Published in final edited form as:

Arch Phys Med Rehabil. 2008 April ; 89(4): 622–629. doi:10.1016/j.apmr.2007.09.053.

Assessing self-care and social function using a computer adaptive testing version of the Pediatric Evaluation of Disability Inventory Accepted for Publication, Archives of Physical Medicine and Rehabilitation

Wendy J. Coster, PhD, OTR/L, Stephen M. Haley, PhD, PT, Pengsheng Ni, MD, MPH, Helene M. Dumas, MS, PT, and Maria A. Fragala-Pinkham, MS, PT

From the Department of Occupational Therapy and Rehabilitation Counseling, Boston University Sargent College, Boston, MA (Coster); Health and Disability Research Institute, Boston University School of Public Health, Boston, MA (Haley, Ni); and Research Center for Children with Special Health Care Needs, Franciscan Hospital for Children, Boston, MA (Dumas, Fragala-Pinkham).

Abstract

Objective—To examine score agreement, validity, precision, and response burden of a prototype computer adaptive testing (CAT) version of the Self-Care and Social Function scales of the Pediatric Evaluation of Disability Inventory (PEDI) compared to the full-length version of these scales.

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

Settings—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); 38 children with disabilities and 35 children without disabilities (cross-validation sample).

Interventions—Not applicable.

Main Outcome Measures—Summary scores from prototype CAT applications of each scale using 15-, 10-, and 5-item stopping rules; scores from the full-length Self-Care and Social Function scales; time (in seconds) to complete assessments and respondent ratings of burden.

Results—Scores from both computer simulations and field administration of the prototype CATs were highly consistent with scores from full-length administration (all r 's between .94 and .99). Using computer simulation of retrospective data, discriminant validity and sensitivity to change of the CATs closely approximated that of the full-length scales, especially when the 15- and 10-item stopping rules were applied. In the cross-validation study the time to administer both CATs was 4 minutes, compared to over 16 minutes to complete the full-length scales.

Conclusions—Self-care and Social Function score estimates from CAT administration are highly comparable to those obtained from full-length scale administration, with small losses in validity and precision and substantial decreases in administration time.

Keywords

outcome assessment (Health Care); pediatrics; rehabilitation

The past decade has seen significant effort directed to improving the measures used to examine health and function in children with disabilities.^{1,2} These efforts reflect the convergence of multiple forces, including increased appreciation that the child's ability to perform important daily activities and to participate in important life situations is the outcome that matters most to families³ and increased emphasis by payers on documentation that services provided have resulted in progress toward these goals. The importance of sound measures of function has been further illustrated by research findings that interventions may be associated with meaningful functional improvement even in the absence of measurable changes in impairments.⁴

Measurement development has also been advanced by the introduction of newer methodologies, in particular those using Item Response Theory (IRT).⁵ These methods have supported clearer construct and item definition and the construction of scales that are sensitive to the smaller degrees of change across time often seen in children with disabilities. Nevertheless, IRT methods alone have been insufficient to address a key challenge for functional assessment: balancing comprehensiveness of coverage against practicality. In order to obtain sufficient coverage of the full range of function across the continuum of development and across degrees of disability, traditional fixed-length instruments tend to be so long as to be impractical for routine use in clinical settings. Alternatively, shorter instruments must sacrifice coverage, either by limiting the number of items (and therefore reducing sensitivity to change) or limiting the age span covered by the instrument (and thereby reducing the ability to track change across the full period of child development using the same instrument).

Recently, computer adaptive testing (CAT) methods have been proposed as a potential solution to this measurement dilemma.^{6–8} Adaptive testing approaches tailor the assessment to the current level of function of the child so that only items that yield useful information (i.e. are neither too hard nor too easy) are administered. In CAT administration, the program uses the response to an initial question to establish a general range of likely function. Subsequent questions are selected through application of algorithms to progressively refine the estimated score to the range of precision established a priori by the examiner. Regardless of the actual items administered all scores are on the same scale, which supports comparisons across time or across groups of individuals with different levels of current functional performance.

Although CAT offers a potential solution to the conflict between comprehensiveness and practicality, the reliability, validity, and acceptability of any application must still be demonstrated through appropriate testing. The purpose of this paper is to present results from a comparison of CAT results to full-length administration of two functional scales for children, one measuring self-care activity performance and the second measuring social function. Although there is some previous work examining CAT applications in the domain of functional mobility,^{9,10} to our knowledge there are no reports of investigation of the feasibility of CAT for measuring these other important domains in children.

