Precision and content range of a parent-reported item bank assessing lower extremity and mobility skills in children with cerebral palsy

GEORGE E GORTON III[1] | KYLE WATSON[2] | CAROLE A TUCKER[3] | FENG TIAN[4] | KATHLEEN MONTPETIT[5] | STEPHEN M HALEY[6] | MARY J MULCAHEY[1]

1 Shriners Hospitals for Children, Springfield, MA, USA. **2** Shriners Hospitals for Children, Philadelphia, PA, USA. **3** Temple University, Philadelphia, PA, USA. **4** Health and Disability Research Institute, School of Public Health, Boston University, Boston, MA, USA. **5** Shriners Hospitals for Children, Montreal, Quebec, Canada. **6** Health and Disability Research Institute, School of Public Health, Boston University, Boston, MA, USA

Correspondence to George E Gorton III, Shriners Hospitals for Children, 516 Carew Street, Springfield, MA 01104, USA. E-mail: ggorton@shrinenet.org

PUBLICATION DATA

Accepted for publication 25th November 2009.

Published online

LIST OF ABBREVIATIONS

CAT	Computer adaptive test
FAQ	Functional Assessment Questionnaire
IRT	Item response theory
LE45	Lower extremity physical functioning and mobility skills 45-item bank
LE85	Lower extremity physical functioning and mobility skills 85-item bank
PODCI	Paediatric Outcomes Data Collection Instrument
WeeFIM	Functional Independence Measure for Children

AIM The aim of this study was to determine the psychometric properties, content range, and measurement precision of a lower extremity physical functioning and mobility skills item bank (LE85) in children with cerebral palsy (CP).

METHOD Lower extremity functioning and mobility skill items were administered to 308 parents of children (169 males, 139 females; mean age 10y 8mo, SD 4y) with spastic CP (145 diplegia, 73 hemiplegia, 89 quadriplegia; [for one person type of CP was unknown]) classified using the Gross Motor Function Classification System (75 level I, 91 level II, 79 level III, 37 level IV, 26 level V). Additional legacy measures were administered to assess concurrent validity. Psychometric characteristics, differential item functioning, content range, and score precision were examined.

RESULTS The LE85 had acceptable psychometric properties. Content range matched the ability range of the sample population and exceeded legacy measures with minimal differential item functioning. The LE85 had good correlation with the Paediatric Outcomes Data Collection Instrument, Functional Independence Measure for Children, Gillette Functional Assessment Questionnaire, and Paediatric Quality of Life Inventory – CP module (range r=0.63–0.86). Precision of the LE85 and 10-item simulated computer adaptive test scores outperformed legacy measures.

INTERPRETATION The LE85 appears to be suitable to administer as a computer adaptive test to measure lower extremity physical functioning and mobility in children with CP.

Improved methods of measuring the functional outcomes of medical, surgical, and therapeutic treatment of children with cerebral palsy (CP) are critical to guiding treatment planning. Existing outcome measures suffer from limited content range, poorly defined measurement precision, poor match between test content and person ability, and can be burdensome to administer.[1–4] When multiple domains of function are of interest, the probing that can be accomplished effectively at one time is limited.

Some challenges inherent in traditional paper-based outcomes testing can be addressed using computer adaptive testing technology.[5] Computer adaptive tests (CATs) are based on item response theory (IRT), which provides a means to calibrate sets of items, termed item banks, based on the information each item contributes towards characterizing the underlying test construct (e.g. physical function ability). After each response, CAT software selects the next question from the item bank that provides the greatest information about the level of ability of that particular person until a reliable estimate of ability is obtained. CATs have been shown to be as precise as, and often more efficient than, traditional tests in measuring outcomes for children with developmental disabilities.[6,7]

The physical function construct often includes skills related to lower extremity functioning and mobility; items are chosen to reflect a continuum of abilities. Measurement is optimal when the difficulty of the items in the bank closely matches the functioning levels of the individuals in the anticipated sample.[8] The extent to which items capture the full range of lower extremity skills and mobility functioning is termed the content range. Content range can be examined by defining how well items measure function along the continuum of the measurement scale, particularly at the high and low ends of the scale.

Precision of the estimates of a person's functional ability level (person score) is defined by the score's mean standard error (SE).[9] Precision is affected by the number of test items, the quality and discrimination ability of items, and the match between item location and person ability levels. Clearly, lower SEs for person scores are desirable in a measurement instrument.

