Published in final edited form as:
Qual Life Res. 2009 May ; 18(4): 461–471. doi:10.1007/s11136-009-9463-5.

Replenishing a computerized adaptive test of patient-reported daily activity functioning

Stephen M. Haley, **Pengsheng Ni**, and **Alan M. Jette**
Boston University School of Public Health, Boston, MA, USA

Wei Tao
ACT, Inc., Iowa City, IA, USA

Richard Moed
CREcare, LLC, Gilford, NH, USA

Doug Meyers
Meyers Healthcare Solutions, Apopka, FL, USA

Larry H. Ludlow
Department of Educational Research, Measurement and Evaluation, Boston College, Boston, MA, USA

Abstract

Purpose—Computerized adaptive testing (CAT) item banks may need to be updated, but before new items can be added, they must be linked to the previous CAT. The purpose of this study was to evaluate 41 pretest items prior to including them into an operational CAT.

Methods—We recruited 6,882 patients with spine, lower extremity, upper extremity, and nonorthopedic impairments who received outpatient rehabilitation in one of 147 clinics across 13 states of the USA. Forty-one new Daily Activity (DA) items were administered along with the Activity Measure for Post-Acute Care Daily Activity CAT (DA-CAT-1) in five separate waves. We compared the scoring consistency with the full item bank, test information function (TIF), person standard errors (SEs), and content range of the DA-CAT-1 to the new CAT (DA-CAT-2) with the pretest items by real data simulations.

Results—We retained 29 of the 41 pretest items. Scores from the DA-CAT-2 were more consistent (ICC = 0.90 versus 0.96) than DA-CAT-1 when compared with the full item bank. TIF and person SEs were improved for persons with higher levels of DA functioning, and ceiling effects were reduced from 16.1% to 6.1%.

Conclusions—Item response theory and online calibration methods were valuable in improving the DA-CAT.

Keywords

Outcomes assessment; Quality of life; Item response theory; Activities of daily living (ADL)

smhaley@bu.edu.

Introduction

Computerized adaptive testing (CAT) is increasingly being proposed for use in routine functional and health-related quality-of-life assessments in outpatient rehabilitation programs [1–4], and is considered the new wave of the future in patient-reported outcome (PRO) assessments [5–10]. CAT employs a simple form of artificial intelligence that selects questions tailored to the patient, shortens or lengthens the test to achieve the desired precision, and scores everyone on a standard metric. Using CAT, respondents are administered items that provide the most information at the current score estimate on a particular PRO domain. Only enough items are administered in order to satisfy preset precision rules, or alternately, a maximum number of items is determined in advance. CAT platforms require the development of a comprehensive and calibrated set of items (item banks) that define each underlying PRO dimension [11].

An item bank is a collection of items that represents a range of performance or difficulty levels for a particular PRO domain [12–14]. Item banks are developed by linking outcome items from different sources, or adding new items so that they can be meaningfully compared together on a common underlying metric. Item response theory (IRT) methods can be used to calibrate items onto a unidimensional scale [9, 15].

Recently, we reported on the prospective use of a functional outcome CAT in a series of 1,815 rehabilitation outpatients who were administered the Activity Measure for Post-Acute Care (AM-PAC) at admission and discharge [4]. In this present study, we wanted to determine if the CAT programs covered the content necessary for a typical out-patient sample and if the psychometric properties of the CAT could be improved by adding additional items to the item bank. By demonstrating that the CAT could be improved incrementally by adding new items to the bank, we could illustrate how to improve systematically the performance of the CAT in a dynamic outpatient rehabilitation setting.

For outpatient programs, we used two of the three AM-PAC domains, namely Basic Mobility and Daily Activities. A third AM-PAC domain (Applied Cognition) was not included because cognitive deficits were infrequent in patients receiving outpatient rehabilitation services. Based on earlier factor analytic and IRT analyses, Basic Mobility and Daily Activity scale domains were identified as unidimensional and distinct [16]. The Basic Mobility domain contains 120 items of essential functional activities such as bending, walking, carrying, and climbing stairs. The item bank for the Daily Activity domain includes 65 distinct personal care, dressing, meal, and instrumental activities of daily living tasks. The content and conceptual approach towards building the Daily Activity item banks have been discussed elsewhere [17].

The Basic Mobility and Daily Activity item banks were developed from a sample of 1,041 post-acute care patients who were actively receiving inpatient and community-based rehabilitation services at the time of assessment [18]. Patients were recruited from one of four post-acute care settings with rehabilitation outpatient services comprising approximately 25% ($n = 237$) of the sample. The remainder of the sample was recruited from inpatient and home-care facilities. Details of the full sampling plan have been published elsewhere [18]. The original post-acute care sample included three major patient groups: (1) 33.2% neurological (e.g., stroke, multiple sclerosis, Parkinson's disease, brain injury, spinal cord injury, neuropathy); (2) 28.4% musculo-skeletal (e.g., fractures, joint replacements, orthopedic surgery, joint or muscular pain); and (3) 38.4% medically complex (e.g., debility resulting from illness, cardiopulmonary conditions, or postsurgical recovery).

Although the findings of use of the AM-PAC-CATs in outpatient rehabilitation programs were generally positive [4], a substantial ceiling effect (approximately 16%) was noted for

the Daily Activity domain. This effect was not entirely unexpected, since the calibration sample included only about a quarter of patients who were from an outpatient rehabilitation setting. Thus, in this study, we sought to improve the scale properties of the Daily Activity domain, and specifically to see if we could reduce ceiling effects. We report only on the changes made in the Daily Activity CAT in this article, even though some Basic Mobility pretest items were also collected from the sample. Thus, the purpose of this article is to present an efficient procedure for rapidly improving the content coverage and psychometric characteristics of an operational CAT using the PRO domain of Daily Activities to illustrate this process.

Methods

Rather than recalibrating the entire Daily Activity item bank, or developing a totally new version of an instrument, the dynamic nature of CAT assessments allowed us to add new items that could be integrated into the scoring metric used in previous versions of the Daily Activity CAT. We seeded new items into the operational Daily Activity CAT assessment after the operational CAT was administered. These new items were not included in the scoring estimate of the operational CAT at the point of service, but were retained for future IRT analyses. Before new items (pretest) could be included into the updated Daily Activity CAT, they had to be calibrated and linked to the original IRT scale underlying the CAT.

