Evidence-based Assessment of Cognitive Functioning in Pediatric Psychology

Jonathan M. Campbell,[1] PhD, Ronald T. Brown,[2] PhD, Sarah E. Cavanagh,[1] MEd, Sarah F. Vess,[1] MEd, and Mathew J. Segall,[1] BA

[1]University of Georgia, and [2]Temple University

Objective To review the evidence base for measures of cognitive functioning frequently used within the field of pediatric psychology. **Methods** From a list of 47 measures identified by the Society of Pediatric Psychology (Division 54) Evidence-Based Assessment Task Force Workgroup, 27 measures were included in the review. Measures were organized, reviewed, and evaluated according to general domains of functioning (e.g., attention/executive functioning, memory). **Results** Twenty-two of 27 measures reviewed demonstrated psychometric properties that met ''*Well-established*'' criteria as set forth by the Assessment Task Force. Psychometric properties were strongest for measures of general cognitive ability and weakest for measures of visual-motor functioning and attention. **Conclusions** We report use of ''*Well-established*'' measures of overall cognitive functioning, nonverbal intelligence, academic achievement, language, and memory and learning. For several specific tests in the domains of visual-motor functioning and attention, additional psychometric data are needed for measures to meet criteria as ''*Well established*.''

Key words assessment; cognitive functioning; evidence-based; neuropsychology; pediatric psychology.

The lead article for the Special Series provides a useful historical overview of the movement occurring within psychology aimed to improve linkages between clinical practice and empirical findings (Cohen et al., 2007). Much of the movement toward grounding psychological practice in science has occurred within the domain of psychotherapy with more recent efforts focusing upon the status of empirical support for psychological assessment (Mash & Hunsley, 2005). The relatively recent focus of the field on empirically supported assessment is particularly relevant for pediatric psychologists, as psychological assessment represents a core domain of expertise, practice, and research. As detailed in other articles contributing to the Special Series, pediatric psychologists engage in psychological assessment across varied domains, such as pain assessment, treatment adherence, and child coping. In addition to evaluating functioning within these domains, pediatric psychologists frequently assess children's cognitive functioning for clinical and research purposes.

Reasons for Assessing Cognitive Functioning of Children with Pediatric Illness

As children with chronic illness continue to survive in greater numbers than ever before, particularly with the advancement of medical technology, pediatric psychologists evaluate children's cognitive functioning for various reasons. First, cognitive assessment may be requested to determine the impact of illness or injury with direct influence on central nervous system (CNS) functioning. For example, children with brain tumors (BT) or traumatic brain injury (TBI) frequently demonstrate cognitive impairments caused directly by these illnesses. Children with BTs often show impairment associated with the localized area affected as well as from consequences of tumor growth, such as seizures resulting from increased intracranial pressure (Moore, 2005). Likewise, children who sustain moderate to severe closed-head TBI may exhibit impairments across domains such as attention, visual-motor functioning, language, and verbal memory (Ewing-Cobbs & Bloom, 2004).

All correspondence concerning this article should be addressed to Prof. Jonathan M. Campbell, PhD, University of Georgia, Department of Educational Psychology and Instructional Technology, 325-J Aderhold Hall, Athens, GA 30602-7143, USA. E-mail: jmcmpbll@uga.edu

Journal of Pediatric Psychology 33(9) pp. 999–1014, 2008
doi:10.1093/jpepsy/jsm138
Advance Access publication January 13, 2008

Second, a variety of pediatric illnesses may impact indirectly on CNS functioning, such as cardiac diseases, hematological disorders, or endocrine dysfunction (Fennell, 2000). For example, children with sickle cell disease (SCD) are susceptible to stroke as well as transient ischemia and silent cerebral infarct (Bonner, Gustafson, Schumacher, & Thompson, 1999). Depending on stroke location, children with SCD may experience impairments in verbal or nonverbal reasoning, language functioning, motor functioning, or visual-motor coordination (Bonner et al., 1999). Similarly, children with insulin-dependent diabetes mellitus (IDDM) may experience attention and memory impairments as the result of severe recurrent hypoglycemia and sustained periods of hyperglycemia (Holmes, Cant, Fox, Lampert, & Greer, 1999). Third, medical treatments may directly impact CNS functioning either acutely, such as resection procedures, or through delayed (or ''late'') effects, such as those associated with prophylactic radiation or chemotherapy (Mulhern & Butler, 2006). For example, children with acute lymphoblastic leukemia (ALL) who undergo prophylactic whole brain radiation treatment (BRT) and/or CNS prophylactic chemotherapy (e.g., intrathecal) are susceptible to neuropsychological sequelae (Armstrong, Blumberg, & Toledano, 1999; Montour-Proulx et al., 2005). For children receiving BRT, research has documented general declines in overall cognitive ability and specific cognitive impairments in the domains of attention, nonverbal reasoning, visual-motor functioning and speeded processing of information (Armstrong et al., 1999). Children with ALL receiving intrathecal chemotherapy have shown declines in nonverbal reasoning in the presence of preserved verbal reasoning (Montour-Proulx et al.). Finally, pediatric psychologists may conduct cognitive assessments to determine the presence of developmental or learning disorders, such as intellectual disability. Although diagnostic questions about the presence of developmental and learning disabilities may be infrequent due to referral to specialized assessment centers or school systems, these assessment questions may still arise.

Importance of Evidence-based Cognitive Assessment for Pediatric Populations

In light of the broad range of direct and indirect impacts of pediatric illness on cognitive functioning, a neuropsychological assessment approach is often recommended to guide rehabilitation and educational programming for many patient groups, such as BT, TBI, ALL, and SCD. Therefore, the typical assessment approach is comprehensive, spanning domains of functioning, such as general cognitive ability, attention/executive functioning, memory, language functioning, and visual-motor skills. In addition to initial comprehensive assessment, regular and frequent re-evaluations are often employed to: (a) document possible declines in functioning, (b) screen for neurological impairment, (c) track recovery, (d) determine the effectiveness of rehabilitation and educational services, and (e) inform and modify treatment. The practice of serial re-evaluation highlights the importance of temporal stability in cognitive assessment with pediatric populations. Due to the importance of cognitive assessment and the necessity of instruments in clinical decision-making (e.g., diagnosing intellectual disability, tracking cognitive recovery), access to and use of evidence-based measures are essential within the field.

With these points serving as backdrop, the purpose of this article is to review the evidence-base for measures of cognitive functioning that pediatric psychologists report using most frequently. Within the context of the Special Issue, we sought to (a) identify the cognitive measures pediatric psychologists report using in practice and research and (b) evaluate the evidence base supporting their use. As such, the scope of the article is limited to appraising measures that constituents report frequently using within the field as opposed to providing a comprehensive review. It also is important to clarify that the purpose of our work was not to produce an ''approved list'' or an exhaustive list of cognitive instruments for use within the field but rather guide clinical and research activities by highlighting psychometric strengths and weaknesses of measures currently endorsed by pediatric psychologists.

We begin the review with a description of how the measures were selected, organized, and reviewed. Second, evaluation criteria are discussed as they pertain to measures of cognitive functioning. Third, specific tests are reviewed within domains of cognitive functioning including strengths and weaknesses associated with each domain. Finally, we conclude with general comments about the current evidence base for measures of cognitive functioning and offer recommendations to guide future research efforts and clinical practice. We describe methods and criteria utilized in the review, in part, to illustrate how the framework and guidelines might be used by professionals who are evaluating measures for their own clinical or research purposes.

Measure Selection

The procedures for identifying cognitive measures are detailed in the lead article prepared by the Evidence-Based

Assessment Task Force Committee [(EBA-TF); Cohen et al., 2007]. Briefly, the EBA-TF and members of the Cognitive Assessment Work Group (CAWG) generated a list of measures of cognitive functioning ($n = 37$) that was submitted to the Society of Pediatric Psychology (SPP), Division 54 listserv, which consists of approximately 325 subscribers. Respondents were asked: (a) to indicate via checkmark if they used or considered using each measure, and (b) to generate additional measures of cognitive functioning not appearing on the survey. A total of 87 listserv subscribers (27%) responded to the survey.