The development of a CAT requires: (1) a large set of items (item pool) examining the functional area of interest; (2) items that scale consistently on a single dimension from low to high functional achievement; and (3) rules to guide starting, stopping, and scoring. Item response theory methods are used to create hierarchically organized item pools, after which software algorithms select items that match the child's estimated functional level. All respondents answer the same first question, which has been selected a priori based on its broad

coverage of the range of function. The response to the first question is used to estimate an initial score and confidence interval and guides selection of a second item within the estimated range. The response to this second question is used to re-estimate the score and confidence interval. The process continues in an iterative fashion until the computer algorithm determines that the stopping rule has been satisfied (either a pre-set number of items or a minimum confidence interval). The stopping rule can be altered to suit the specific purpose of measurement, e.g. a larger confidence interval may be acceptable for large population studies, whereas a narrow confidence interval might be important for the precision required in a clinical trial.

In the present study we created prototype CATs using the Self-Care and Social Function Functional Skills items from the Pediatric Evaluation of Disability Inventory (PEDI).¹¹ Two phases of testing were conducted using the prototype CATs: computer simulation studies of retrospective data and a prospective validation study. In addition to examining the accuracy and precision of the CATs compared to the standard fixed-form assessment, we also examined perceived respondent burden for each method.

Methods

Samples

Analytic sample—We used an existing database of 881 children who had complete data on the 73-item Self-Care and the 65-item Social Function Functional Skills scales of the PEDI. This retrospective analytic sample included two 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 youth (ages 6–7 years) who had received inpatient, outpatient, or school-based rehabilitation services at Franciscan Hospital for Children, Boston, MA. Of the 469 clinical cases, 249 had longitudinal data appropriate for sensitivity analyses for the Self-Care scale and 200 had data for the Social Function scale.

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 presented in Table 1. The sample size of 881 is acceptable for initial calibration work for a prototype CAT.¹²

Cross-validation sample—We recruited a convenience sample of 73 children and youth for the prospective cross-validation study. Thirty-eight children with disabilities, ages 1 year to 17 years, were recruited from the clinical programs (inpatient, outpatient, early intervention, and hospital-based school) at Franciscan Hospital for Children. Ethnic representation corresponding to the current United States 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 each of the following four impairment groups: congenital or inherited disease, growth and maturation disorders, acquired conditions, and traumatic injuries. Thirty-five 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 two Field-Test Coordinators.

Instrument

The PEDI¹¹ is a comprehensive functional assessment instrument that measures both capability and performance of functional activities. The Self-Care and the Social Function Functional Skills scales were used in the present investigation. Results of a CAT application

for the Mobility domain of the PEDI have been reported elsewhere.¹⁰ The Self-Care domain includes 73 activities involved in eating and drinking, grooming, dressing, and toileting tasks, which are assessed with a series of items using a dichotomous 'capable' or 'unable' scoring criterion. The Social Function domain includes 65 items related to communication (expression and comprehension), problem solving, interactions with peers and adults, and safety at home and in the community. Several studies have supported the reliability and validity of the PEDI scales in a wide variety of clinical samples.^{13,14} Evidence of construct validity has been obtained by demonstrating the ability of the PEDI to correctly identify children with and without disabilities,¹⁵ and to discriminate between different types of acquired brain injury.^{16,17} Studies also have reported successful outcome monitoring using the PEDI in children with cerebral palsy,^{18,19} myelodysplasia,^{20,21} osteogenesis imperfecta,²² and traumatic brain injury.^{23–27} The ability of the the PEDI Functional Skills scales to detect meaningful clinical changes has also been demonstrated.²⁸ Because the development of the PEDI scales and construction of summary scores are based on Rasch Rating Scale methodology,^{29–31} these scales provide an excellent starting point for the development of prototype CATs.

Development of the CAT

Unidimensionality and local independence—IRT and CAT methods assume certain measurement properties of item sets that purport to represent a functional construct (latent variable). These include the assumptions of unidimensionality, local independence, and stability of item parameters across groups (e.g., clinical versus normative samples). Item sets that violate these assumptions may be less effective in modeling the latent variable 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 one another.^{32,33} We tested the latent structure of the Self-Care and Social Function items in a series of confirmatory factor analyses³⁴ and evaluated item loadings and residual correlations between items using MPlus software.³⁵ We used weighted least squares means and variance adjusted estimation methods, which are more precise when analyzing moderate-size samples with skewed categorical data.^{34,36} To determine the extent to which a unidimensional model adequately represented scale structure we considered the eigenvalues associated with each factor extracted; item loadings on the primary factor; and results from overall model fit tests. To ensure adequate sample size for estimation of model parameters we combined the normative and clinical PEDI samples. Assuming the item parameters are similar across groups, combining the samples enhances generalizability of results across both groups and provides a greater number of persons at the moderate to low end of the scale to enhance precision of estimated scores in this region.

