Published in final edited form as:
J Pediatr Orthop. 2010 ; 30(1): 71. doi:10.1097/BPO.0b013e3181c85453.

Validity and Reliability of Physical Functioning Computer Adaptive Tests for Children with Cerebral Palsy

Abstract

Background—The purpose of this study was to assess the concurrent validity and reliability of scores from four new parent-report computer adapted testing (CAT) programs developed to measure physical functioning of children with cerebral palsy (CP). The Shriners Hospitals for Children (SHC) CP-CAT battery includes upper extremity skills (UE), lower extremity and mobility skills (LE), activity (ACT), and global physical health (GPH).

Methods—This was a prospective study of 91 children with CP who were tested cross-sectionally and 27 children with CP who were administered the CP-CAT programs twice within approximately a one-month interval. We examined concurrent validity of the four SHC CP-CAT programs by Pearson correlations with comparative parent-report instruments. Scale reliability was tested by developing estimates of marginal reliability; test-retest reliability was assessed by intraclass correlations.

Results—Pearson correlations were moderate to high in matching content domains of the CATs with the comparison measures. Marginal reliability estimates were always better for the CAT program than the comparative instruments. Average test-retest reliability using Intraclass correlations across the four CATs was $ICC_{3,1}$ = 0.91 with a range of 0.88–0.94.

Conclusions—We found the CAT scores to be related to expected domains from external instruments, to have good scale reliability and to have stable scores as determined by test-retest reliability. These results support using parent-report CATs in the assessment of physical functioning in children with CP.

Level of Evidence—This is a level II prospective study designed to establish the validity and reliability of computer adaptive testing as an evaluation method.

Keywords

computerized adaptive testing; assessment; outcomes; cerebral palsy

INTRODUCTION

In 2006, the Shriners Hospitals for Children (SHC) embarked on a major effort to evaluate new technology for assessing physical functioning in children with cerebral palsy (CP). Children with cerebral palsy make up the largest volume of patients receiving care within the SHC system. By developing a series of parent-reported computer adaptive testing (CAT) programs for children with CP, SHC has begun to develop a contemporary and efficient method to collect functional outcome information on the children in their system. One end-point of clinical research and outcome management is the development of efficient assessments of functional activities observed either by the parent or caregiver. These functional instruments hold great promise for clinical, research and surveillance purposes.[1] The use of functional outcome measures is consistent with an overall expansion of outcomes within the SHC organization from technical measures such as radiographs or gait analysis, to more global functional domains that can provide information about how the child completes daily activities.[2]

The search for the "one best" functional instrument for use in the SHC research and clinical activities for children with CP across all ages has been elusive.[1, 2] A common functional outcome instrument would yield important research and clinical benefits to SHC and the children and families that are served. On the clinical side, a common functional outcome instrument would enable clinicians to monitor the impact of medical, surgical and rehabilitation interventions on an ongoing basis across severity levels and facilities.[3, 4] Enhanced monitoring should result in better program decisions and improved service delivery for children and help in the identification of evidence-based clinical pathways. On the research side, a uniform approach towards measuring functional outcomes should improve the comparability of measures across facilities and research projects. To date, it is not possible to compare scores across different measures directly because a common metric linking instruments does not exist.[5] The lack of comparability limits the generalizability of study results and may slow down the adoption of promising interventions from research into future clinical practice. In addition, a uniform outcome instrument for use in the SHC research and clinical programs for CP would minimize a number of practical barriers. Since the spectrum of severity and physical functioning limitations in children with CP is broad, the specific content domains of interest may not be found within any one instrument. Often, many instruments are currently needed to match content, different levels of severity, age groups and body systems that are affected (e.g., upper and lower extremity). Using an assortment of instruments may create redundancy and wasted effort on the part of the family and the testing staff and complicates credentialing and training for a large clinical staff. In many instances, clinicians simply lack the time for administration of "the best" instrument due to time constraints or length of the instrument.

The advent of contemporary measurement technology and computerized adaptive testing (CAT) for health care applications has offered an alternative to traditional, fixed length instruments. CAT platforms are built from a set of coordinated items (item banks) that define a common dimension. Each test administration is adapted to the unique ability level of each child. The basic notion of an adapted test is to mimic what an experienced clinician would do. A clinician learns most when he/she directs questions at the child's approximate level of proficiency. Administering functional items that are either too easy or too hard provides little information.