We used a real-data simulation approach to compare the performance characteristics of the operational CAT (DA-CAT-1) to the expanded item bank in the CAT (DA-CAT-2) with the pretest items that were retained [19]. Real-data simulation (computer simulation) is based on the actual computer adaptive testing and pretest response data. As items were selected for administration in the simulation, responses were taken from the actual data set. The validity of this real-data simulation approach for studying CAT estimated scores assumes that persons respond in much the same way to items regardless of their context; that is, items that precede or follow, or short versus long forms, would not impact on a person's responses to items.

Subjects

We recruited 6,682 patients who were recently admitted to an outpatient rehabilitation program in 147 outpatient clinics across 13 states of the USA that were operated by Select Physical Therapy and NovaCare, a division of clinics owned and operated by Select Medical Corporation. Patients were mainly seeking treatment for orthopedic or sports injuries. Because we are only reporting on the Daily Activity pretest items in this article, most of the subjects had upper extremity impairments, spine or other conditions; few subjects are included here with only lower extremity impairments. Spine impairments included impairments of the cervical, thoracic or lumbosacral region of the spine. Upper extremity impairments included conditions of the shoulder, elbow, hand, and wrist. Lower extremity impairments were conditions of the hip, knee, foot, and ankle. Other conditions included neurological, medical, and unspecified major traumatic impairments. See Table 1 for full demographics of the sample. The Institutional Review Board at Boston University Medical Center approved all data transfer procedures to protect the identity of individual subjects.

Daily Activity item bank construction

The 65-item Daily Activity item bank was used in the original Daily Activity CAT. Items were phrased, “How much difficulty do you currently have (without help from another person or device) with the following activities …?” A polytomous response choice included “none,” “a little,” “a lot,” and “unable.” We framed the activity questions in a general fashion without specific attribution to health, medical conditions or disabling factors. Details

of the stop rules, person estimation, and content balancing routines for the DA-CAT-1 are provided in Jette et al. [4]. Based on input from clinical experts, we developed 41 new Daily Activity items to be incorporated into the DA-CAT-2. The content items were primarily in the areas of work or sports medicine, more complex upper extremity tasks, and daily tasks that required neck, shoulder, and trunk movements: content areas that were not adequately covered by the original item bank. The initial two waves included some Basic Mobility items (not reported here), thus a full set of ten new Daily Activity items were not tested. The third wave included mainly instrumental activities of daily living (ADL) items. The fourth wave was comprised of items that were oriented to shoulder conditions. The final wave included items specifically directed to patients with cervical conditions. The new items were intended to plug content gaps and extend the content range of the original item bank, to improve applicability to a general rehabilitation outpatient program, and to enrich functional content for patients with shoulder and cervical conditions.

Data collection procedures

Subjects completed the self-report DA-CAT-1 on a tablet computer provided to them in the waiting room prior to their initial outpatient visit. After the DA-CAT-1 was completed, patients were asked if they would answer up to ten additional questions to help the clinic improve its outcome assessments in the future. Only 1% of the patients refused to complete the supplementary items.

Because of the time demands for data collection in a busy clinic environment, not every patient answered questions on the full set of 41 pretest items, as this was felt to create a burden for both patients and staff. We conducted the item calibration work in five waves, in which cohorts of patients (without overlap) were asked to respond to no more than ten new items after they completed the operational Daily Activity CAT assessment. The pretest items were tested for inclusion in the Daily Activity item bank for subsequent use in the revised CAT (DA-CAT-2). See Table 2 for the sample sizes per pretest wave and the Daily Activity items that were part of each wave. Subject demographic information, surgical status, and major impairment were all available from administrative data collected routinely by each outpatient clinic, and combined with the CAT scores. An office staff member was available to the subjects during the administration process to answer any questions.

Analyses

We conducted the analyses in three phases: (1) we examined the fit of pretest items into the original Daily Activity item bank; (2) using real-data simulations (see “Computer simulation studies” section below) we directly compared the psychometrics of the original CAT version (DA-CAT-1) to the new CAT version with the pretest items (DA-CAT-2), as we were particularly interested in testing new items that might increase the ceiling of the Daily Activity item bank; and (3) we applied a series of criteria [scoring consistency with full item pool, test information function (TIF), person standard errors (SE), and content range, including ceiling effects] to determine if the pretest Daily Activity items embedded into the revised CAT (DA-CAT-2) improved the operational CAT (DA-CAT-1).

Item parameters and fit—In order to determine if an item was to be included in the next version of the CAT (DA-CAT-2), we examined unidimensionality and local independence, item fit, and differential item functioning (DIF). We tested the latent structure of the Daily Activity items for each separate wave in a series of confirmatory factor analyses (CFA) and evaluated item loadings and residual correlations between items using MPlus software [20]. Due to missing data, we were unable to conduct a full CFA across all pretest items, and performed the CFA at each wave. We used weighted least-square methods for factor analysis of categorical data because traditional factor analysis could overestimate the

number of factors and underestimate the factor loadings when analyzing skewed categorical data. As criteria, we used traditional fit statistics CFI, TLI values greater than 0.9, and RMSEA less than 0.1. To assess potential scale drift, we examined the correlations of the person scores from the full item bank to the CAT scores at each wave, and the original person score distribution mean and its change with respect to the new items added in the final CATs.

We applied Stone's approach to assess item fit [21–23]. Rather than using the point estimate of the person score, as is the practice in many fit statistics, information from the full posterior distribution of the person score was used to construct the fit statistics. The continuous θ scale is approximated by a set of discrete points, then the posterior probabilities at given discrete θ can be estimated. The pseudocounts of the number of persons at each ability group are then calculated. Then the Pearson chi-square goodness-of-fit statistic comparing the pseudo-observed and expected score distribution are calculated. The *P*-value is calculated from the comparison between the actual statistics and a simulated statistics distribution. The simulated person response data were based on the person ability and item parameters estimated from the actual data. The actual statistics are then compared with the simulated statistics distribution based on 100 simulated samples. Statistical significance was set at $\alpha < 0.05$.