In the Self-Care domain, one factor explained 87.9% of the item variance and all the factor loadings were very high (range from 0.778 – 0.974). The Comparative Fit Index (CFI) value of 0.995 indicated very good fit and can be interpreted as an indicator that 99% of covariance in the data is reproducible by the model. This conclusion was supported by the Tucker-Lewis Index (TLI) value of 0.997, also indicating good fit. The Root Mean Square Error of Approximation (RMSEA) of 0.078 is in the acceptable range. In the Social Function domain one factor explained 87.8% of the item variance. All the factor loadings were very high, ranging from 0.770 to 0.987 and the fit indexes also supported the one-factor model (CFI=0.994, TLI=0.997, RMSEA=0.104).

The requirement of local independence means that scale items must be independent, or unrelated, to each other at a given score level. One indicator that items share more than the latent trait is high residual correlations. High residual correlations (greater than ± 0.2) were observed between 9 pairs of items on the Self-Care scale and 24 pairs on the Social Function

scale.³⁷ These correlations likely reflect the structure of the PEDI, which groups similar items into skill sets that have an implicit hierarchical relation to each other. For example, the item “Eats all textures of table food” implies accomplishment of the previous item “Eats cut up/chunky/diced foods” and thus the response to the more challenging item is not independent of the response to the easier item. This violation of model assumptions may affect the estimation of test information and item discrimination parameters, but cannot be rectified in an existing database.

Item calibrations—The item parameters for each scale were estimated using the Rasch model, which estimates the item difficulty parameters.^{38–40} The Rasch model was selected as the best solution for this phase of the project because of simplicity in interpretation and flexibility about the underlying form of the population or trait distributions. The item parameters and fit statistics were calculated using ConQuest,⁴¹ which is based on marginal maximum likelihood estimation. We evaluated fit using the fit statistics for each item based on the comparison of expected and observed value. To maximize 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).¹¹

In the Self-Care domain there were 4 items that did not fit the model including “allows nose to be wiped” (INFIT=1.52), “removes socks and unfastened shoes” (INFIT=1.60), “manages tangles and part hair” (INFIT=1.72) and “brushes or combs hair” (INFIT=1.68). Those items were removed from the item set to be used for the CAT prototype. In the Social Function domain only one item did not fit the model: “if upset because of a problem, child must be helped immediately or behavior deteriorates” (INFIT=1.81). Because of the important content reflected in this item we chose to keep it in the item pool. We estimated the individual scores using weighted maximum likelihood (WML)⁴² estimation. WML is preferable to the Expected a Posteriori (EAP) methods because it adjusts the first-order bias. The individual scores were standardized to a mean 50 and standard deviation of 10.

Differential item functioning—In item response theory, the child’s score on an item should depend entirely on the latent variable being measured. Significant differential item function (DIF) indicates that variables other than the latent variable, such as diagnosis, age, or gender, are likely influencing the response.⁴³ We used logistic regression to determine the extent to which item responses to the self-care and social function items differed by clinical diagnosis or age. The diagnosis variable was treated dichotomously (clinical; typical) while age was treated as a continuous variable. If diagnosis or age produced significant model coefficients and the child variable explained more than 2% of variance, considering the total score, then an item was considered to exhibit DIF. A Bonferroni corrected p-value was applied for significance testing ($p < 0.05/73 = 0.000685$ for the Self-Care domain; $p < 0.05/65$ items = 0.00077 for the Social Function domain). We also assessed the amount of model variance explained by the group variables.

One of the 73 Self-Care items (“remove socks and unfastened shoes”) exhibited DIF by diagnosis. This item also showed misfit on the previous analyses thus supporting the decision to remove this item. Sixteen of the 65 social function items exhibited DIF by diagnosis or age. There were 2 items that functioned differently for both diagnosis and age: “If upset because of a problem, child must be helped immediately or behavior deteriorates” and “Explores and functions in familiar community settings without supervision.” Because the problematic items represent important content, we did not remove them. However, these items are clearly candidates for future revision.

Development of the CAT program—We based the Self-care and Social Function CAT algorithms on the HDRI™ software developed at the Health and Disability Research Institute. The CATs were designed to be completed by a child's clinician or parent and can be administered from a stand-alone computer. We programmed the CATs to use weighted maximum likelihood (WML) score estimation.⁷ We selected the items "Puts on pants with an elastic waist" and "Provides names and descriptive information about family members" to be the first items administered to all respondents for the Self-care and Social Function CATs respectively. These items were chosen because their difficulty parameters were in the middle of the range, they did not exhibit DIF, and the content seemed appropriate for most respondents. The response to the first item is fed into the 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. In order to be able to compare results from the simulation and cross-validation studies we used a fixed-stopping rule of 15 items in the present project. However, we expected that only a few respondents would need to complete that many items to attain desirable levels of precision.