An adaptive test first asks questions in the middle of the ability range, and then directs questions to an appropriate level based on the responses without asking unnecessary questions. This allows for fewer items to be administered, while gaining precise information regarding a child's placement along a continuum of functional ability. CAT applications require large item banks in any one functional domain, contain items that consistently scale along a dimension of low to high functional proficiency, and have rules guiding starting, stopping and scoring procedures.[6] A strategy of matching items to respondents has been used to achieve short and precise educational and psychological tests for decades.[7]

Before the CAT programs can be used confidently for monitoring patients or for clinical research, it is necessary that their reliability and validity are established. The use of the CAT platform is expected to lead to a uniform system of assessment, reduced testing burden, improved precision of scores and enhanced sensitivity towards identifying progress related to medical, surgical and rehabilitation interventions. SHC has built four CATs to assess the physical functioning of children with CP, which include the content domains of upper extremity skills (UE)[8] lower extremity skills and mobility (LE),[9] activity (ACT)[10] and global physical health (GPH).[11] In previous work, we have shown that a physical functioning CAT was both feasible and efficient when used at the SHC.[12] In this study, we examine the concurrent validity and reliability of the four CATs developed for children with CP in comparison to fixed-length parent-reported outcome instruments (legacy measures) typically used at the SHC. Our goal

was to examine the psychometric properties of the CAT programs when used prospectively in three SHC facilities.

METHODS

We examined concurrent validity using a cross-sectional design. We collected parent-report data on a convenience sample of 91 children with CP across three SHC hospital outpatient clinics (Montreal, Philadelphia, Springfield, MA). Inclusion criteria were parents or caregivers of children with a known diagnosis of CP, ages 2–20 years. Participants were excluded if their child had received surgical or pharmacological interventions within the past six months. For the test-retest reliability sample, parent-report data were collected on a convenience sample of 27 children with CP on two occasions approximately one-month apart. There were 18 children who participated in both the concurrent validity and test-retest reliability samples. Demographic characteristics of the samples are presented in Table 1.

Items from the four CAT programs were administered to parents using a PC-based tablet. Both English and Spanish CAT versions were available. In many cases, parents completed the survey during their child's clinic visit or therapy session at one of the SHC outpatient clinics. Parents who were unable to complete the CAT programs during the clinic visit completed them at home using a web-based interface. Trained therapy and research staff were available to answer any questions regarding the study protocol or the interpretation of items. All items were completed by a parent or caregiver. For concurrent validity, the CAT was administered along with the following legacy measures: Pediatric Outcomes Data Collection Instrument (PODCI),[13] the Functional Independence Measure for Children (Wee-FIM ™),[14] the Pediatric Quality of Life Inventory Cerebral Palsy Version (PedsQL-CP™),[15] and the Functional Assessment Questionnaire (FAQ)[16] Due to the time constraints within outpatient clinical appointments, the series of legacy measures were not completed for every child. For the test-retest reliability sample, the four CATs were repeated with an average time interval of about a month (mean = 31.8 days; SD=14.4; range = 11–74). A large majority (n=20) of the participants in the test-retest sample completed one or both CAT test-retest sessions via the Internet.

The LE-CAT has an item bank of 85 items with content reflecting basic mobility, transfers and ambulation skills. The UE-CAT consists of 46 items reflective of skills in self-care, writing, manipulation of objects, and use of environmental control devices (e.g. switches). The ACT-CAT item bank consists of 45 items incorporating activities of daily living (ADL), instrumental-ADLs, and sports, play and recreation activities. The GPH-CAT consists of 37 items pertaining to pain, fatigue, drooling, and joint stiffness. The LE, UE and ACT-CATs were developed based on a unidimensional model, while the GPH-CAT was built on a multidimensional bi-factor model. Each CAT stopping rule was pre-set at 15 items. The development and refinement of the item banks and CAT programs have been described in detail elsewhere.[16, 17]

To serve as concurrent validity comparisons, the PODCI[13] (n=66), Wee-FIM [14] (n=42), PedsQL-CP [15] (n=60) and the FAQ[16] (n=61) were also completed on a sub-set of children. The PODCI was developed specifically to assess changes following pediatric orthopedic interventions for a broad range of diagnoses. The Wee-FIM is a standard outcome measure traditionally used in many of the SHCs. The FAQ is a parent report questionnaire developed at Gillette Children's Specialty Healthcare that covers a broad range of walking and gross motor activities. We used the FAQ 22-item survey that encompasses a variety of high-level mobility and activity skills. The PedsQL-CP is adapted from the generic PedsQL, and developed specifically for children with CP. The PedsQL-CP overall score can be considered as a measure of the global health of children with CP.