DIF was assessed using logistic regression, with the dependent variable as the item score, and the independent variables were the background variables (such as age, gender, etc.), the ability level (total test score), and the background variable and ability interaction [24]. In a DIF study, if the background effect is significant and the interaction effect is not, then the item has uniform DIF; on the other hand, if the interaction effect is significant, the item has nonuniform DIF. The analytic strategy was to successively add ability level, background variables, and interaction terms into the model. The model comparison is based on the likelihood ratio test. We used Bonferroni corrected *P*-values for significance testing. Pseudo-R^2 change was used to quantify the effect size of both uniform and nonuniform DIF. Zumbo [24] proposed a cutoff value of 0.13 for the R^2 change. This corresponds to a medium effect size [25] where small, medium, and large effect sizes are 0.02, 0.13, and 0.26, respectively. Local dependence was determined by calculating the residual correlation between item pairs after controlling for the trait estimate. For each person, the expected value of each item is calculated, and then the residual is calculated as the difference between the observed value and the expected value for each item. The residual correlation is the correlation between item pairs across all the subjects. Local dependence was defined as the residual correlation between item pairs greater than 0.2. This is a common procedure and criterion for screening out local dependence [26–28].

Because we calibrated items based on an existing CAT, which uses a small number of items per assessment, and new pretest items that require item parameter estimates equated to the existing CAT, we used an online pretest calibration method. Numerous methods have been proposed for online calibrations [29, 30]. They differ in the number of parameters estimated, their estimation procedures and the number of estimation cycles, their functional form, and whether ability or item estimates are fixed during various stages of the procedure. For our study, we employed Stocking's method A [31]. This was an appropriate choice for this project since the person scores were estimated from the CAT assessments, often restricted to seven items or fewer. We also performed the “one EM cycle” approach of Wainer and Mislevy [32]. The results were similar to Stocking's method A in the first wave analysis, thus we chose to use Stocking's method A because of its greater simplicity and efficiency. We experimented with using the multiple EM cycle, but chose not to use this method because some item discrimination parameters appeared overly inflated and because a small

number of poorly discriminating items may markedly affect the item calibrations, especially the posterior distribution.

The IRT modeling of the items was conducted using the generalized partial credit model (GPCM) [33]. The GPCM is an extension to polytomous items of the two-parameter logistic model for dichotomous items. In the model each item has a slope parameter describing the item's ability to discriminate between subjects with different levels of ability and a set of threshold parameters describing the “difficulty” of the item [34]. Actually, the IRT models are formally equivalent to nonlinear mixed models; for instance, the GPCM model could be considered the random-effects adjacent-categories logit model [35].

The GPCM uses two parameters—item category/boundary locations and discrimination—in estimating item locations and person scores. In the GPCM, there is no equality constraint on the threshold parameters, thus each item can have unique category characteristic curves. The model estimates item difficulty parameters as logit, equal-interval logarithm-transformed units of measurement [36]. We used the GPCM in this study since our original Daily Activity calibrations were based on GPCM, and thus we continued to use this IRT model in the recalibration with the pretest items.

Starting with the GPCM, Stocking's method A fixes the ability estimates computed from the operational CAT (DA-CAT-1) item responses. The fixed ability estimates were based on the existing items in the bank which allowed us to calibrate the new items. These fixed ability estimates then produced the new item parameters (slope and location) for the pretest items. This simple fixing of the ability estimates results in pretest item calibrations on the same scale as the original operational item estimates.

Computer simulation studies—Once we decided which pretest items to retain and had obtained our final item calibrations, we conducted a series of real-data CAT simulations with the DA-CAT-2 and DA-CAT-1 and compared results. In the DA-CAT-2, we used the new item bank with the chosen pretest items to simulate a new CAT program. Responses are fed into the CAT program based on how persons answered items from the previously completed functional assessment. In the present research, we used an empirical simulation approach for investigating the merits of the new pretest items by using the complete set of the actual item responses of patients reporting about their daily activity functioning. As items were selected for administration in the DA-CAT-2, responses were extracted from the patient's actual dataset. For the CAT simulations, we used identical procedures as for the operational CAT, including a stop rule of no more than seven items, and a content balancing algorithm that allowed items to be selected based on both content specifications and maximum information function for the first four items of the Daily Activity scale. That is, the item with the maximum information at the current score level within one of four separate content areas was chosen. After the first four items, the selection of the rest of the items was based solely on maximizing the information value at the current score level [4]. The CAT was limited by a seven-item stop rule. We based the CAT algorithms used in this study on software developed at the Health and Disability Research Institute, HDRI™, Boston University.

Impact of pretest items (DA-CAT-2 versus DA-CAT-1)—We examined the impact of the pretest items by studying patient scoring differences between DA-CAT-1 and DA-CAT-2. We used four criteria: (1) scoring consistency with full item pool, (2) TIF, (3) precision using average SEs for person scores, and (4) content range. We defined consistency as the level of correspondence [using an intraclass correlation coefficient (ICC)] between either the CAT-based DA-CAT-1 or the DA-CAT-2 and the IRT criterion score (best possible score estimate based on the full item bank). We defined precision as the

average SE associated with person scores [37]. TIF was defined as a summary of information provided by individual items in the instrument and it identifies where along an underlying scale items have their best level of discrimination and precision [38]. TIF values are closely related to the SE of the person ability estimates for Daily Activities. Specifically, the SE of the person ability estimate of Daily Activity is the inverse of the square root of the negative second derivative of the log likelihood (observed information), which is the same as Fisher's expected information in GPCM [39].

The location on the Daily Activity scale where the TIF curve peaks indicates the portion on the scale best measured by that instrument. When the TIF is peaked at or around the same range on the scale as the peak of the patients' ability distribution, the instrument is appropriate for the ability levels of the population being measured. The content range of the DA-CAT-1 and DA-CAT-2 was based on estimated locations of the item-response categories that represent the lowest and highest level of ability [40]. The term “content range” refers to the breadth of the functional levels captured by the items, and does not imply an assessment of content validity. The ceiling effect was the point at which score estimates exceeded the highest estimated item-response category (a four-point rating scale was used, so each item had three category threshold parameters). The floor effect was likewise defined as the point at which Daily Activity measures fell beyond the lowest point of the estimated item-response category.

Estimated AM-PAC scores for each subject in the sample were converted to a transformed score, which is a simple linear translation that expressed scores as deviations from a measure of central tendency. In this study, we used a mean of 50 and a standard deviation of 10 (T-scale) [6]. By using norm-based scoring instead of the more traditional 0–100 scale, we are able to raise the ceiling or lower the floor of the scale in the future by adding and calibrating new items, and the placement (and scoring) of the item thresholds in relation to the average does not change.

