Development of a Parent-Report Cognitive Function Item Bank Using Item Response Theory and Exploration of its Clinical Utility in Computerized Adaptive Testing

Jin-Shei Lai,[1,2] PhD, OTR/L, Zeeshan Butt,[1,3,4] PhD, Frank Zelko,[5] PhD, David Cella,[1,6] PhD, Kevin R. Krull,[7] PhD, Mark W. Kieran,[8] MD, PhD, and Stewart Goldman,[5] MD

[1]*Medical Social Sciences, Feinberg School of Medicine, Northwestern University,* [2]*Pediatrics, Feinberg School of Medicine, Northwestern University,* [3]*Comprehensive Transplant Center, Northwestern University Transplant Outcomes Research Collaborative (NUTORC),* [4]*Institute for Healthcare Studies, Northwestern University,* [5]*Children's Memorial Hospital,* [6]*Psychiatry and Behavior Science, Feinberg School of Medicine, Northwestern University,* [7]*St. Jude Children's Research Hospital, and* [8]*Children's Hospital Boston and Dana-Farber Cancer Institute, Harvard Medical School*

All correspondence concerning this article should be addressed to Jin-Shei Lai, PhD, OTR/L, Research Associate Professor, Department of Medical Social Sciences, Feinberg School of Medicine at Northwestern University, 710 N Lake Shore Drive, #724, Chicago, IL 60611, USA. E-mail: js-lai@northwestern.edu

Received July 27, 2010; revisions received January 4, 2011; accepted January 10, 2011

Objective The purpose of this study is to report the reliability, validity, and clinical utility of a parent-report perceived cognitive function (pedsPCF) item bank. **Methods** From the U.S. general population, 1,409 parents of children aged 7–17 years completed 45 pedsPCF items. Their psychometric properties were evaluated using Item Response Theory (IRT) approaches. Receiver operating characteristic (ROC) curves and discriminant function analysis were used to predict clinical problems on child behavior checklist (CBCL) scales. A computerized adaptive testing (CAT) simulation was used to evaluate clinical utility. **Results** The final 43-item pedsPCF item bank demonstrates no item bias, has acceptable IRT parameters, and provides good prediction of related clinical problems. CAT simulation resulted in correlations of 0.98 between CAT and the full-length pedsPCF. **Conclusions** The pedsPCF has sound psychometric properties, U.S. general population norms, and a brief-yet-precise CAT version is available. Future work will evaluate pedsPCF in other clinical populations in which cognitive function is important.

Key words assessment; cancer and oncology; cognitive assessment; computer applications/eHealth; neuropsychology; quality of life.

Introduction

Cognitive decrements are a common concern in pediatric illnesses such as acute lymphoblastic leukemia (ALL) and brain tumors, as a function of illness itself and also secondary to treatment. These decrements can negatively impact quality of life in survivors and their families (Cardarelli et al., 2006; Lai, Goldman, Cella, & Carlson, 2003; National Cancer Policy Board, Hewitt, Weiner, & Simone, 2003; Zeltzer et al., 1997). Routine assessment of cognitive dysfunction is critical, as symptoms may first appear years after treatment (Ris, Packer, Goldwein, Jones-Wallace, & Boyett, 2001; Sonderkaer et al., 2003). Clinicians and researchers typically rely on standardized neuropsychological testing batteries to assess cognitive function in these populations. However, administration of such batteries is time-consuming, labor-intensive and therefore costly, and impractical within the context of routine medical follow-up. Furthermore, when repeated over

Journal of Pediatric Psychology 36(7) pp. 766–779, 2011
doi:10.1093/jpepsy/jsr005
Advance Access publication March 4, 2011

time, the reliability and validity of neuropsychological testing can be compromised by practice effects (Iuvone et al., 2002; Jacobson & Truax, 1991; Mollica, Maruff, Collie, & Vance, 2005; Temkin, Heaton, Grant, & Dikmen, 1999). Alternate methods for valid and efficient identification of cognitive dysfunction in cancer patients and other pediatric populations are clearly needed. A method will have maximal impact if it (a) minimizes demands upon limited clinical and financial resources, (b) can be administered repeatedly and maintain validity, and (c) is sensitive to early stages of cognitive deterioration.

One potential method for identifying cognitive dysfunction is via observer report. This approach has long been used in clinical assessment of behavior and adaptive functioning (Achenbach & Edelbrock, 1983; Sparrow, Balla, Cicchetti, & Doll, 2005). More recently, this approach has also been applied to the assessment of executive functioning (Gioia, Isquith, Guy, & Kenworthy, 2000). The potential of observer report as a method for monitoring cognitive outcome in cancer is suggested by recent studies, which have found significant correlations between cognitive complaints and neuro-imaging findings (de Groot et al., 2001; Ferguson, McDonald, Saykin, & Ahles, 2007; Mahone, Martin, Kates, Hay, & Horska, 2009). Such studies led us to speculate that observer-reported cognitive function [i.e., Perceived Cognitive Function (PCF)] could be a useful screening tool for adverse cognitive outcomes and possibly even a predictor of structural CNS changes over time. However, to date, little effort has been devoted to the precise and accurate measurement of PCF in clinical populations such as cancer survivors.

To address this need, we have developed a psychometrically sound PCF measurement system based on Item Response Theory (IRT). The starting point for such a system is the development of one or more group(s) of calibrated statements or questions (i.e., an *Item Bank*) that an informant responds to, thus characterizing the individual under scrutiny. A system that samples cognitive functions across the lifespan would consist of several developmentally appropriate item banks linked by common items which allow monitoring PCF change over time, with items appropriate for children through young adulthood and across the disease continuum. An important advantage of the development of such a system under IRT is that all items are calibrated onto the same measurement continuum. This feature allows scores from different informant reports at different times to be compared regardless of whether the same items are administered.

This measurement concept has recently been adopted, more generally, by the field of health-related quality of life/symptom management. For example, the Patient Reported Outcomes Measurement Information System (PROMIS; www.nihpromis.org), part of the NIH Roadmap initiative, has developed 11-item banks measuring common concerns of people with various chronic diseases. One advantage of the use of such item banks is that the entire item set does not have to be administered to all respondents via traditional questionnaires or interviews for valid assessment. One novel application of a comprehensive, IRT-optimized item bank is computerized adaptive testing (CAT). CAT enables precise estimation of a trait while simultaneously minimizing response burden to informants (Cella, Gershon, Lai, & Choi, 2007; Choi, Reise, Pilkonis, Hays, & Cella, 2009). In CAT, items administered are selected on the basis of the informant's previous item responses, using a preset computerized algorithm based on individual item information functions, defined as the reciprocal of the standard error in IRT models (Weiss & Kingsbury, 1984). With this approach, each assessment is individualized based on the symptom level of the patient at that point in time. Furthermore, CAT algorithms allow for the possibility that the same informant could respond to different items over time, depending on developmental and/or symptom changes, while still allowing for comparability of scores for the patient being rated. Since only the most psychometrically informative items are presented to the respondent, a high level of measurement precision can be achieved using few items.