DOI: 10.1111/j.1469-8749.2010.03615.x

The psychometric properties of a lower extremity physical functioning and mobility skills 45-item bank (LE45) have been described previously.[10] The LE45 has been shown to be unidimensional, with acceptable item fit statistics based on calibration responses from 190 parents of children and adolescents with CP. Content range across a broad range of ability with a negligible ceiling effect and no floor effect was demonstrated. Scores based on the LE45 discriminated across severity levels and type of CP and correlated well with existing measures in routine clinical use (legacy measures). Using simulated CAT administration, scores derived from answering fewer items specifically targeted at the respondent's ability level showed excellent correlation with scores derived from the full 45-item set.

The LE45 did not include items related to the use of wheelchair or assistive devices. The purpose of this study was to determine the psychometric properties of an expanded item bank consisting of the LE45 items supplemented by additional wheelchair and assistive device items, to further examine content range, and to determine the measurement precision of the expanded item bank.

METHOD

Participant selection

A convenience sample of 308 parent respondents of children with spastic CP between the ages of 2 and 20 years (mean 10y 8mo, SD 4y), who had not undergone surgical interventions, botulinum toxin injections, or serial casting within the last 6 months, was recruited from the clinical programmes of four paediatric hospitals (Shriners Hospitals for Children in Montreal, Philadelphia, and Springfield, and the Franciscans Hospital for Children, Boston). Approval was obtained from each hospital's local institutional review board, and appropriate consent and authorization were obtained from respondents before participation.

Participants comprised 145 individuals with diplegia, 73 with hemiplegia, and 89 with quadriplegia (for one person CP type was unknown). Functional mobility, as classified by the Gross Motor Function Classification System (GMFCS),[11] included 75 children at level I, 91 at level II, 79 at level III, 37 at level IV, and 26 children at level V. Functional dexterity, as classified by the Manual Ability Classification System (MACS),[12] included 93 children at level I, 117 at level II, 53 at level III, 18 at level IV, and 19 children at level V (for eight persons the MACS level was unknown). Participant demographics are shown in Table I.

Lower extremity item-bank development

The lower extremity and mobility skills item bank was developed through review of existing assessments of function and mobility, taking into consideration medical, surgical, and therapeutic interventions directed at improving basic mobility and lower extremity function in children with CP across a span of functional levels. The content of the final item bank includes items reflecting transfers, basic mobility, and ambulation skills with and without assistive devices. Initial reports of item-bank development and calibration are described in detail elsewhere.[10,13–15]

What this paper adds

- LE85 measures lower extremity physical functioning and mobility in children with CP, including those using assistive devices.
- Psychometric properties are acceptable.
- Item response theory-derived scores correlate highly with legacy measures.
- Excellent precision and content over a broad range of ability when administered as a 10-item simulated computer adaptive test.

Item calibration

Sample size estimation for calibration is dependent on the IRT model used and whether acceptable levels of error and item fit are achieved. Because of the relatively small sample size in this study, we calculated item parameters and ability estimates with a one-parameter graded response model with fixed slope. Calibration continued until the SEs of ability estimates for modelled items decreased to an acceptable level and item fit statistics were satisfactory.

The lower extremity and mobility skill items were administered to parents, along with items from three other item banks, using a computer tablet with custom-written software that also provided introductory information and instructions. Items explored capacity (what the child was able to do) using the stem 'my child can…'. The recall timeframe was within the past 7 days. Responses utilized a 5-point rating scale focusing on difficulty (unable to do, can do with much difficulty, with

Table I: Demographics of children of parent respondents

Age range	2–20y
Mean age (SD)	10y 8mo (4y)
Age groups, *n* (%)	
<5y	17 (5.5)
5–9y	114 (37.0)
10–14y	123 (39.9)
15–19y	50 (16.2)
≥20y	4 (1.3)
Sex, *n* (%)	
Female	139 (45)
Male	169 (55)
Ethnicity	
Non-Hispanic or Latino	281 (91.2)
Hispanic or Latino	25 (8.1)
Other	2 (0.6)
Race, *n* (%)	
Caucasian	260 (84.4)
African–American	27 (8.8)
Asian	9 (2.9)
Other	12 (3.9)
Gross Motor Functional Classification System level, *n* (%)	
I	75 (24.4)
II	91 (29.6)
III	79 (25.7)
IV	37 (12.0)
V	26 (8.4)
Type (*n*=307), *n* (%)	
Diplegia	145 (47.2)
Hemiplegia	73 (23.8)
Quadriplegia	89 (29.0)

Values are based on *n*=308, unless otherwise specified.

some difficulty, with a little difficulty, and without any difficulty). Filter questions were used to eliminate irrelevant items. For example, parents answered the 15 walker items only if their child used a walker.