Accuracy of the CAT

Computer simulations—We evaluated the IRT-based algorithms for each CAT using computer simulation methods for the analytic sample. The simulations compare the psychometric merits of alternative strategies for programming assessments. In these simulations, responses to items selected by the CAT software were obtained for cases in the analytic data set and "fed" to the computer to simulate the conditions of an actual CAT assessment. As in an actual CAT, the simulation uses the IRT model to select the best item to administer next, i.e., the one with the highest information function given the current score level, re-estimates the domain score and confidence interval (CI), and decides whether or not to continue testing. In the present study, in order to be able to compare results from the simulation and cross-validation studies, we used a fixed-stopping rule of 15 items. We developed three CAT scores in the simulations to reflect 3 potential item-stopping rules (SC or SF-CAT-15, SC or SF-CAT-10 and SC or SF-CAT-5). These simulated scores were compared to a "gold standard" – the actual IRT latent trait score (Self-Care or Social Function) estimated by the full model.

Cross-validation field test—The Self-Care and Social Function CATs and full-length scales for each domain were completed on a sample of children with disabilities from the FCH clinical programs in the same manner typically used in that setting, that is by clinical observation or through parent interview conducted by the child's physical therapist. For children without disabilities, we administered both instruments via interview with the parent or the parent's designee (in some cases the child's teacher or day care worker). The CAT was completed using the pre-set 15-item stopping rule to enable comparison to scores from the full length scale. We provided formal training in the administration of the CAT to the physical therapists. The clinical staff was already familiar with the full length Self-Care and Social Function scales as they are used regularly in the programs at the facility. For most children, both the CAT and full length scale were completed during one session and the maximum time interval between test modes for an individual child was 2 days. For both groups (children with and without disabilities), the order of assessment type was counterbalanced to avoid an order effect. Following administration, we obtained written feedback from the physical therapist and/or parent respondent about the relative merits or limitations of both modes of administration. We collected the actual time (to the closest minute) required for administration of the full length scale in 73% of the cases; each CAT had an internal clock to track the amount of time and the number of items needed to meet pre-set levels of precision. Demographic information

(ethnicity, sex, age, and diagnosis when applicable) was collected for each child. All procedures were approved by the Institutional Review Boards at Boston University and Franciscan Hospital for Children.

Data Analysis

Pearson correlations were calculated between each of the CAT scores and the optimal IRT-based latent trait score (full length scale) to assess the extent to which simulated CAT scores were consistent with scores from the full length form. The ability of each CAT version to discriminate between groups of children on the basis of diagnosis (normative versus clinical) as compared to the full length scale was evaluated by comparing average scores and relative validity coefficients (RV) based on F-ratios, as in previous studies.⁴⁴ RV is the ratio of the F statistic for the measure in question divided by that for the best measure. The full length scale for each domain was established as the “gold standard” and the “Relative Validity” ratio was set to 1.0. The comparability of simulated CAT-based estimates in measuring change over time was examined within a sub-sample of the analytic clinical sample (N=249 for Self-Care; N = 200 for Social Function) who had been administered each PEDI scale more than once during their rehabilitation program. Average scores and relative validity coefficients based on F-ratios were compared. To compare the relative precision of the CAT scores to scores from the full length scales we plotted the confidence intervals in relation to the person ability scores. A series of paired *t*-tests was used to examine differences in the amount of time needed for each CAT (internal clock) and full length scale (timing by test administrators) in the cross-validation study.

Results

Score Agreement

As seen in Table 2, the descriptive statistics for scores from the 10- and 15-item simulation CAT were quite similar to those for the full item pool score for both the Self-Care and Social Function domains. The mean score of the 5-item CAT was higher than the full item pool score while the variance and range of the 5-item CAT score were smaller. The Pearson correlations between CAT scores and the full item pool scores were quite strong even in the 5-item simulation indicating that the CAT scores accurately captured the information in the original scales.

Score Precision

Examination of the standard errors and corresponding confidence intervals of different scores showed that the CAT-15 and CAT-10 had a similar pattern however standard errors of CAT-5-item were larger across all ranges. As expected, CAT-15 and CAT-10 standard errors somewhat larger than those from the full-length version because fewer items were used to calculate the overall score. These patterns are illustrated in Figure 1. For all methods, the standard errors were greater at extreme score ranges.

Validity

Discriminant accuracy of the 15- and 10-item CAT was very similar for both the Self-Care and Social Function domains, although the relative validity (RV) coefficients for the Social Function CATs were much closer to the RV of the full item pool. The coefficient for the 5-item Self-Care CAT simulation was considerably lower than for the 10- and 15-item CATs however the difference was not as pronounced for the Social Function 5-item CAT. (Table 3).

Table 4 summarizes the results of the responsiveness comparisons. RCI reflects the likelihood that the change in score from admission to discharge is due to real change rather chance

variation. An RCI value greater than 1.96 suggests it is unlikely ($p < 0.05$) the difference from admission to discharge is not reflecting real change.⁴⁵ For both Self-Care and Social Function, only the CAT-15 and full item pools had values that met this criterion. The relative validity (RV) ratios in both domains followed a similar pattern with the 15-item CAT having highest values followed relatively closely by CAT-10 and with CAT-5 values the lowest.