We analyzed the concurrent validity of the CATs and legacy measures using Pearson correlation coefficients. Specifically, Pearson correlations coefficients were calculated between scores from the LE-CAT and the PODCI basic mobility core sub-scale (11 items), the Wee-FIM transfers and locomotion sub-scale (5 items), the Wee-FIM motor scale (13 items) and the 22-item FAQ; between the UE-CAT and the PODCI upper extremity sub-scale (8 items) and the Wee-FIM self-care sub-scale (5 items); between the ACT-CAT and the PODCI sports sub-scale (12 items) and the PedsQL-CP daily activity sub-scale (9 items); and between the GPH-CAT and the PedsQL-CP overall measure (35 items). Intraclass correlation coefficients were calculated for the four CP CATs across the two test-retest reliability occasions. To examine internal consistency, we used marginal reliability calculations that are specific to item response theory (IRT) and allow us to compare legacy forms with the CATs. Marginal reliabilities are similar to cronbach alpha used in classical measurement theory in that it is a measure of how well items within a domain relate to each other. Marginal reliabilities can only be calculated on unidimensional models, and therefore were only calculated for LE-CAT, UE-CAT, and ACT-CAT.[18, 19] We used a bi-factor model on the GPH-CAT, which precluded our ability to place legacy measures on the same scale and make marginal reliability comparisons between the GPH-CAT with legacy measures.

RESULTS

Pearson correlations were moderate to high in matching content domains of the LE, UE, ACT and GPH CATs to the legacy measures. Correlations ranged from $r = 0.59$ to 0.91 with an average of $r = 0.81$. See Table 2.

Average test-retest reliability using Intraclass correlations across the four CATs was 0.91 with a range of 0.88–0.94. The individual CAT ICCs and 95% confidence intervals (CI) are LE-CAT = 0.96 (CI_{95} = 0.92 – 0.99); UE-CAT = 0.86 (CI_{95} = 0.71 – 0.94); ACT-CAT = 0.88 (CI_{95} = 0.72 – 0.95) and GPH-CAT = 0.94 (CI_{95} = 0.84 – 0.98). In all cases, marginal reliability estimates of similar content all favored the CATs versus the legacy measures. The marginal reliability for the three CATs tested was 0.97, 0.95, and 0.95 for the LE-CAT, UE-CAT, and ACT CAT, respectively (Table 3). Marginal reliabilities for each of the legacy measures were lower ranging from 0.71 to 0.91.

DISCUSSION

In addition to previous work in testing CATs in pediatric clinical environments,[12] this study adds to the general findings that CAT programs are valid and reliable for clinical use. The series of four CAT programs developed for the SHC show strong concurrent validity with common parent-reported legacy measures that are routinely used in clinical care and research. These correlations suggest that the SHC CATs represent much of the content that has been used in previous fixed-format instruments at the SHC. Moreover, due to the fundamental principles of CAT, items are more accurately matched to children's functioning thereby providing for a more meaningful assessment as compared to traditional measures. The "bottom line" benefit to CAT may reside in the potential for CAT to be more efficient than traditional measures.

The International Classification of Functioning, Disability and Health[20] recommends that outcome measures are selected such that all aspects of health and disability are examined; function, participation, and aspects of quality of life. For example, the PODCI is one of the few outcome tools which measures function, participation (subscale sports) and some aspects of quality of life (subscales pain and global satisfaction). The correlations of the CATs to the parallel subscales of the PODCI clearly indicate that these new CAT instruments can cover all important content domains in one efficient system. Because the CATs have a large item bank

behind them and only questions that are near a child's functional level are given,, the CATs can be more precise as well as take up 40% less to time complete than traditional paper and pencil questionnaires.[12] In this sample, the average time to complete a CAT was just over 4 minutes.

We did find two areas of the legacy measures that did not correlate as well as expected with CAT scores. There was a poor correlation (r=0.59) between the PedsQL-CP and GHP-CAT overall scores. The GPH CAT has mainly fatigue and pain items, whereas the PedsQL-CP covers a much wider spectrum of content. Second, the FAQ-22 did not correlate with the LE-CAT (r = 0.78) as well as expected. This may be due to the concentration of some very difficult ambulation and mobility items on the FAQ and fairly low levels of precision for children who have low mobility functioning, while the LE-CAT has a broader range of mobility content. Despite this two lower than expected correlations, overall the correlations were moderate to good and therefore the CATs are appropriate for clinical and research use.