A subset of parents also completed the Paediatric Outcomes Data Collection Instrument (PODCI;[16] n=176), the Functional Independence Measure for Children (WeeFIM;[17] n=120), the Functional Assessment Questionnaire (FAQ;[18] n=170), and the Paediatric Quality of Life Inventory CP version[19] (n=81) to assess concurrent validity. The PODCI Transfers and Basic Mobility scale, WeeFIM Transfers and Locomotion subscale, FAQ Walking Scale and 22 Skill Items, and Paediatric Quality of Life Inventory CP version overall score, daily activity, school activity, and move and balance subscales were compared with scores from simulated CAT administration of the lower extremity physical functioning and mobility skills 85-item bank (LE85) as evidence of concurrent validity. The study relied on parental report and did not include independent assessment by a physical therapist. Gross motor function and manual dexterity were classified by study personnel using the GMFCS and MACS.

Statistical analysis procedures

IRT methods were used to refine the LE45 previously described,[10] as well as to incorporate 43 items related to wheelchair and assistive device use, resulting in a final item bank of 85 items (LE85). IRT methods assume that item banks meet criteria for unidimensionality and local independence. Unidimensionality was confirmed using exploratory and confirmatory factor analysis. Local independence was confirmed by examining the residuals of inter-item correlations between pairs of items using Mplus software (v5.21; Muthen & Muthen, Los Angeles, CA, USA).[20] Fit statistics (comparative fit index, Tucker–Lewis index, and root mean square error of approximation) for a one-factor model using confirmatory factor analysis in Mplus were examined. Item parameters and fit statistics were estimated with PARSCALE[21] using the graded response model, with slope restricted to a fixed value across all items. We evaluated item fit using the likelihood ratio χ^2 statistic ($p<0.05$ suggest item misfit). Ability scores were estimated using weighted maximum likelihood estimation[22] and were standardized to a mean of 50 and standard deviation of 10 (T-scale) to create a more interpretable score.

All respondents answered core items; assistive device items were answered only by respondents whose children used specific devices. Common item equating was used to set parameters of device and core items on the same metric to permit direct comparison. Person ability was first defined using the core lower extremity items with no assistive device questions. The core item difficulty estimates were then used as 'anchors' for the difficulty estimates of 15 items related to crutch or cane use, 13 items related to walker use, and 15 items related to wheelchair use relative to the core anchor items. In the same way, 11 PODCI items, five WeeFIM items, and 22 FAQ items were cocalibrated with the core lower extremity items as anchor items. This resulted in one ability–difficulty scale, with all people and items on the same metric.

Differential item functioning

Item responses should depend only on the latent variable of interest (lower extremity physical functioning and mobility). If the response is dependent upon other factors, biased estimates can occur, with items showing different measurement characteristics (differential item functioning; DIF) based on the other factors.[23] DIF was examined using ordinal logistic regression to evaluate whether the pattern of item responses was influenced by factors such as sex, age, type of CP, and GMFCS level. If the likelihood ratio test was statistically significant and the R^2 change was greater than 0.035, then DIF was assumed to be present.

Content range

To produce reliable and precise estimates for a wide range of person abilities, the range of difficulty of the items in the scale must be consistent with the range in abilities of the intended population. The best person ability estimates will occur when the content of the items is at or near the ability level of the person. Items that are much too easy or much too hard add little information to the estimate of a person's ability level. The content range of the final item bank was examined using item maps, item difficulty and person ability plotted simultaneously on the same scale. For each item, the estimated ability level for each of the item's five responses was used in the item map.

CAT real data simulation

Simulation is often used for investigating the merits of administering an item bank within a CAT framework based on the calibration data set. We simulated CAT administration using software developed at the Health and Disability Research Institute (Boston, MA, USA). The CAT software[6] includes options for item selection, score estimation using weighted likelihood,[22] and stopping rules based on number of items, level of precision, or both. As items were 'administered', responses were taken from the calibration data set. The first item administered had appropriate content and mid-range difficulty. Iteratively, the next items administered provided the greatest information at the estimated score based on the previous item(s). Stop rules, determining when the CAT ended, were based on the number of items (5, 10, and 15 items) administered, and did not specify target score precision. We report results from a 10-item CAT simulation (CAT-10) in comparison with the full item bank.