Respondents identified 10 additional measures not included in the initial list, resulting in a total of 47 assessment instruments available for review. Respondent selections ranged from 0 to 40 across the 47 measures ($M = 14.95$; $SD = 11.34$; $Mdn = 13$), with the Wechsler Intelligence Scale for Children, Third Edition (WISC-III; Wechsler, 1991) receiving the greatest frequency of responses ($n = 40$). We selected and reviewed measures that met or exceeded the median nomination value of 13 ($n = 27$; Appendix A), which left 20 measures identified by listserv respondents that were not reviewed (Appendix B). The criterion we employed resulted in a selective review process that ensued from a convenience sample; therefore, the reader is cautioned against inferring that nominated but nonreviewed tests are not empirically supported.

Several measures identified in the initial EBA-TF survey have undergone revision between the time that the survey was constructed in 2001 and the completion of the current review. As a result, respondents nominated several measures that had undergone revision. For example, the fourth edition of the WISC has been published (WISC-IV; Wechsler, 2003), the Kaufman Assessment Battery for Children (KABC) now exists in revised form (KABC-II; Kaufman & Kaufman, 2004), the fifth edition of the Stanford–Binet has been published (S-B 5; Roid, 2003), and the Bayley Scales of Infant Development is now in its third edition (BSID-III; Bayley, 2006). In response to these circumstances, CAWG reviewers decided to present psychometric data for the most recent version of the measure as well as summarize the evidence-base for the prior version. Reviewers reasoned that the revised versions represented significant overlap with their predecessors with perhaps the S-B 5 the most notable exception.

Organization of the Review

CAWG reviewers organized the review according to seven domains as follows: (a) general intelligence (i.e., IQ); (b) nonverbal intelligence; (c) achievement; (d) attention/executive functioning; (e) memory and learning; (f) visual-motor and motor functioning; and (g) language. Although the cognitive domains correspond generally with those assessed via neuropsychological evaluation, we realize that other groupings might be proposed, such as conceptual reasoning, domains divided, such as motor functioning and visual-motor construction, or others added, such as somatosensory functioning. Within each domain, the measures identified in the survey are reviewed as a group and strengths and weaknesses about the group are identified. After each domain is reviewed, we conclude with general comments about the evidence base for the entire group of cognitive measures.

Assessment Criteria

The EBA-TF created a general set of guidelines for CAWG to use to describe the degree of empirical support for measures (Cohen et al., 2007). The EBA-TF guidelines correspond with similar rubrics published by APA Division 12 (Chambless & Hollon, 1998) and Division 54 (Spirito, 1999) to describe the empirical support for psychological and related therapies. The EBA-TF guidelines preserve several of the defining features of the empirically supported treatment (EST; Spirito) guidelines. For example, both EST and EBA-TF criteria highlight the importance of manualized procedures to allow for replication of findings (i.e., EST Criterion III; EBA-TF Criterion II). Similarly, EST and EBA-TF criteria require verification of findings across different investigators or investigative teams (i.e., EST Criterion V; EBA-TF Criterion I) to reach the *Well-established* threshold.

Finally, both EST and EBA-TF guidelines require a degree of subjectivity in judging the quality of the empirical data base, whether the appropriateness of experimental design in the case of ESTs or psychometric properties of measures in the case of EBAs. For example, EST guidelines require that both between group and single case experiments demonstrate ''good'' design features, such as possessing adequate statistical power. Similarly, the EBA-TF criteria require that ''good'' psychometric data (i.e., reliability and validity statistics) be published in order for a measure to qualify as a *Well-established* assessment. Of course, opinions vary as to what constitutes ''good'' psychometric evidence, particularly in the case of evaluating test validity as tests may be used for varied purposes. We operationally defined psychometric guidelines identified by the EBA-TF (i.e., Criteria III) after consulting several works describing technically adequate measurement (Strauss, Sherman, & Spreen, 2006). For reliability evidence, we determined that internal consistency reliability ≥.80 was ‘‘good’’ and temporal stability reliability ≥.70 was ‘‘good.’’ For validity evidence, we summarized different types of data and in the case of evaluating validity coefficients (e.g., correlation between the test and similar measures) no specific thresholds were established. Indeed, creating a criterion for what constitutes ‘‘good’’ validity is particularly challenging as judgments of validity are tied to the purpose of the assessment. For example, a test may demonstrate diagnostic utility but be poorly suited for predicting adaptive outcomes.

Defining criteria for ‘‘good’’ psychometric properties for some psychological tests, such as appropriate normative sampling and reliability values, proved to be challenging. For example, the value of norm-referencing for some tests has been debated within the clinical neuropsychology literature. Authors questioning the value of norm-referencing have argued that norm-referenced standard scores: (a) may not be useful in describing clinical symptoms, (b) are antithetical toward the original purpose of tests designed to discriminate between individuals who are designated to have normal functioning and those with brain damage, and (c) are problematic for tests that produce nonnormal distributions for typically developing children and clinical populations (Lezak, Howieson, Loring, 2004; Retzlaff & Gibertini, 2000).

Review Procedures and Works Consulted

As a general procedure, we first consulted the test manual or appropriate text for each test followed by a search for studies examining the measure, particularly its use with pediatric populations. We conducted literature searches using the *Social Sciences Citation Index* and *MedLine* to locate relevant articles for each measure and consulted three compendia: (a) *Neuropsychological Evaluation of the Child* (Baron, 2004), (b) *A Compendium of Neuropsychological Tests* (Strauss et al., 2006), and (c) *Handbook of Normative Data for Neuropsychological Assessment* (Mitrushina, Boone, Razani, & D’Elia, 2005). The texts were consulted, in part, due to the fractional normative information presented for several tests in the literature and manuals.

Review and Critique of Measures

Our reviews focused on internal consistency and temporal stability reliability for measures identified in the survey; we also examined inter-rater reliability and reliability reported from diverse samples. Validity data were organized according to three broad categories: (a) criterion-concurrent validity (i.e., statistical relationships between the measure with measures of similar constructs), (b) criterion-group validity (i.e., comparing test performance between patient groups and controls), and (c) construct-structural/factor analytic validity (i.e., support for the theoretical model underlying the measure). We also reviewed additional validity evidence when available, such as predictive validity (i.e., criterion validity where test performance predicts later outcome) and divergent validity (i.e., low statistical relationships between the test and measures of dissimilar constructs). A summary table of our findings is accessible at the following web address: www.societyofpediatricpsychology.org.

General Intelligence

We reviewed eight measures of general intelligence representing some of the most well-known assessment instruments in psychology, such as the Wechsler intelligence tests. Psychometric data for measures of general intellectual functioning were strong for the measures selected by Division 54 listserv subscribers; each measure was rated as ‘‘*Well established*.’’ Internal consistency reliabilities for overall test composite scores (e.g., Wechsler Full Scale IQ) ranged from .93 to .98. Index and cluster scores (e.g., Wechsler Performance IQ,) also consistently met or exceeded .90. At the subtest level, internal consistency reliability ranged from .69 to .95, showing greater variability than index and cluster scores.

For overall IQ scores, temporal stability reliability ranged from .72 to .96 across all scales. For index and cluster scores, temporal stability reliability ranged from .60 to .97 across scales with evidence for greater stability for measures of verbal functioning. At the subtest level, temporal stability reliability revealed wide variability with values ranging from .50 to .94. Generally, there was evidence for less stability for measures of timed task performance (e.g., WISC-IV Processing Speed Index $M = .86$), when compared to verbal comprehension measures (e.g., WISC-IV Verbal Comprehension $M = .93$). Due to the possibility for unreliable measurement over time, interpretation of single subtest scores from test batteries should be undertaken cautiously and only after the subtest in question has been identified as reaching a minimal stability coefficient for the appropriate age group.

An impressive amount of validity data is presented in test manuals for the measures reviewed. Factor analytic evidence is typically presented as well as numerous criterion-related validity studies. The most recent revisions also typically report numerous criterion-group analyses to validate the measure. In addition, the Wechsler series includes reliability data for special population groups to demonstrate reliable measurement within clinical groups. As a group, the general intellectual measures have been subjected to strong standardization procedures and include nationally representative samples with the notable exception of the McCarthy Scales' outdated norms. As measures have been revised, manuals report an increasing amount of evidence supporting their use. Test manuals for recent versions of measures report extensive reliability evidence typically separated by age group and averaged across the normative sample. The amount of validity evidence presented also has grown, with test manuals reporting data relevant to content, construct, and criterion validity of tests.