Results

Item parameters

We eliminated 12 of the 41 pretest items due to combinations of lack of unidimensionality based on the CFA at each wave, poor item fit, local dependence or, in one case, because we could not get an item estimate to converge. Items eliminated include: “Doing a push-up” (item fit and local dependence), “Lifting 25 pounds from the ground to above shoulder height–,” “Lifting 5 pounds from a table to above shoulder height–” (local dependence), “Digging a hole in the ground with a shovel,” “Setting a watch,” “Flossing your teeth,” “Throwing a ball–,” “Sleeping on your `painful' shoulder,” “Sit in chair holding a book while reading,” “Checking the blind spot while driving a car,” “Holding head steady while driving a car in traffic on a local road” (unidimensionality and item fit), and “Coughing” (no converge). Table 3 lists the 29 pretest items retained. A few items had less than ideal fit, but were retained for their location on the scale or the relevance of their content to the Daily Activity scale. The overall level of item and person misfit was greater than the chance range (1–2 items) expected under the criterion of $\alpha < 0.05$. Since the item fits are based on the estimated scores from just seven items (CAT stop rule) the item fits may be biased, and therefore we chose at this time to maintain the full item bank and monitor the fit in subsequent analyses. None of the retained items showed any DIF by gender or age; some of the items had been calibrated on female samples only (make-up and use of hairdryer). We did not have available to us race or ethnic information, so we could not perform a DIF analysis on these variables. Correlations between CAT scores and the full item bank were consistently at the $r = 0.81$ level. The original person score distribution mean was 54.1 and standard deviation 10.6; the person score distribution after adding the new items was 55.3

and standard deviation 12.9. These means are very similar and, along with the correlation results, suggest that the CAT scores did not drift across successive waves.

Impact of pretest items (DA-CAT-2 versus DA-CAT-1)

The scoring consistency (comparison of person scores from full item pool to CAT) was improved when using the pretest items in DA-CAT-2 versus the original CAT without the pretest items. The ICC between DA-CAT-1 and the full set of items (best possible estimate) was 0.90, while the ICC between the full item pool and DA-CAT-2 was 0.96.

We built items to improve the entire Daily Activity scale, but primarily we focused on the more advanced performance levels of Daily Activity. We used TIF to see if we were able to shift greater information to the higher performance end of the scale. Figure 1 depicts the TIFs for the DA-CAT-1 and DA-CAT-2. Note that the TIF for DA-CAT-2 is shifted to the right (toward the higher-performance end of the Daily Activity scale) of the original TIF for DA-CAT-1. However, the magnitude of the TIF is dependent on the number of items, so it would be expected that the TIF for DA-CAT-2 would be greater, although not necessarily changed in position as we note in Fig. 1.

In Fig. 2, we have represented the improvement (decrease) in person score standard error (SE) when comparing DA-CAT-2 with DA-CAT-1. The decrease in standard error occurs primarily at the high-performance end of the scale. As depicted in Fig. 2, the major differences in SE take place at the higher-performance end of the Daily Activity scale. The effective content range of DA-CAT-2 was improved over DA-CAT-1 (Table 4). When scoring persons with the DA-CAT-2, we obtained about a 10% reduction of ceiling effects seen in DA-CAT-1 (16.11% versus 6.13%).

Discussion

One of the proposed advantages of the CAT methodology is to improve measurement of important PRO concepts dynamically, that is, to change or replace items iteratively as new samples of patients require more relevant assessments or as better items become available [41, 42]. This process is completed regularly in educational performance tests, in which sample items are routinely given to students so that new items can be evaluated for inclusion in future versions. These updates are also needed in health care applications, although many of the conditions for testing new items in a busy clinical environment are not as amenable to collecting pretest items as in a “captured audience” in a classroom setting.

Based on previous work, we suspected that a sample of patients typically seen in an outpatient rehabilitation setting might encounter ceiling effects with the Daily Activity scale. The original calibrations were drawn from a diverse sample of inpatient and community-based post-acute care patients [18], many of whom had substantial disability. Initial review of the data indicated that, indeed, ceiling effects were present [4], and therefore clinics were interested in examining new items that would help make the Daily Activity CAT scale more relevant to their patients and to avoid reaching a maximum score.

Although a series of 41 new Daily Activity items were written and tested, we found that 29 could be included in the next version of the Daily Activity item bank. A number of the new items contained important content, but were apparently not written clearly and were removed for the new version of the item bank. We did not conduct any cognitive testing or other qualitative analyses on the new pretest items and this may have contributed to the lack of clarity of some items. We did not find any particular pattern or theme to items that had significant fit problems, although some items included multiple tasks; for example, the item “Sit in chair holding a book while reading” was removed due to poor item fit, presumably

because it was not clear if we were asking about a person's difficulty in sitting, holding a book or actually reading. This item may be the focus of revision in the future. Other items showed too much local dependence, such as a series of items assessing lifting (100, 50, 10 pounds). Some items were retained because of the importance of their content and scale location; others were removed from this version. In the future, we may develop a series of testlet-type items [43, 44], in which the amounts of weight are reworked as response choices in one item. To remove items purely on a statistical basis without assessing the impact on the content validity and coverage of the scale is, in our view, undesirable. In the final analysis, our decision to retain items was made both on content and statistical criteria.

We chose to use a simple form of online calibration methodology, as we found no real advantages for using more complex methods. We did explore using a sparse-matrix concurrent calibration approach, and with one and multiple EM cycles. However, we found very similar results in wave 1 to the results presented herein. More work in the future should be directed at whether one of these methods is preferable under different data-acquisition conditions. Additionally, we used the GPCM for estimating item parameters since we had conducted all previous analyses using this model. Other IRT models might have been selected here and may have fit the data better; however, it made most sense for us to maintain the original IRT model from the initial calibration sample.

Both the consistency of the person score estimates with the full item bank and the precision of the DA-CAT-2 were superior to those of DA-CAT-1 without the pretest items. We note that the precision gains are at the high-performance levels of Daily Activity, which was where improvement was needed. The precision gains should be interpreted in light of the fact that additional items with high discrimination will improve precision (decrease standard errors). However, the precision improvements are not random, but are clearly at the point of the scale (greater ability in Daily Activities) that is needed for this outpatient sample. Furthermore, when comparing the performance of the two CAT versions, the respondent burden in the DA-CAT-2 had the same number of items (seven) as in the DA-CAT-1; there were just more items available for the CAT to choose from the item bank.