Score precision

Score precision quantifies how well an item bank estimates each ability level across a range of ability levels. We examined the precision of the LE85 item bank and of scores from the simulated CAT-10 by calculating the SEs at each ability level. Additionally, we calculated precision for the five WeeFIM items, 11 PODCI items, and 22 FAQ items for comparison. To provide a context for the size of the SE, an SE of 3.2 corresponds approximately to a reliability coefficient of 0.90, and an SE of 2.2 corresponds to a reliability coefficient of 0.95.[9,24]

RESULTS

Initially, an LE45 with no assistive device questions was reported.[11] Through analysis of our expanded calibration sample, three items were removed based on poor item fit ('When in a seated position, my child can turn his/her head towards objects or sounds', 'My child can roll onto his/her side', and 'Which statement best describes your child's balance and falls?'). The remaining 42 core items met criteria for unidimensionality and local independence and underwent subsequent IRT analyses. A comparative fit index of 0.981, a Tucker–Lewis index of 0.996, and a root mean square error of approximation of 0.088 indicate acceptable fit to a unidimensional model. No item had a residual inter-item correlation of more than 0.20, demonstrating local independence. The final bank has 85 items, consisting of 42 core items, 15 items related to crutch or cane use, 13 items related to walker use, and 15 items related to wheelchair use. Assistive device items explore a range of environments and situations common to daily life. Filter questions were used to determine inclusion of appropriate assistive devices items for respondents.

Item fit

Acceptable item fit statistics were found for the LE85. Two walker items and nine core items demonstrated some misfit but were kept after expert review because of the importance of their content. Estimated item difficulties for the LE85 are shown in Appendix S1 (supporting information published online). A subset of the 42 core items is shown in Figure 1, portraying average item difficulties along the spectrum of lower extremity physical functioning and mobility.

Concurrent validity

Scores based on the LE85 show excellent correlation with the FAQ walking scale (r=0.82, 95% confidence interval (CI) 0.77–0.87, p<0.001, n=170) and average score of the 22-item skill set (r=−0.86, 95% CI −0.8 to −0.91, p<0.001, n=90), as well as the PODCI transfers and basic mobility subscore (r=0.85, 95% CI 0.81–0.89, p<0.001, n=176) and the WeeFIM lower extremity subscore (r=0.83, 95% CI 0.77–0.88, p<0.001, n=120). Good correlation is also shown with the Paediatric Quality of Life Inventory CP version overall score (r=−0.76, 95% CI −0.65 to −0.84, p<0.001, n=81) and the daily activity (r=−0.68, 95% CI −0.55 to −0.79, p<0.001, n=81), school activity (r=−0.63, 95% CI −0.48 to −0.75, p<0.001, n=81), and movement and balance (r=−0.73, 95% CI −0.62 to −0.82, p<0.0011, n=81) subscores.

Differential item functioning

No DIF was found for any of the 42 core items, for the crutch or cane items, or for the walker device items. One wheelchair item showed DIF for age ('transfer between a seat and his/her wheelchair'), one for GMFCS level ('wheeling, moves in line