Nonverbal intelligence

We reviewed three measures of nonverbal intelligence, the Columbia Mental Maturity Scale (CMMS; Burgemeister, Hollander, & Lorge, 1972), Leiter International Performance Scale-Revised (Leiter-R; Roid & Miller, 1995, 1997), and the Raven Progressive Matrices (RPM; Raven, Raven, & Court, 1998) and each received ''*Well-established*'' ratings. As a group, internal consistency and temporal stability ranged from .85 to .93 for total (CMMS; RPM) or FSIQ scores (Leiter-R). In addition to the FSIQ score, the Leiter-R yields composite and subtest scores with median reliability values that range from .80 to .85. Concurrent validity has been established for each measure as evidenced by relationships with measures of nonverbal and general intelligence with correlations typically falling in the .5–.8 range. The Leiter-R has been subjected to CFA that supports a multi-factorial structure of the scale; the RPM often is considered to be a highly ''g''-saturated measure, but also has been found to have a hierarchical factor structure with ''g'' and three lower-order factors (Lynn, Allik, & Irwing, 2004). Little validity research has been conducted as of late with the CMMS, which is likely due to the fact that the measure was normed in 1970 to match 1968 US census data.

From a clinical perspective, nonverbal intellectual functioning may be assessed to estimate general intelligence for children with significant language or motor impairments (e.g., developmental language disorders), children who cannot understand English, or for children with hearing impairments. From the measures identified by the listserv, the CMMS cannot be recommended for clinical assessment due to the outdated norms, despite the CMMS meeting ''*Well-established*'' assessment criteria. Based on data presented in the test manual, the Leiter-R appears to be well-normed; however, the normative sampling procedures for the RPM lack important detail, i.e., the number of children sampled are not provided (Strauss et al., 2006).

Academic Achievement

We reviewed four measures in the Achievement category: (Peabody Individual Achievement Test-Revised [PIAT-R; (Markwardt, 1998)], Wechsler Individual Achievement Test, Second Edition [WIAT-II; (Wechsler, 2002)], Wide Range Achievement Test 3 [WRAT-3; (Wilkinson, 1993)], Woodcock-Johnson III Tests of Achievement [WJ-III; (McGrew & Woodcock, 2001)] all of which received ''*Well-established*'' ratings. Internal consistency and test–retest reliability for each of the instruments have been demonstrated to be high, although low subtest reliability has been reported for some subtests of the WIAT-II, PIAT-R, and WJ-III. This is not surprising given that there are few items available on some of the subtests, particularly for younger children. Construct validity for the PIAT-R, WIAT-II, and the WRAT-3 including convergent and discriminant validity has been demonstrated for the majority of subtests with the exception of the WJ-III, which features moderate reliability in some studies. Also of interest is the fact that achievement scales and measures of intelligence (e.g., WISC–III FSIQ) are only moderately related, thereby suggesting that academic achievement is not a proxy for intellectual functioning. Normative data are ample for various age groups for each of the achievement measures reviewed.

In selecting an achievement measure either for clinical use or for a dependent measure in an investigation, the psychologist must first have an understanding of the purpose for employing the achievement measure. For example, the WRAT-3 primarily is a screening instrument that broadly examines reading decoding, basic spelling, and computational arithmetical abilities. In contrast, the PIAT-R, WIAT-II, and the WJ-III assess a broader array of achievement across many domains including reading comprehension, applied mathematics, and written language. Thus, for the purpose of clinical assessment, the WJ-III, WIAT-II, and PIAT-R, when compared to the WRAT-3, provide more comprehensive assessment across academic skill areas.

Attention/Executive Functioning

Routine assessment of attention/executive functioning has been suggested for some pediatric populations, such as children with SCD due to the increased risk of frontal lobe impairment via stroke and silent infarct (Brown et al., 2000). Listserv respondents identified two instruments proposed to measure attention and executive functioning [Conners Continuous Performance Task-II (CPT II; Conners & MHS Staff, 2000); Trail Making Test (TMT; Reitan & Wolfson, 1993)] and both received ''*Approaching well-established assessment*'' ratings. Problems exist for TMT child norms, although reliability and validity data exist for some pediatric populations, such as children with SCD. Despite incomplete normative data, the TMT is widely used and has been employed in a number of published studies thereby receiving the designation of ''*Approaching well-established assessment*.'' The CPT II has high split-half reliability for both omission ($r = .95$) and commission ($r = .94$) error scores, and test–retest reliability has been fairly high for participants with attention-deficit hyperactivity disorder (ADHD; $r = .89$) and those with neurological impairments ($r = 92$). However, the test–retest reliability for CPT II index scores has ranged from.05 to.92. Adequate construct validity has been established for both measures. For example, studies examining the factorial validity of the Halstead–Reitan Battery (HRB) found support for the TMT as a measure of visual attention.

With regard to convergent validity, the CPT II has correlated with other assessments of attention and overactivity including behavioral ratings, and the CPT II and TMT have garnered evidence in support of criterion-group validity. Test authors report that the CPT-II has differentiated those samples with ADHD and other neurological disorders from typical peers (Conners & MHS Staff, 2000). In contrast, however, McGee, Clark, and Symons (2000) found poor criterion-group validity between children with ADHD and reading disorders. The TMT has been found to discriminate between children with achievement deficits and learning disabilities from their normally developing peers. When compared with TMT Trails A, Trails B tends to be more sensitive in discriminating children with neurological and learning impairments from their normally developing peers.

In summary, the reliability and validity data that have been reported for the CPT II are generally more consistent than for the TMT, although some of the subscales have been shown to be less reliable than others. Psychometric support for the TMT consists largely of criterion-group validity, particularly detection of brain dysfunction versus controls. The TMT has a short-term memory component, which allows for discrimination between clinical populations characterized by neurological impairments and typically developing children. In deciding between the two instruments either for research or for clinical activities, the CPT II has a greater track record with regard to reliability data; the TMT needs greater support with respect to reliability with pediatric populations. The CPT II assesses vigilance without any higher order learning, whereas the TMT assesses both attention and short-term memory that may explain its consistent track record in identifying children with neurological impairments. Nonetheless, both instruments have been employed widely in the clinical and research literatures.

Memory and Learning

Assessment of memory and learning in pediatric populations may be warranted for reasons outlined earlier in the review. Recently, the importance of memory has been highlighted for children with IDDM, particularly in view of the predictive relationship between verbal memory skills and self-care behaviors for adolescents (Soutor, Chen, Streisand, Kaplowitz, & Holmes, 2004). In the areas of memory and learning, respondents identified two instruments, the California Verbal Learning Test-Children's Version (CVLT-C; Delis, Kramer, Kaplan, & Ober, 1994), and the Wide Range Assessment of Memory and Learning (WRAML; Sheslow & Adams, 1990); both measures were rated as ''*Well established*.''

Reliability for both instruments is fairly high, albeit variable, with internal consistency reliabilities ranging from .72 to .96; the highest reliability has been reported for the verbal memory and general memory indexes for both instruments. Test–retest reliability is more variable for the CVLT-C (.31–.90) than for the WRAML (.61–.84). Criterion validity for both instruments has been documented with the CVLT-C demonstrating sensitivity to memory impairments among children with TBI and females with ALL. The CVLT-C differentiates children with dyslexia from their normally developing peers, and the WRAML has been found to discriminate those children with comorbid learning disabilities and ADHD from those with ADHD alone. In addition, the WRAML has shown fairly high convergent validity with various memory indices found on general intelligence tests, and has been found to discriminate children with severe TBI from those with mild to moderate TBI. Finally, the factorial validity of each scale has been established with at least one pediatric group: the WRAML with children with diabetes (Lynch, Chen, & Holmes, 2004) and the

CVLT-C with children with epilepsy (Griffiths et al., 2006). In making decisions as to which measure to use, both instruments yield similar psychometric data, with the CVLT-C yielding slightly better prediction with regard to academic outcomes for children with TBI, an area frequently studied in the pediatric psychology literature. Nonetheless, the WRAML seems to be a more useful instrument in the examination of both verbal and visual memory. Thus, the decision as to which of these instruments to use largely depends on the questions pertaining to visual or verbal memory that have been posed by the investigator or the clinician.