To address the need for effective and efficient assessment of cognitive functioning in pediatric cancer, we developed an item bank measuring perceived cognitive function (pedsPCF) reported by parents of children aged 7–17 years. This article describes the psychometric properties of this newly developed measure. We completed a simulation study to demonstrate the potential precision and efficiency of the item bank in CAT, in comparison with the full-length fixed item pedsPCF. We also describe the development of U.S. general population-based norms for the pedsPCF CAT, which will allow for meaningful future comparisons in chronic illness populations, including but not limited to childhood ALL and brain tumor.

Methods

The pedsPCF item bank

The development of the pedsPCF item bank, reported in detail elsewhere (Lai, Zelko et al., 2011), was based upon the input of experts in pediatric neuro-oncology as well as patients of the Falk Brain Tumor Center, Children Memorial Hospital in Chicago and their parents. In brief, items were developed by interviews (of children, parents, teachers, and clinicians), literature review, and search of a

database of more than 7,000 items that our department (Medical Social Sciences at Northwestern University) has maintained. The initial item pool consisted of 145 questions. A series of steps were taken prior to field testing, including item review within the research team, review by clinicians and teachers, a consensus meeting of consultants and members of the research team, and think-aloud interviews (DeWalt, Rothrock, Yount, Stone, & PROMIS Cooperative Group, 2007) with children. Forty-five items from the initial 145-question pool were retained for field testing. One of the 45 items was dropped following a series of tests to assess item bank unidimensionality. Unidimensionality of the remaining 44 items was observed (Lai Zelko et al., 2011) supporting the use of a single PCF score to estimate the underlying trait of perceived cognitive functioning. Depending on item content, items are measured using either a 5-point frequency rating scale (5 = none of the time, 1 = all of the time) or a 5-point intensity rating scale (5 = not at all, 1 = very much). In order to be consistent with the PROMIS measurement system, higher scores represent better functioning while lower scores indicate more complaints.

Participants

This study was approved by Institutional Review Boards at all participating sites.

Data were collected from 1,409 parents and children/adolescents drawn from the U.S. general population (51.8% aged 7–12 years; 48.2% aged 13–17 years) by an Internet survey company, Toluna (www.toluna.com; formerly Greenfield, http://www.greenfield.com). Specifically, Toluna sent e-mail invitations to potential participants from their database (i.e., panel members) to recruit them for field testing. Potential participants were screened via Internet to ensure their eligibility (i.e., English speaking and have a child aged 7–18 years). Participants (one parent from each family) signed an online consent and completed a survey of demographic information, the pedsPCF items, and the Child Behavior Checklist [CBCL(Achenbach, 1991)] with reference to an eligible child in the household who was available and agreed to participate. Following a parent's ratings of his/her child, the child completed a self-report version of the pedsPCF items (data not reported here). All procedures were conducted on-line and recruitment was terminated when the preset goal was reached. The reasons that panel members declined to participate are not obtained by the company. To address concerns about ceiling effects, which can result in unstable item parameters in IRT analysis, we overrecruited parents with children with reported neurological conditions.

The mean age of participating children was 12.3 years ($SD = 3.0$; range: 7–17 years), with 56.8% male and 83.0% White. The average days of school missed in the past year was 2.0 ($SD = 3.2$). Of them, 29.5% were described as having previously received mental health services (51.5% social work services at school and 70.4% clinical/outpatient counseling). Most children (85.7%) attended a mainstream classroom; 28.6% were enrolled in individualized education programs. The average parent age was 40 years ($SD = 8$; range: 25–65 years); 83% were White and 68% were married. For paternal informants (40% of respondents), 25.2% were high school graduates or less, 31.1% some college and 43.7% had a college degree or higher. For maternal informants (60% of respondents), 37.3% were high school graduates or less, 37.6% some college and 25.1% had a college degree or higher. Furthermore, 319 (22.6%) parents indicated that they had been told by a physician or a health professional that their child had a neurologic condition, including epilepsy (15.0%), traumatic brain injury (3.4%), cerebral palsy (2.6%) and brain tumor (1.6%). Additionally, 26.4% of the parents indicated that their child had attentional deficit/hyperactivity disorder (ADHD) and 16.3% that their child had repeated a grade. Of those diagnosed as having ADHD, 30.1% had been given ADHD medication in the past 3 months. As rated by their parents, most children had either very good (43.2%) or excellent (24.1%) quality of life.

Analysis

Forty-four items that demonstrated adequate unidimensionality from previous analyses were included in the present analyses, which were carried out to achieve three main aims: (a) to calibrate items onto the PCF measurement continuum using IRT models, (b) to evaluate the clinical utility of the measure using receiver operating characteristics (ROC) curves, and (c) to demonstrate an application of the pedsPCF in a CAT simulation study.

Item parameter estimation using item response theory

Fundamental IRT framework

A fundamental feature of IRT is that the psychometric properties of individual items are related to the estimated amount of a patient's ''latent trait'' that is sampled. This feature allowed us to estimate the likelihood of a parent's specific response to an item, given the overall parent rating of the child's cognitive function and knowledge of individual items' psychometric properties (Bjorner, Chang, Thissen, & Reeve, 2007; Hambleton, Swaminathan, & Rogers, 1991). Numerous models fall into the IRT family.

The 1-parameter logistic (1-PL)/Rasch model and the 2-PL model are most commonly used in health-related quality of life research. The 1-PL/Rasch is based upon the assumption that the probability of a respondent endorsing a particular response of an item depends solely upon his/her PCF level relative to the location of that item, that is, item difficulty, on the PCF continuum. In addition to the difficulty parameter, the 2-PL model also estimates discrimination parameter (or the ''slope parameter'') to describe how well an item discriminates among individuals on different points of the PCF continuum. Item slope can be interpreted as describing how an item may be related to the trait measured by the scale. Desired values are between 1 and 5, where higher values provide more information about a respondent than less discriminating items at the same location. An item with a low discrimination value usually indicates that the item may not define the same construct as the rest of the items in the scale. On the other hand, an item with an extremely high discrimination value is typically the result of a skewed distribution.

There is no universal agreement regarding the ideal IRT model for evaluating multiple response categories for the difficulty parameter. We chose to use the Graded Response Model (Samejima, van der Liden, & Hambleton, 1996) as implemented using MULTILOG computer software (Thissen, 2003) to be consistent with the PROMIS methodology. Specifically, Samejima's GRM reports difficulty parameters by estimating rating scale thresholds (b_k). Threshold parameter b_k is the point on the latent trait where an participant has a 50% chance of responding positively to an specific item response category. A response category with a higher threshold value is less often endorsed. In this study, the discrimination parameter describes the strength of an item's discrimination between people at different PCF levels below and above the threshold b_k, indicating the degree of association between item responses and the PCF latent trait. Each threshold parameter ranges from negative to positive infinity. The values reported here reflect the degree of PCF where the *most probable response* occurs in a given category or higher. GRM calibrated scores are reported in *z*-score units (expected mean = 0 and $SD = 1$). Threshold parameter values can be used as a *reference* illustrating whether the items are distributed throughout the PCF measurement continuum. Discrimination and threshold parameters are used to calculate the item information function, which indicates the PCF range over which an item is most useful for distinguishing among individuals. Unique to IRT, the information function is the reciprocal of the standard error (*SE*) function and varies along the continuum and can be converted into reliability function *at each point on the continuum*.

