© Health Research and Educational Trust
DOI: 10.1111/j.1475-6773.2005.00446.x

Dynamic Assessment of Health Outcomes: Time to Let the CAT Out of the Bag?

Karon F. Cook, Kimberly J. O'Malley, and Toni S. Roddey

Background. The use of item response theory (IRT) to measure self-reported outcomes has burgeoned in recent years. Perhaps the most important application of IRT is computer-adaptive testing (CAT), a measurement approach in which the selection of items is tailored for each respondent.
Objective. To provide an introduction to the use of CAT in the measurement of health outcomes, describe several IRT models that can be used as the basis of CAT, and discuss practical issues associated with the use of adaptive scaling in research settings.
Principal Points. The development of a CAT requires several steps that are not required in the development of a traditional measure including identification of "starting" and "stopping" rules. CAT's most attractive advantage is its efficiency. Greater measurement precision can be achieved with fewer items. Disadvantages of CAT include the high cost and level of technical expertise required to develop a CAT.
Conclusions. Researchers, clinicians, and patients benefit from the availability of psychometrically rigorous measures that are not burdensome. CAT outcome measures hold substantial promise in this regard, but their development is not without challenges.

Key Words. Measurement, quality of life, psychometrics, reliability

Measures of patient-centered outcomes are critical to health services research since they supply information on which clinical and health policy decisions may be made. Patient-level data is used to track important health variables including burden of illness, patient-satisfaction, and other health-related quality of life (HRQoL) outcomes. Increasingly, HRQoL is being used as a primary outcome in clinical trials. HRQoL variables often also serve as elements for economic and public policy analyses.

In the last three decades, the use of item response theory (IRT) to measure self-reported outcomes has burgeoned. IRT is often called "modern psychometric theory" to distinguish it from classical test theory (CTT). Health outcome scientists have used IRT to evaluate the properties of existing measures (Kirisci, Tarter, and Hsu 1994; Kirisci, Moss, and Tarter 1996), develop new measures (Velozo and Peterson 2001; Cook, Roddey, and Gartsman 2003), and equate various item pools to a common mathematical metric (Cella, Llyod, and Wright 1996; Baker, Rounds, and Zevon 2000; McHorney and Cohen 2000; Wolfe 2000). Arguably the most useful application of IRT is computer-adaptive testing (CAT).

The purpose of this paper is to describe methods for developing CATs and discuss CAT's advantages and disadvantages. Although a full technical discussion is beyond the scope of this paper, we provide an overview of adaptive testing, introduce several IRT models used in developing CATs, and discuss practical issues associated with implementing CAT.

DEFINITIONS

The lexicon of psychometrics includes many words that have everyday as well as technical meanings. One of the more confusing terms is the word, "scale." Depending on the context, a scale can be: (1) a device for measuring mass (e.g., bathroom scales), (2) a collection of items developed to measure a quantity of interest (e.g., the SF-36 physical functioning scale), or (3) a system of units and numbers that define a mathematical metric (e.g., feet and inches are units of the *imperial* scale; centimeters and meters are units of the *metric* scale). To avoid confusion, we will refrain from using the word "scale." When we refer to a collection of items developed to measure a quantity of interest, we will use the word, "measure." When we refer to a system of units and numbers that define a mathematical metric, we will use the term, "metric."

Four other psychometric terms often are confused—*construct*, *trait level*, *latent*, and *continuum*. The word "construct" refers to the dimension that the measure is intended to assess. The first step in measurement is to define (i.e., construct) this dimension. Measured constructs also are referred to as "traits."

"Trait level" refers to a person's "true" amount (measured without error) of the specified construct. A person's observed score on an outcome measure

Address correspondence to Karon F. Cook, Ph.D., Veterans Affairs Measurement Excellence Training Resource Information Center (METRIC), Houston Center for Quality Care & Utilization Studies, 801 Cortlandt St. Houston, TX 77007. Dr. Cook is also with the Department of Rehabilitation Medicine, School of Medicine, University of Washington, Seattle, WA. Kimberly J. O'Malley, Ph.D., is with Pearson Educational Measurement, Austin, TX. Toni S. Roddey, P.T., Ph.D., O.C.S., F.A.A.O.M.P.T., is with Department of Physical Therapy, Texas Woman's University, Houston, TX.

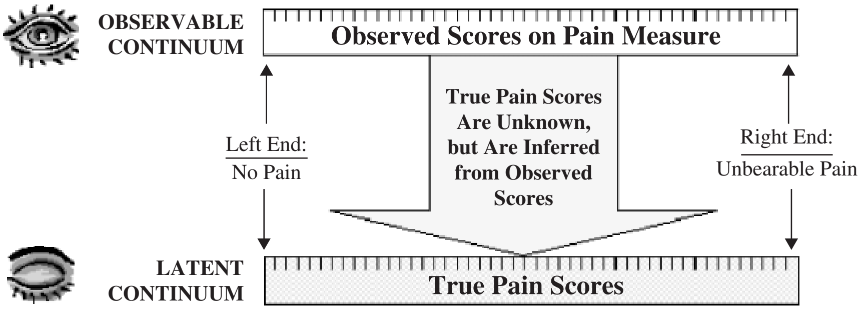
estimates his or her true score. The accuracy of this estimate depends on how close the person's measured score (observed score) is to the person's actual trait level.