Visual-motor and Motor Functioning

We reviewed six measures endorsed by listserv respondents designed to assess a variety of aspects of visual-motor functioning, such as fine motor dexterity (e.g., pegboard tasks), strength, speed (e.g., finger tapping), and graphomotor skills. The Beery–Buktenica Developmental Test of Visual Motor Integration (VMI; Beery & Beery, 2004), Bender Visual Motor Gestalt Test (Bender; Bender, 1946) and its updated version, the Bender-II (Brannigan & Decker, 2003), Finger Tapping Test (FTT; Reitan & Wolfson, 1993), the Grip Strength Test (GST; Reitan & Wolfson), the Grooved Pegboard Test (GPT; Trites, 1989), and Rey-Osterrieth Complex Figure Text (ROCF; Rey & Osterreith, 1993) were reviewed. From the group of measures, three measures were rated as ''*Well established*,'' whereas the remaining measures were rated as ''*Approaching well-established*.''

Three measures reviewed (i.e., GST, GPT, and FTT) lacked adequate norms for children. Normative data for the GST and FTT have been presented based on unpublished ''meta-norms,'' which were produced by combining data presented in 20 articles published between 1969 and 1989 (Baron, 2004; Strauss et al., 2006). FTT meta-norms were created from the performance of 1,591 primarily middle- to upper-class Caucasian children with high average IQ (Strauss et al.). Similarly, GST meta-norms were created using data from 822 children by combining males and females within age groups. Combining male and female data for both FTT for GST performance is problematic due to observed sex differences favoring males over females for both tests (Roselli, Ardila, Bateman, & Guzman, 2001). Other normative data for the GST have been culled from several other articles and presented in published texts in child neuropsychology; however, these data are outdated and several age groups are not represented (e.g., ages 9–10; Strauss et al., p. 1055). No ''meta-norms'' appear to exist for the GPT; however, the test manual presents normative data for children reprinted from varied sources that predate 1987. Therefore, available norms for the FTT, GST, and GPT are outdated, not representative of the US population, and, for some age groups, do not sample an adequate number of children (e.g., $n = 23$ 10-year-olds for the GST; $n = 11$ 6-year-olds for the GPT).

Due to the administration of ROCF copy and recall trials, the measure is described as a task of visual organization, visual-motor construction (copy) and visual memory (recall), among others. As such, multiple scoring criteria exist for the ROCF (Knight, 2003), including two systems recently published that feature detailed scoring criteria and reasonably adequate norms for children and adolescents ($N = 454$, Bernstein & Waber, 1996; $N = 505$, Meyers & Meyers, 1996). The Bender–Gestalt test has been recently re-normed with an impressive normative sample using detailed scoring procedures and has shown good psychometric properties including reliability, concurrent validity, criterion-group validity, and construct validity.

Language

In the area of language, only the Peabody Picture Vocabulary Test, Third Edition (PPVT-III; Dunn & Dunn, 1997) was identified by numerous respondents and found to be a ''*Well-established*'' instrument. The PPVT-III is a brief measure of single word receptive vocabulary, and according to the test manual, may be used as a means of estimating verbal cognitive ability. As such, the PPVT has been used as a proxy for cognitive functioning with pediatric populations, such as children with spina bifida (Rose, Holmbeck, Coakley, & Franks, 2004). Internal consistency ($r = 95$) and test–retest ($r = 92$) reliability are high, although construct validity is variable ranging from .40 to .87 when validating the PPVT-III among children with autism. For typically developing children, however, convergent validity with WISC-III Verbal IQ is high. Finally, for young African-American children construct validity also has been demonstrated to be fairly low with a widely employed test of intelligence. Thus, our workgroup provided only one endorsement of an assessment of language among pediatric populations and the reliability and convergent validity for the PPVT-III among typically developing children has been demonstrated to be fairly high. Nonetheless, the astute investigator or practitioner should proceed with caution in using this measure with young African-American children and special populations due to concerns with regard to construct validity.

Listserv respondents identified a limited number of language instruments, which may be due to findings that language functions are fairly well-preserved among children with chronic illness. The limited number of language measures may also reflect that pediatric psychologists are more frequently asked to assess short-term change in cognitive functions, such as attention/executive functioning, memory, and motor functions.

The NEPSY: A Comprehensive Neuropsychological Battery

Finally, respondents identified one comprehensive assessment battery, the NEPSY (Korkman, Kirk, & Kemp, 1998), which was designed to assess a variety of cognitive functions: attention/executive, language, sensorimotor, visuospatial, and memory/learning. Given the NEPSY's assessment of multiple areas, reviewers opted to review the NEPSY in a separate category. The NEPSY was developed, in part, in response to the weak standardization and norming properties described for other cognitive measures used with children. The NEPSY standardization sample consisted of 1,000 children representative of 1995 US census data.

The NEPSY was classified as a ''*Well-established*'' assessment by the CAWG task force. Internal consistency reliability ranged from .70 to .91 for the five domain scores, and subtest scores ranged from .50 to .91. Temporal stability ranged from .67 to .90 for domains and .42–.89 for subtests. The test manual provides concurrent and criterion-group validity in support of the NEPSY; however, the hypothesized factor structure of the NEPSY was not evaluated and some of the concurrent validity findings are weak. For example, the NEPSY Sensorimotor Domain was only slightly correlated with the Bayley-II Psychomotor Development Index. Independent evaluations have provided initial support for the NEPSY. For example, NEPSY performance differs among children with neurological impairment, children with academic problems, children with autism, and comparison controls (Hooper, Poon, Marcus, & Fine, 2006; Schmitt & Wodrich, 2004). In contrast, however, Stinnett, Oehler-Stinnett, Fuqua, and Palmer (2002) provide evidence for a single factor for the NEPSY as opposed to the proposed five-factor structure outlined by the author. Clearly, efforts to validate, revise, and improve the NEPSY should continue; however, the NEPSY meets ''*Well-established*'' criteria.

Conclusions, Limitations, and Recommendations

All told, ~81% (22 of 27) of the measures identified by pediatric psychologists and reviewed by CAWG met EBA-TF criteria for ''*Well-established*'' classification and, in general, the group of cognitive measures endorsed by EBA-TF survey respondents demonstrates good psychometric properties. The most notable exceptions are several outdated measures of intelligence (e.g., CMMS) and several subscales from the HRB. From the HRB, measures of motor strength, speed, and dexterity suffer from incomplete or outdated normative data; the TMT, a measure of attention and executive functioning, also suffers from problematic normative data for children.

In general, recently published measures of cognitive functioning and achievement featured comprehensive norms, strong reliability support, and detailed validity findings. For all measures, independent psychometric evaluation consisted primarily of validation as opposed to reliability analysis. Validation most often consisted of criterion-related validity, such as concurrent validity between the measure and a similar measure or criterion-group analyses. Criterion-group comparisons most often consisted of a typically developing group versus pediatric comparison group as opposed to comparison of two pediatric groups. With the exception of the measures of cognitive functioning, few studies examined structural validity. Despite psychometric strengths, we acknowledge that ''*Well-established*'' measures are imperfect and future improvements are needed. For example, measures fall short in documenting meaningful change in test performance over time, which the committee acknowledged as a key purpose for assessing cognitive functioning with pediatric populations.

The splintered and outdated normative data for several tests brings the discussion of the value and limitations of norming into the foreground, an issue that has been debated within the clinical neuropsychology literature. A guiding principle of psychological assessment within typically developing populations is norm-referenced measurement, whereby an individual's performance may be compared against a meaningful reference group, such as a nationally representative sample of typical peers. Several traditional measures, such as the TMT, lack adequate norms to make useful comparisons with typical peers; however, some have argued that such comparisons may be of limited value, particularly for children who demonstrate significant impairment (e.g., use of a norm-referenced vocabulary test to assess a child with aphasia; Lezak et al., 2004). One possible outcome in light of this debate is the creation of condition-specific norm groups, such as those published for adaptive functioning for children with autism (Carter et al., 1998), for the purpose of more precisely describing functioning and tracking progress.