Cross-Validation Study

Results from administration of the prototype CATs and previous results from simulation studies were very similar. With administration of 10 or more items, the results from the CAT were very close to scores obtained with the full item pool in terms of precision. Correlations between prototype CAT scores and scores generated from the total item pool were only very slightly lower than the correlations obtained previously with the simulated CATs. (Table 5).

There were 38 children in the clinical group (mean age 8.7 years, range = 1.23~17.7) and 35 typical children (mean age 4.09 years, range = 0.42~7.5) in the sample. A general linear model that included age, group (1: clinical group, 0: typical group), and the interaction of age and group was used for analysis. Results showed a positive main effect of age indicating scores increased with chronological age. However, in the typical group the increase slope was much steeper than in the clinical group. There was no main effect of group, but there was a significant by group interaction (i.e. whether age had an effect depended on which group the child was in). These results may reflect the fact the most of the children in the clinical group were older so the expected age effect would be much less.

Comparing the response burden of the CAT administration to the paper form (full item pool), 81% of respondents said the paper version was more burdensome compared to 3% who found the CAT more burdensome. In fact, the average total time to administer both CATs was 3.9 minutes, compared to 16.49 minutes to complete both long forms (difference significant at $p < .0001$). In addition, 84% of respondents answered that the paper version asked more irrelevant questions than the CAT while only 4% gave the opposite response. An equal percent (37–38%) selected the CAT or the paper version as providing more meaningful information. Finally, 70% answered that they would be more likely to use the CAT in the future, compared to 6% who preferred the long paper form and 23% who said they would be equally likely to use either.

Discussion

The results of our analyses indicate that CAT models built from the PEDI self-care and social function item pools can provide accurate and valid estimates of children's functional capabilities while substantially reducing administrative burden compared to the full-length instruments. These results are consistent with previous research with CAT models for functional mobility¹⁰ and confirm that effective and efficient models can be developed for other domains of function important to children and families. Results from the field study were highly similar to those from the simulation studies in spite of the smaller number of participants in the cross-validation sample. These findings suggest that simulations may provide very good approximations of actual CAT administration.

Most disabling conditions in children affect self-care skill acquisition or performance and/or social development. There are also a number of significant clinical disorders that may affect these functional domains almost exclusively, such as autism spectrum disorders, emotional disorders, and intellectual disabilities, and others such as traumatic brain injury that may have significant impact across all three of the areas examined by the PEDI. Thus, it is important that measures developed to document outcomes of rehabilitation services examine content in each of these areas in order to provide an accurate and comprehensive picture of function and disability. The results from the present study are encouraging in that they demonstrate that the

goal of comprehensive coverage may be achievable without loss of precision or excessive administrative burden. Although further research is clearly needed, the results suggest that the PEDI-CAT offers the possibility of an outcome measure that could be usefully applied across diverse populations of children with disabilities.

As was found previously for the mobility CAT, the present results suggest that very little sensitivity to change or ability to discriminate across known groups is lost as long as the CAT program has between 10 and 15 items. However, the 5-item CATS were notably less accurate and sensitive and therefore would not be recommended for most purposes. In a CAT model using a stopping rule based on a desired level of score precision, it is quite possible that the scores of some individuals might be estimated with fewer than 10 items. One of the advantages of CAT is that it allows users to specify the level of score precision necessary for their current purpose. Thus, in individual assessment, where high precision is desirable, a 15-item stopping rule or a criterion reflecting a smaller degree of measurement error could be applied. On the other hand, for large scale studies where efficiency of administration is essential and less precision is required, even the 5-item CAT may be acceptable.

It is noteworthy that even the 15-item CAT substantially reduced the administration time required to complete both scales to an average of 4 minutes (combined). In contrast, completion of the entire PEDI questionnaire through parent interview typically takes between 30 and 45 minutes. The brief administration time of the CAT makes it far more feasible to conduct regular assessment of a child's functional status and may support alternative methods for administration such as telephone follow-up interviews that are not practical with the longer survey format. Parent respondents may also respond more positively to the assessment in the CAT format because they are asked fewer questions that are clearly irrelevant for their child.

The present analyses also identified a number of areas where further revision of the item pools would be appropriate. There were a substantial number of item pairs in the Social Function pool that did not meet the criterion for local independence as well as a smaller number in the Self-Care pool. This finding likely reflects the hierarchical organization of the 5-item sets within each original scale and suggests that some of these items should be dropped or re-worded to capture more distinct aspects of function in their respective areas. Further exploration should also be undertaken to understand the possible reasons for differential item function (DIF) by group in 16 of the Social Function items so that this problem can be addressed either by rewriting or dropping the items. Although such revisions would likely improve performance of the PEDI-CAT, our results suggest that the CAT is robust even when some items that violate scaling assumptions are retained. More direct investigation of the impact of various violations of Rasch and IRT assumptions on the performance of CAT algorithms would be extremely useful to guide future measurement efforts.