The word "latent" means hidden or dormant. A latent construct is one that cannot be seen but can be inferred based on observations of persons' behavior. For example, pain cannot be observed directly, but can be inferred by observing behaviors (e.g., grimaces, guarding, responses to items about pain). In educational and psychological assessment, a measure is defined as a "sample of behavior." The sample of behavior observed is the response category choices made by persons responding to items. Parenthetically, this is a reason single-item measures typically yield the least reliable scores—they estimate a person's trait level based on the smallest sample—a sample of one.

The word, "continuum," has been defined as "a coherent whole characterized as a collection, sequence, or progression of values or elements varying by minute degrees" (Merriam-Webster Inc. 2003) Using pain as an example, we describe two continua, one latent, the other observable (Figure 1). The latent continuum represents persons' actual levels of pain. Scores on this continuum are *true* scores. True pain scores cannot be observed directly nor can they be known with certainty. Thus, it is *hypothesized* that an unobservable pain continuum exists on which there is a "progression of values or elements varying by minute degrees."

The upper continuum in Figure 1 represents the continuum of observed scores. Observed scores *indicate* persons' true levels of pain. The relationship between observed and true pain scores cannot be determined exactly, but

Figure 1: Relationship between Observed Scores on a Measure of Physical Function and True Scores That Are Hypothesized to Exist on a Latent Continuum


OBSERVABLE CONTINUUM
Observed Scores on Pain Measure
True Pain Scores Are Unknown, but Are Inferred from Observed Scores
Left End: No Pain
Right End: Unbearable Pain
LATENT CONTINUUM
True Pain Scores

psychometric analyses reduce our uncertainty about this relationship (e.g., criterion validity, test/retest reliability).

Putting all this terminology together, the study of health and medical outcomes concerns itself with the measurement of latent constructs that are hypothesized to be ordered along an underlying and unobservable continuum.

What Is CAT?

A CAT is a computer-administered test (measure) in which, after the first item, presentation of items is determined by persons' responses to previous ones. A measure that is computer administered *only* is not computer *adaptive.* In the former, the patient responds by computer to an exact equivalent of the measure's paper-and-pencil counterpart. Only the delivery mode changes. Persons respond to each item, and their scores are computed exactly as they would have been in the paper-and-pencil version. A measure is adaptive only when the selection of items is *adapted* to prior estimates of the respondent's level of the construct being measured.

The Math Behind the CAT

CTT requires that participants respond to every item of a measure. If a person skips an item or chooses not to answer an item, the missing response must be imputed. Persons' trait levels are estimated by manipulating item scores mathematically. In most cases, the mathematics is the simple summing of item scores.

With an IRT model, imputation of missing responses is unnecessary. Whereas in CTT, trait levels are estimated by asking the question, "Given a person's total score, what is the respondent's level on the trait being measured?", the fundamental estimation question in IRT is, "Given what is known about the items and the persons' responses, what is the respondent's most likely level of the trait being measured?" Note that CTT concerns itself with *total scores* and IRT with *item responses.*

Item Parameters

IRT methods model how persons with given trait levels will respond to items that have specified characteristics. The "specified characteristics" may include item difficulty, item discrimination, and item "guessing." The guessing parameter (sometimes called pseudo-guessing) has limited relevance in outcomes measurement. It estimates the probability of getting an item correct purely by chance and applies to knowledge-based assessments (e.g., knowledge of cancer warning signs) in which items are scored as correct or incorrect.

The meaning of item difficulty is intuitive for traits such as physical function. A shoulder function item that asks respondents if they can throw a ball 20 yards overhand is more difficult than one that asks about picking up a 1 lb can. Another way to say this is to note that, for the first item, more shoulder function is required for an affirmative response than is required for an affirmative response to the second. Item difficulty of psychosocial variables can be thought of as difficulty to endorse. Consider two items from the Center for Epidemiologic Studies depression measure (CES-D): (1) "I felt that everything I did was an effort" and (2) "I felt hopeless about the future." The second item is more difficult to endorse and, thus, would have the higher estimated item difficulty in an IRT calibration.

The item discrimination parameter models the rate of increase in the probability of endorsing an item as trait level increases. The parameter indicates the strength of association between an item and the construct being measured. Highly discriminating items improve a measure's ability to demarcate fine gradations among persons with similar levels of the measured trait, that is, to discriminate among them.