Attempts to address inadequate norms have yielded ''meta-norms,'' that is, normative data synthesized from individual normative studies that typically span many years. Soukup, Ingram, Grady, and Schiess (1998) have argued against the creation of meta-norms when there are significant inconsistencies found across the smaller norming samples, such as those for the TMT. Creating meta-norms from smaller studies may not account for differences in test performance for sex and IQ. It also is unclear how culling information in this manner over time might affect measurement outcomes. For example, as IQ tests undergo revision, norms become more stringent over time, the so-called ''Flynn effect'' (Flynn, 1984). In addition to being outdated and incomplete, meta-norms cited in the child neuropsychology literature are not readily available, although recently published texts are making these more accessible (Baron, 2004).

As evidenced by criterion-group validity findings, cognitive measures appear to be sensitive to broad CNS impairment, particularly in differentiating between groups with known cognitive or learning disorders and typical peers. In general, however, these tests perform less favorably in distinguishing between groups of children with cognitive disorders. That is, the cognitive measures reviewed show adequate sensitivity, but limited specificity. Poor specificity is a problem when working with children with chronic illness because specific cognitive functions, such as motor coordination or attention, are more frequently affected as opposed to global impairments in general cognitive functioning. With respect to research endeavors, the strongest measures from a psychometric standpoint (e.g., IQ and ACH) are not typically identified as dependent variables in intervention studies.

As a group, many cognitive measures fail to describe the functional implications of test performance, and, in most cases, yield data that do not demonstrate ecological validity (Silver, 2000). Ecological validity has been defined as consisting of two parts: *similarity* of a test to naturally occurring environmental demands, and *prediction* of behavior within real-world environments based on test performance (Franzen & Wilhelm, 1996). Good examples of this limitation are computer-administered measures of sustained attention or vigilance, such as the CPT-II, which are routinely criticized for lacking ecological validity. For example, Epstein et al. (2003) failed to find expected relationships among CPT-II omission errors and symptoms of inattention and associations between commission errors and hyperactive/impulsive symptoms.

Evaluating the ecological validity of cognitive test performance is a complex endeavor. For example, obtaining a child's best performance under optimal conditions often is a goal of assessment to guide modifications to learning environments outside of the testing situation that ''match'' the child's strengths and weaknesses (Silver, 2000). Therefore, utilizing tests that possess adequate construct validity and low ecological validity may be justified in order to guide intervention planning. In addition to predicting outcomes from a compromised yet developing CNS, as in the case of late effects, establishing ecological validity is further complicated by the influence of environmental variables on cognitive outcomes, such as the contributions of family adjustment on TBI outcomes (Ewing-Cobbs & Bloom, 2004). Potential responses to the ecological validity problem include synthesis of data from multiple sources (e.g., cognitive testing and behavioral ratings; Silver), emphasis of ecological validity at test creation (e.g., ''everyday'' memory tasks vs. recalling word lists), or employng novel approaches such as virtual reality to enhance environmental similarity (Rizzo, Schultheis, Kerns, & Mateer, 2004).

Within the field, a growing number of studies examine how children with illnesses with indirect CNS effects perform on measures identified in this review, which should continue. Most of the recently revised measures (e.g., WISC-IV) include criterion-group validity data comparing typical groups with children with learning disabilities, developmental disabilities, and other cognitive problems, such as ADHD; however, children with medical illness are less represented. Our review reveals that commonly examined pediatric groups are children with BT, TBI, seizure disorders, and other illnesses with direct CNS effects.

We offer several recommendations in response to the limitations identified in our review. As highlighted in the review, we need to correct the norming problem for several measures used in the pediatric literature, such as the popular and time-honored TMT. One straightforward solution to this problem is to create a set of updated norms with a nationally representative sample of children, similar to what has been accomplished with the BGT. Although tedious, this work is important and may not be accomplished without coordinating the efforts of a team of investigators. Another possibility is for the field to use similar measures that feature adequate norms, such as the Comprehensive Trail Making Test (Reynolds, 2002).

Related to general normative expectations, the field would benefit from further documentation of how groups of children with specific illnesses perform on cognitive measures. One possible solution is to consider if and how data might be synthesized across collaborative research sites, such as the Stroke Prevention Trial in Sickle Cell Disease (STOP). Combined with creation of illness-specific norms for measures, future validation of cognitive tests should aim to link test data with salient functional implications for groups. A good example of this strategy exists for children with autism. Carter et al. (1998) provide autism-specific normative data for the Vineland Adaptive Behavior Scales, allowing for detailed measurement of important functional domains for children with autism, such as communication skills. Similarly, it would be useful to know what level of performance to expect from a child with illness and subsequently translate deviations from expected performance to functional limitations and outcomes.

Another recommendation that may facilitate achieving the aforementioned goals is to synthesize the current empirical support for measures with specific disorders. For example, Mash and Hunsley (2005) outlined a detailed review strategy for specific instruments, whereby norming procedures, reliability, and validity are evaluated for specific groups of children. In contrast to the test-centered review completed for this special series, an illness-focused review with this level of detail might prove more useful to pediatric psychologists when selecting cognitive instruments for clinical or research purposes. A series of reviews may also focus work on the gaps remaining within a specific illness group. For childhood survivors of ALL and BT, work of this sort has begun via proposed practice standards and research batteries for psychological evaluation (Mulhern, Armstrong, & Thompson, 1998; Mulhern & Butler, 2006).

In addition to accounting for diversity across illness groups, the potential cultural non-equivalence of tests across diverse subgroups of children is an area that clearly warrants additional consideration. Although outside the scope of the review, the documented and potential for cultural nonequivalence of many cognitive tests is particularly relevant for pediatric psychologists working with US populations as well as those working in countries outside of the US. Within the current EBA review template, future reviewers are advised to incorporate evidence of cultural equivalence as an additional criterion for evaluating the empirical support for assessments.

We also found a need to create more refined classification criteria. For example, there was no explicit criterion to address problems that our work group encountered regarding poor or questionable norming procedures. Consistent with arguments outlining limitations of norm-referenced measures (Retzlaff & Gibertini, 2000), several measures demonstrated concurrent and criterion-group validation in the presence of poor norms, and it was not clear how to reconcile these findings. Similarly, the EBA guidelines seemed generous in allowing assessments to meet ''*Well-established*'' criteria in the presence of a single supporting investigation when it was possible for a majority of studies to report contrary findings. One possible modification to the EBA guidelines is to include evaluation criterion for adequate norming procedures and require multiple documentation of adequate psychometrics.

Our review documents that pediatric psychologists use ''*Well-established*'' assessments in clinical practice and research within each cognitive domain, with the exception of measures of attention/executive functioning. Most of the measures identified in the review (81.5%) meet ''*Well-established*'' criteria; however, a small group of traditionally used measures of motor functioning were found to be inadequately normed. As cognitive measures have undergone revision, test authors and publishers have reported an increasing amount of standardization, reliability, and validity data supporting test usage. Despite the positive ratings for many instruments, however, there are gaps in the pediatric literature both in terms of the typical expectations for many pediatric illness groups and the functional implications of test performance. The EBA-TF classification criteria were found to be a useful starting point for evaluating empirical support for cognitive measures in pediatric psychology, but the broad-based categories proved somewhat limiting in the process.

Two important limitations deserve mention with respect to how well our conclusions may or may not generalize to pediatric psychologists. First, participants for the survey were recruited using the Division 54 listserv, which resulted in unknown sample characteristics. For example, Division 54 listserv subscribers do not have to be trained as pediatric psychologists and there may be a number of pediatric psychologists who do not subscribe to the listserv. For subscribers, some addresses may have been defunct, and some addresses may have blocked the attached survey due to its large size. Second, only 27% of Division 54 listserv subscribers responded to the survey. All told, participant recruitment methods and subsequent response rates limit the generalizability of conclusions reached in our review.

Despite these limitations, we hope that our review provides some direction for the selection of cognitive instruments in both research and clinical practice in pediatric psychology. At the very least, it is hoped that this review provides future direction in both the development and validation of cognitive instruments for use with children with chronic illness, developmental disabilities, and other pediatric conditions.

Conflicts of interest: None declared.