By comparing information function/reliability and person scores, we can describe whether the pedsPCF precisely measures perceived cognitive functioning for individuals who might require more medical attention. We used the IRT-based information function to estimate reliability and error functions at both the scale and item level, to allow examination of precision levels along the PCF continuum. An information function is cumulative, such that the information function of an entire scale is the summation of the information functions for the items that comprise the scale. High information functions correspond to high precision of the estimates (i.e., low *SE*). To facilitate effective communication with end-users, we transformed the *SE* function to a reliability function and used it to describe the pedsPCF item bank characteristics. A low reliability estimate ($r < .7$) may or may not be a concern depending on where it is identified. Low reliability is a particular concern when it occurs at PCF levels indicating cognitive complaints requiring medical attention; in those instances items may need to be added to enhance reliability (Lai et al., 2007; Lai, Cella, Peterman, Barocas, & Goldman, 2005).

Use of IRT in the present study

In the present study, IRT was used to examine how well the items fit the PCF latent trait being measured, to calibrate items on the PCF continuum, compare item discrimination power, and calculate and graph the information function at both item and scale levels. We also compared item characteristics and the stability of item parameters using differential item functioning (DIF) between child gender, child age, parent education levels, and parent gender.

We also estimated a patient's level of PCF on the trait defined by this group of items (Reise & Waller, 2009; Thissen, Orlando, Thissen, & Wainer, 2001). Prior to these analyses, we used S-X^2 and S-G^2 (Orlando & Thissen, 2003) fit indices to evaluate how well items fit a unidimensional model by comparing observed to expected frequencies as defined by the 2-PL IRT model. Poorly fitting items (i.e., $p < .01$) were candidates for removal.

Differential item functioning

DIF is a condition when an item performs differently for participants from two different subgroups, after controlling for differences between subgroups on the latent trait. DIF

analyses are particularly critical when developing a new measurement system or revising an existing one for different disease groups, to ensure consistency in measurement properties across groups to minimize potential item bias. We evaluated DIF, or item bias, on the following variables: child age (ages 7–12 years vs. 13–17 years) and gender, and parent gender and education (''high school graduate or less'' vs. ''at least some college''). Two DIF methods were chosen: nonparametric-based Mantel chi-squared DIF statistic (Mantel, 1963; Zwick & Thayer, 1996) using DIFAS computer software (Penfield, 2006) and IRT-based ordinal logistic regression (OLR) (Crane, Gibbons, Jolley, & van Belle, 2006) as implemented in LORDIF freeware (Choi, Gibbons, & Crane, 2008).

The null hypothesis for the Mantel chi-squared statistic states that when members of the two groups are matched on a sum score, they tend to show the same item scores. The OLR method assesses both uniform and nonuniform DIF. Uniform DIF is analogous to a significant group effect, conditional on the latent trait (e.g., PCF); nonuniform DIF is equivalent to a significant interaction of group and trait. There is no consensus as to which DIF approach is the best (Lai, Teresi, & Gershon, 2005). For this study, we followed the approach of the PROMIS project by employing at least two DIF procedures on more than one variable, with results from the different DIFs compared to make a final recommendation.(Lai, Cella et al., in press).

Items that showed DIF (criterion: $p < .001$) on more than one condition using both DIF methods implied potential measurement bias and were removed from the final pedsPCF item bank. Decisions regarding inclusion/exclusion of items that showed DIF on only one condition by either or both methods or more than one condition by either DIF method were made on an item-by-item basis by the study team. The final pedsPCF item bank was comprised of items remaining after DIF analysis and fit statistics, with final item calibrations estimated using the 2-PL IRT GRM model as described earlier.

Clinical utility of the pedsPCF item bank

All PCF raw scores were converted to IRT-based scaled scores for the following analyses, with a higher score representing better function and a lower score representing more complaints (dysfunction). We used analysis of variance (ANOVA) to compare pedsPCF for children across different levels of health-related quality of life (QOL), which was evaluated using the question ''*How do you rate your child's quality of life in general*'' (1 = poor; 5 = excellent). We also evaluated the ability of the PedsPCF to discriminate the following six clinical groups as described by parents: (a) a normal group ($n = 857$) without neurological diagnosis (epilepsy, traumatic brain injury, cerebral palsy, brain tumor) or attention deficit disorder (ADD); (b) a group with no neurological diagnosis but positive for ADD ($n = 233$); and groups with (c) epilepsy ($n = 211$); (d) traumatic brain injury (TBI; $n = 48$); (e) cerebral palsy (CP; $n = 37$); and (f) brain tumor (BT; $n = 23$). While previous analyses showed that pedsPCF items defined a single construct, three subdomains were suggested under the overall PCF index: memory retrieval, attention/concentration and working memory (Lai, Zelko et al., 2011). We, therefore, also conducted ANOVA on these individual subdomains, in addition to the overall PCF index.

ROC were used to evaluate how well the pedsPCF item bank predicted children within the normal (vs. borderline or clinical) range of three selected CBCL scales (attention; social; thought problems). ROC curves are a graphic display of sensitivity versus specificity (1-specificity is used as the *x*-axis) for predicting participant membership in dichotomized groups (i.e., yes/no, or normal vs. outside of normal range). Specificity quantifies the ability of a test to identify ''true negatives'' (values for 1-specificity used for the *x*-axis indicate the proportion of false positives). In our analysis, ''cognitive difficulties'' is defined as ''positive.'' High specificity indicates a low false positive rate; that is, a high percentage of children without cognitive difficulty are properly identified having no cognitive difficulty. On the other hand, sensitivity quantifies the ability of a test to identify ''true positives.'' High sensitivity indicates a low false negative rate; that is, a high percentage of children with cognitive difficulty are properly identified as having cognitive difficulty. Investigators can determine cutoff points based on their desired levels of sensitivity and specificity. Though both indices are important, adjusting the pedsPCF cutoff value to increase sensitivity will usually result in a corresponding reduction in specificity. The ROC curve demonstrates this trade-off for all possible cutoff values. The maximum sensitivity (=1) always accompanies the lowest specificity (=0). In the current study, sensitivity and specificity were reported at the cutoff values where the maximum accuracy rate of predicting children's membership (i.e., normal vs. elevated in CBCL problems scales) was reported by the discriminant function analysis. Area under the curve (AUC) of the ROC curve was used to evaluate the validity of the prediction model. A general guideline for AUC is: excellent when AUC is between 0.9 and 1.0; good when between 0.80

and 0.90; fair when between 0.7 and 0.8 and poor or fail when AUC < 0.60.

Advanced application of the pedsPCF item bank: computerized adaptive testing

We explored the ability of the pedsPCF item set to support CAT assessment by conducting a post hoc simulation study using Firestar software (Choi, 2009). In this simulation, Firestar generated respondents with predefined PCF scores equally distributed on the pedsPCF measurement continuum, from least to greatest cognitive complaints. These ''virtual'' respondents all first completed the item with the maximum expected information over the prior distribution; the initial PCF score was estimated; an item with the maximum information function on the prior estimated PCF score was chosen as the subsequent item to be administered; and the PCF score was re-estimated based on the participant's response to that item. This iterative estimation process continued till the stopping rule was met: *SE* of measurement <0.3 or number of items exceeded 20, whichever came first. To minimize selection bias while maintaining efficiency of the CAT administration, the initial item was randomly selected by the simulation program from one of the three most informative items. In this study, the three candidate items were: ''It is hard for your child to concentrate in school,'' ''Your child has trouble paying attention to the teacher,'' and ''Your child has to work really hard to pay attention or he/she makes mistakes.'' We set the minimum number of items to be administered at 4. We then compared simulated PCF scores obtained from CAT with scores based on completion of all pedsPCF items.