In a previous study with the mobility CAT,¹⁰ clinician respondents reported that they often used the context of completing the full-length PEDI in a parent interview to establish rapport and initiate discussion with families around the needs of their child. In the present study, when asked which version they found most informative, approximately equal percentages selected the CAT and the full-length version. These findings suggest that factors other than the time required for administration may be important determinants of clinicians' acceptance and use of assessments. These factors need to be considered carefully in future CAT work so that the CAT interface, interpretative supports, and reports are optimally designed to meet the needs of clinicians and families seeking information about a child's functioning for various purposes.

Conclusions

The results of the present study confirm that computerized adaptive testing methods can be applied successfully in two important domains of children's functioning that have not been examined previously. Although the content of the self-care and social function item pools was substantially different from the previously examined mobility domain, the results of the simulation and cross-validation studies were very similar. Thus, application of CAT methodology can substantially reduce the time required for administration without significant loss of precision or sensitivity to change. Although further work is recommended to refine the item pools in these two domains, the results suggest that the CAT approach offers a valid and viable solution to the long-standing conflict between the need for accuracy in clinical assessment and the equal need for practicality of administration.

Acknowledgements

Supported by the National Center on Medical Rehabilitation Research /NICHHD/NIH (grant nos. R43 HD42388-01, K02 HD45354-01A1).

References

1. Msall M. Tools for measuring daily activities in children: promoting independence and developing a language for child disability. *Pediatr* 2002;317–319.
2. Lollar D, Simeonsson R, Nanda U. Measures of outcome in children and youth. *Arch Phys Med Rehabil* 2000;81:S46–S51. [PubMed: 11128903]
3. Butler C. Outcomes that matter [editorial]. *Dev Med Child Neurol* 1995;37:753–754. [PubMed: 7589857]
4. Nordmark E, Jamlo GG, Hagglund G. Comparison of the Gross Motor Function Measure and Paediatric Evaluation of Disability Inventory in assessing motor function in children undergoing selective dorsal rhizotomy. *Dev Med Child Neurol* 2000;42:245–252. [PubMed: 10795563]
5. Hays R, Morales L, Reise S. Item response theory and health outcomes measurement in the 21st century. *Med Care* 2000;38:II-28–II-42. [PubMed: 10982088]
6. Ware J, Bjorner J, Kosinski M. Practical implications of item response theory and computerized adaptive testing. *Med Care* 2000;38:II 73–II 82.
7. Wainer, H.; Dorans, N.; Flaugher, R., et al. Computerized adaptive testing: A primer. Vol. 2nd ed.. Mahwah, NJ: Erlbaum; 2000.
8. Revicki DA, Cella DF. Health status assessment for the twenty-first century: item response theory, item banking and computer adaptive testing. *Qual Life Res* 1997;6:595–600. [PubMed: 9330558]
9. Dijkers M. A computer adaptive testing simulation applied to the FIM instrument motor component. *Arch Phys Med Rehabil* 2003;84:384–393. [PubMed: 12638107]
10. Haley SM, Raczek AE, Coster WJ, Dumas HM, Fragala-Pinkham MA. Assessing mobility in children using a computer adaptive testing version of the Pediatric Evaluation of Disability Inventory. *Arch Phys Med Rehabil* 2005;86:932–939. [PubMed: 15895339]
11. Haley, SM.; Coster, WJ.; Ludlow, LH., et al. Development, Standardization and Administration Manual. Boston, MA: Trustees of Boston University; 1992. Pediatric Evaluation of Disability Inventory.
12. Embretson, SE.; Reise, SP. Item response theory for psychologists. Mahwah, NJ: Laurence Erlbaum; 2000.
13. Wright FV, Boschen KA. The Pediatric Evaluation of Disability Inventory (PEDI): validation of a new functional assessment outcome instrument. *Can J Rehabil* 1993;7:41–42.
14. Nichols DS, Case-Smith J. Reliability and validity of the Pediatric Evaluation of Disability Inventory. *Pediatr Phys Ther* 1996;8:15–24.
15. Feldman AB, Haley SM, Coryell J. Concurrent and construct validity of the Pediatric Evaluation of Disability Inventory. *Phys Ther* 1990;70:602–610. [PubMed: 2217539]