Received December 31, 2006; revision received and accepted December 20, 2007

Appendix A. Measures Reviewed

Bayley Scales of Infant and Toddler Development, Third Edition (Bayley-III)
Central reference. Bayley, N. (2006). *Bayley Scales of Infant and Toddler Development (3rd Edition) Technical Manual*. San Antonio, TX: Harcourt Assessment.
Purpose of measure. The Bayley-III assesses cognitive and developmental functioning.
Address for manual and measure. The Psychological Corporation, A brand of Harcourt Assessment, Inc., 19500 Bulverde Road, San Antonio, TX 78259, USA; www.PsychCorp.com
EBA classification. The Bayley-III is a well-established assessment.

Beery-Buktenica Developmental Test of Visual-Motor Integration, 5th Edition (VMI).
Central reference. Beery, K. E., & Beery, N. A. (2004). *The Beery-Buktenica Developmental Test of Visual-Motor Integration (5th edition) Administration, Scoring, and Teaching Manual*. Minneapolis, MN: NCS Pearson, Inc.
Purpose of measure. The VMI is designed to assess visual-motor functioning.
Address for manual and measure. NCS Pearson, Inc., PO Box 1416, Minneapolis, MN, USA; www.pearsonassessment.com
EBA classification. The VMI is a well-established assessment.

Bender Visual Motor Gestalt Test (Bender)
Central references. Bender, L. (1946). *Instructions for the use of Visual Motor Gestalt Test*. Alexandria, VA: The American Orthopsychiatric Association, Inc.
Brannigan, G. G., & Decker, S. L. (2003). *Bender Visual-Motor Gestalt Test* (2nd ed.). Itasca, IL: Riverside Publishing.
Koppitz, E. M. (1975). *The Bender Gestalt Test for Young Children. Volume II: Research and Application 1963–1973*. Orlando, FL: Grune & Stratton, Inc.
Purpose of measure. The Bender measures visual-motor functioning and visual-perceptual skills.
Address for manual and measure. Original Bender: American Orthopsychiatric Association, 2001 N. Beauregard Street, 12th Floor, Alexandria, VA 22311, USA; Amerortho@aol.com.
Bender-II: Riverside Publishing Company, 425 Spring Lake Drive, Itasca, IL 60143-2079, USA; www.riversidepublishing.com
EBA classification. The Bender is a well-established assessment.

California Verbal Learning Test Children's Version (CVLT-C)
Central reference. Delis, D., Kramer, J. H., Kaplan, E., & Ober, B. A. (1994). *California Verbal Learning Test*. San Antonio, TX: The Psychological Corporation.
Purpose of measure. The CVLT-C is a measure of verbal learning and memory.
Address for manual and measure. Psychological Corporation, 1950 Bulverde Road, San Antonio, TX 78259, USA; www.harcourtassessment.com
EBA classification. The CVLT-C is a well-established assessment.

Columbia Mental Maturity Scale (CMMS)
Central reference. Burgemeister, B. B., Hollander, L. H., & Lorge, I. (1972). *Columbia Mental Maturity Scale: Guide for administering and interpreting*. US: Harcourt Brace Jovanovich, Inc.
Purpose of measure. The CMMS is a measure of general reasoning ability.
Address for manual and measure. Education Measurement Division, The Psychological Corporation, 757 Third Avenue, New York, NY 10017, USA.
EBA classification. The CMMS is a well-established assessment.

Conners' Continuous Performance Test II (CPT II)
Central references. Conners, C. K., & MHS Staff (2000). *Conners' Continuous Performance Test II Computer Program for Windows Technical Guide and Software Manual*. Toronto, ON: Multi-Health Systems, Inc.
Purpose of measure. The CPT II is a measure of vigilance and sustained attention.
Address for manual and measure. Multi-Health Systems, Inc., PO Box 950, North Tonawanda, NY 14120-0950, USA; www.mhs.com
EBA classification. The CPT-II is approaching a well-established assessment.

Halstead–Reitan Finger Tapping Test (FTT)
Central reference. Reitan, R. M., & Wolfson, D. (1993). *The Halstead-Reitan Neuropsychological Test Battery: Theory and clinical interpretation* (2nd ed). Tucson, AZ: Neuropsychological Press.
Purpose of measure. To measure maximal motor speed of the index finger of each hand.

Address for manual and measure. Reitan Neuropsychology Laboratory, 2920 4th Street, Tuscon, AZ 85775-1336, USA (www.reitanlabs.com).

EBA classification. The HFTT is approaching a well-established assessment.

Halstead–Reitan Grip Strength Test

Central reference. Reitan, R. M., & Wolfson, D. (1993). *The Halstead-Reitan Neuropsychological Test Battery: Theory and clinical interpretation* (2nd ed.). Tucson, AZ: Neuropsychological Press.

Purpose of measure. To measure the motor strength of each hand.

Address for manual and measure. Reitan Neuropsychology Laboratory, 2920 4th Street, Tuscon, AZ 85775-1336, USA (www.reitanlabs.com).

EBA classification. The Grip Strength test is approaching a well-established assessment.

Kaufman Assessment Battery for Children, 2nd Edition (KABC-II)

Central references. Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman assessment battery for children (2nd edition) manual*. Circle Pines, MN: AGS.

Purpose of measure. The KABC-II is an individually administered measure of mental processing and cognitive ability for children and adolescents.

Address for manual and measure. AGS Publishing, 4201 Woodland Road, Circle Pines, MN 55014-1796, USA; www.agsnest.com

EBA classification. The KABC-II is a well-established assessment.

Lafayette Grooved Pegboard (GPT)

Central references. Trites, R. L. (1989). *Lafayette Grooved Pegboard Task. Instruction/Owner's Manual. Lafayette*, IN: Lafayette Instrument Company.

Purpose of measure. The GPT is a measure of eye-hand coordination and motor speed.

Address for manual and measure. Lafayette Instrument, 3700 Sagamore Parkway N. PO Box 5729, Lafayette, IN 47903, USA; Lic@licmef.com

EBA classification. The GPT is approaching a well-established assessment.

Leiter International Performance Scale-Revised (Leiter-R)

Central reference. Roid, G. H., & Miller, L. J. (1995, 1997). *Leiter International Performance Scale-Revised*. Wood Dale, IL: Stoelting Co.

Purpose of measure. The Leiter-R is a measure of cognitive functioning, particularly nonverbal intelligence.

Address for manual and measure. Stoeling Company, 620 Wheat Lane, Wood Dale, IL 60191, USA.

EBA classification. The Leiter-R is a well-established assessment.

McCarthy Scales of Children's Abilities (MSCA)

Central reference. McCarthy, D. (1972). *McCarthy Scales of children's abilities*. New York: The Psychological Corporation.

Purpose of measure. The MSCA is measure of various cognitive abilities.

Address for manual and measure. Riverside Publishing Company, 425 Spring Lake Drive, Itasca, IL 60143–2079, USA; www.stanford-binet.com

EBA classification. The MSCA is a well-established assessment.

NEPSY: A Developmental Neuropsychological Assessment

Central references. Korkman, M., Kirk, U., & Kemp. S. (1998). *NEPSY: A Developmental Neuropsychological Assessment.* San Antonio, TX: The Psychological Corporation.

Purpose of measure. The NEPSY is a measure designed to assess neuropsychological functioning in five domains.

Address for manual and measure. Harcourt Assessment, Inc., 19500 Bulverde Road, San Antonio, TX 78259, USA (www.harcourtassessment.com).

EBA classification. The NEPSY is a well-established assessment.

Peabody Individual Achievement Test-revised (PIAT-R)

Central references. Markwardt, F. C. (1989). *Peabody individual achievement test-revised*. Circle Pines, MN: American Guidance Service.

Markwardt, F. C. (1998). *Peabody individual achievement test-revised normative update*. Circle Pines, MN: American Guidance Service.

Purpose of measure. The PIAT-R is a measure of academic achievement.

Address for manual and measure. American Guidance Service, Inc., 4201 Woodland Road, Circle Pines, MN 55014-1796, USA; www.agsnet.com

EBA classification. The PIAT-R is a well-established assessment.

Peabody Picture Vocabulary Test, Third Edition (PPVT-III)

Central references. Dunn, L. M., & Dunn, L. M. (1997). *Examiner's manual for the Peabody Picture Vocabulary Test - Third Edition*. Circle Pines, MN: American Guidance Service.

Purpose of measure. The PPVT-III is a measure of single-word receptive vocabulary.