The major finding in this study of a reduction in ceiling effects using the CAT with the pretest items in the item bank is promising. In our experience, ceiling effects are a notable challenge in most IRT scales and CAT applications [4, 45], and we achieved about a 10% reduction with the incorporation of the new pretest items. Constant monitoring of this effect is important so that the scale can be as sensitive as possible to health and functional changes throughout the entire continuum of Daily Activities. We still, however, have rather large standard errors in the DA-CAT-2 at person score estimates greater than 75. Although these SE are less in the DA-CAT-2 than in DA-CAT-1, they still need to be reduced by improving the item content and discrimination at the higher end of the scale.

In the fifth wave of new items, clinicians suggested that we incorporate body-region-specific items for cervical patients. Six of these ten pretest items were retained in the DA-CAT-2. As more of these items are evaluated and retained in the CAT, it may be possible to set a filter on the CAT to administer mainly Daily Activity items that require some form of neck stabilization or movements for patients with specific cervical conditions. The advantage of this type of content organization is that a metric of Daily Activity functioning is theoretically available with any combination of items from the item bank, and that cervical patients could be administered items that are most relevant to their condition, yet their score could be compared with other patients who were administered different subsets of items with the CAT. This is an exciting possibility worthy of future research. In future studies, options include administering additional items to improve precision based on a predetermined standard.

Our analyses and results suggest a potential approach towards evaluating new items for inclusion in updating PRO CATs. By using online calibration techniques and computer simulations of real data, we were able to provide the best estimate as to how this new item bank and CAT would function. However, full implementation of the new CAT version and its monitoring over time will be the final determinant of whether we achieved meaningful improvements in scoring consistency with the full item pool, precision, and a reduction of ceiling effects as the simulation analyses suggest.

We note a number of limitations to the study. Due to practical concerns in the clinic with patient response burden and potential interruptions in patient availability, we were unable to administer all of the pretest items to each patient. A compromise was reached to administer no more than ten new items per patient. We found a very high compliance rate (99%), thus patients seemed to be willing to take a minimal amount of extra time to participate; however, the clinic wanted to acquire the CAT data first. Patients then volunteered to answer the pretest items. Since patients knew that the pretest items were not part of their own health and functional data, there may have been less interest in answering these items with the same care as items administered in the operational CAT. Since we were unable to administer the items randomly or co-vertly within the CAT session, this may have led to serial order effects or bias. However, we did not find excessive person-fit problems in the data, thus it appears that the bias may have been relatively small. Replication with an improved item administration design is needed to minimize these potential biases.

The operational CAT was set up with a stop rule of seven items. Although this worked quite well for the purposes of group monitoring and sensitivity to change within and across sites [4], it did create missing data. We tried to offset this limitation by recruiting fairly large sample sizes in each prospective wave of data collection; however, replication is needed to confirm the stability of these findings. To maximize sample size, we sampled data from the initial intake (admission) to the outpatient episode, which is usually available at a much higher rate than assessments performed at or near discharge. As more data are available, we plan to validate these calibrations and CAT results with data collected at discharge from outpatient rehabilitation.

We believe these results reveal that, by continually monitoring outcomes and the quality of PRO measurement, outcome measurement can be rapidly improved for many general and specific patient populations. The simulation results described in this article are a first step in optimizing the measurement of important PRO concepts in future CAT applications.

Acknowledgments

Select Medical Corporation purchased the Outpatient Rehabilitation Division of HealthSouth Corporation on May 1, 2007 and the individual clinics that participated in this study are now known as “Select Physical Therapy and NovaCare.” We would like to thank all of the Select Physical Therapy and NovaCare clinical sites who participated in our study by providing the data used in this study.

Sources of support: Select Medical Corporation and in part by an Independent Scientist Award (K02 HD45354-01) to Dr. Haley.

Abbreviations

ADL	Activities of daily living
AM-PAC	Activity Measure for Post-Acute Care
CAT	Computerized adaptive testing

CFA	Confirmatory factor analyses
CFI	Comparative fit index
DA-CAT-1	Original Daily Activity item bank in the CAT
DA-CAT-2	New expanded item bank with pretest items in the new CAT version
DIF	Differential item functioning
GPCM	Generalized partial credit model
ICC	Intraclass correlation coefficient
IRT	Item response theory
PRO	Patient-reported outcome
RMSEA	Root-mean-square error of approximation
SE	Standard errors
TIF	Test information function
TLI	Tucker–Lewis index

References

1. Hart DL, Cook KF, Mioduski JE, Teal CR, Crane PK. Simulated computerized adaptive test for patients with shoulder impairments was efficient and produced valid measures of function. Journal of Clinical Epidemiology. 2006; 59(3):290–298. doi:10.1016/j.jclinepi.2005.08.006. [PubMed: 16488360]
2. Hart DL, Mioduski JE, Stratford PW. Simulated computerized adaptive tests for measuring functional status were efficient with good discriminant validity in patients with hip, knee, or foot/ankle impairments. Journal of Clinical Epidemiology. 2005; 58(6):629–638. doi:10.1016/j.jclinepi.2004.12.004. [PubMed: 15878477]
3. Hart D, Mioduski J, Werenke M, Stratford P. Simulated computerized adaptive test for patients with lumbar spine impairments was efficient and produced valid measures of function. Journal of Clinical Epidemiology. 2006; 59:947–956. doi:10.1016/j.jclinepi.2005.10.017. [PubMed: 16895818]
4. Jette A, Haley S, Tao W, Ni P, Moed R, Meyers D, et al. Prospective evaluation of the AM-PAC-CAT in outpatient rehabilitation settings. Physical Therapy. 2007; 87:385–398. [PubMed: 17311888]
5. Jette AM, Haley SM. Contemporary measurement techniques for rehabilitation outcomes assessment. Journal of Rehabilitation Medicine. 2005; 37(6):339–345. doi:10.1080/16501970500302793. [PubMed: 16287664]
6. Cella D, Gershon R, Lai J-S, Choi S. The future of outcomes measurement: Item banking, tailored short forms, and computerized adaptive assessment. Quality of Life Research. 2007; 16:133–141. doi:10.1007/s11136-007-9204-6. [PubMed: 17401637]
7. Fries J, Bruce B, Cella D. The promise of PROMIS: Using item response theory to improve assessment of patient-reported outcomes. Clinical and Experimental Rheumatology. 2005; 23(5 (suppl 39)):S53–S57. [PubMed: 16273785]
8. Cella D, Young S, Rothrock N, Gershon R, Cook K, Reeve B, et al. The patient-reported outcomes measurement information system (PROMIS): Progress of an NIH Roadmap Cooperative Group during its first two years. Medical Care. 2007; 45(5):S3–S11. doi:10.1097/01.mlr.0000258615.42478.55. [PubMed: 17443116]
9. Hambleton, RK. Applications of item response theory to improve health outcomes assessment: Developing item banks, linking instruments, and computer-adaptive testing. In: Lipscomb, J.; Gotay, CC.; Snyder, C., editors. Outcomes assessment in cancer. Cambridge University Press; Cambridge, UK: 2005. p. 445-464.