IRT Estimation of Trait Level

To get a feel for how an IRT mathematical model estimates a person's level on a latent construct, consider the following example. Suppose you observe the responses of two persons to several items that measure shoulder function. The first person is asked to rate her difficulty using her involved arm to: (1) put a gallon of milk on a waist-high shelf, (2) throw a softball overhand 20 yards, and (3) reach into the backseat of a car and pull a heavy object into the front seat. The respondent indicates that she has "no difficulty," "little difficulty," and "some difficulty" with these three tasks, respectively. A second person also responds to three items. He is asked to rate his difficulty using his involved arm to: (1) turn on a light switch, (2) push a 1 lb. object across a table while seated, and (3) place a soup can on a waist-high shelf. His responses are the same as the first respondent's—"no difficulty," "little difficulty," and "some difficulty," respectively. Summing each person's item scores yields the same score for each person. However, using only what you intuit about the relative difficulty of the tasks and the responses that each person gave to each item, you could guess (estimate) that the first person has higher shoulder function than the second. Conceptually, this is similar to how persons' scores are estimated using an IRT model.

Estimation of Trait Level Using CAT

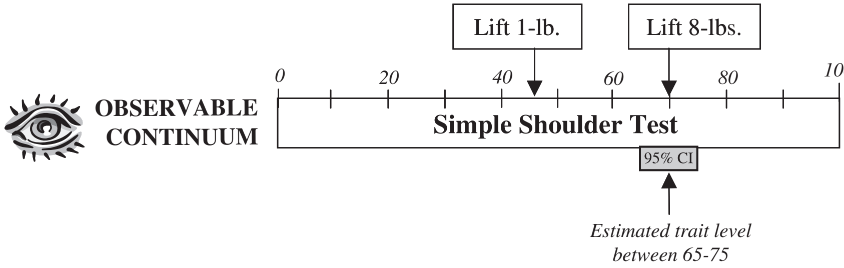
IRTs probability-based estimation strategy is what makes CAT possible. Because it is unnecessary for persons to respond to the same items, the CAT algorithm can pick and chose items that yield the most information about persons' trait levels. Generally the most informative items are those whose difficulty is comparable with the person's trait level. For example, from a set of items that can be answered "yes" or "no," the most informative items for a given person are those to which respondents, based on their trait level, are approximately equally likely to answer "yes" as to answer "no."

A specific example may clarify why items matched to persons' trait levels provide more measurement information. In previous work (Cook, Gartsman, and Roddey 2001), we used an IRT model to calibrate the items of the Simple Shoulder Test (SST) (Lippitt, Harryman, and Matsen 1993), a measure of shoulder function that has a yes/no response format. Scores were calibrated to a range of 0–100 with higher scores indicating greater shoulder function. Two of the items of the SST ask respondents if they can lift an object to shoulder level without bending their elbow. One item asks about lifting an object weighing 1 lb.; the other asks about lifting an object weighing 8 lbs. Persons with an IRT-calibrated score higher than 43 were equally likely to answer "yes" as to answer "no" to the question about a 1 lb. weight. Persons with an IRT-calibrated score of 70 were equally likely to answer "yes" as "no" to the question about an 8 lb. weight. Figure 2 portrays the locations of these items on the latent continuum.

Item Banks

Item banks are sets of items that, after thorough evaluation of their clarity, content, sensitivity, and other psychometric properties, have been calibrated to an IRT model. Suppose you have a shoulder function item bank, and the two items from the SST described above are in that bank. Suppose further that a woman taking the CAT has answered three items thus far. Based on her responses to the initial items, the computer algorithm estimates at a 95 percent confidence interval that her shoulder function is somewhere between 65 and 75. The woman's response to the next item should help us hone in on a more

Figure 2: Relative Positions of Two Items of the Simple Shoulder Test and an Estimated Trait Level Estimated to Be between 65 and 75 at the 95 Percent Confidence Interval (CI)



precise estimate of her shoulder function. Which of the two items described above would the computer algorithm choose to administer next? The CAT algorithm would select the one that asks about her ability to lift an 8 lb., not a 1 lb., weight. The information that she can lift a 1 lb. object to shoulder level helps little in obtaining a more precise estimate of her shoulder function.

If the woman in the example above had been responding to a classical measure of shoulder function, she would be asked to respond to all the items because, in a CTT-based measure, even those with a massive rotator-cuff tear have to answer questions like, “Can you throw a ball overhand 20 yards?” A CAT version would present relatively easy items to persons with very low function and harder items to those with higher function.

HOW DOES CAT WORK?

The process CATs use to select items and estimate trait level is iterative. Based on continuing feedback, the CAT gradually “hones in” on a trait-level estimate. Continuing with the shoulder function example, the CAT algorithm obtains an initial, gross estimate of shoulder function based on the person’s response to the initial item. If the patient replies “great difficulty” to the first shoulder function item, the computer algorithm selects an easier item to administer next. If the patient replies “no difficulty,” the CAT chooses a more difficult item. The person’s response to the second item is used to update the estimate of the person’s shoulder function, and a new item, matched to the updated estimate, is presented. This continues until a prespecified “stopping rule” is reached.

