Assessing Mobility in Children Using a Computer Adaptive Testing Version of the Pediatric Evaluation of Disability Inventory

Stephen M. Haley, PhD, PT, Anastasia E. Raczek, MEd, Wendy J. Coster, PhD, OTR, Helene M. Dumas, MS, PT, Maria A. Fragala-Pinkham, MS, PT


Objective: To assess score agreement, validity, precision, and response burden of a prototype computerized adaptive testing (CAT) version of the Mobility Functional Skills Scale (Mob-CAT) of the Pediatric Evaluation of Disability Inventory (PEDI) as compared with the full 59-item version (Mob-59).

Design: Computer simulation analysis of cross-sectional and longitudinal retrospective data; and cross-sectional prospective study.

Setting: Pediatric rehabilitation hospital, including inpatient acute rehabilitation, day school program, outpatient clinics, community-based day care, preschool, and children’s homes.

Participants: Four hundred sixty-nine children with disabilities and 412 children with no disabilities (analytic sample); 41 children without disabilities and 39 with disabilities (cross-validation sample).

Interventions: Not applicable.

Main Outcome Measures: Summary scores from a prototype Mob-CAT application and versions using 15-, 10-, and 5-item stopping rules; scores from the Mob-59; and number of items and time (in seconds) to administer assessments.

Results: Mob-CAT scores from both computer simulations (intraclass correlation coefficient [ICC] range, .94–.99) and field administrations (ICC = .98) were in high agreement with scores from the Mob-59. Using computer simulations of retrospective data, discriminant validity, and sensitivity to change of the Mob-CAT closely approximated that of the Mob-59, especially when using the 15- and 10-item stopping rule versions of the Mob-CAT. The Mob-CAT used no more than 15% of the items for any single administration, and required 20% of the time needed to administer the Mob-59.

Conclusions: Comparable score estimates for the PEDI mobility scale can be obtained from CAT administrations, with losses in validity and precision for shorter forms, but with a considerable reduction in administration time.

Key Words: Outcome assessment (health care); Pediatrics; Rehabilitation.

© 2005 by American Congress of Rehabilitation Medicine and the American Academy of Physical Medicine and Rehabilitation

LIKE OTHER ARENAS of health care, pediatric rehabilitation is under increasing pressure to document functional changes that occur while children are receiving services. Because resources for the continuation of services often depend on documentation of consistent improvement in the attainment of new functional skills, measures used to monitor progress must provide meaningful information about current status or changes in function. To date, there has been little consistency in how disability assessments for children are developed or administered. Ideally, measures of disability should be sensitive to changes, easy to administer, and able to be completed with the smallest possible effort by a clinician, child, and/or caregiver.

Selecting a functional assessment to cover a wide age range and diverse functional content presents clinicians and researchers with serious trade-offs. Difficult choices must be made between comprehensiveness and practicality. Measurement precision is optimal when the content of functional items and the level of the child’s abilities are closely matched. However, in heterogeneous groups, such as seen in many pediatric rehabilitation settings, an optimal set of items that fits most children is difficult to achieve. In efforts to make instruments practical, the range and specificity of content is often compromised, leading to large amounts of measurement noise and floor and ceiling effects.

Researchers, clinicians, and families have an important stake in the development of standardized disability assessments that are tailored or “adapted” to the child’s functional level. Most fixed-length standardized functional assessments in pediatric rehabilitation require the administration of all items to produce a summary score, regardless of the usefulness of the information. Theoretically, adaptive assessments, which ask only those questions that are relevant to the child’s functional level, could reduce the amount of time required to complete functional assessments. In practice, most clinicians intuitively adapt histories and tests to the child’s apparent functional level. For example, a question about community ambulation is unnecessary if the child is just beginning to take steps in the parallel bars, whereas a question about the ability to negotiate 3 to 5 steps is pointless if the child is currently walking up and down 3 flights of stairs. Administering functional items that are either too easy or too hard provides little information, and may waste valuable time. In an adaptive test, each administration is adapted uniquely to the child’s ability level at that time. As the child progresses and new test administrations are needed, new items are selected that best reflect the child’s more advanced functional level.