10. Fayers P. Applying item response theory and computer adaptive testing: The challenges for health outcomes assessment. Quality of Life Research. 2007; 16(1):187–194. doi:10.1007/s11136-007-9197-1. [PubMed: 17417722]
11. Wainer, H. Computerized adaptive testing: A primer. Lawrence Erlbaum Associates; Mahwah, NJ: 2000.
12. Hambleton, R.; Swaminathan, H. Item Banking. In: Hambleton, R.; Swaminathan, H., editors. Item response theory: Principles and applications. Kluwer Nijoff Publishing; Boston, MA: 1985. p. 255-279.
13. Revicki DA, Cella DF. Health status assessment for the twenty-first century: Item response theory, item banking and computer adaptive testing. Quality of Life Research. 1997; 6:595–600. doi:10.1023/A:1018420418455. [PubMed: 9330558]
14. Bode RK, Lai JS, Cella D, Heinemann AW. Issues in the development of an item bank. Archives of Physical Medicine and Rehabilitation. 2003; 84(2):S52–S60. doi:10.1053/apmr. 2003.50247. [PubMed: 12692772]
15. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Medical Care. 2000; 38(9s):II-28–II-42. doi:10.1097/00005650-20000 9002-00007. [PubMed: 10982088]
16. Haley SM, Coster WJ, Andres PL, Ludlow LH, Ni PS, Bond TLY, et al. Activity outcome measurement for post-acute care. Medical Care. 2004; 42(1):I-49–I-61. doi:10.1097/01.mlr.0000103520.43902.6c. [PubMed: 14707755]
17. Coster WJ, Haley SM, Andres PL, Ludlow LH, Bond TLY, Ni PS. Refining the conceptual basis for rehabilitation outcome measurement: personal care and instrumental activities domain. Medical Care. 2004; 42(Suppl 1):I-62–I-72. doi:10.1097/01.mlr.0000103521.84103.21. [PubMed: 14707756]
18. Haley SM, Ni P, Hambleton RK, Slavin MD, Jette AM. Computer adaptive testing improves accuracy and precision of scores over random item selection in a physical functioning item bank. Journal of Clinical Epidemiology. 2006; 59(2):1174–1182. doi:10.1016/j.jclinepi.2006.02.010. [PubMed: 17027428]
19. Sands, WA.; Waters, BK.; McBride, JR. Computerized adaptive testing: From inquiry to operation. American Psychological Association; Washington DC: 1997.
20. Muthen, B.; Muthen, L. Mplus User's Guide. Muthen & Muthen; Los Angeles: 2001.
21. Stone C. Empirical power and type I error rates for an IRT fit statistic that considers the precision of ability estimates. Educational and Psychological Measurement. 2003; 63:566–583. doi:10.1177/0013164402251034.
22. Stone CA. Monte Carlo based null distribution for an alternative goodness-of-fit test statistic in IRT models. Journal of Educational Measurement. 2000; 37:58–75. doi:10.1111/j.1745-3984.2000.tb01076.x.
23. Stone CA, Zhang B. Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. Journal of Educational Measurement. 2003; 40:331–352. doi:10.1111/j.1745-3984.2003.tb01150.x.
24. Zumbo, B. A Handbook on the theory and methods of differential item functioning (DIF). Directorate of Human Resources Research and Evaluation; Ottawa, ON: 1999.
25. Cohen, J. Statistical power analysis for the behavioral sciences. 2nd ed.. Erlbaum; Hillsdale, NJ: 1988.
26. Chen W, Thissen D. Local dependence indexes for item pairs using item response theory 289. Journal of Educational and Behavioral Statistics. 1997; 22:265–289.
27. Tate R. A comparison of selected empirical methods for assessing the structure of responses to test items. Applied Psychological Measurement. 2003; 27(3):159–203. doi:10.1177/014662 1603027003001.
28. Yen WM. Scaling performance assessments: strategies for managing local item dependence. Journal of Educational Measurement. 1993; 30(3):187–213. doi:10.1111/j.1745-3984.1993.tb00423.x.