Results

IRT-related analyses

Logistic regression results showed that seven items demonstrated DIF on parent's gender, one on child's age, and two on parent's education. Mantel chi-squared DIF statistics identified 10 items with DIF on parent's gender, 2 on child's gender, and 3 on child's age. No item showed DIF on all variables using both DIF approaches. The item ''*Your child has to contact his/her friends for homework he/she forgets*'' showed a poor fit ($p < .001$) on both S-X^2 and S-G^2, with DIF on more than 2 variables (parent gender, parent education, and child age) using both DIF methods and was therefore removed, resulting in a total of 43 items in the final item bank.

GRM results showed acceptable discrimination power for all remaining items, ranging from 2.03 to 4.35 (expected values between 1 and 5). Item threshold parameters ranged from −3.2 to 1.0. Based on the inspection of the PCF IRT-scaled scores, which ranged from −3.2 to 1.7, we concluded that our sample was sufficiently assessed by the items, with the possible exception being those with negligible complaints (i.e., better cognitive function). The pedsPCF score distribution is shown in Figure 1, with IRT-scaled scores converted to U.S. general population-based T-scores (mean = 50; $SD = 10$) to enhance interpretation of results. Also shown in Figure 1, we compared the reliability function to the T-scores distribution and found that all PCF scores based upon parent reports in this study were measured reliably ($r > .7$). Specifically, 91.9% ($n = 1{,}323$) had reliabilities at or above 0.9; 6.1% ($n = 85$) had reliability indices between 0.8 (included) and 0.9, and only one participant had a reliability index between 0.7 and 0.8.

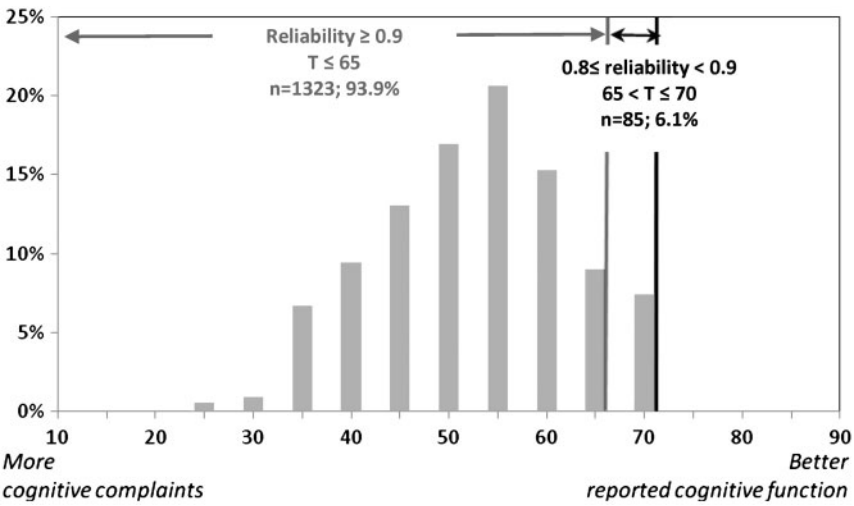
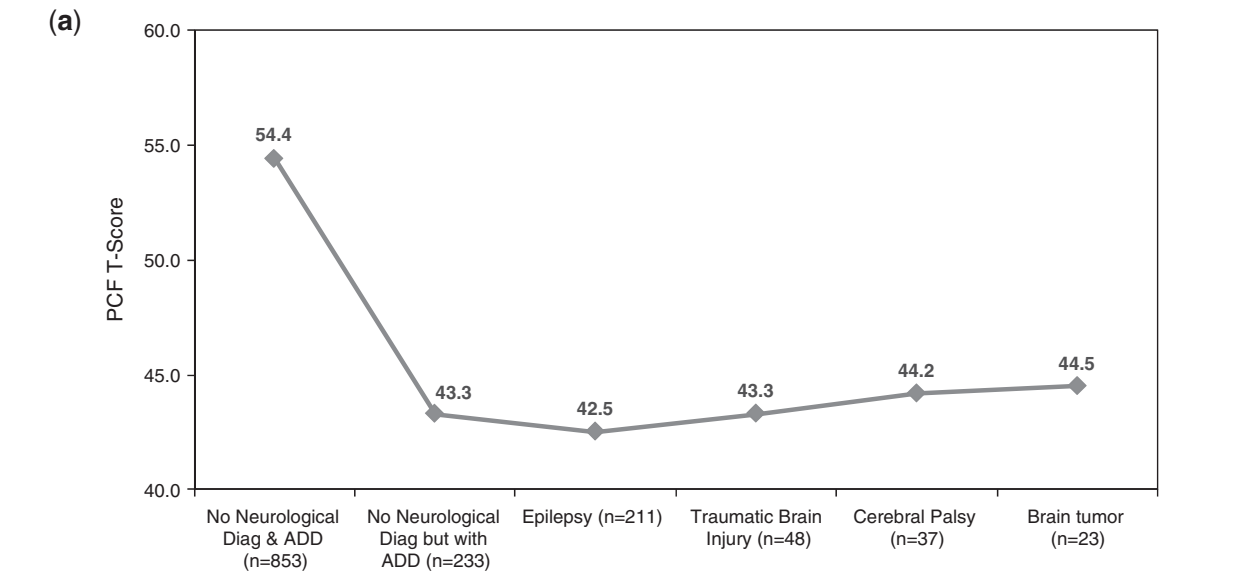


Figure 1. Sample distribution (in T-score matrix, where mean = 50 and $SD = 10$).

Participants with reliability indices below 0.9 all had negligible cognitive complaints, which would typically not indicate the need for medical attention. For the 86 subjects with pedsPCF score reliability less than 0.9, 7% ($n = 6$) had a neurological condition, 2.3% ($n = 2$) had ADHD, and CBCL scores fell outside of the normal range, 2.3% ($n = 2$), 10.5% ($n = 9$) and 1.2% ($n = 1$) for the social, thought and attentional problems scales, respectively. We therefore concluded that the pedsPCF showed satisfactory psychometric properties throughout the meaningful clinical range, and that it could provide reliable PCF estimates.

Clinical utility of the pedsPCF item bank

ANOVA results showed that the pedsPCF could differentiate children described by parents as with or without neurological diagnoses or ADHD, $F(5, 1403) = 124.14$, $p < .001$, $R^2 = .31$. Specifically, post hoc Tukey tests showed that children without neurological diagnoses and without ADHD had significantly better PCF scores than other groups (Figure 2a) with differences ranging between one-half and two-thirds *SD*. Considering each of the three subdomain areas separately, significant ($p < .001$) ANOVA results were found for each area, with $F(5, 1403) = 132.6$ ($R^2 = .32$), $F(5, 1403) = 121.5$ ($R^2 = .30$), and $F(5, 1394) = 85.9$ ($R^2 = .23$), for memory retrieval, attention/concentration, and working memory, respectively. Post hoc Tukey (Figure 2b) analyses showed that the normal group reported significantly better PCF than the other five groups in all three subdomain areas. The ADHD group had significantly better memory retrieval ratings (4.2 T-score units) than the epilepsy group and poorer working memory (4.7 T-score units) than the cerebral palsy group.