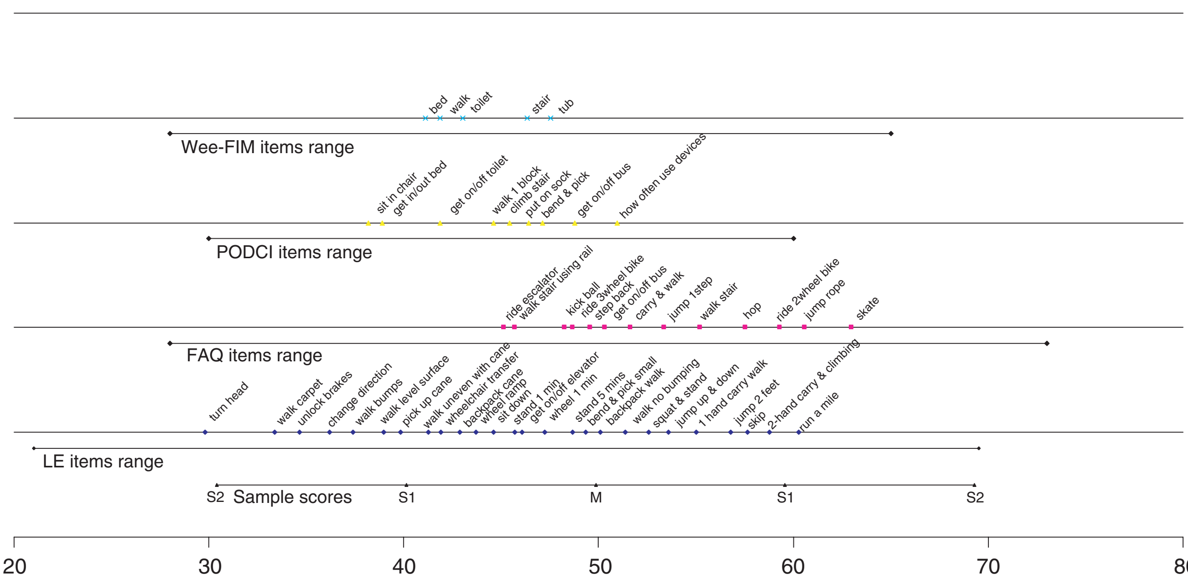

Figure 1: Selected items from the lower extremity physical functioning and mobility skills 85-item bank (LE85) are shown in parallel with items from the Functional Assessment Questionnaire (FAQ) 22-item skill set, Paediatric Outcomes Data Collection Instrument (PODCI) Transfers and Basic Mobility scale, and Functional Independence Measure for Children (WeeFIM) Transfers and Locomotion subscale. The location of each item represents the mean difficulty level of the item; respondents of this ability level will have a 50% likelihood of responding such that they are able to complete the item. Items are arranged in order of difficulty along a scale normalized to a mean of 50 and standard deviation of 10, where 50 represents the mean ability level of the sample population as well as the mean difficulty level of the items. The content range includes the difficulty levels of each response option and is shown as a line beneath each item map. The content range stretches from the least difficult response option of the least difficult item to the most difficult response option of the most difficult item.

without bumping into other people'), and two for both MACS and GMFCS level ('wheel across level outdoor surfaces', 'wheeling, can change direction'). Because of the limited number of respondents for the wheelchair questions and the importance of the content, these items were retained.

Scale coverage

The LE85 items range from simple, such as 'lying on back, turns head from side to side', to complex mobility items, such as 'runs a mile without stopping'. The items cover the full range of person ability with good distribution of items throughout the entire range. No area within the ability range of the sample population has limited content. Device items provided supplemental content at the easier end of the ability range, as seen in Figure 1. Comparatively, the range of item difficulty provided by the five WeeFIM items, the 11 PODCI items, or the 22 FAQ items is limited (Fig. 1). The five WeeFIM and the 11 PODCI items fell within the middle to the easier end of the ability range. The 22 FAQ items were at the higher end of the ability range.

Score precision and standard error

SEs of person ability scores based on the LE85 and the 10-item CAT outperformed the 22-item FAQ skill set, 11-item PODCI set, and five-item WeeFIM set in terms of range of abilities for which the score is at an excellent level of reliability (Fig. 2). The LE85 item bank demonstrated a SE of less than 2.2 over a wide range of person ability levels spanning nearly two standard deviations above and below the mean ability of the population. The 10-item CAT based on the 85-item set performed similarly, with slightly reduced range but excellent precision across the entire target ability range of the sample population.