From the Health and Disability Research Institute, Boston University, Boston, MA (Haley, Coster); QualyMetrix Inc, Lincoln, RI (Raczek); and the Research Center for Children with Special Health Care Needs, Franciscan Hospital for Children, Boston, MA (Dumas, Fragala-Pinkham).

Supported in part by the National Institute of Child Health and Human Development, National Center on Medical Rehabilitation Research (grant nos. R43 HD42388-01, K02 HD45354-01A1).

No commercial party having a direct financial interest in the results of the research supporting this article has or will confer a benefit on the author(s) or on any organization with which the author(s) is/are associated.

Reprint requests to Stephen M. Haley, PhD, PT, Health and Disability Research Institute, Boston University, 53 Bay State Rd, Boston, MA 02215, e-mail: smhaley@bu.edu.

Recently, the general health and rehabilitation fields have begun to test and apply methods such as computer adaptive testing (CAT) that offer a potential solution to the conflict between practicality and precision. CAT employs a simple form of artificial intelligence that selects questions directly tailored to the child, shortens or lengthens the test to achieve the desired precision, and scores each test on a standard metric so that results can be compared across time points and across groups of children. Although immediately appealing in terms of reduced respondent burden, CATs must also meet satisfactory reliability and validity standards for acceptance in clinical and research applications. Thus, the purpose of this report is to describe the score agreement, validity, precision, and response burden of a tailored approach to the assessment of functional mobility in children, by comparing CAT to a full-length administrative format.

CAT applications require: (1) a large set of items for each functional area of interest (item pools), (2) items that scale consistently along a dimension of low to high functional proficiency, and (3) rules that guide starting, stopping, and scoring procedures. Item response theory (IRT) methods are used to create hierarchically ordered item pools, and then software algorithms select items to match the child’s functional level. The logic of CAT begins with administration of a global item that is selected a priori based on its broad coverage of the range of function. All respondents answer the same first question, an advantage for purposes of standardization. On the basis of the response to the first item, a score and confidence interval (CI) are estimated, then the next optimal item is presented and a response is recorded. With administration of the next item, the score is reestimated along with a new CI. The computer algorithm determines whether the stopping rule has been satisfied. If satisfied, the assessment of that functional domain ends. If not satisfied, new items are administered in an iterative fashion until the stopping rule is satisfied. By altering the stopping rule, it is possible to match the level of score precision to the specific purpose of measurement.

Our objective in the present study was to apply the same logic used successfully to date in headache and adult functional assessment areas to test the potential of CAT assessment of pediatric physical disability. In previous work, we demonstrated that CATs are vastly more efficient than traditional fixed forms of assessments. In the present study, we report on a prototype CAT program using the Mobility Functional Skills Scale (Mob-CAT) of the Pediatric Evaluation of Disability Inventory (PEDI). We expected that the advantage of a CAT-based assessment for pediatric rehabilitation would be reduced respondent burden, with only small losses in score accuracy, precision, and validity as compared with the full-length assessment. We examined these questions in both computer simulation studies of retrospective data and a prospective cross-validation study.

**METHODS**

**Samples**

**Analytic sample.** We used an existing database of 881 children who completed the 59-item Mobility Functional Skills Scale (Mob-59) of the PEDI. This retrospective, analytic sample included 2 groups: (1) a normative sample of 412 healthy children between the ages of 6 months and 7.5 years that was also used to create the initial standardization and normative scoring of the PEDI, and (2) a clinical sample of 469 children and adolescents (ages 6–17y) who had received inpatient, outpatient, or school-based rehabilitation services at Franciscan Hospital for Children (FHC), Boston, MA. Of the 469 clinical cases, 263 had longitudinal data appropriate for sensitivity analyses.

Approximately 48% of the children in the clinical sample had congenital or inherited diseases, 21% had growth and maturation disorders, 16% had acquired conditions, and 15% were diagnosed with traumatic injuries. Demographic characteristics of the analytic sample are in Table 1. The sample size of 881 is acceptable for initial calibration work for a prototype CAT.