	No Neurological Diag & ADD	No Neurological Diag but with ADD	Epilepsy	Traumatic Brain Injury	Cerebral Palsy	Brain Tumor
Memory retrieval	54.35	45.35	41.18	43.16	42.75	42.93
Attention	54.3	42.76	43.11	42.97	44.17	45.26
Working memory	53.74	43.31	43.74	45.55	47.99	44.78

Figure 2. Comparisons of PCF T-scores across six clinical groups. (**a**) Overall PCF index and (**b**) by subdomains: memory retrieval, attention, working memory.

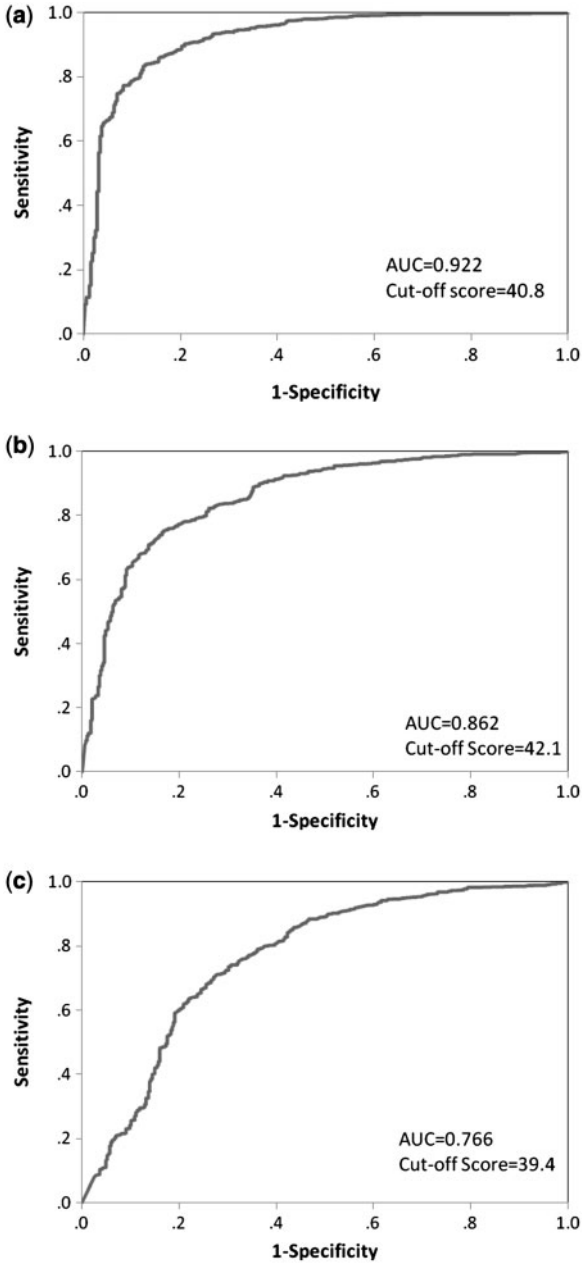


Figure 3. ROC and cutoff scores of the CBCL problem scales. (**a**) Attentional problem scale, (**b**) social problem scale, and (**c**) thought problem scale.

In terms of quality of life, pedsPCF significantly differentiated parent-reported child's QOL, $F(4, 1304) = 99.5$, $p < .001$ ($R^2 = .22$) in all comparisons, except for the comparison between ''fair'' and ''poor'' QOL groups.

ROC curves are presented in Figure 3a–c to demonstrate the ability of the pedsPCF to discriminate children with and without problems as identified by the CBCL. AUC were all greater than 0.7 (summarized in Table 1). The pedsPCF best predicted ratings on the CBCL attention scale (AUC = 0.92), followed by scales sampling social problems (AUC = 0.86), and thought problems (AUC = 0.77). We used discriminant function analysis to estimate cutoff scores, which were all around 1 *SD* (i.e., 10 T-score points) below the mean ($t = 50$). All specificities were greater than 0.9 while sensitivities ranged from 0.40 (thought problems) to 0.68 (attention problems).

Table I. *Discriminative Statistics for pedsPCF Item Bank*

	AUC	Cutoff T-score	Sensitivity	Specificity
CBCL-attention problems	0.922	40.8	0.683	0.947
CBCL-social problems	0.863	42.1	0.587	0.920
CBCL-thought problems	0.766	39.4	0.399	0.930

Samples with T-scores equal to or higher than the cutoff scores tend to be those within normal range of attention, social, and thought problem scales.

CAT simulation

The average number of items administered to virtual respondents by the Firestar CAT simulation software was 6.1 ($SD = 3.2$). Not surprisingly, more items were required at the higher end (i.e., better reported cognitive function) of the PCF continuum, where relatively low reliabilities were found. The correlation between CAT scores and full-item bank scores was 0.98, indicating that CAT based upon the PCF item bank can produce results that are very similar to those obtained with administration of the entire 43 PCF item set (Figure 4).

Discussion

The purpose of this project was to develop a tool to screen children at risk for adverse cognitive outcomes, so that cognitive deficits can be identified as early as possible to facilitate appropriate and timely referral for formal cognitive assessments and intervention services. Such a screening tool should be especially beneficial for patients such as childhood cancer/brain tumor survivors, given that adverse cognitive symptoms may be a late effect of their treatment and emerge over time. As previously discussed, we believe that a psychometrically sound PCF measure has the potential to serve as such a screening tool, given its ease of administration, low cost, and accumulating evidence that PCF deficits are significantly associated with external clinical criteria such as abnormal neuroimaging findings. Further research on the association between current pedsPCF, formal neuropsychological tests, and neuro-imaging findings should be conducted to examine their interrelationships. If significant correlations are found, the pedsPCF is supported as an effective screening

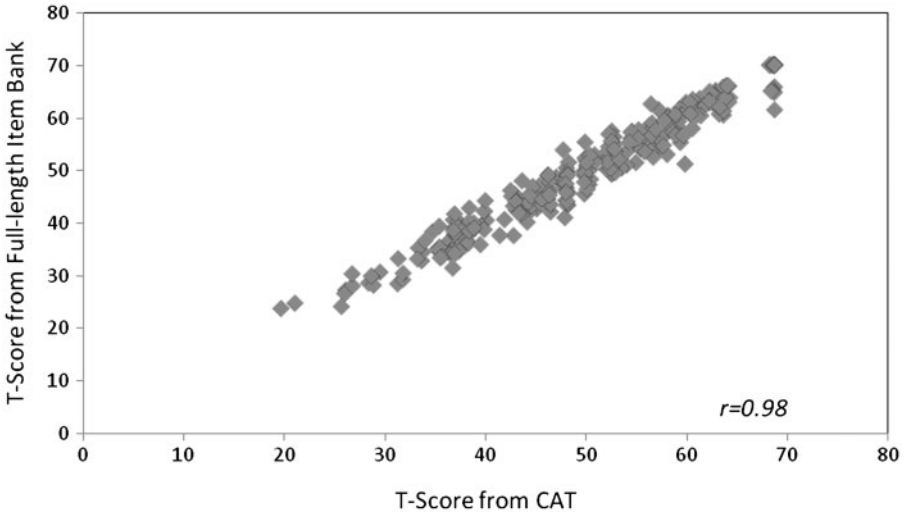


Figure 4. Comparison between CAT-based and full-length pedsPCF item bank scores.

tool offering ease of administration and low financial burden.