ADVANTAGES OF CAT

The contribution of a CAT approach is increased "measurement efficiency." We define measurement efficiency as the ratio of a measure's psychometric soundness (e.g., reliability, validity) to the response burden the measure imposes (e.g., time and attention required to respond to the items). The achievement testing literature confirms the increased measurement efficiency of CATs for dichotomously scored items (right/wrong). Vispoel, Wang, and Bleiler (1997) found an adaptive test matched the reliability of each of two traditional tests of music memory with 57 percent fewer items. Bowers (1991) found that an adaptive test required one-third as many items as a traditional test to measure mastery/nonmastery of melodic interval identification. In a comparison of traditional versus adaptive testing of vocabulary, 13 items provided higher levels of reliability and concurrent validity than a 40-item traditional test (Vispoel 1993). Similar results have been obtained in personality assessment (Waller and Reise 1989; Reise and Henson 2000; Weiss 2003).

Increased measurement efficiency of CAT is beneficial to the field of achievement and psychological testing, but in the assessment of outcomes in patient populations, decreasing response burden may be not only beneficial, but merciful. It is intuitive that cancer patients should not be asked to complete a lengthy self-report measure of fatigue. The focus of this supplement is the population of veterans who receive their health care in the Veterans Affairs health care system. This population, on average, is in poorer health than the general population and may benefit substantially from the decreased response burden of CAT-based health outcomes assessment (Agha et al. 2000).

In traditional measures, the lack of items targeting low levels of function or high levels of symptoms results in imprecise measurement of the most affected patients. These floor and ceiling effects are reduced with a CAT that has an extensive item bank. Persons very high or very low on the trait being measured receive items that target their level of trait and yield more precise estimates.

Another advantage of CAT is flexibility. Not only do CAT measures adapt to the trait level of the respondent, they can be adapted to specific measurement contexts. When precise trait-level estimates are needed, a standard error of measurement (SEM) stopping rule might be chosen that ensures 95 percent confidence (± 2 SEM) that a respondent's actual trait level falls within a range of only a few points. When this high level of accuracy is not needed, the stopping rule can be relaxed. The CAT asks fewer items, stops sooner, and respondent burden is less.

How Are CATs Developed?

Initial steps in developing a CAT are similar to those necessary for developing a traditional measure, however a larger number of items is needed for CAT item banks. Developmental steps include specification of the construct to be measured, creation of a developmental pool of items, evaluation and modification of the item pool, and assessment of the psychometric properties of the newly developed measure. If the measure will be a CAT, additional steps must be taken. The developmental items must be administered to a large sample of persons who are representative of the population of interest. These responses are used to tests the assumptions of IRT and to calibrate the items to an IRT model. Items should also be evaluated for differential item functioning (DIF). DIF is present when an item functions differently in different subgroups (e.g., males and females). A decision must be made regarding misfitting items and items that exhibit DIF. DIF items may be dropped or they may be calibrated separately for each subgroup. Items that misfit the chosen IRT model may be dropped or, when appropriate, a less restrictive IRT model chosen. In developing a CAT, it also is necessary to identify starting and stopping rules.

Choosing an IRT Model

Most CATs are developed on the basis of unidimensional IRT models. Unidimensional IRT models assume that how patients respond to a measure's items depends upon how much they have of a single, latent construct. There are IRT models in which two or more latent constructs are assessed simultaneously (see work by Gardner, Kelleher, and Pajer 2002). In most research settings, however, obtaining distinct scores for each latent construct engenders greater conceptual clarity and score interpretability. Also, multidimensional IRT models require large sample sizes and are impractical in many clinical and research settings. For the current discussion, therefore, we limit ourselves to the more widely applicable, unidimensional IRT models.

The simplest of the unidimensional IRT models are the one parameter-logistic (1 p-l) models, also called "Rasch models" (Rasch 1960; Andrich 1978; Masters 1982; Wright and Masters 1982). The single item parameter estimated in a Rasch model is item difficulty. In the two parameter-logistic (2 p-l) models, both item difficulty and item discrimination are estimated (Samejima 1969; Muraki 1992). A three parameter-logistic (3 p-l) model estimates, additionally, a pseudo-chance or guessing parameter.

An additional distinction is drawn among IRT models based on number of responses options. Measures may be comprised of items with a dichotomous response format (e.g., yes/no) or a polytomous response format (e.g., never/sometimes/always). The first IRT models were developed to calibrate dichotomous items, but extensions of these models later were developed for items with polytomous responses. Table 1 categorizes several IRT models by their item response format and by the number of item parameters estimated. Dodd, De Ayala, and Koch (1995) present a detailed description and discussion of polytomous IRT models used in the context of CAT.

The selection of an IRT model should be supported by careful consideration of the measurement application. If a new measure is being developed, the Rasch model often may be the better choice because of its unique and desirable measurement properties (e.g., raw score is a sufficient statistic for estimating trait level). Also, because there are fewer item parameters in a Rasch model, stable parameter estimation may be achieved with smaller sample sizes. Before settling on a Rasch model for a scale's calibration, however, the scale developer should verify that the selection of homogenously discriminating items has not deleteriously impacted the content coverage of the items. If adequate content coverage results in substantial variation in the discrimination of the items, a two-parameter IRT model should be chosen.

