Macmillan Publishing; 1988.
- Andrich DA. A rating formulation for ordered response categories. *Psychometrika* 1978;43:561-73.
- BIGSTEPS. Rasch-model computer program. Chicago: MESA Press; 1997.
- Cook KF, Gartsman GM, Roddey TS, Olson SL. The measurement level and traits-specific reliability of 4 scales of shoulder functioning: an empiric investigation. *Arch Phys Med Rehabil* 2001;82:1558-65.
- Cook KF, Roddey TS, Gartsman GM, Olson SL. Development and psychometric evaluation of the Flexilevel Scale of Shoulder Function. *Med Care* 2003;41:823-35.
- Cook KF, Roddey TS, Olson SL, et al. Reliability by surgical status of self-reported outcomes in patients who have shoulder pathologies. *J Orthop Sports Phys Ther* 2002;32:336-46.
- Gartsman GM. Arthroscopic acromioplasty for lesions of the rotator cuff. *J Bone Joint Surg Am* 1990;72:169-80.
- Hambleton R, Swaminathan H. Item response theory: principles and applications. Norwell (MA): Kluwer Academic; 1985.
- Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. *Med Care* 2000;38:1128-42.
- Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. *J Clin Epidemiol* 2000;53:459-68.
- Jenkinson C, Fitzpatrick R, Garratt A, Peto V, Stewart-Brown S. Can item response theory reduce patient burden when measuring health status in neurological disorders? Results from Rasch analysis of the SF-36 physical functioning scale (PF-10). *J Neurol Neurosurg Psychiatry* 2001;71:220-4.
- Leggin BG, Shaffer MA, Neuman RM, et al. Shoulder outcome measurement. In: Iannotti JP, Williams GR, editors. Disorders of the shoulder: diagnoses and management. Baltimore: Lippincott Williams & Wilkins; 1999. p. 1023-40.
- Lippitt SB, Harryman DT, Matsen FA. A practical tool for evaluating function: the simple shoulder test. In: Matsen FA, Fu FH, Hawkins RJ, editors. The shoulder: a balance of mobility and stability. Rosemont (IL): American Academy of Orthopedic Surgeons; 1993. p. 501-30.
- Lohr KN. Health outcomes methodology symposium: summary and recommendations. *Med Care* 2000;38:1194-208.
- Masters GN. A Rasch model for partial credit scoring. *Psychometrika* 1982;47:149-74.
- McHorney CA. Generic health measurement: past accomplishments and a measurement paradigm for the 21st century. *Ann Intern Med* 1997;127:743-50.
- 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.
- Richards RR, Bigliani LU, Gartsman GM, Iannotti JP, Zuckerman JD. A standardized method for the assessment of shoulder function. *J Shoulder Elbow Surg* 1994;3:347-52.
- Roach KE, Budiman-Mak E, Songsiridej N, Lertratanakul Y. Development of a shoulder pain and disability index. *Arthritis Care Res* 1991;4:143-9.
- Roddey TS, Olson SL, Cook KF, Gartsman GM, Hanten W. Comparison of the University of California-Los Angeles Shoulder Scale and the Simple Shoulder Test with the shoulder pain and disability index: single-administration reliability and validity. *Phys Ther* 2000;80:759-68.
- 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;34:220-33.
- Ware JE Jr. The status of health assessment 1994. *Annu Rev Public Health* 1995;16:327-54.
- 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;38:1173-82.
- Ware JE Jr, Kosinski M, Bjorner JB, et al. Applications of computerized adaptive testing (CAT) to the assessment of headache impact. *Qual Life Res* 2003;12:935-52.