In this article, we examined the psychometric properties of a newly developed 43-item pedsPCF item bank. We believe that the items included in the pedsPCF allow effective sampling of the neurocognitive functioning of pediatric cancer survivors, as they were developed via an extensive process using various informant resources and refined via expert reviews and interviews with children and parents (Lai, Zelko et al., 2011). The end result of this process was not a measure of global cognitive functioning but rather an item set sampling facets of cognitive functioning such as attention, memory retrieval, and working memory which appear to be particularly sensitive to disease- and treatment-related mental status changes in this population (Krull et al., 2008). Though we expect the current pedsPCF item set to have utility within the setting of a traditional medical follow-up clinic, expansion and refinements of the pedsPCF will continue in order to broaden its utilization to not only clinical settings but also educational settings. In IRT-related analyses, the pedsPCF item bank demonstrated negligible measurement bias by child age, child gender, parent gender, and parent education. It produced reliable PCF scores and discriminated children with and without elevated levels of attention, social and thought problems. CAT simulation results indicate that our items adequately assesses along the PCF measurement continuum, with a high correlation found between the full-length pedsPCF item bank and CAT.

To improve the clinical utility of the pedsPCF, we estimated cutoff scores separating clinical groups using discriminant function analysis, using dichotomized CBCL attention, social problems, and thought problem scales. To facilitate their application in clinical settings, these cutoff points are reported in a T-score metric (mean = 50, $SD = 10$) making it easy for investigators and parents to gage how discrepant their child's PCF is from the general population. A 1 *SD* difference on the pedsPCF best discriminated children with attentional problems (the highest AUC and best sensitivity and specificity). The high specificities of the three problem scales suggest that the pedsPCF correctly identified children within the normal range as reported by parents, with few false positives. The pedsPCF performed moderately well (sensitivity = 0.68) at identifying children with reported attentional problems, but was less able to identify children with elevated parent-report social (sensitivity = 0.59) and thought problems (sensitivity = 0.40), with high numbers of false negatives. As expected, low sensitivities were found on the ''social'' and ''thought'' problem scales, which relate more to psychiatric concerns than pure cognitive concerns. These results support the divergent validity of the pedsPCF.

Though we would have preferred higher sensitivity for the CBCL attention scale, we feel the current value sufficient to support the convergent validity of the pedsPCF, particularly given the excellent AUC obtained (0.92). We therefore conclude that the pedsPCF holds promise as an effective screening tool for children with cognitive difficulties, especially attention problems. However, future studies are needed to compare the pedsPCF with other cognition specific measures/scales such as the BRIEF and formal neuropsychological indices of attention and executive skills. Based on the current data, we feel it reasonable to recommend that, as a general rule, children with pedsPCF T-score scores at or below 40 (i.e., 1 *SD* below the

U.S. general population norm) should seek more comprehensive evaluation. To enhance its utility as a clinical screening tool, futures studies should establish cutoff scores related to other cognitive measures. We dichotomized CBCL in this study for analytic reasons; future studies evaluating the relationship between pedsPCF and other continuous indices of cognitive dysfunction are also warranted.

The pedsPCF has been developed to capture cognitive functioning as observed in children's everyday lives. For example, children with brain tumors experience impairments which not only reduce current processing speed and efficiency, but also interfere with future development in the child's ability to acquire new skills and information.(Mulhern et al., 2001; Packer et al., 2003; Schatz, Kramer, Ablin, & Matthay, 2000) The impact of cognitive impairments can become more apparent over time, as the discrepancy between emerging skills of the brain-tumor survivor and his/her peers increases. In some instances, early-stage cognitive impairments can be difficult to discern as functional impairments in traditional, office-based assessment due to the sterile and highly controlled nature of the testing environment. Observer-reported cognition, standardized neuropsychological tests, and neuroimaging techniques all offer unique contributions to the identification of children with cognitive difficulties. We propose that observer-reported cognition is a practical and efficient way to identify early signs of cognitive impairments and enable timely referral for more comprehensive cognitive assessment. The relationship between PCF and formal neuropsychological tests are inconclusive based on current literature, in part due to inconsistent methodology across studies as well as the lack of comprehensive and psychologically sound PCF measurement tool. We hope the pedsPCF developed in this study will facilitate exploration of the relationship between perceived cognitive functioning and other presumably more objective indicators of brain dysfunction such as structural and functional neuroimaging techniques and brain pathology. Understanding these relationships can facilitate a better symptom monitoring program by establishing clinically meaningful cutoff scores and identifying patients who require further, more detailed neurocognitive assessment.

CAT is a relative new concept in healthcare. Wouters (Wouters et al., 2009) investigated whether the Cambridge Cognitive Examination, a widely used screening test for dementia, could be administered as a CAT. They concluded that such tailored testing provided a much more efficient screening for dementia than traditional assessment involving the administration of a full test battery. In the current study, we explored the potential utility of this approach with the pedsPCF by conducting a CAT simulation study. In this simulation, scores from full-length item bank administrations correlated highly with those produced from a CAT engine, supporting the functional equivalence of the two approaches. Comparable measurement precision was obtained under CAT with responses to only six items on average, indicating that effective screening can be accomplished in less than 2 min. These results suggest that the pedsPCF CAT can provide brief-yet-precise measures that are individualized—because items administered are selected based on the informant's responses to previous items—and which can be easily implemented in busy clinical settings. Another benefit of such an approach is that patients can complete different items over time yet rating results can still be meaningfully compared to prior years because the items are all scaled on the same metric. (Cella et al., 2007).

We believe that the efficient yet precise format of the pedsPCF CAT can facilitate patient care, allowing for routine assessments in everyday clinic settings. We envision a number of possible options for its administration. It can be administered by a standalone computer in clinic, and it can also be made available via Internet. Children and parents can complete the pedsPCF at home or anywhere with internet access via web-accessible electronic devices such as computers or cellular phones. Healthcare providers can review pedsPCF scores during or prior to patient visits by accessing a data center on the internet or a hardcopy printed by families. More comprehensive evaluations may be warranted when scores fall below a cutoff point or when a pattern of decline over time is seen.