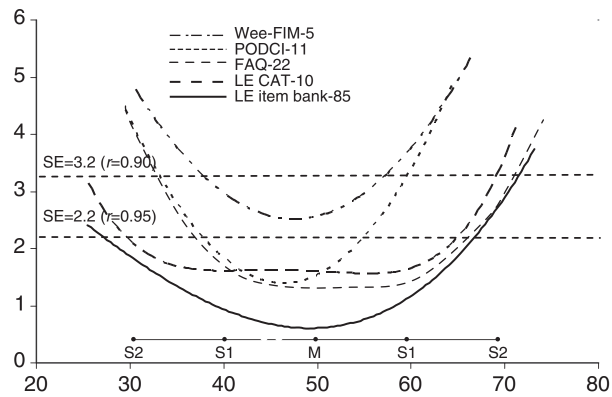

Figure 2: The precision (standard error [SE]) of the lower extremity physical functioning and mobility skills 85-item bank (LE85) is shown (solid line) compared with a simulated 10-item version (wide dashed line), the Functional Assessment Questionnaire (FAQ) 22-item skill set (medium dashed line), 11 Paediatric Outcomes Data Collection Instrument (PODCI) items (narrow dashed line), and five Functional Independence Measure for Children (WeeFIM) items (dash dotted line) along the spectrum of difficulty normalized to a mean of 50 and a standard deviation of 10. An SE of 3.2 logits is equivalent to a traditional reliability coefficient (*r*) of 0.90, and an SE of 2.2 corresponds to a reliability coefficient of 0.95. Each instrument varies, based on its content, in the content range and level of precision that can be achieved. Instruments with more items, such as the LE85, can achieve greater precision than instruments with fewer items. Administering the LE85 in an adaptive format (LE-CAT-10) retains the high precision and wide content range of the full bank in a more efficient manner, using only a few well-selected questions.

DISCUSSION

Our previous work on a lower extremity physical functioning and mobility skills item bank[10] showed good preliminary psychometrics but contained insufficient responses to calibrate wheelchair or assistive device items. The current study examined content range and measurement precision of an expanded 85-item bank and calibrated the supplemental device items. The LE85 item bank has excellent correlation with legacy measures and has acceptable item fit statistics, unidimensionality, and local independence. Content range matched the ability range of the sample population. Precision of the LE85 and a 10-item simulated CAT outperformed legacy measures over a broad range of ability.

Children with CP who use assistive devices frequently view independent ambulation without assistive devices as a goal. Progression in assistive device use follows a fairly typical sequence. Children with significant lower extremity involvement and limited mobility skills may require the use of a wheelchair. A highly supportive adaptive walker or gait trainer is considered a higher functioning level. With continued progression, a walker may be used with development to crutches or cane before independent ambulation without assistive devices. Inclusion of wheelchair and other items using assistive devices increased the content range of the item bank and provided the opportunity to capture this aspect of progressive ambulation.

Simulated 10-item CATs based on the LE85 exhibited high concurrent validity with existing measures. Precision across the breadth of ability was better than existing measures, including the WeeFIM, PODCI, and FAQ. The LE85 exhibited higher precision than existing measures in the lower difficulty range, suggesting that it has improved performance in capturing ability level in more involved children. Precision in the mid-range, nearest the mean of the sample population, was very high, even when administered with as few as 10 items as a CAT. Precision estimates were based on a simulated CAT administration, potentially overestimating precision. Future work will prospectively evaluate the performance of different-length versions of the CAT in this population.

Our expanded LE85 demonstrates excellent performance when implemented as a simulated CAT, with improved content coverage and precision when compared with similar existing measures of lower extremity physical functioning and mobility skills. Future research is planned to assess the responsiveness of the CAT in children with CP before and after interventions, and to generalize the measure for other clinical populations. Our long-term goal is to create a system of CATs to measure the functional status and outcomes of children and adolescents with CP based on a set of item banks that assess the constructs of global physical health, upper extremity skills and dexterity, lower extremity skills and mobility, and activity in a clinical setting as well as for clinical research.

ACKNOWLEDGEMENTS

This study was supported by Shriners Hospitals for Children (grant 8957) and an Independent Scientist Award to Dr Haley (National Centre on Medical Rehabilitation Research/National Institute of Child Health and Human Development/National Institutes of Health, grant K02 HD45354-01A1).

ONLINE MATERIAL

The following information is available for this article online:
Appendix SI: Items and item difficulties.