**Cross-validation sample.** We recruited a convenience sample of 80 children and adolescents for the prospective cross-validation study. Thirty-nine children with disabilities, ages 6 months to 18 years, were recruited from the clinical programs (inpatient, outpatient, early intervention, hospital-based school) at FHC. Ethnic representation corresponding to the current US Census was targeted for recruitment; however, respondents who did not speak English as a primary language were excluded because of the prohibitive cost of translating and interpreting. Children were further selectively recruited to assure representation of each of the following 4 impairment groups: congenital or inherited disease, growth and maturation disorders, acquired conditions, and traumatic injuries. Forty-one children without disabilities, ages 6 months to 7.5 years, were recruited through the Franciscan Family Child Care Center and the home communities of the 2 field-test coordinators (HMD, MAF-P).

**Mobility Scale**

The PEDI is a comprehensive functional assessment instrument that measures both capability and performance of functional activities, including mobility. The mobility domain includes activities such as floor mobility, ambulation, transfers, and movement in different home and community environments. The Mobility Functional Scale assesses the capability of 59 mobility activities using a dichotomous capable or unable scoring criterion. Several studies have supported the reliability and validity of the PEDI mobility scale in a wide variety of clinical samples. Construct validity of the PEDI mobility scale has been established by demonstrating the ability of the PEDI to correctly identify children with and without disabilities, and to discriminate between different types of acquired brain injury. Studies have also reported successful outcome monitoring using the PEDI in children with cerebral palsy, myelodysplasia, osteogenesis imperfecta, and traumatic brain injury. Recent work has established the responsiveness of the PEDI mobility scale. Because the development and construction of the PEDI mobility scale and summary scores are based on Rasch rating scale methodology, the scale provides an excellent starting point for the development of a prototype CAT.

Table 1: Demographic Characteristics of Samples

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Analytic Sample</th>
<th>Cross-Validation Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age range (y)</td>
<td>6mo–17y</td>
<td>6mo–18y</td>
</tr>
<tr>
<td>Female (%)</td>
<td>45.2</td>
<td>39.4</td>
</tr>
<tr>
<td>Hispanic or Latino (%)</td>
<td>8.3</td>
<td>10.0</td>
</tr>
<tr>
<td>Asian (%)</td>
<td>1.5</td>
<td>11.3</td>
</tr>
<tr>
<td>Other (%)</td>
<td>5.8</td>
<td>3.8</td>
</tr>
<tr>
<td>Black or African American (%)</td>
<td>14.6</td>
<td>11.3</td>
</tr>
<tr>
<td>White (%)</td>
<td>68.8</td>
<td>63.7</td>
</tr>
<tr>
<td>Total sample size (n)</td>
<td>881</td>
<td>80</td>
</tr>
</tbody>
</table>

Arch Phys Med Rehabil Vol 86, May 2005
Development of the Mob-CAT

Unidimensionality and local independence. Both IRT and CAT systems assume basic measurement properties of item sets that make up a functional construct (latent variable). These assumptions include unidimensionality, local independence, and stability of item parameters between groups (eg, clinical vs normative and age groups). Item sets that violate these assumptions may be more or less effective in achieving appropriate modeling of the data and may limit the accuracy of a CAT instrument.

A key assumption of the latent variable models that serve as the basis for CAT is that all items in a scale measure a single, unitary concept; that is, the items are unidimensional. The latent variable alone should explain how items are related to each other.\(^{41,42}\) We tested the latent structure of the Mob-59 items in a series of confirmatory factor analyses\(^{43}\) and evaluated item loadings and residual correlations between items using MPLus software.\(^{44,45}\) 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. Three pieces of evidence were reviewed to determine the extent to which a unidimensional model adequately represented scale structure: (1) eigenvalues associated with each factor extracted, (2) item loadings on the primary factor, and (3) results from overall model fit tests.