Some limitations of the current study are noted. We have focused in the current report on parent-reported pedsPCF but not on child self-ratings, as we are aware of potential limitations (due to the emergence of metacognitive awareness) of the reliability and validity of self-report in children. However, we do not dismiss the potential value of child self-report, and we plan on evaluating the psychometric properties of a self-report version of pedsPCF in the future. Though data used for this study were drawn from a U.S. pediatric general population, they were not stratified for geographic and demographic factors, and children with neurological conditions were oversampled. As a result, the current sample cannot be considered to represent a true cross-section of children nationwide. Furthermore, households of lower socioeconomic status (SES) and those living in rural areas may be underrepresented in the current sample, due to disproportionately limited access to the Internet, which was the medium of recruitment. Though we believe that the potential impact of sampling bias is minimized by the size of our sample, generalizability of the current results will clearly be broadened by future

efforts to study disadvantaged subpopulations. We focused on the general population in this study with the intention of establishing U.S. general population-based norms for comparison with clinical samples in the future. Another limitation of this study is that neurological conditions were ascertained on the basis of parent report, and not confirmed by healthcare providers. Further studies are necessary for validation of the pedsPCF by using clinical populations with confirmed clinical diagnoses. Finally, all of the scales used in the current report were completed by parents. Thus, we are unable to rule out common method variance or response bias influence in the current results. Information from additional informants such as teachers will help to overcome this limitation. While the current pedsPCF items were developed with input from educators, we have not yet developed a teacher's version of this instrument. We intend to develop and validate a PCF item set for teacher report, in an effort to obtain as a complete picture of children's cognitive functioning as possible.

In conclusion, the 43-item pedsPCF exhibited sound psychometric properties and reliably assessed participants with moderate to severe cognitive difficulties. U.S. general population norms and a brief-yet-precise CAT version are available. The current results support the application of the pedsPCF in research. Future studies should be conducted to evaluate the relationship among PCF, formal neuropsychological tests, and neuroimaging findings. Additionally, further validation in clinical populations is warranted before the pedsPCF can be expected to play a role in clinical decision making.

Acknowledgment

The sampling method of this study is based on a study first reported in Lai et al. (2011).

Funding

National Cancer Institute (Grant number: R01CA125671; PI to J-S.L.). National Center for Research Resources (Grant UL1RR025741 PI: Phillip Greenland, MD); National Institutes of Health (to Z.B.).

Conflicts of interest: None declared.

References

Achenbach, T. M. (1991). *Integrative guide to the 1991 CBCL/4-18, YSR, and TRF profiles*. Burlington, VT: University of Vermont, Department of Psychology.

Achenbach, T. M., & Edelbrock, C. (1983). *Manual for the Child Behavior Checklist and Revised Child Behavior Profile*. Burlington, VT: University of Vermont, Department of Psychiatry.

Bjorner, J. B., Chang, C. H., Thissen, D., & Reeve, B. B. (2007). Developing tailored instruments: Item banking and computerized adaptive assessment. *Quality of Life Research, 16*, 95–108.

Cardarelli, C., Cereda, C., Masiero, L., Viscardi, E., Faggin, R., Laverda, A., . . . Perilongo, G. (2006). Evaluation of health status and health-related quality of life in a cohort of Italian children following treatment for a primary brain tumor. *Pediatric Blood and Cancer, 46*(5), 637–644.

Cella, D., Gershon, R., Lai, J. S., & Choi, S. (2007). The future of outcomes measurement: Item banking, tailored short-forms, and computerized adaptive assessment. *Quality of Life Research, 16*(Suppl 1), 133–141.

Choi, S. (2009). Firestar: Computerized Adaptive Testing Simulation Program for Polytomous Item Response Theory Models. *Applied Psychological Measurement, 33*(8), 644–645.

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2008). *Development of Freeware for an Iterative Hybrid Ordinal Logistic Regression/Item Response Theory Differential Item Functioning Framework, and an Illustration of Analyses of DIF Related to Age, Gender, and Education on the PROMIS Anxiety Scale*. Paper presented at the Second PROMIS Conference: Improving Measurement of Patient-Reported Outcomes-New Tools and the Science Behind Them, Bethesda, MD.

Choi, S. W., Reise, S. P., Pilkonis, P. A., Hays, R. D., & Cella, D. (2009). Efficiency of static and computer adaptive short forms compared to full length measures of depressive symptoms. *Quality of Life Research, 19*(1), 125–136.

Crane, P. K., Gibbons, L. E., Jolley, L., & van Belle, G. (2006). Differential Item Functioning Analysis With Ordinal Logistic Regression Techniques: DIFdetect and difwithpar. *Medical Care, 44*(11 Suppl 3), S115–S123.

de Groot, J. C., de Leeuw, F. E., Oudkerk, M., Hofman, A., Jolles, J., & Breteler, M. M. (2001). Cerebral white matter lesions and subjective cognitive dysfunction: The Rotterdam Scan Study. *Neurology, 56*(11), 1539–1545.

DeWalt, D. A., Rothrock, N., Yount, S., & Stone, A. A.PROMIS Cooperative Group (2007). Evaluation of

Item Candidates: The PROMIS Qualitative Item Review. *Medical Care, 45*(5 Suppl 1), S12–S21.

Ferguson, R. J., McDonald, B. C., Saykin, A. J., & Ahles, T. A. (2007). Brain structure and function differences in monozygotic twins: Possible effects of breast cancer chemotherapy. *Journal of Clinical Oncology, 25*(25), 3866–3870.

Gioia, G. A., Isquith, P., Guy, S., & Kenworthy, L. (2000). *Behavior rating inventory of executive function (BRIEF): Professional Manual*. Odessa, FL: Psychological Assessment Resources.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: SAGE Publications, Inc.

Iuvone, L., Mariotti, P., Colosimo, C., Guzzetta, F., Ruggiero, A., & Riccardi, R. (2002). Long-term cognitive outcome, brain computed tomography scan, and magnetic resonance imaging in children cured for acute lymphoblastic leukemia. *Cancer, 95*(12), 2562–2570.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*(1), 12–19.

Krull, K. R., Okcu, M. F., Potter, B., Jain, N., Dreyer, Z., Kamdar, K., & Brouwers, P. (2008). Screening for Neurocognitive Impairment in Pediatric Cancer Long-Term Survivors. *Journal of Clinical Oncology, 26*(25), 4138–4143.

Lai, J. S., Cella, D., Choi, S. W., Junghaenel, D. U., Christodolou, C., & Gershon, R. (in press). Development of the PROMIS Fatigue Item Bank, computerized adaptive testing and short-forms. *Archives of Physical Medicine and Rehabilitation.*

Lai, J. S., Cella, D., Kupst, M. J., Holm, S., Kelly, M. E., Bode, R. K., & Goldman, S. (2007). Measuring fatigue for children with cancer: Development and validation of the pediatric Functional Assessment of Chronic Illness Therapy-Fatigue (pedsFACIT-F). *Journal of Pediatric Hematology/Oncology, 29*(7), 471–479.

Lai, J. S., Cella, D., Peterman, A., Barocas, J., & Goldman, S. (2005). Anorexia/cachexia related quality of life for children with cancer: Testing the psychometric properties of the Pediatric Functional Assessment of Anorexia/Cachexia Therapy (peds-FAACT). *Cancer, 104*(7), 1531–1539.

Lai, J. S., Goldman, S., Cella, D., & Carlson, A. (2003). Quality of Life for childhood brain tumor survivors. *Quality of Life Research, 12*, 792.

Lai, J. S., Teresi, J. A., & Gershon, R. (2005). Procedures for the analysis of Differential Item Functioning (DIF) for small sample sizes. *Evaluation and the Health Professions, 28*, 283–294.