Our test-retest reliability results were very promising, particularly since the data were collected in some real world conditions that might have affected the results negatively. For example, many of the test-retest reliability cases were conducted with the Spanish version of the instrument, which has not been as fully tested as the English version. Also, some participants used a mixed mode of administration. In about one-half the cases, the only feasible method to get test-retest data within a month's interval was to administer one of the tests during a clinic visit, and have one of the tests completed at home by Internet. Although all aspects of the test administration looked identical between PC tablet in the clinic and the web-based program, setting differences may have influenced the accuracy of the data. Despite these issues, the CAT scores were remarkably stable. Due to the relatively small sample, we are unable to separate out possible bias effects of language version and setting, but consideration should be given to address these issues in future work. These findings do suggest that home Internet use for administration may be feasible. The potential benefits of performing at-home assessments instead of burdening the family to conduct the assessments during busy clinical visits are potentially significant.

Scale reliability of the CAT was determined by marginal reliability. Marginal reliability is an index of how precise the instrument is overall, similar to a cronbach alpha used in classical measurement theory with higher reliabilityestimates correlating with smaller standard errors around person scores. For all three CATs tested, the marginal reliability was higher than the corresponding legacy measures. This was true even when the number of items in the legacy measures was greater than the number of items used for the CATs, which limits the computer to 15 questions.

This study demonstrated the validity and reliability of four new CAT programs developed at the SHC to measure the physical functioning of children with CP. It is worth noting again that these data come from prospective studies in which the CATs were used in busy clinical and research environments and at home via the Internet. We believe that these findings are very promising and should lead to the further testing and utilization of these CAT products in future clinical research endeavors.

Acknowledgments

Supported by the Shriners Hospital for Children Foundation (grant no. 8957) and an Independent Scientist award to Dr Haley (National Center on Medical Rehabilitation Research/NICHD/NIH, grant no. K02 HD45354-01A1).

References

1. McCarthy ML, Silberstein CE, Atkins EA, et al. Comparing reliability and validity of pediatric instruments for measuring health and well-being of children with spastic cerebral palsy. Develop Med Child Neurol 2002;44(7):468–476. [PubMed: 12162384]
2. Oeffinger DJ, Tylkowski CM, Rayens MK, et al. Gross Motor Function Classification System and outcome tools for assessing ambulatory cerebral palsy: a multicenter study. Develop Med Child Neurol 2004;46(5):311–319. [PubMed: 15132261]
3. Finkenflugel HJM, van Maanen V, Schut W, et al. Appreciation of community-based rehabilitation by caregivers of children with a disability. Disabil Rehabil 1996;18(5):255–260. [PubMed: 8743304]
4. Msall ME. Tools for measuring daily activities in children: promoting independence and developing a language for child disability. Pediatr 2002:317–319.
5. McHorney CA, Cohen AS. Equating health status measures with item response theory: illustrations with functional status items. Med Care 2000;38(Suppl II):II-43–II-59. [PubMed: 10982089]
6. Wainer, H. Computerized Adaptive Testing: A Primer. Mahwah, NJ: Lawrence Erlbaum Associates; 2000.
7. van der Linden WJ. The changing conception of measurement in education and psychology. Appl Psychol Meas 1986;10(4):325–332.
8. Tucker CA, Montpetit K, Bilodeau N, et al. Assessment of children with cerebral palsy using a parent–report computer adaptive test: I. Upper extremity skills. Develop Med Child Neurol. in press.
9. Tucker CA, Gorton GE, Watson K, et al. Assessment of children with cerebral palsy using a parent–report computer adaptive test: II. Lower extremity and mobility skills. Develop Med Child Neurol. in press.
10. Haley S, Fragala-Pinkham M, Dumas H, et al. Evaluation of an item bank for a computerized adaptive assessment of physical activity in children with cerebral palsy. Phys Ther 2009;89:589–598. [PubMed: 19423642]
11. Haley SM, Ni P, Dumas HD, et al. Measuring global physical health in children with cerebral palsy: illustration of a bi-factor model and computerized adaptive testing. Qual Life Res 2009;18:359–365. [PubMed: 19221892]
12. Mulcahey MJ, Haley SM, Duffy T, et al. Measuring Physical Functioning in Children With Spinal Impairments With Computerized Adaptive Testing. J Pediatr Orthop April/May 2008;28(3):330–335.
13. Daltroy LH, Cats-Baril WL, Katz JN, et al. The North American Spine Society Outcome Assessment Instrument: reliability and validity tests. Spine 1996;21(6):741–749. [PubMed: 8882698]
14. Guide for the Functional Independence Measure for Children(WeeFIM) of the Uniform Data System for Medical Rehabilitation, Version 4.0—Community/Outpatient. Buffalo, NY: State University of New York at Buffalo; 1993.
15. Varni JW, Burwinkle TM, Berrin SJ, et al. The PedsQL in pediatric cerebral palsy: reliability, validity, and sensitivity of the Generic Core Scales and Cerebral Palsy Module. Develop Med Child Neurol 2006;48(6):442–449. [PubMed: 16700934]
16. Tucker CA, Haley SM, Watson K, et al. Physical function for children and youth with cerebral palsy: Item bank development for computer adaptive testing. J Pediatr Rehabil Med 2008;1:237–244. [PubMed: 19779597]
17. Dumas H, Watson K, Fragala-Pinkham M, et al. Using Cognitive Interviewing for Test Items to Assess Physical Function in Children with Cerebral Palsy. Pediat Phys Ther 2008;20(4):356–362.
18. Green B, Bock R, Humphreys L, et al. Technical guidelines for assessing computerized adaptive tests. J Ed Measur 1983;21:347–360.
19. Wang W-C, Chen P-H. Implementation and measurement efficiency of multidimensional computerized adaptive testing. Appl Psychol Meas 2004;28:295–316.
20. World Health Organization. International Classification of Functioning, Disability and Health (ICF). Geneva: World Health Organization; 2001.