16. Fragala MA, Haley SM, Dumas HM, Rabin JP. Classifying mobility recovery in children and youth with brain injury during hospital-based rehabilitation. *Brain Inj* 2002;16:149–160. [PubMed: 11839109]
17. Dumas HM, Haley SM, Ludlow LH, Rabin JP. Functional recovery in pediatric brain injury during inpatient rehabilitation. *Am J Phys Med Rehabil* 2002;81:661–669. [PubMed: 12172518]
18. Ostensjo S, Strinnholm M, Carlsson M, Dahl M. Everyday functioning in young children with cerebral palsy: functional skills, caregiver assistance, and modifications of the environment. *Develop Med Child Neurol* 2003;45:603–612. [PubMed: 12948327]
19. Ketelaar M, Vermeer A, Hart H, van Petegem-van Beek E, Helders PJ. Effects of a functional therapy program on motor abilities of children with cerebral palsy. *Phys Ther* 2001;81:1534–1545. [PubMed: 11688590]
20. Norrlin S, Strinnholm M, Carlsson M, Dahl M. Factors of significance for mobility in children with myelomeningocele. *Acta Paediatr* 2003;92:204–210. [PubMed: 12710647]
21. Tsai P, Yang T, Chan R, et al. Functional investigation in children with spina bifida-measured by the Pediatric Evaluation of Disability Inventory (PEDI). *Child's Nerv Syst* 2002;18:48–53. [PubMed: 11935244]
22. Engelbert RHH, Custers JWH, van der Net J, et al. Functional outcome in osteogenesis imperfecta: Disability profiles using the PEDI. *Pediatr Phys Ther* 1997;9:18–22.
23. Haley SM, Dumas HM, Ludlow LH. Mobility outcomes of children and adolescents in an inpatient rehabilitation program: Variation by diagnostic and practice pattern groups. *Phys Ther* 2001;81:1425–1436. [PubMed: 11509072]
24. Kothari DH, Haley SM, Gill-Body KM, Dumas HM. Measuring functional change in children with acquired brain injury: Comparison of normative and disease-specific scoring models using the Pediatric Evaluation of Disability Inventory (PEDI). *Phys Ther* 2003;83:776–785. [PubMed: 12940765]
25. Dumas H, Haley S, Rabin J. Short term durability and improvement of function in traumatic brain injury: a pilot study using the Pediatric Evaluation of Disability Inventory (PEDI) classification levels. *Brain Inj* 2001;15:891–902. [PubMed: 11595085]
26. Dumas HM, Haley SM, Bedell GM, Hull EM. Social function changes in children and adolescents with acquired brain injury during inpatient rehabilitation. *Ped Rehab* 2001;4:177–185.
27. Dumas HM, Haley SM, Fragala MA, Steva BJ. Self-care recovery of children with brain injury: descriptive analysis using the Pediatric Evaluation of Disability Inventory (PEDI) functional classification levels. *Phys Occup Ther Ped* 2001;21:17–27.
28. Iyer LV, Haley SM, Watkins MP, Dumas HM. Establishing minimal clinically important differences for scores on the Pediatric Evaluation of Disability Inventory for inpatient rehabilitation. *Phys Ther* 2003;83:888–898. [PubMed: 14519060]
29. Ludlow L, Haley S. New directions in pediatric rehabilitation measurement: The growing challenge. *J Outcome Meas* 2000;4:482–490. [PubMed: 11272597]
30. Ludlow L, Haley S. Effect of context in rating of mobility activities in children with disabilities: An assessment using the Pediatric Evaluation of Disability Inventory. *Educ Psychol Meas* 1996;56:122–129.
31. Haley SM, Ludlow LH, Coster WJ. Pediatric Evaluation of Disability Inventory: Clinical interpretation of summary scores using Rasch rating scale methodology. *Phys Med Rehabil Clin N Am* 1993;4:529–540.
32. Hambleton, RK.; Swaminathan, H.; Rogers, HJ. Fundamentals of item response theory. Newbury Park, California: Sage Publications; 1991.
33. Van der Linden, W.; Hambleton, R. Handbook of modern item response theory. Berlin: Springer; 1997.
34. Mislevy RJ. Recent developments in the factor analysis of categorical variables. *J Educ Stat* 1986;11:3–31.
35. Muthen, B.; Muthen, L. Mplus user's guide. Los Angeles: Muthen & Muthen; 1998.
36. Beauducel A, Herzberg PY. On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Struct Equat Model* 2006;13(2):186–203.

37. Tjur T. A connection between Rasch's item analysis model and a multiplicative Poisson model. *Scand J Stat* 1982;9:23–30.
38. Fischer, G.; Molenaar, I. Rasch models: Foundations, recent developments, and applications. Berlin: Springer-Verlag; 1995.
39. Andrich, D. Rasch models for measurement. Beverly Hills, CA: Sage Publications; 1998.
40. Masters GN. A Rasch model for partial credit scoring. *Psychometrika* 1982;47:149–174.
41. Wu, ML.; Adams, RJ. ConQuest (computer software and manual). Melbourne, Australia: Australian Council for Educational Research; 1998.
42. Warm TA. Weighted likelihood estimation of ability in item response theory. *Psychometrika* 1989;54:427–450.
43. Hariharan S, Rogers HJ. Detecting differential item functioning using logistic regression procedures. *J Educ Meas* 1990;27:361–370.
44. McHorney CA, Ware JE, Lu JF, Sherbourne CD. The MOS 36-item short-form health survey (SF-36): III. Tests of data quality, scaling assumptions and reliability across diverse patient groups. *Med Care* 1994;32:40–66. [PubMed: 8277801]
45. Jacobson NS, Truax P. Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *J Consul Clin Psychol* 1991;59(1):12–19.