Lai, J. S., Zelko, F., Butt, Z., Cella, D., Kieran, M., & Krull, K. (2011). Perceived cognitive function reported by parents of the United States pediatric population. *Child's Nervous System, 27*(2), 285–293.

Mahone, E. M., Martin, R., Kates, W. R., Hay, T., & Horska, A. (2009). Neuroimaging correlates of parent ratings of working memory in typically developing children. *Journal of the International Neuropsychological Society, 15*(1), 31–41.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58*, 690–700.

Mollica, C. M., Maruff, P., Collie, A., & Vance, A. (2005). Repeated assessment of cognition in children and the measurement of performance change. *Child Neuropsychol, 11*(3), 303–310.

Mulhern, R. K., Palmer, S. L., Reddick, W. E., Glass, J. O., Kun, L. E., Taylor, J., ... Gajjar, A. (2001). Risks of young age for selected neurocognitive deficits in medulloblastoma are associated with white matter loss. *Journal of Clinical Oncology, 19*(2), 472–479.

National Cancer Policy Board, Hewitt, M., Weiner, S. L., & Simone, J. V. (2003). *Childhood cancer survivorship: Improving care and quality of life*. Washington, DC: National Academies Press.

Orlando, M., & Thissen, D. (2003). Further examination of the performance of S-X^2, an item fit index for dichotomous item response theory models. *Applied Psychological Measurement, 27*, 289–298.

Packer, R. J., Gurney, J. G., Punyko, J. A., Donaldson, S. S., Inskip, P. D., Stovall, M., ... Robison, L. L. (2003). Long-term neurologic and neurosensory sequelae in adult survivors of a childhood brain tumor: Childhood cancer survivor study. *Journal of Clinical Oncology, 21*(17), 3255–3261.

Penfield, R. D. (2006). DIFAS: Differential Item Functioning Analysis System. *Applied Psychological Measurement, 29*, 150–151.

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology, 5*, 27–48.

Ris, M. D., Packer, R., Goldwein, J., Jones-Wallace, D., & Boyett, J. M. (2001). Intellectual outcome after reduced-dose radiation therapy plus adjuvant chemotherapy for medulloblastoma: A Children's Cancer

Group study. *Journal of Clinical Oncology, 19*(15), 3470–3476.

Samejima, F., van der Liden, W. J., & Hambleton, R. (1996). The graded response model. In W. J. van der Linden (Ed.), *Handbook of modern item response theory* (pp. 85–100). New York, New York: Springer.

Schatz, J., Kramer, J. H., Ablin, A., & Matthay, K. K. (2000). Processing speed, working memory, and IQ: A developmental model of cognitive deficits following cranial radiation therapy. *Neuropsychology, 14*(2), 189–200.

Sonderkaer, S., Schmiegelow, M., Carstensen, H., Nielsen, L. B., Muller, J., & Schmiegelow, K. (2003). Long-term neurological outcome of childhood brain tumors treated by surgery only. *Journal of Clinical Oncology, 21*(7), 1347–1351.

Sparrow, S. S., Balla, D. A., Cicchetti, D. V., & Doll, E. A. (2005). *Vineland-II, Vineland adaptive behavior scales: Survey forms manual.* Minneapolis, MN; Circle Pines, Minn.: NCS Pearson, Inc.; AGS Publishing.

Temkin, N. R., Heaton, R. K., Grant, I., & Dikmen, S. S. (1999). Detecting significant change in neuropsychological test performance: A comparison of four models. *Journal of the International Neuropsychological Society, 5*(4), 357–369.

Thissen, D. (2003). MULTILOG (Version Windows 7.0). Lincolnwood, IL: Scientific Software International, Inc.

Thissen, D., Orlando, M., Thissen, D., & Wainer, H. (2001). Item response Theory for items scored in two categories. In D. Thissen (Ed.), *Test scoring* (pp. 74–140). Mahwah, NJ: Lawrence Erlbaum Associates.

Weiss, D. J., & Kingsbury, G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*(4), 361–375.

Wouters, H., de Koning, I., Zwinderman, A. H., van Gool, W. A., Schmand, B., Buiter, M., & Lindeboom, R. (2009). Adaptive cognitive testing in cerebrovascular disease and vascular dementia. *Dementia and Geriatric Cognitive Disorders, 28*(5), 486–492.

Zeltzer, L. K., Chen, E., Weiss, R., Guo, M. D., Robison, L. L., Meadows, A. T., . . . Byrne, J. (1997). Comparison of psychologic outcome in adult survivors of childhood acute lymphoblastic leukemia versus sibling controls: A cooperative Children's Cancer Group and National Institutes of Health study. *Journal of Clinical Oncology, 15*(2), 547–556.

Zwick, R., & Thayer, D. T. (1996). Evaluating the magnitude of differential item functioning in polytomous items. *Journal of Educational and Behavioral Statistics, 21*(3), 187–201.

Appendix A

Table IA. *Parent-reported Pediatric Perceived Cognitive Function Item Bank (pedsPCF)*

Item stem	Rating scale[a]
• It is hard for your child to find his/her way to a place that he/she has visited several times before	F
• Your child has trouble remembering where he/she put things, like his/her watch or his/her homework	F
• Your child has trouble remembering the names of people he/she has just met	F
• It is hard for your child to take notes in class	I
• It is hard for your child to learn new things	I
• It is hard for your child to understand pictures that show how to make something	I
• It is hard for your child to pay attention to something boring he/she has to do	I
• It is hard for your child to pay attention to one thing for more than 5–10 min	F
• Your child has trouble recalling the names of things	F
• Your child has trouble keeping track of what he/she is doing if he/she gets interrupted	F
• It is hard for your child to do more than one thing at a time	F
• Your child forgets what his/her parents or teachers ask him/her to do	F
• Your child walks into a room and forgets what he/she wanted to get or do	F
• Your child has trouble remembering the names of people he/she knows	F
• It is hard for your child to add or subtract numbers in his/her head	I
• Your child has trouble remembering the date or day of the week	F
• When your child has a big project to do, he/she has trouble deciding where to start	F
• Your child has a hard time keeping track of his/her homework	I and F[b]
• Your child forgets to bring things to and from school that he/she needs for homework	I and F[b]
• Your child forgets what he/she is going to say	I and F[b]
• Your child has to read things several times to understand them	I and F[b]
• Your child reacts slower than most people his/her age when he/she plays games	I and F[b]
• It is hard for your child to find the right words to say what he/she means	I and F[b]
• It takes your child longer than other people to get his/her school work done	I and F[b]
• Your child forgets things easily	I and F[b]
• Your child has to use written lists more often than other people his/her age so he/she will not forget things	I and F[b]
• Your child has trouble remembering to do things like school projects or chores	I and F[b]
• It is hard for your child to concentrate in school	I and F[b]
• Your child has trouble paying attention to the teacher	I and F[b]
• Your child has to work really hard to pay attention or he/she makes mistakes	I and F[b]

[a]F (Frequency): 1 = None of the time; 2 = A little of the time; 3 = Some of the time; 4 = Most of the time; 5 = All of the time; I (Intensity): 1 = Not at all; 2 = A little bit; 3 = Somewhat; 4 = Quite a bit; 5 = Very much.

[b]No local dependency is found for items that share the same item stem but are measured by both frequency and intensity types of rating scales. Therefore, both ratings are retained.