Table 1

Demographic Data for Concurrent Validity and Test-retest Reliability Samples

	Total	Concurrent Validity				Test-retest
	N = 91	**PEDSQL N=60**	**PODCI N=66**	**Wee-FIM N=42**	**FAQ N=61**	**N = 27**
Age Range	2–20	2–20	2–20	3–20	2–20	2–19
Mean Age yrs; (sd)	9.9 (4.12)	9.2 (3.6)	9.6 (3.9)	10.0 (3.8)	9.1 (3.8)	10.2 (3.8)
Female count; (%)	39 (42.9)	26 (43.3)	30 (45.5)	17 (40.5)	26 (42.7)	8 (29.6)
Ethnicity count; (%) Hispanic or Latino	38 (41.8)	27(45.0)	28 (42.4)	13 (31.9)	25 (41.0)	7 (26.9)
Race Asian count; (%)	1 (1.1)	1 (1.7)	1 (1.5)	1 (2.49)	1 (1.6)	
African American count; (%)	5 (5.5)	2 (3.3)	2 (3.0)	1 (2.49)	2 (3.3)	2 (7.4)
Caucasian count; (%)	54 (59.3)	34 (56.7)	39 (59.1)	28 (66.7)	35 (57.4)	18 (66.7)
American Indian/Alaskan Native count; (%)	3 (3.3)	2 (3.3)	3 (4.5)		2 (3.3)	
Other/unknown count; (%)	28 (30.8)	21 (35.0)	21 (31.8)	12 (28.6)	21 (34.4)	7 (25.9)
Gross Motor Functional Classification System (GMFCS) count; (%)	(n=86)					(n=25)
I	14 (16.3)	9 (15.0)	11(16.7)	7 (16.7)	10 (16.3)	6 (24.0)
II	15 (17.4)	8 (13.3)	11(16.7)	10 (23.8)	9 (14.8)	5 (20.0)
III	32 (37.2)	22 (36.7)	23 (34.8)	13 (31.0)	22 (36.1)	7 (28.0)
IV	13 (15.1)	10 (16.7)	9 (13.6)	7 (16.7)	9 (14.8)	4 (16.0)
V	12 (14.0)	11 (18.3)	12 (18.2)	5 (11.9)	11 (18.0)	3 (12.0)

Table 2

Pearson Correlation Coefficients between CAT and Legacy Measures

	Correlation coefficients			
	LE- CAT	**UE CAT**	**ACT-CAT**	**GPH CAT**
PODCI basic mobility (11 items)	0.88			
Wee-FIM motor (13 items)	0.89			
Wee-FIM transfers and locomotion (5 items)	0.91			
FAQ (22 items)	0.78			
PODCI upper extremity (8 items)		0.85		
Wee-FIM self-care (5 items)		0.82		
PedsQL-CP activity (9 items)			0.80	
PODCI sports (12 items)			0.83	
PedsQL-CP overall (35 items)				0.59

Table 3

Marginal Reliability Estimates for CATs and Selected Legacy Measures

	N	Marginal Reliability
LE-CAT	91	0.97
FAQ-22	61	0.82
Wee-FIM mobility and transfers	42	0.88
Wee-FIM motor	42	0.85
PODCI basic mobility	66	0.87
UE-CAT	91	0.96
PODCI upper	66	0.91
Wee-FIM self-care	42	0.74
ACT-CAT	91	0.95
PODCI Sports	66	0.81
PEDSQL Daily Activity	55	0.71