We combined the normative and clinical PEDI samples (analytic sample) to ensure adequate sample size for estimation of model parameters. If the item parameters were similar between normative and clinical groups, combining samples would have the practical effect of improving generalizability of the model parameters across both normative and clinical groups, and providing adequate numbers of cases in the moderate to low end of the scale where good precision is especially needed (clinical group). In a categorical factor analysis of all Mob-59 items, 1 factor explained a substantial proportion (84\%) of item variance. With the exception of 1 item (Sits if supported), loadings on the primary factor were all quite high, ranging from .82 to .99. Results of overall model fit tests supported the interpretation that Mob-59 items form a unidimensional construct. The Comparative Fit Index value of .993 indicated very good fit, and can generally be interpreted as an indicator that 99\% of covariance in the data is reproducible by the hypothesized model. Tucker-Lewis Index (TLI) values close to 1.0 indicate good fit, and the mobility scale achieved a TLI equal to .997.

A second requirement, local independence, requires scale items to be independent, or unrelated to each other, at a given score level. Meeting this assumption was a more significant challenge for the Mob-59 scale. One indicator that items share a TLI equal to .997.

A second requirement, local independence, requires scale items to be independent, or unrelated to each other, at a given score level. Meeting this assumption was a more significant challenge for the Mob-59 scale. One indicator that items share more than the latent trait is the residual correlation. High residual correlations (\(\geq 0.2\)) were observed among several pairs of items.\(^{46}\) The PEDI groups similar items into skills-based blocks, such as toilet transfers, car transfers, and indoor and outdoor locomotion. For instance, separate items within the outdoor locomotion block appraise distances of 50 (15m), 100 (30m), and 150 feet (45m), and of course, answers to the item “150 feet” are not independent of the answer to “100 feet.” This violation of model assumptions can affect the estimation of test information and item discrimination parameters,\(^{46}\) but of course cannot be rectified in an existing database.

Item calibrations. IRT modeling of the Mob-59 items was completed using the Rasch model, which estimates item difficulty.\(^{47,49}\) The Rasch model was selected as the best solution for this phase of our research because of its simplicity in interpretation, and flexibility about the underlying form of population or trait distributions. Based on sample size and the distribution of item difficulty, data for the total analytic sample were used to generate item calibrations. Note that the original item calibration and instrument standardization for the PEDI was conducted using the normative sample alone (N=412).\(^{24}\) Initial item parameters and tests of overall model fit for the Mob-59 were generated using OPLM software\(^{41,42}\) and conditional maximum likelihood estimation. We evaluated the fit to the model using multiple fit statistics for each item based on the comparison of expected and observed values of item slope across the distribution of the latent variable; Bonferroni-adjusted \(P\) values were used for significance testing. Ten items did not fit the model optimally. Four of the 10 assessed “indoor locomotion” skills and several of the ill-fitting items tapped the use of upper body for mobility (eg, getting in and out of chairs). These items are potential candidates for revision or removal from the scale in future work.

Nonetheless, we proceeded with a final model based on all 59 items, in which adequate fit was demonstrated for both the overall model and for persons. The final step was to standardize item parameters by rescaling the item slopes and thresholds to have a mean of 0 and standard deviation (SD) of 1 for the normative group (N=412). This is a standard metric used in most of the IRT estimation programs. We estimated IRT-based scores for the Mob-59 at each time period using weighted maximum likelihood (WML) estimation.\(^{57}\) The final IRT-based Mob-59 scores were standardized to a mean of 50 and SD of 10 in the normative group.

Differential item functioning. We examined the extent of differential item functioning (DIF) on the Mob-59 items to determine the extent to which item responses differed between children by clinical diagnosis or age. We generated logistic regression models for each mobility item on the total scale score and used Bonferroni-adjusted \(P\) values for significance testing (.05/59 items =\(P < .001\)). We also assessed the amount of model variance explained by the age and clinical diagnosis, taking into account total mobility score. If diagnosis or age produced significant model coefficients for a Mob-59 item, and the child variable explained more than 2\% of variance, considering total score, the item was considered to exhibit DIF.\(^{52}\) None of the 59 items functioned differently for both clinical diagnosis and age, although none were removed for this application. These items include: “Sits if supported,” “Sits on/off toilet,” “Climbs on/off adult-sized toilet,” “Gets on/off toilet, not needing arms,” “Gets in/out of chair, not needing arms,” “Manages seat belt,” “Sits if supported in tub or sink,” “Moves within room,” and “Moves between rooms.” Because some of these skills, such as toileting, are indeed a function of developmental age, stature, as well as mobility level, it seems consistent that age-related DIF is present. The current PEDI takes children’s development into account through the use of age-specific norms. Factors that might influence clinical diagnosis-related DIF would be adaptations used in toilet transfers, if completed, and differences in types of mobility (ambulation vs wheelchair use) between clinical and normative samples. Future efforts to improve or extend the range of the PEDI will have to consider the effects of DIF for transfer and mobility items. We have found summary scores on the PEDI to be quite robust, even though item calibrations between normative and clinical groups have been shown to differ.\(^{36}\)