29. Morgan, D.; Way, W.; Augemberg, K. A comparison of online calibrations methods for a CAT. Paper presented at the National Council on Measurement on Education; San Francisco, CA. 2006, April 8;
30. Ban, J-C.; Hanson, B.; Wang, T.; Yi, Q.; Harris, D. A comparative study of online pretest item calibration/scaling methods in CAT. American Educational Research Association; Washington, DC: 2000.
31. Stocking M, Swanson L. Optimal design of item banks for computerized adaptive tests. Applied Psychological Measurement. 1998; 22(3):271–279. doi:10.1177/01466216980223007.
32. Wainer, H.; Mislevy, R. Item response theory, item calibration, and proficiency estimation. In: Wainer, H., editor. Computer adaptive testing: A primer. Lawrence Erlbaum; Hillsdale, NJ: 1990. p. 65-102.
33. Muraki, E.; Bock, RD. PARSCALE: IRT item analysis and test scoring for rating-scale data. Scientific Software International; Chicago: 1997.
34. van der Linden, W.; Hambleton, R. Handbook of modern item response theory. Springer; Berlin: 1997.
35. Rijmen F, Tuerlinckz F, De Boeck P, Kuppens P. A nonlinear mixed model framework for item response theory. Psychological Methods. 2003; 8:185–205. doi:10.1037/1082-989X.8.2. 185. [PubMed: 12924814]
36. Ludlow LH, Haley SM. Rasch model logits: Interpretation, use, and transformation. Educational and Psychological Measurement. 1995; 55(6):967–975. doi:10.1177/0013164495055006 005.
37. Luecht, R. Computer-adaptive testing. In: Everett, B.; Howell, D., editors. Encyclopedia of statistics in behavioral science. Wiley; New York: 2004.
38. Samejima F. Some critical observations of the test information function as a measure of local accuracy in ability estimation. Psychometrika. 1994; 59(3):307–329. doi:10.1007/BF022 96127.
39. Donoghue JR. An empirical examination of the IRT information of polytomously scored reading items under the generalized partial credit model. Journal of Educational Measurement. 1994; 31(4):295–311. doi:10.1111/j.1745-3984.1994.tb00448.x.
40. Lai J-S, Cella D, Dineen K, Bode R, Von Roenn JH, Gershon RC. An item bank was created to improve the measurement of cancer-related fatigue. Journal of Clinical Epidemiology. 2005; 58:190–197. doi:10.1016/j.jclinepi.2003. 07.016. [PubMed: 15680754]
41. Ware JE Jr. Gandek B, Sinclair SJ, Bjorner B. Item response theory in computer adaptive testing: Implications for outcomes measurement in rehabilitation. Rehabilitation Psychology. 2005; 50(1): 71–78. doi:10.1037/0090-5550.50.1.71.
42. Ware JE Jr. Conceptualization and measurement of health-related quality of life: comments on an evolving field. Archives of Physical Medicine and Rehabilitation. 2003; 84:S43–S51. doi:10.1053/apmr.2003.50246. [PubMed: 12692771]
43. Wainer H, Kiely G. Item clusters and computerized adaptive testing: A case for testlets. Journal of Educational Measurement. 1987; 24:185–201. doi:10.1111/j.1745-3984.1987.tb00274.x.
44. Lee G, Brennan RL, Frisbie DA. Incorporating the testlet concept in test score anaylses. Educational Measurement: Issues and Practice. 2000; 19(4):9–15. doi:10.1111/j.1745-3992.2000.tb00041.x.
45. Haley SM, Coster WJ, Andres PL, Kosinski M, Ni PS. Score comparability of short-forms and computerized adaptive testing: Simulation study with the Activity Measure for Post-Acute Care (AM-PAC). Archives of Physical Medicine and Rehabilitation. 2004; 85:661–666. doi:10.1016/j.apmr.2003.08.097. [PubMed: 15083444]

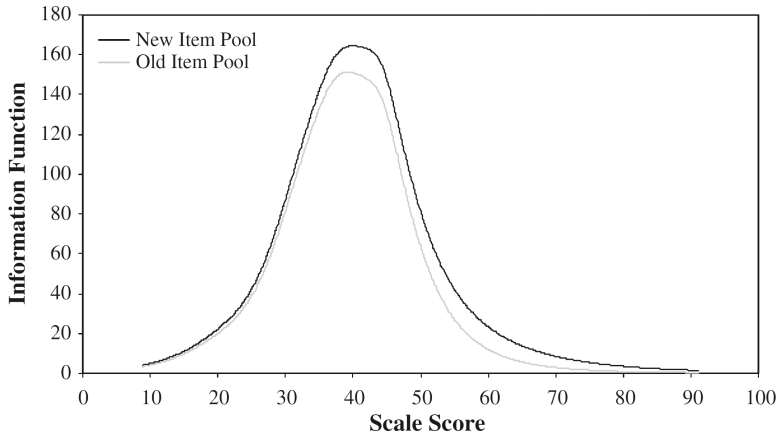

Fig. 1.
Information functions for DA-CAT-1 and DA-CAT-2

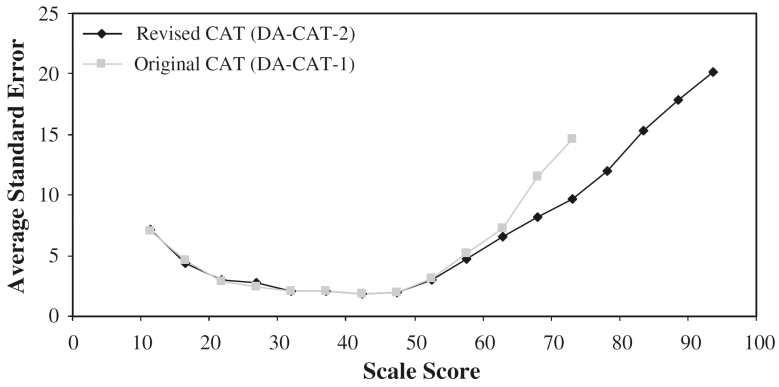

Fig. 2.
Comparison of person score standard errors between CAT versions

Table 1

Demographics of sample per wave

	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5	Total
N	1733	1951	1317	1054	827	6882
Mean age (SD), years	49.6 (16.7)	49.9 (16.7)	48.5 (17.7)	52 (17.9)	52.1 (19.2)	50.15 (17.44)
Gender (female)	57.9%	60.5%	59.8%	57.6%	58.28%	58.96%
Mean no. (SD) visits	10.2 (8.7)	9.6 (7.7)	9.3 (8.2)	10.1 (8.4)	8.2(5.3)	9.59 (7.95)
Surgery (yes)	20.5%	23.3%	23.22%	21.6%	19/9%	21.93%
Body part						
Lower	46 (2.7%)	41 (2.1%)	35 (2.7%)	25 (2.4%)	21 (2.5%)	168 (2.44%)[a]
Upper	648 (37.4%)	745 (38.2%)	496 (37.7%)	436 (41.4%)	325 (39.3%)	2650 (38.51%)
Spine	868 (50.1%)	977 (50.1%)	631 (47.9%)	478 (45.4%)	386 (46.7%)	3340 (48.53%)
Other	171 (9.9%)	188 (9.6%)	155 (11.8%)	115 (10.9%)	95 (11.5%)	724 (10.52%)
Mean (SD) duration of episode (days)	35.9 (34.7)	33.7 (28.8)	33.3 (29.5)	35.4 (29.8)	26.8 (17.7)	33.61 (29.76)
Mean acuity (SD) (days)	189.2 (638.5)	224.4 (867.3)	283.3 (937.7)	271.6 (878.1)	327 (1803.5)	244.6 (977.62)
Mean Daily Activity CAT score	54.4 (10.3)	54.2 (10.4)	53.7 (10.6)	53.1 (10.2)	54.5 (12.1)	54 (10.6)