STARTING RULES

A CAT must start somewhere. One starting rule is to present the same first item for every respondent. If, as is most often the case, we assume that nothing

Table 1: Selected Item Response Theory (IRT) Models Classified by Item Parameter and Item Response Format

	Item Difficulty Modeled (Rasch Models)	*Item Difficulty and Discrimination Modeled*	*Item Difficulty, Discrimination, and Guessing Modeled*
Dichotomous responses (e.g., yes/no)	One parameter-logistic model (Rasch G)	Two-parameter-logistic model (Birnbaum; Hambleton and Swaminathan)	Three-parameter logistic (Hambleton and Swaminathan)
Polytomous responses (e.g., never/ sometimes/ always)	Andrich's rating scale model (Andrich) Partial credit model (Masters GN)	Graded response model (Samejima) Generalized partial credit model (Muraki)	None

is known about the respondent's trait level before the CAT is administered, the first item presented will be one of medium difficulty. An option to administering the same item to each patient is to administer an item randomly chosen from a set of items all of which are of medium difficulty. An alternative is to select the difficulty of the starting item based on prior information (e.g., disease severity, score on previously administered outcome measure).

Stopping Rules

With traditional measures, assessment stops when the respondent completes all items. With CAT measures, participants respond only to a subset of the items in the item bank, and a stopping rule must be identified. Some stopping rules specify the number of items to be administered, e.g., stop assessment once patient responds to 10 items. An advantage to stopping after a specified number of items is that response burden is constant across the study sample. If in a study, participants are asked to respond to numerous measures, a stopping rule of five items per measure might be selected to standardize and minimize total response burden. In other situations, a stopping rule of 10, 20, or more items could be specified.

An alternative stopping rule is to cease once a specified SEM is reached. The SEM quantifies the variance in estimated trait level that would be expected if a measure was administered repeatedly to an individual, without the individual remembering his or her responses to previous administrations. The better the estimate of patients' trait level, the smaller the SEM (Anastasi 1988; Beckerman et al. 1996; Kramer and Ng 1996). Based on the probabilities associated with a normal distribution, SEMs can be used to draw confidence intervals around trait-level estimates. The 95 percent confidence interval, for example, is the observed score ± 2 SEM.

The characteristic of IRT models that supports an SEM-based stopping rule is their estimation of standard errors conditioned on trait level. That is, a distinct SEM is estimated for each point along the measurement continuum. This contrasts with CTT in which measurement error estimates summarize a measure's reliability across all levels of the trait being measured. CTTs averaging strategy can be problematic, since measures tend not to be equally reliable across trait levels (Cook, Gartsman, and Roddey 2001). Often a measure best assesses trait level in the middle of the measurement continuum. In CAT assessments, SEMs are updated for each person after each response.

Typically, the magnitude of the SEM diminishes as more items are administered, and the CAT algorithm hones in on an estimate of a respondent's trait level (Wainer 1990).

THE UNIDIMENSIONALITY ASSUMPTION

As stated earlier, the most commonly used IRT models assume one latent construct explains how patients respond to items (Hambleton and Swaminathan 1985). In their introductory text on IRT, Hambleton and Swaminathan (1985) discuss several statistical strategies for evaluating unidimensionality. It is doubtful that a measure is ever perfectly unidimensional (Reckase 1985; Harrison 1986), and "essential unidimensionality" has been accepted as a more practical criterion, meaning that idiosyncratic, methodological, or trivial dimensions are ignored (Stout 1987, 1990; Junker 1991).

There is no consensus regarding what constitutes essential unidimensionality, and, within the context of health outcomes research, these issues seldom have been investigated. The practical challenge is to identify the impact of various degrees of multidimensionality on IRT calibrations. Psychometric researchers in education and psychology have found multidimensionality can deleteriously impact IRT calibrations (Nandakumar 1994). The impact of inaccurate calibrations is particularly problematic in CAT assessments. Folk and Green simulated two-dimensional data and compared score estimates based on adaptive and nonadaptive assessments (Folk and Green 1989). They found that multidimensionality caused greater problems in adaptive assessments. In their data, estimated scores based on nonadaptive assessment represented a composite of the two dimensions measured. In the CAT assessments, scores were closely related to one dimension or the other, not both.

In recent years, there has been increased interest among outcomes researchers in cocalibrating, in a single IRT run, the items from two or more measures (Fisher et al. 1995; Bjorner, Kosinski, and Ware 2003). When IRT assumptions are met, such cocalibrations allow scores of two or more scales to be compared directly using a common mathematical metric. Multidimensionality in the data, however, can substantially impact the credibility of the results.

It is not clear how much multidimensionality is tolerated by unidimensional IRT models in general and CAT applications in particular. Calibration of a pool of activities of daily living and instrumental activities of daily living items might be relatively unaffected by mild multidimensionality. A unidimensional IRT calibration of pain, social function, and fatigue items, however, would more likely have significant multidimensionality and a resulting degradation in the trustworthiness of the scores. A CAT assessment based on such a calibration would be even more suspect.