CAT program. We based the Mob-CAT algorithms on the DYNHA software\(^\text{5}^{5}\) developed at QualityMetric.\(^{15,22}\) The Mob-CAT was designed to be completed by a child’s clinician or parent and can be administered from a personal computer.
programmed the Mob-CAT to use WML score estimation. We selected the item “Climbs or scoots in/out of tub” to be the first Mob-59 item administered to all respondents because its difficulty parameter was in the middle of the range, it did not exhibit DIF, and content seemed appropriate for most respondents. The response to the first item is fed into the DYNHA engine, and the application calculates a probable score, as well as a person-specific measure of how precise that score is. If the score is not estimated with sufficient precision, according to internal guidelines, additional questions are selected and administered until either the precision standard is reached or the defined maximum number of items has been administered. We defined the maximum number of items to be administered as 15. However, we believed that only a few respondents would need to complete that many items.

Rules for stopping the Mob-CAT were developed based on score precision. The mobility score distribution was divided into 5 intervals, and precision cutoffs of 95% CIs were developed for each category. For scores estimated to be less than 20 in a standardized T scale (mean ± SD, 50±10), a precision cutoff of 9 must be reached before the assessment will stop. Scores in the range of 20 to 30 must have a precision estimate of 6; scores ranging from 30 to 50 have a precision of 5; scores from 50 to 57, 6 points; and scores greater than or equal to 58, 12 points.

Mob-CAT Versus Mob-59

Computer simulations. We evaluated the IRT-based algorithms for Mob-CAT using real-data computer simulation methods for the analytic sample. The simulations compare the psychometric merits of alternative strategies for programming assessments. In these simulations using cases in the analytic dataset, responses to questions selected by the CAT software were entered into the computer to simulate the conditions of an actual CAT assessment. As in the actual Mob-CAT, the simulation uses the IRT model to select the best item to administer next; that is, the item with the highest information function given the current score level, reestimates the domain score and CI, and decides whether to continue testing. We developed 4 Mob-CAT–like scores in the simulations: a score directly comparable to a full CAT interaction based on meeting the specified precision levels (Mob-CAT) and 3 scores based on item-stopping rules, regardless of precision: a 15-item assessment (Mob-CAT-15), a 10-item assessment (Mob-CAT-10), and a 5-item assessment (Mob-CAT-5). These simulated scores were compared with a criterion standard—the actual IRT mobility latent trait score (Mob-59) estimated by the full model. The computer simulations of the Mob-CAT include an assessment of the number of items required to meet preset standards of measurement precision.

Cross-validation field test. The Mob-CAT and Mob-59 were administered to a sample of children with disabilities from the FHC clinical programs in the same manner typically used in that setting; that is, by clinical observation or through parent interview by the child’s physical therapist. For children without disabilities, we administered the Mob-CAT and Mob-59 via interview with the parent or the parent’s designee (in some cases the child’s teacher or day care worker). Only 1 model of Mob-CAT (using the preset levels of precision) was completed for comparison to the Mob-59 in the field study. We provided formal training in the administration of the Mob-CAT to the physical therapists. Fortunately, the clinical staff was already familiar with the Mob-59, because it is used in the FHC programs. For most children, both Mob-59 and Mob-CAT were completed during 1 session, and the maximum time interval between test modes for an individual child was 2 days. For both groups (children with disabilities and without disabilities), the order of assessment type was randomly alternated to avoid an order effect. After administration, we obtained written feedback from the physical therapist and/or parent respondent about the relative merits or limitations of both modes of administration. In this exit interview questionnaire, we asked 4 basic questions about preferences (Mob-59 vs Mob-CAT). Which version: (1) had more irrelevant questions, (2) seemed more burdensome, (3) provided more meaningful information, and (4) would you prefer in the future? We also allowed participants to provide written free-form comments on their experiences with both versions. We collected the actual time (to the closest minute) required for administration of the Mob-59 in 73% of the cases; the Mob-CAT had an internal clock to track the amount of time and the number of items needed to meet preset levels of precision. Demographic information (ethnicity, sex, age, diagnosis when applicable) was collected for each child. All procedures were approved by the institutional review boards at Boston University and FHC.