Address for manual and measure. American Guidance Service, Inc., 4201 Woodland Road, Circle Pine, MN 55014-1796, USA; www.agsnet.com

EBA classification. The PPVT-III is a well-established assessment.

Raven Progressive Matrices (RPM)

Central reference. Raven, J., Raven, J. C., & Court, J. H. (1998). *Raven manual: Section 1. General overview*. Oxford: Oxford Psychologists Press Ltd.

Purpose of measure. The RPM tests measure general cognitive ability.

Address for manual and measure. Harcourt Assessment, Inc., 19500 Bulverde Road San Antonio, TX 78259, USA; www.harcourtassessment.com

EBA classification. The RPM is a well-established assessment.

Rey-Osterrieth Complex Figure Test (ROCF)

Central references. Mitrushina, M., Boone, K. B., Razani, J., & D'Elia, L. F. (2005). *Handbook of normative data for neuropsychological assessment* (2nd ed.). New York: Oxford University Press.

Rey, A., & Osterreith, P. A. (1993). Translations of excerpts from Andre Rey's ''Psychological examination of traumatic encephalopathy'' and P. A. Osterrieth's ''The Complex Figure Copy Test.'' *Clinical Neuropsychologist,* 7, 4–21.

Purpose of measure. The ROCF assesses visuospatial constructional ability and visual memory. The ROCF also allows assessment of organization, planning, and problem-solving skills.

Address for manual and measure. Psychological Assessment Resources, Inc., 16204 N. Florida Avenue, Lutz, FL 33549, USA; www.parinc.com

EBA classification. The ROCF is a well-established assessment.

Stanford-Binet Intelligence Scales, Fifth Edition (SB-5)

Central references. Roid, G. H. (2003). *Stanford-Binet Intelligence Scales* (5th ed.). Itasca, IL: Riverside Publishing.

Purpose of measure. The SB-5 is a measure of intelligence and cognitive abilities.

Address for manual and measure. Riverside Publishing Company, 425 Spring Lake Drive, Itasca, IL 60143–2079, USA; www.stanford-binet.com

EBA classification. The SB-5 is a well-established assessment.

Trail Making Test (TMT)

Central references. Reitan, R. M., & Wolfson, D. (1993). *The Halstead-Reitan Neuropsychological Test Battery: Theory and clinical interpretation* (2nd ed.). Tucson, AZ: Neuropsychology Press.

Purpose of measure. The (TMT) is a measure of attention and executive functioning.

Address for manual and measure. Reitan Neuropsychology Laboratory, 2920 4th Street, Tuscon, AZ 85775-1336, USA (www.reitanlabs.com)

EBA classification. The TMT is approaching a well-established assessment.

Wechsler Adult Intelligence Scale, Third Edition (WAIS-III)

Central references. Wechsler, D. (1997). *Administration and scoring manual for the Wechsler Adult Intelligence Scale-Third Edition*. San Antonio: The Psychological Corporation.

Purpose of measure. The WAIS-III is a measure of general intellectual ability.

Address for manual and measure. Harcourt Assessment, Inc., 19500 Bulverde Road San Antonio, TX 78259, USA; www.harcourtassessment.com

EBA classification. The WAIS-III is a well-established assessment.

Wechsler Individual Achievement Test, Second Edition (WIAT-II)

Central references. Wechsler, D. (2002). *Wechsler Individual Achievement Test - Second edition. Examiner's manual*. San Antonio: The Psychological Corporation.

Purpose of measure. The WIAT-II is test of academic achievement in multiple domains.

Address for manual and measure. The Psychological Corporation, 19500 Bulverde Road, San Antonio, TX 78259, USA; www.PsychCorp.com

EBA classification. The WIAT-II is a well-established assessment.

Wechsler Intelligence Scale for Children, Fourth Edition (WISC-IV)

Central references. Wechsler, D. (2003). *Administration and scoring manual for the Wechsler Intelligence Scale for Children* (4th ed.). San Antonio: The Psychological Corporation.

Purpose of measure. The WISC-IV is measure of intellectual functioning.

Address for manual and measure. The Psychological Corporation, 19500 Bulverde Road, San Antonio, TX 78259, USA; www.PsychCorp.com

EBA classification. The WISC-IV is a well-established assessment.

Wechsler Preschool and Primary Scale of Intelligence-Third Edition (WPPSI-III)

Central reference. Wechsler, D (2002). *Manual for the Wechsler Preschool and Primary Scale of Intelligence, Third Edition (WWPSI-III)*. San Antonio, TX: The Psychological Corporation.

Purpose of measure. The WPPSI-III is an individually administered test of intellectual functioning for young children.

Address for manual and measure. The Psychological Corporation, 19500 Bulverde Road, San Antonio, TX 78259, USA; www.PsychCorp.com

EBA classification. The WPPSI-III is a well-established assessment.

Wide Range Achievement Test 3 (WRAT-3)

Central reference. Wilkinson, G. S. (1993). *The Wide Range Achievement Test: Administration Manual*. Wilmington, Delaware: Wide Range, Inc.

Purpose of measure. The WRAT-3 is a measure of academic achievement.

Address for manual and measure. Wide Range, Inc., 15 Ashley Place, Suite 1A, Wilmington, DE 19804-1314, USA.

EBA classification. The WRAT-3 is a well-established assessment.

Wide Range Assessment of Memory and Learning (WRAML)

Central references. Sheslow, D., & Adams, W. (1990). *Wide range assessment of memory and learning*. Wilmington, DE: Jastak

Purpose of measure. The WRAML is designed to assess the child's ability to actively memorize and learn a variety of verbal and visual information.

Address for manual and measure. Jastak Associates, Inc., 1526 Gilpin Avenue, Wilmington, DE 19806, USA; www.widerange.com

EBA classification. The WRAML is a well-established assessment.

Woodcock-Johnson III Tests of Achievement (WJ III Ach)

Central references. McGrew, K. S., & Woodcock, R. W. (2001). *Technical Manual. Woodcock-Johnson III*. Itasca, IL: Riverside Publishing.

Purpose of measure. The WJ III ACH is a measure of academic achievement.

Address for manual and measure. The Riverside Publishing Company, 425 Spring Lake Drive, Itasca, IL 60143-2079, USA; www.woodcock-johnson.com

EBA classification. The WJ-III Ach is a well-established assessment.

Woodcock-Johnson III Tests of Cognitive Abilities (WJ-III Cog)

Central references. McGrew, K. S., & Woodcock, R. W. (2001). *Technical Manual: Woodcock-Johnson III*. Itasca, IL: Riverside Publishing.

Purpose of measure. The WJ-III Cog is a measure of cognitive abilities.

Address for manual and measure. Riverside Publishing Company, 425 Spring Lake Drive, Itasca, IL 60143–2079, USA; www.woodcock-johnson.com

EBA classification. The WJ-III Cog is a well-established assessment.

Appendix B. List of measures identified by listserv respondents not included in the review

Intellectual functioning

Battelle Development Inventory
Differential Ability Scales (DAS)

Nonverbal intellectual functioning

Comprehensive Test of Nonverbal Intelligence (CTONI)
Hiskey–Nebraska Tests of Learning Aptitude
Matrix Analogies Test
Test of Nonverbal Intelligence-Third Edition (TONI-3)
Universal Nonverbal Intelligence Test (UNIT)

Memory and learning

Children's Memory Scale (CMS)
Rey Auditory Verbal Learning Test

Language

Expressive Vocabulary Test (EVT)
Oral and Written Language Scales (OWLS)

Visual-motor and motor functioning

Bruininks–Oseretsky
Motor Free Test of Visual Perception
Peabody Motor Scale
Purdue Pegboard
Test of Visual Perceptual Skills

Attention/Executive functioning

Stroop Color–Word Test

Achievement

Boehm Test of Basic Concepts-Revised
Bracken Basic Concept Scale-Revised (BBCS-R)
Key-Math Revised

Note. Measures presented above received fewer than the median number of responses from survey respondents. Tests included in Appendix B should not be viewed as psychometrically unsound simply as a result of appearing in the table.

References

Armstrong, F. D., Blumberg, M. J., & Toledano, S. R. (1999). Neurobehavioral issues in childhood cancer. *School Psychology Review, 28*, 194–203.

Baron, I. S. (2004). *Neuropsychological evaluation of the child*. New York: Oxford University Press.