Defining and confirming unidimensionality requires both empirical and theoretical work. Should pain be considered a single dimension or a tridimensional composite of pain frequency, intensity, and duration? Are upper body pain and lower body pain different dimensions, or are they components of a single dimension? Conceptually, the construct should "hold together" as a unitary trait in the context of a specified theory. As in all measurement contexts, interpretation of empirical findings must be made in the context of theory, reason, and logic.

The Future of CAT

CAT has been applied frequently in medical licensure testing (Fields 1992; Ruiz et al. 1995; Bergstrom 1996), but infrequently in the assessment of health outcomes (Ware, Bjorner, and Kosinski 2000; Gardner, Kelleher, and Pajer 2002). We predict the next several years will witness an impressive increase in the number of CATs developed for measuring patient outcomes. However, formidable challenges must be met for the quality of CATs to keep pace with this expected increase in the quantity of CATs. It is likely that software will be developed that makes it relatively simple to move from a data set of items and item responses to an fully operational CAT. Unfortunately, user-friendly software is also ignorance friendly. The onus will be on those who develop CATs to ensure that IRTs assumptions are met and that there is good fit to the chosen IRT model; the software will not make this distinction. The quality of CAT depends on the properties of the item bank. To be effective, CAT item banks must be of adequate size and breadth so that floor and ceiling effects are avoided, and they should be free of differential item function in relevant subpopulations.

Disadvantages of CAT

A significant barrier to implementing CAT-based assessment is the relatively high start-up costs. Even though individual patients do not complete a large number of items, the item bank should be large. If the CAT bank is substantial, the algorithm can be "choosy" in selecting items to present individual respondents. Items in the bank should cover all levels of the construct that occur in the population that will be measured. A CAT-based shoulder function measure would need to include items targeting low levels of function (e.g., ability to move the arm), high levels of function (e.g., ability to throw a baseball with speed and accuracy), and all levels in between. Developing a wide-ranging and large set of items is both time-consuming and resource demanding.

Another substantial expense is the requisite computer programming. Most CATs are developed "in house." The expense of programming, therefore, is incurred every time a measure is developed. If software existed that generalized the process of developing CATs, costs would drop precipitously. There are a number of software programs that calibrate item pools using various IRT models. What is needed is software that takes the next step and, using calibrated items as input, constructs CAT measures. At the time of this writing, no such software existed. However, in 2004, the National Cancer Institute issued a call under the National Institutes of Health Small Business Innovation Research (SBIR) Program (SBIR 211: Developing Item Response Theory Software for Outcomes and Behavioral Measurement). The objective of the initiative was the development of user-friendly software that incorporated IRT and classical measurement approaches. CAT development was one of several recommended extensions of the proposed software. The software developed under this contract promises to substantially improve capacity for applying IRT models, perhaps including the development of CATs.

Another expense in implementing CAT is the "delivery device." Because CATs are administered by computer, either respondents or an interviewer must interface with a computer. The computer could be a dedicated desktop, a laptop, a tablet personal computer, a kiosk with one or more stations, or a personal digital assistant (PDA). In addition to being expensive, these interfaces have other drawbacks. If desktop computers are used, a lockable room or rooms must be dedicated to measuring the outcome(s). Kiosks setup in public areas consumes substantial floor space and have limited portability. PDAs and laptops, on the other hand, are so portable that risk of theft accompanies their use.

The necessity of interfacing with a computer could be off-putting to patients unfamiliar or uncomfortable with computer technology. Studies of patients' reactions to computer-based outcomes assessment, however, have been positive, even among patients without prior experience with a computer (Buxton, White, and Osoba 1998; Sutherland et al. 2001; Hahn et al. 2003, 2004). Further, with the current and widespread use of computers in homes and in public arenas, the number of people who feel intimidated by computer interfaces has diminished. However, research is needed in subpopulations such as the veteran population to assess the extent to which a computer interface is a barrier to patient reported outcomes.

Persons contemplating the development of a CAT should weigh the considerable costs against the benefits to be gained with CAT-based assessments. The development process requires a team of experts including stakeholders, experts in the content area, psychometricians with training in IRT, and computer programmers with expertise in software development. This level of expertise is seldom available within a single organization and, therefore, often will require support from industry or government.

DISCUSSION

In the current outcomes-aware milieu, researchers, clinicians, and patients all benefit from the development of psychometrically rigorous measures that are not burdensome. Outcome measures developed to be administered adaptively hold great promise. As is the case for any outcome measure, the context in which an instrument will be used must be considered. CATs will not be appropriate with every population and in every setting, and their advantages should be weighed against their several drawbacks. Nevertheless, the unique properties and considerable advantages of CAT measures make it likely that their use will increase substantially in coming years.

ACKNOWLEDGMENTS

The research reported here was supported by the Department of Veterans Affairs, Veterans Health Administration, Health Services Research and Development Service, Measurement Excellence and Training Resource Information Center (METRIC; RES 02-235). Dr. Cook is a METRIC core investigator at the Houston Center for Quality of Care & Utilization Studies and Senior Research Scientist at the University of Washington, Department of Rehabilitation Medicine.