Analyses

Intraclass correlation coefficients (ICC(3,1)) between each of the CAT scores and the optimal IRT-based latent trait score (Mob-59) were calculated to assess the extent to which simulated CAT scores accurately reproduced the full-length form. The ability of Mob-CAT versions to discriminate between groups of children on the basis of diagnosis (normative vs clinical) as compared with Mob-59 was evaluated by comparing average scores and relative validity (RV) coefficients based on F ratios, as in previous studies. RV is the ratio of the F statistic for the measure in question divided by that for the best measure. The Mob-59 was established as the criterion standard and the RV ratio was set to 1.0. The comparability of simulated Mob-CAT–based estimates in measuring change over time was examined within a subsample of the analytic clinical sample (n=263), who had been administered the PEDI mobility scale more than once during their rehabilitation program. Average scores and RV coefficients based on F ratios were compared. To compare the relative precision of Mob-CAT scores with Mob-59 scores, we plotted the CIs in relation to the person’s ability scores. Finally, we used a series of 1-sample t tests to examine the difference between the number of items required in the Mob-CAT (simulation and field test) versus the Mob-59, and a series of paired t tests to examine differences in the amount of time needed for the Mob-CAT (internal clock) and Mob-59 (timing by test administrators) in the cross-validation study.

RESULTS

Score Agreement

ICCs between simulated score estimates of all of the CATs and the Mob-59 indicated a high degree of correspondence. As
seen in Table 2, the simulated Mob-CAT, the Mob-CAT-15, and the Mob-CAT-10 reproduce Mob-59 scores, with little information lost in going from a 15- to a 10-item CAT. The accuracy of the Mob-CAT-5 relative to the Mob-59 form is less than all of the other Mob-CAT forms.

Validity

The Mob-CAT-15 and the Mob-CAT-10 performed similarly to Mob-59 in discriminating between clinical and normative groups (Table 3). Discriminant accuracy of the Mob-CAT (using the preset precision rule) was slightly less than that of the Mob-CAT-15 (obtained by administering 15 items, eg, likely more than the preset precision Mob-CAT would require). Only the 5-item CAT simulation failed to approximate the Mob-59 in this test of known-groups validity.

Table 4 summarizes the results of the responsiveness comparisons. All of the CAT forms, except for the CAT-5, detected a similar magnitude of change from the first to second visit. The Mob-CAT simulation, CAT-15, and CAT-10 almost exactly reproduced the original Mob-59 change scores. However, the 5-item CAT form did not achieve the same level of precision in measuring change.

Score Precision

The 95% CIs around score estimates for the Mob-59 were, as expected, the smallest among the comparisons. For the CATs, the CIs became larger, because fewer items were used to estimate the overall score. The breadth of CIs across different score levels were higher (lower precision) in the 15-, 10-, and 5-item versions of the CAT than in the Mob-59 (Fig 1). For all forms, 95% CIs were greatest at extreme score ranges.

Time Burden

The mean number of items ± SD required for the Mob-CAT simulation was 8.2 ± 2.7 for the total analytic sample, which is less than 14% of the number of items in the Mob-59. The normative group required fewer items to meet the stopping rules (7.6 ± 3.1, or 13% of items) than the clinical group (8.6 ± 1.9, or 15% of the items). Because the mobility items tended to measure the mid-to-lower end of the scale, less precision was possible for respondents at the higher end, such as the normative group. The lower mean number of items administered by the normative group may also reflect the approach we have taken in building these CAT programs, in which we apply more relaxed CI requirements at the very extremes of the range. It is also notable that the smaller SD in the clinical group reflects less variation in the number of items required for a precise estimate of ability.