[a]The low value of lower extremity group is because the Daily Activity domain was primarily administered to upper/spine/other groups

Table 2

Sample sizes and pretest item list administered at each data collection wave (some items abbreviated)

Wave 1 (*n* = 1733)	**Wave 2 (*n* = 1951)**	**Wave 3 (*n* = 1317)**	**Wave 4 (*n* = 1054)**	**Wave 5 (*n* = 827)**
Items (*k* = 4)	**Items (*k* = 7)**	**Items (*k* = 10)**	**Items (*k* = 10)**	**Items (*k* = 10)**
Lift 100 pounds	Lift 25 pounds to table	Painting a door	Reaching overhead into a cabinet to get something off a shelf	Sit in chair holding a book while reading
Lift heavy object overhead	Lift 25 pounds above shoulder	Digging a hole in the ground with a shovel	Changing a light bulb above your head	Hold a telephone between ear and shoulder
One push-up	Lift 10 pounds to table	Hanging wash on a line at eye level or above…	Reaching into the back pocket of a pair of pants	Checking the blind spot while driving a car
Five push-ups	Lift 10 pounds above shoulder	Washing indoor windows	Washing your lower back	Turning head quickly to look behind you
	Lift 5 pounds to a table	Opening a stuck window	Fastening a necklace behind your neck	Look up at clouds in the sky
	Lift 5 pounds above shoulder	Setting a watch	Throwing a ball	Watch a movie in a movie theatre
	Hang curtains	Moving a sofa to clean under it…	Working with your hands overhead for 2–5 min	Look at a computer screen for more than 15 min
		Flossing your teeth	Reaching behind you to get your seatbelt	Gargle with mouthwash with head tilted back
		Blow dry your hair	Sleeping on your `painful' shoulder	Coughing
		Fastening clothing behind your back–	Carrying two plastic grocery bags with handles at your side for 50 feet	Hold head steady while driving a car in traffic on a local road

Table 3

Item parameters of 29 pretest items added to DA-CAT-2

New items retained	Item calibration, slope and range			Item fit		Comments
	Mean calibration50/10**	Slope	Range 50/10**	$x^{2\#}$	P^*	
Lifting 100 lbs or more…	72.22	0.501	58.24–88.08	57.04	0.02	Retained for content and to reduce ceiling
Lifting a heavy object overhead–	65.63	0.628	52.07–81.9	49.97	0.14	
Doing five push-ups	64.70	0.570	55.15–75.73	52.14	0.14	
Moving a sofa to clean under it–	57.32	0.666	45.9–70.59	31.68	0.09	
Lifting 25 pounds from the ground to a table–	57.01	0.679	43.84–71.62	57.06	0.05	Retained for content and to reduce ceiling
Working with your hands overhead for 2–5 min	54.17	0.862	43.84–66.47	35.01	0.52	
Opening a stuck window	53.85	1.089	44.87–63.38	49.96	0.10	
Lifting 10 pounds from a table to above shoulder height…	53.61	0.711	42.81–66.47	61.47	0.08	
Hanging curtains…	52.19	0.869	44.87–61.33	36.75	0.95	
Fastening clothing behind your back…	51.16	0.881	42.81–60.3	49.40	0.00	Retained for content
Carrying two plastic grocery bags with handles at your side for 50 feet	50.89	1.153	42.81–61.33	47.78	0.30	
Changing a light bulb above your head	50.78	0.918	41.78–61.33	60.70	0.01	Retained for content
Fastening a necklace behind your neck	49.5	1.085	43.84–56.18	98.54	0.00	Retained for content; female item only
Reaching overhead into a cabinet to get something off a shelf	48.88	0.891	36.64–61.33	42.35	0.10	
Lifting 10 pounds from the ground to a table…	48.87	0.745	38.69–61.33	44.13	0.16	
Hanging wash on a line at eye level or above…	48.34	0.922	39.72–58.24	27.31	0.27	
Washing indoor windows	48.27	0.999	39.72–58.24	38.00	0.06	
Painting a door	47.42	1.095	40.75–55.15	55.66	0.00	Retained for content
Reaching behind you to get your seatbelt	45.93	1.007	35.61–59.27	26.95	0.53	
Washing your lower back	45.81	1.010	38.69–55.15	26.76	0.55	
Reaching into the back pocket of a pair of pants	43.66	1.069	35.61–52.07	39.60	0.11	
Lifting 5 pounds from the ground to a table…	43.26	0.833	35.61–52.07	29.48	0.43	
Blow dry your hair	41.89	0.995	34.58–50.01	41.83	0.00	Retained for content
Hold a telephone between ear and shoulder	38.75	0.508	24.08–52.79	20.09	0.24	
Turning head quickly to look behind you	36.64	0.421	13.9–57.11	19.26	0.25	
Gargle with mouthwash with head tilted back	27.97	0.502	15.24–40.65	18.44	0.33	
Look at a computer screen for more than 15 min	23.46	0.397	7.42–61.12	17.18	0.29	

New items retained	Item calibration, slope and range			Item fit		Comments
	Mean calibration50/10**	Slope	Range 50/10**	$x^{2\#}$	P^{*}	
Watch a movie in a movie theatre	22.64	0.350	7.73–62.66	19.60	0.20	
Look up at clouds in the sky	21.36	0.412	0–42.5	17.92	0.35	

$P<0.05$ indicated statistical significant

$^{*}P$ is the simulated P-value calculated from the simulated person response data using the person ability and item parameters from the actual data; actual statistics were compared with the simulated statistics distribution based on 100 simulated samples.

Table 4

Comparison of scoring range between CAT versions with (DA-CAT-2) and without (DA-CAT-1) pretest items

	Number of items	Scale score range	Floor effects (%)	Ceiling effects (%)
DA-CAT-1	65	9.89–78.71	0	16.11
DA-CAT-2	65 + 29	9.89–99.00	0	6.13