REFERENCES

Agha, Z., R. P. Lofgren, J. V. VanRuiswyk, and P. M. Layde. 2000. "Are Patients at Veterans Affairs Medical Centers Sicker? A Comparative Analysis of Health Status and Medical Resource Use." *Archives of Internal Medicine* 160 (21): 3252–7.

Anastasi, A. 1988. *Psychological Testing.* New York: Macmillan Publishing Company.

Andrich, D. A. 1978. "A Rating Formulation for Ordered Response Categories." *Psychometrika* 43: 561–73.

Baker, J. G., J. B. Rounds, and M. A. Zevon. 2000. "A Comparison of Graded Response and Rasch Partial Credit Models with Subjective Well-Being." *Journal of Educational and Behavioral Statistics* 25: 253–70.

Beckerman, H., T. W. Vogelaar, G. J. Lankhorst, and A. L. Verbeek. 1996. "A Criterion for Stability of the Motor Function of the Lower Extremity in Stroke Patients Using the Fugl–Meyer Assessment Scale." *Scandinavian Journal of Rehabilitation Medicine* 28: 3–7.

Bergstrom, B. A. 1996. "Computerized Adaptive Testing for the National Certification Examination." *Journal of American Association of Nurse Anesthetists* 64: 119–24.

Birnbaum, A. 1968. "Some Latent Trait Models and Their Use in Inferring an Examinee's Ability." In *Statistical Theories of Mental Test Scores,* edited by F. M. Lord and M. R. Novick. Reading, MA: Addison-Wesley.

Bjorner, J. B., M. Kosinski, and J. E. Ware Jr. 2003. "Using Item Response Theory to Calibrate the Headache Impact Test (HIT) to the Metric of Traditional Headache Scales." *Quality of Life Research* 12 (8): 981–1002.

Bowers, D. R. 1991. "Computer-Based Adaptive Testing in Music Research and Instruction." *Psychomusicology* 10: 49–63.

Buxton, J., M. White, and D. Osoba. 1998. "Patients' Experiences Using a Computerized Program with a Touch-Sensitive Video Monitor for the Assessment of Health-Related Quality of Life." *Quality of Life Research* 7 (6): 513–9.

Cella, D. F., S. R. Llyod, and B. D. Wright. 1996. "Cross-Cultural Instrument Equating: Current Research and Future Directions." In *Quality of Life and Pharmacoeconomics in Clinical Trials,* 2nd ed., edited by B. Spilker, pp. 705–15. Philadelphia: Lippincott-Raven Publishers.

Cook, K. F., G. M. Gartsman, T. S. Roddey, and S. L. Olson. 2001. "The Measurement Level and Trait-Specific Reliability of 4 Scales of Shoulder Functioning: An Empiric Investigation." *Archives of Physical Medicine and Rehabilitation* 82: 1558–65.

Cook, K. F., T. S. Roddey, G. M. Gartsman, and S. Olson. 2003. "Development and Psychometric Evaluation of the Flexilevel Scale of Shoulder Function (FLEX-SF)." *Medical Care* 41 (7): 823–35.

Dodd, B. G., R. J. De Ayala, and W. R. Koch. 1995. "Computerized Adaptive Testing with Polytomous Items." *Applied Psychological Measurement* 19: 5–22.

Fields, F. A. 1992. "Computerized Adaptive Testing for NCLEX-PN." *Journal of Practical Nursing* 42: 8–10.

Fisher, W. P. Jr., R. F. Harvey, P. Taylor, K. M. Kilgore, and C. K. Kelly. 1995. "Rehabits: A Common Language of Functional Assessment." *Archives of Physical Medicine and Rehabilitation* 76 (2): 113–22.

Folk, V. G., and B. F. Green. 1989. "Adaptive Estimation When the Unidimensionality Assumption of IRT Is Violated." *Applied Psychological Measurement* 13: 373–89.

Gardner, W., K. J. Kelleher, and K. A. Pajer. 2002. "Multidimensional Adaptive Testing for Mental Health Problems in Primary Care." *Medical Care* 40: 812–23.

Hahn, E. A., D. Cella, D. G. Dobrez, G. Shiomoto, S. G. Taylor, A. G. Galvez, P. Diaz, V. Valenzuela, H. L. Chiang, S. Khan, S. A. Hudgens, and H. Du. 2003. "Quality of Life Assessment for Low Literacy Latinos: A New Multimedia Program for Self-Administration." *Journal of Oncology Management* 12 (5): 9–12.

Hahn, E. A., D. Cella, D. Dobrez, G. Shiomoto, E. Marcus, S. G. Taylor, M. Vohra, C. G. Chang, B. D. Wright, J. M. Linacre, B. D. Weiss, V. Valenzuela, H. L. Chiang, and K. Webster. 2004. "The Talking Touchscreen: A New Approach to Outcomes Assessment in Low Literacy." *Psycho-Oncology* 13: 86–95.

Hambleton, R. K., and H. Swaminathan. 1985. *Item Response Theory: Principles and Applications.* Norwell, MA: Kluwer Academic Publishers.

Harrison, D. A. 1986. "Robustness of IRT Parameter Estimation to Violations of the Unidimensionality Assumption." *Journal of Educational Statistics* 11: 91–115.