Table 5 summarizes the relative burden for 58 respondents in the field trial for whom information on time to administer was available. Overall, the Mob-CAT form yielded large decreases in respondent burden compared with the Mob-59, requiring about 15% the number of items and 21% the administration time of the full-length survey. The average time to complete the Mob-CAT was similar for the groups with and without disability (=1 min 13s). The difference between the number of items and amount of time required to complete the 2 different formats was significant in both clinical and normative groups by paired sample t tests (P < .001).

**DISCUSSION**

The results of these simulation and field study analyses suggest that CAT models yield accurate, valid, and efficient estimates of overall functional mobility scores in children. We are encouraged that the results of the field study are quite similar to the results of the simulation studies, particularly with respect to accuracy. Most previous CAT studies have only conducted simulations using existing datasets that include responses to all items in an instrument or item pool. The computer simulations are conducted under the assumption that answers to a subset of those items selected using CAT would be identical to the answers given if they were embedded in a self- or clinician-based administration. We have assumed that such simulations are likely very good approximations of actual CAT administrations, but the present study provides important evidence to support the quality of the CAT simulation algorithms.

The simulation studies suggest that very little sensitivity to change or ability to discriminate across known groups is lost with a Mob-CAT program that has at least 10 items. One
potential clinical concern related to CATs is that the fewer number of items administered may contribute to a loss of sensitivity to change or to the instrument’s ability to discriminate among groups. Our data from simulations suggest that a significant loss in validity of the scores from all items to a CAT format is not likely. It must be noted that we were not able to achieve acceptable levels of sensitivity or discrimination with the 5-item CAT; thus, a 5-item stopping rule is not recommended for these purposes. Although the initial data from the computer simulation studies are encouraging, these studies should clearly be replicated over repeated administrations in real-time field studies.

The precision estimates of the different CAT forms as depicted in figure 1 indicate an increase in CIs (loss of precision) around an individual score between the full-item version and CAT, and a gradual increase in CIs from the 15- to the 5-item CAT versions. Therefore, if high precision is needed for an individual score, 15 items or more may be needed. This loss of precision of the CAT (especially with the 10- and 15-item versions) at the individual level, however, did not substantially affect the group responsiveness, as noted above. The prespecified stopping rules of the Mob-CAT that we applied in this study could be modified for individual-level assessment to achieve smaller CIs. The CAT administers the minimum number of test items needed for each individual, depending on where his/her responses fit on the scale and how much precision in scoring is needed. The practical implications of CAT are apparent in the amount of time saved in administering the Mob-CAT versus the Mob-59 in the field study. Amazingly, the Mob-CAT took an average of less than 2 minutes to complete and required no more than 15 items. Using experienced clinicians who use the Mob-59 routinely, the Mob-59 took on average about 5 to 6 minutes. For clinicians who are not as familiar with the Mob-59, we anticipate the time requirements to be close to 10 minutes.24 (See table 5 for a comparison of times for the Mob-CAT and Mob-59.)

This time frame applied to both methods of administration used in the study, namely, having clinicians summarize a child’s function by running the CAT not during the point of clinical contact, and having clinicians interview parents at the point of contact. This efficiency of testing should have tremendous appeal for clinical programs that are trying to collect outcome data in a cost-effective manner. Based on the positive findings in the present study, we recommended further research with CAT-based dynamic functional assessments among children in rehabilitation settings.

We found that the original PEDI Mobility Functional Skills Scale, although clearly unidimensional, had some scale properties that were not ideal for CAT applications. A number of items violated assumptions of local independence, goodness of fit, and DIF. Even though we chose not to eliminate items based on these criteria for the CAT modeling, we found the 1-parameter Rasch scaling and resulting CAT programs to be quite robust. It is not clear to what extent scaling assumptions can be violated and still yield accurate CAT programs. In the development of the original PEDI, we erred on the side of including content that seemed to be critical for a meaningful clinical scale, even though items violated known scaling assumptions. In the future, we suggest conducting more research to determine how violations of IRT and Rasch scaling affect CAT algorithms.