Figure 1.
Plot of standard errors of individual subject scores based on 5-, 10- and 15-item simulated CAT compared to full item pool (Self-Care domain)

Standard Error **Figure 2.**

Plot of standard errors of individual subject scores from 5-, 10- and 15-item prototype CAT compared to full item pool (Self-Care domain)

Table 1

Demographic characteristics of samples

	Analytic Sample	Cross-validation Sample
Age Range	6 mos. – 17 yrs.	6 mos. – 18 yrs.
% Female	45.2	49.3
% Hispanic or Latino	9.3	5.5
% Asian	1.5	5.5
% Other	5.8	4.1
% Black or African American	14.6	2.7
% White	68.8	82.2
Total Sample Size	881	73

Table 2
Comparison of scores from simulated CAT and full item pool

	Self-care				Social Function			
	Mean	SD	Range	Correlation	Mean	SD	Range	Correlation
Full item pool								
CAT-15	49.91	8.89	31.45–62.78	-	50.60	9.61	27.50–67.33	-
CAT-10	49.95	8.85	31.58–62.48	0.99	50.66	9.71	27.52–67.21	0.99
CAT-5	49.94	8.77	31.85–62.17	0.99	50.63	9.68	27.62–67.18	0.99
	50.14	7.96	37.03–61.49	0.97	50.98	7.89	38.65–62.26	0.95

Table 3
Between-group discrimination (normative vs clinical) by simulated CAT and full item pool for Self-Care and Social Function scales

	Normative Group		Clinical Group		Group Difference		
	Mean	SD	Mean	SD	F	RV	
Self-Care	N=412		N=446				
Full item pool	52.59	7.66	47.44	9.24	79.39*	1.00	
CAT-15	52.50	7.54	47.60	9.31	72.08*	0.91	
CAT-10	52.48	7.52	47.60	9.20	71.74*	0.90	
CAT-5	52.32	7.41	48.12	7.92	64.64*	0.81	
Social Function	N=412		N=399				
Full item pool	52.82	7.90	48.31	10.64	47.22*	1.00	
CAT-15	52.86	7.94	48.38	10.79	45.72*	0.97	
CAT-10	52.85	7.93	48.34	10.76	46.39*	0.98	
CAT-5	52.76	7.21	49.15	8.14	44.89*	0.95	

* p<0.001

Table 4
Sensitivity to change of simulated CAT and full item pools for Self-Care and Social Function domains.

	Visit 1			Visit 2			Change			RCI			RV
	Mean	SD		Mean	SD		Mean	SD		Mean	SD	F	
Self-Care (N=249)													
Full item pool	46.13	10.36		51.85	10.41		5.73	7.47		2.59	2.92	146.41*	1
CAT-15	46.39	10.53		51.97	10.40		5.57	7.51		2.10	2.52	137.12*	0.94
CAT-10	46.38	10.34		51.91	10.39		5.54	7.54		1.80	2.22	134.10*	0.92
CAT-5	47.27	8.66		52.07	9.27		4.81	6.67		1.17	1.48	129.28*	0.88
Social Function (N=200)													
Full item pool	46.81	12.58		51.56	11.43		4.75	7.21		2.47	3.27	86.68*	1.00
CAT-15	46.93	12.83		51.83	11.62		4.90	7.55		2.13	2.96	84.27*	0.97
CAT-10	46.91	12.80		51.68	11.73		4.78	7.51		1.77	2.52	81.00*	0.93
CAT-5	48.30	9.09		51.62	8.67		3.32	5.58		0.85	1.33	70.90*	0.82

* :<0.001

Table 5
Comparison of scores from prototype CAT and full item pool

	Self-care				Social Function			
	Mean	SD	Range	Correlation	Mean	SD	Range	Correlation
Full item pool	52.32	7.61	35.33–62.79	-	55.55	8.86	34.99–67.23	-
Actual CAT-15	52.45	7.52	35.56–62.49	0.99	55.59	9.31	33.78–67.21	0.98
Actual CAT-10	52.39	7.79	34.53–62.19	0.98	55.53	9.40	33.78–67.18	0.98
Actual CAT-5	51.83	7.79	37.08–61.52	0.95	54.73	8.18	37.88–62.18	0.94