Bernstein, J. H., & Waber, D. P. (1996). *Developmental scoring system for the Rey-Osterrieth complex figure*. Professional manual. Lutz, FL: Psychological Assessment Resources, Inc.

Brown, R. T., Davis, P. C., Lambert, R., Hsu, L., Hopkins, K., & Eckman, J. (2000). Neurocognitive functioning and magnetic resonance imaging in children with sickle cell disease. *Journal of Pediatric Psychology, 25*, 503–513.

Bonner, M. J., Gustafson, K. E., Schumacher, E., & Thompson, J. R. (1999). The impact of sickle cell disease on cognitive functioning and learning. *School Psychology Review, 28*, 182–193.

Carter, A. S., Volkmar, F. R., Sparrow, S. S., Wang, J., Lord, C., Dawson, G., et al. (1998). The Vineland Adaptive Behavior Scales: Supplementary norms for individuals with autism. *Journal of Autism and Developmental Disorders, 28*, 287–302.

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology, 66*, 7–18.

Cohen, L. L., LaGreca, A. M., Blount, R. L., Kazak, A. E., Holmbeck, G. N., & Lemanek, K. L. (2007). Introduction: Evidence-based assessment in pediatric psychology. *Journal of Pediatric Psychology,* Journal of Pediatric Psychology. doi:10.1093/jpepsy/jsj115.

Epstein, J. N., Erkanli, A., Conners, C. K., Klaric, J., Costello, J., & Angold, A. (2003). Relations between continuous performance test performance measures and ADHD behaviors. *Journal of Abnormal Child Psychology, 31*, 543–554.

Ewing-Cobbs, L., & Bloom, D. R. (2004). Traumatic brain injury: Neuropsychological, psychiatric and educational issues. In R. T. Brown (Ed.), *Handbook of pediatric psychology in school settings* (pp. 313–331). Mahwah, NJ: Lawrence Erlbaum.

Fennell, E. B. (2000). Issues in child neuropsychological assessment. In R. D. Vanderploeg (Ed.), *Clinician's guide to neuropsychological assessment* (2nd ed., pp. 357–381). Mahwah, NJ: Lawrence Erlbaum.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95*, 29–51.

Franzen, M. D., & Wilhelm, K. L. (1996). Conceptual foundations of ecological validity in neuropsychological assessment. In R. J. Sbordone, & C. J. Long (Eds.), *Ecological validity of neuropsychological testing* (pp. 91–112). Delray Beach, FL: GR Press.

Griffiths, S. Y., Sherman, E. M. S., Slick, D. J., Lautzenhiser, A., Westerveld, M., & Zaroff, C. M. (2006). The factor structure of the CVLT-C in pediatric epilepsy. *Child Neuropsychology, 12*, 191–203.

Holmes, C. S., Cant, M. C., Fox, M. A., Lampert, N. L., & Greer, T. (1999). Disease and demographic risk factors for disruptive cognitive functioning in children with insulin-dependent diabetes mellitus (IDDM). *School Psychology Review, 28*, 215–227.

Hooper, S. R., Poon, K. K., Marcus, L., & Fine, C. (2006). Neuropsychological characteristics of school-age children with high-functioning autism: Performance on the NEPSY. *Child Neuropsychology, 12*, 299–305.

Knight, J. A. (Ed.). (2003). *The handbook of Rey-Osterrieth complex figure usage: Clinical and research applications*. Lutz, FL: PAR, Inc.

Lezak, M. L., Howieson, D. B., & Loring, D. W. (2004). *Neuropsychological assessment* (4th ed.), New York: Oxford University Press.

Lynch, P. A., Chen, R., & Holmes, C. S. (2004). Factor structure of the Wide Range Assessment of Memory and Learning (WRAML) in children with insulin dependent diabetes mellitus (IDDM). *Child Neuropsychology, 10*, 306–317.

Lynn, R., Allik, J., & Irwing, P. (2004). Sex differences on three factors identified in Raven's standard progressive matrices. *Intelligence, 32*, 411–424.

Mash, E. J., & Hunsley, J. (2005). Evidence-based assessment of child and adolescent disorders: Issues and challenges. *Journal of Clinical Child and Adolescent Psychology, 34*, 632–379.

McGee, R. A., Clark, S. E., & Symons, D. K. (2000). Does the Conners' Continuous Performance Test aid in ADHD diagnosis? *Journal of Abnormal Child Psychology, 28*, 415–424.

Meyers, J. E., & Meyers, K. R. (1996). *Rey complex figure test and recognition trial. Supplemental norms for children and adolescents.* Lutz, FL: Psychological Assessment Resources, Inc.

Mitrushina, M., Boone, K. B., Razani, J., & D'Elia, L. F. (2005). *Handbook of normative data for neuropsychological assessment* (2nd ed.), New York: Oxford University Press.

Montour-Proulx, I., Kuehn, S. M., Keene, D. L., Barrowman, N. J., Hsu, E., Matzinger, M., et al. (2005). Cognitive changes in children treated for acute lymphoblastic leukemia with chemotherapy only according to the Pediatric Oncology Group 9605 Protocol. *Journal of Child Neurology, 20*, 129–133.

Moore, B. D. (2005). Neurocognitive outcomes in survivors of childhood cancer. *Journal of Pediatric Psychology, 30*, 51–63.

Mulhern, R. K., Armstrong, F. D., & Thompson, S. J. (1998). Function-specific neuropsychological assessment. *Medical and Pediatric Oncology Supplement, 1*, 34–40.

Mulhern, R. K., & Butler, R. W. (2006). Neuropsychological late effects. In R. T. Brown (Ed.), *Childhood cancer and sickle cell disease: A* *biopsychosocial approach* (pp. 262–278). New York: Oxford University Press.

Retzlaff, P. D., & Gibertini, M. (2000). Neuropsychometric issues and problems. In R. D. Vanderploeg (Ed.), *Clinician's guide to neuropsychological assessment* (2nd ed., pp. 277–299). Mahwah, NJ: Lawrence Erlbaum.

Reynolds, C. R. (2002). *Comprehensive trail-making test. Examiner's manual*. Austin, TX: PRO-ED, Inc.

Rizzo, A. A., Schultheis, M., Kerns, K. A., & Mateer, C. (2004). Analysis of assets for virtual reality applications in neuropsychology. *Neuropsychological Rehabilitation, 14*, 207–239.

Rose, B. M., Holmbeck, G. N., Coakley, R. M., & Franks, E. A. (2004). Mediator and moderator effects in developmental and behavioral pediatric research. *Journal of Developmental and Behavioral Pediatrics, 25*, 58–67.

Roselli, M., Ardila, A., Bateman, J. R., & Guzman, M. (2001). Neuropsychological test scores, academic performance, and developmental disorders in Spanish-speaking children. *Developmental Neuropsychology, 20*, 355–373.

Schmitt, A. J., & Wodrich, D. L. (2004). Validation of a developmental neuropsychological assessment (NEPSY) through comparison of neurological, scholastic concerns, and control groups. *Archives of Clinical Neuropsychology, 19*, 1077–1093.

Silver, C. H. (2000). Ecological validity of neuropsychological assessment in childhood traumatic brain injury. *Journal of Head Trauma Rehabilitation, 15*, 973–988.

Soukup, V. M., Ingram, F., Grady, J. J., & Schiess, M. C. (1998). Trail making test: Issues in normative data selection. *Applied Neuropsychology, 5*, 65–73.

Soutor, S. A., Chen, R., Streisand, R., Kaplowitz, P., & Holmes, C. S. (2004). Memory matters: Developmental differences in predictors of diabetes care behaviors. *Journal of Pediatric Psychology, 29*, 493–505.

Spirito, A. (1999). Introduction to empirically supported treatments in pediatric psychology. *Journal of Pediatric Psychology, 24*, 87–90.

Stinnett, T. A., Oehler-Stinnett, J., Fuqua, D. R., & Palmer, L. S. (2002). Examination of the underlying structure of the NEPSY: A developmental neuropsychological assessment. *Journal of Psychoeducational Assessment, 20*, 66–82.

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary* (3rd ed.), New York: Oxford University Press.

Wechsler, D. (1991). *Manual for the Wechsler Intelligence Scale for Children-Third Edition*. San Antonio: The Psychological Corporation.