Will the CAT program be accepted by parents and clinicians for routine assessment of children’s functional status? Based on an exit interview, 82% of the physical therapist and caregiver participants said that the Mob-59 was more burdensome and asked more irrelevant questions than the Mob-CAT. Even though the Mob-CAT was more efficient, 40% thought the Mob-59 version provided more meaningful information in a clinical situation. Although we can only surmise the reasons that some therapists and parents preferred the Mob-59, users may feel that more items assist in goal planning and developing specific treatment objectives, or that more items would help detect change. Several respondents noted that the Mob-59 was a good “tool for discussion with parents.” Although a few respondents noted that the Mob-59 seemed long and redundant, most respondents were quite positive about the Mob-59, because it provided more detailed functional information on the child than did the CAT version. More than 1 participant said that the longer version “seems like it would be more accurate,” in contrast to our analytic results that have demonstrated the score comparability of these 2 formats. In many cases, the full version of any assessment will be more accurate than a sample of items; however, the CAT technology appears to obtain large gains in efficiency for a relatively small loss in accuracy. The small loss in accuracy is due to the selection of items that are best for any 1 individual, rather than using the same items for everyone as in short-fixed format instruments. Overall, 62% of the respondents preferred the CAT for future administrations.

![Fig 1. Plot of 95% CIs around individual patient scores based on the 5-, 10-, and 15-item Mob-CAT versus all 59 items (Mob-59) on the PEDI Functional Mobility Skills Scale.](image)

**Table 5: Respondent Burden PEDI Mobility Scale Versions, a Cross-Validation Sample**

<table>
<thead>
<tr>
<th>Scale Versions</th>
<th>Norm Sample (n=41)</th>
<th>Clinical Sample (n=39)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean ± SD</td>
<td>Range</td>
</tr>
<tr>
<td>Mob-59 Time (min)</td>
<td>6.33±3.36</td>
<td>2.0–10.0</td>
</tr>
<tr>
<td>Mob-CAT Time (min)</td>
<td>1.21±0.58</td>
<td>0.3–3.1</td>
</tr>
<tr>
<td>No. of items</td>
<td>8.72±2.71</td>
<td>5.0–15.0</td>
</tr>
</tbody>
</table>

*Difference between 2 forms significant at P<.001.
From these exit comments, however, it is clear that both parents and clinicians will have to be assured that CAT technology can provide accurate and meaningful results. Depending on the intended purpose, perhaps assessments should provide users with options to conduct the full item test, sample items from selected content areas, or use the CAT technology to get an estimate of mobility functioning by using the least number of items. Future efforts are needed, not only to improve CAT programs, but, more important, to make the CAT interfaces and reports sensitive to the concerns of clinicians and parents who want to use the information to understand their child’s functional mobility status.

CONCLUSIONS

Score accuracy, validity, precision, and response burden were evaluated for simulated and real-time CAT administrations compared with a standard scale of pediatric functional mobility. We demonstrated good to excellent CAT score comparability and validity, but with reduced precision of individual scores in comparison to the full-length form. Only the 5-item stopping rule CAT did not meet acceptable levels of agreement, validity, or precision. The time savings of the CAT were apparent through both the computer simulation and the field studies. Clinician and parent participants generally favored the CAT approach versus the long-form, but a number of respondents expressed concern about the accuracy and clinical utility of the CAT scores. Future research is needed to examine how CAT programs work in practice to monitor functional recovery over time, to examine levels of precision needed for individual versus group analyses, and to develop interfaces and reports that support the clinical acceptability of CAT programs.

Acknowledgment: We extend our appreciation to John Ware for his support and guidance on this project. We also acknowledge Mark Kosinski for his advice in the early stages of this project.

References


 Suppliers
a. Muthen & Muthen, 3463 Stoner Ave, Los Angeles, CA 90066.
b. CITO, PO Box 1034, 6801 MG Arnhem, Netherlands.