Junker, B. W. 1991. "Essential Independence and Likelihood-Based Ability Estimation for Polytomous Items." *Psychometrika* 56: 255–78.

Kirisci, L., H. B. Moss, and R. E. Tarter. 1996. "Psychometric Evaluation of the Situational Confidence Questionnaire in Adolescents: Fitting a Graded Item Response Model." *Addictive Behaviors* 21 (3): 303–17.

Kirisci, L., R. E. Tarter, and T. C. Hsu. 1994. "Fitting a Two-Parameter Logistic Item Response Model to Clarify the Psychometric Properties of the Drug Use Screening Inventory for Adolescent Alcohol and Drug Abusers." *Alcoholism: Clinical and Experimental Research* 18 (6): 1335–41.

Kramer, J. F., and L. R. Ng. 1996. "Static and Dynamic Strength of the Shoulder Rotators in Healthy, 45-to 75-Year-Old Men and Women." *Journal of Orthopedic and Sports Physical Therapy* 24: 11–8.

Merriam-Webster Inc. 2003. "Merriam-Webster Online: The Language Center" [accessed June 6, 2003]. Available at http://www.m-w.com/home.htm

Lippitt, S. B., D. T. Harryman, and F. A. Matsen. 1993. "A Practical Tool for Evaluating Function: The Simple Shoulder Test." In *The Shoulder: A Balance of Mobility and Stability*, edited by F. A. Matsen, F. H. Fu, and R. J. Hawkins, pp. 501–30. Rosemont, IL: The American Academy of Orthopedic Surgeons.

Masters, G. N. 1982. "A Rasch Model for Partial Credit Scoring." *Psychometrika* 47: 149–74.

McHorney, C. A., and A. S. Cohen. 2000. "Equating Health Status Measures with Item Response Theory: Illustrations with Functional Status Items." *Medical Care* 38 (suppl): II43–59.

Muraki, E. 1992. "A Generalized Partial Credit Model: Application of an EM Algorithm." *Applied Psychological Measurement* 16: 159–76.

Nandakumar, R. 1994. "Assessing Dimensionality of a Set of Item Responses—Comparison of Different Approaches." *Journal of Educational Measurement* 31: 17–35.

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

Reckase, M. D. 1985. "The Difficulty of Test Items That Measure More Than One Ability." *Applied Psychological Measurement* 9: 401–12.

Reise, S. P., and J. M. Henson. 2000. "Computerization and Adaptive Administration of the NEO PI-R." *Assessment* 7: 347–64.

Ruiz, B., P. A. Fitz, C. Lewis, and C. Reidy. 1995. "Computer-Adaptive Testing: A New Breed of Assessment." *Journal of the American Dietetic Association* 95: 1326–7.

Samejima, F. 1969. "Estimation of Latent Ability Using a Response Pattern of Graded Scores." *Psychometrika* 17 (monograph supplement).

Stout, W. F. 1987. "A Nonparametric Approach for Assessing Latent Trait Unidimensionality." *Psychometrika* 52: 589–617.

———. 1990. "A New Item Response Theory Modeling Approach with Applications to Unidimensionality Assessment and Ability Estimation." *Psychometrika* 55: 293–325.

Sutherland, L. A., M. Campbell, K. Ornstein, B. Wildemuth, and D. Lobach. 2001. "Development of an Adaptive Multimedia Program to Collect Patient Health Data." *American Journal of Preventive Medicine* 21 (4): 320–4.

Velozo, C. A., and E. W. Peterson. 2001. "Developing Meaningful Fear of Falling Measures for Community Dwelling Elderly." *American Journal of Physical Medicine and Rehabilitation* 80: 662–73.

Vispoel, W. P. 1993. "Computerized Adaptive and Fixed-Item Versions of the ITED Vocabulary Subtest." *Educational and Psychological Measurement* 53: 779–89.

Vispoel, W. P., T. Wang, and T. Bleiler. 1997. "Computerized Adaptive and Fixed-Item Testing of Music Listening Skill: A Comparison of Efficiency, Precision, and Concurrent Validity." *Journal of Educational Measurement* 34: 43–63.

Wainer, H. 1990. *Computerized Adaptive Testing: A Primer.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Waller, N. G., and S. P. Reise. 1989. "Computerized Adaptive Personality Assessment: An Illustration with the Absorption Scale." *Journal of Personality & Social Psychology* 57: 1051–8.

Ware, J. E. Jr., J. B. Bjorner, and M. Kosinski. 2000. "Practical Implications of Item Response Theory and Computerized Adaptive Testing: A Brief Summary of Ongoing Studies of Widely Used Headache Impact Scales." *Medical Care* 38 (suppl): II73–82.

Weiss, D. J. 2003. "Polytomous CAT." [Written Communication March 12, 2003.]

Wolfe, E. W. 2000. "Equating and Item Banking with the Rasch Model." *Journal of Applied Measurement* 1: 409–34.

Wright, B. D., and G. N. Masters. 1982. *Rating Scale Analysis.* Chicago: Mesa Press.