REFERENCES

1. Bagley AM, Gorton G, Oeffinger D, et al. Outcome assessments in children with cerebral palsy, part II: discriminatory ability of outcome tools. *Dev Med Child Neurol* 2007; **49:** 181–6.
2. Oeffinger D, Gorton G, Bagley A, et al. Outcome assessments in children with cerebral palsy, part I: descriptive characteristics of GMFCS Levels I to III. *Dev Med Child Neurol* 2007; **49:** 172–80.
3. Oeffinger DJ, Tylkowski CM, Rayens MK, et al. Gross Motor Function Classification System and outcome tools for assessing ambulatory cerebral palsy: a multicenter study. *Dev Med Child Neurol* 2004; **46:** 311–9.
4. Sullivan E, Barnes D, Linton JL, et al. Relationships among functional outcome measures used for assessing children with ambulatory CP. *Dev Med Child Neurol* 2007; **49:** 338–44.
5. Wainer H. Computerized Adaptive Testing: A Primer, 2nd edn. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.
6. Haley SM, Ni P, Ludlow LH, Fragala-Pinkham MA. Measurement precision and efficiency of multidimensional computer adaptive testing of physical functioning using the paediatric evaluation of disability inventory. *Arch Phys Med Rehabil* 2006; **87:** 1223–9.
7. Haley SM, Raczek AE, Coster WJ, Dumas HM, Fragala-Pinkham MA. Assessing mobility in children using a computer adaptive testing version of the paediatric evaluation of disability inventory. *Arch Phys Med Rehabil* 2005; **86:** 932–9.
8. Lai JS, Cella D, Chang CH, Bode RK, Heinemann AW. Item banking to improve, shorten and computerize self-reported fatigue: an illustration of steps to create a core item bank from the FACIT-Fatigue Scale. *Qual Life Res* 2003; **12:** 485–501.
9. Hambleton RK, Swaminathan H, Rogers H. Fundamentals of Item Response Theory. Newbury Park, CA: Sage Publications, 1991.
10. Tucker CA, Gorton GE, Watson K, et al. Development of a parent-report computer-adaptive test to assess physical functioning in children with cerebral palsy I: lower-extremity and mobility skills. *Dev Med Child Neurol* 2009; **51:** 717–24.
11. Palisano R, Rosenbaum P, Walter S, Russell D, Wood E, Galuppi B. Development and reliability of a system to classify gross motor function in children with cerebral palsy. *Dev Med Child Neurol* 1997; **39:** 214–23.
12. Eliasson AC, Krumlinde-Sundholm L, Rosblad B, et al. The Manual Ability Classification System (MACS) for children with cerebral palsy: scale development and evidence of validity and reliability. *Dev Med Child Neurol* 2006; **48:** 549–54.
13. Dumas HM, Watson K, Fragala-Pinkham MA, et al. Using cognitive interviewing for test items to assess physical function in children with cerebral palsy. *Pediatr Phys Ther* 2008; **20:** 356–62.
14. Tucker CA, Haley SM, Dumas HM, et al. Physical function for children and youth with cerebral palsy: item bank development for computer adaptive testing. *J Pediatr Rehabil Med* 2008; **1:** 245–53.
15. Tucker CA, Montpetit K, Bilodeau N, et al. Development of a parent-report computer-adaptive test to assess physical functioning in children with cerebral palsy II: upper-extremity skills. *Dev Med Child Neurol* 2009; **51:** 725–31.
16. Daltroy LH, Liang MH, Fossel AH, Goldberg MJ. The POSNA paediatric musculoskeletal functional health questionnaire: report on reliability, validity, and sensitivity to change. Paediatric Outcomes Instrument Development Group. Paediatric Orthopaedic Society of North America. *J Pediatr Orthop* 1998; **18:** 561–71.
17. Ottenbacher KJ, Msall ME, Lyon NR, Duffy LC, Granger CV, Braun S. Interrater agreement and stability of the Functional Independence Measure for Children (WeeFIM): use in children with developmental disabilities. *Arch Phys Med Rehabil* 1997; **78:** 1309–15.
18. Novacheck TF, Stout JL, Tervo R. Reliability and validity of the Gillette Functional Assessment Questionnaire as an outcome measure in children with walking disabilities. *J Pediatr Orthop* 2000; **20:** 75–81.
19. Varni JW, Burwinkle TM, Berrin SJ, et al. The PedsQL in paediatric cerebral palsy: reliability, validity, and sensitivity of the Generic Core Scales and Cerebral Palsy Module. *Dev Med Child Neurol* 2006; **48:** 442–9.
20. Muthen BO, Muthen L. Mplus User's Guide. Los Angeles: Muthen & Muthen, 1998.
21. Muraki E, Bock RD. Parscale: IRT Item Analysis and Test Scoring for Rating-Scale Data. Chicago: Scientific Software International, 1997.
22. Warm TA. Weighted likelihood estimation of ability in item response theory. *Psychometrika* 1989; **54:** 427–50.
23. Bond TG, Fox CM. Applying the Rasch Model: Fundamental Measurement in the Human Sciences. Mahwah, NJ: Lawrence Erlbaum Associates, 2007.
24. Hambleton RK, Jones RW. An NCME instructional module on comparison of classical test theory and item response theory and their applications to test development. *Educ Meas Issues Pract* 1993; **12:** 38–47.