Published in final edited form as:
Monogr Soc Res Child Dev. 2013 August ; 78(4): 1–149. doi:10.1111/mono.12031.

I. NIH TOOLBOX COGNITION BATTERY (CB): INTRODUCTION AND PEDIATRIC DATA*

Sandra Weintraub, **Patricia J. Bauer**, **Philip David Zelazo**, **Kathleen Wallner-Allen**, **Sureyya S. Dikmen**, **Robert K. Heaton**, **David S. Tulsky**, **Jerry Slotkin**, **David L. Blitz**, **Noelle E. Carlozzi**, **Richard J. Havlik**, **Jennifer L. Beaumont**, **Dan Mungas**, **Jennifer J. Manly**, **Beth G. Borosh**, **Cindy J. Nowinski**, and **Richard C. Gershon**

Abstract

This monograph presents the pediatric portion of the National Institutes of Health (NIH) Toolbox Cognition Battery (CB) of the NIH Toolbox for the Assessment of Neurological and Behavioral Function. The NIH Toolbox is an initiative of the Neuroscience Blueprint, a collaborative framework through which 16 NIH Institutes, Centers, and Offices jointly support neuroscience-related research, to accelerate discoveries and reduce the burden of nervous system disorders. The CB is one of four modules that measure cognitive, emotional, sensory, and motor health across the lifespan. The CB is unique in its continuity across childhood, adolescence, early adulthood, and old age, and in order to help create a common currency among disparate studies, it is also available at low cost to researchers for use in large-scale longitudinal and epidemiologic studies. This chapter describes the evolution of the CB; methods for selecting cognitive subdomains and instruments; the rationale for test design; and a validation study in children and adolescents, ages 3–15 years. Subsequent chapters feature detailed discussions of each test measure and its psychometric properties (Chapters 2–6), the factor structure of the test battery (Chapter 7), the effects of age and education on composite test scores (Chapter 8), and a final summary and discussion (Chapter 9). As the chapters in this monograph demonstrate, the CB has excellent psychometric properties, and the validation study provided evidence for the increasing differentiation of cognitive abilities with age.

The NIH Toolbox was conceived as an instrument for the systematic collection of data on cognitive, emotional, sensory, and motor health across disparate studies. It was intended to provide a brief assessment tool for large-scale epidemiologic and longitudinal studies for projects in which neurologic function may not necessarily constitute the primary focus but in which its assessment could be useful and also allow cross-study comparisons. The NIH Toolbox was designed as part of the NIH Blueprint initiative in the neurosciences, involving 16 different institutes.[1] The Request for Applications from the NIH specified that the NIH Toolbox instruments: (1) include measures relevant to development and health across the life span from ages 3 to 85 years; (2) assess the full range of normal functioning (i.e., the instruments are not intended to screen for disease); (3) cover several different subdomains

[*] This monograph was accepted under the editorship of W. Andrew Collins.

Corresponding author: Sandra Weintraub, Cognitive Neurology and Alzheimer's Disease Center, Northwestern Feinberg School of Medicine, Chicago, IL, sweintraub@northwestern.edu.

[1] These institutes are: National Institute on Aging (NIA), National Institute on Mental Health (NIMH), National Institute on Drug Abuse (NIDA), National Institute on Neurological Disorders & Stroke (NINDS), National Institute on Environmental Health Sciences (NIEHS), National Institute on Deafness and Other Communicative Disorders (NIDCD), National Eye Institute (NEI), National Institute on Child Health and Human Development (NICHD), National Institute of Nursing Research (NINR), National Institute on General Medical Sciences (NIGMS), Office of Behavioral and Social Sciences Research (OBSSR), National Center on Complementary and Alternative Medicine (NCCAM), National Institute on Alcohol Abuse and Alcoholism (NIAAA), National Center on Research Resources (NCRR).

within each of four domains (cognitive, emotional, sensory, and motor) essential to health and life adaptation; (4) be brief and easy to administer and score; (5) be freely available to researchers; and (6) be modifiable to accommodate advances in science.

This monograph focuses on the development of the instruments in the NIH Toolbox that assess cognitive function—the Cognition Battery (CB). An initial challenge in meeting the mandate of the NIH Toolbox was the selection of particular subdomains for assessment. Cognition includes many essential subdomains, some of which require lengthy and complex methods of assessment, so difficult decisions needed to be made regarding which to include at the cost of others. A systematic, research-based process was followed to select the subdomains for assessment. The process resulted in a focus on (1) executive function and attention, (2) episodic memory, (3) language, (4) working memory, and (5) processing speed. In the first section of this chapter, we describe this process, explain how instruments were selected for the subdomains, and outline the steps taken to ensure the usability of the instruments for diverse populations.

A second significant challenge was creating a single set of measures that is valid and appropriate across the entire 3- to 85-year age range. The difficulty of achieving this goal was obvious at the outset from the lack of such measures in most areas of cognition (despite the need for such measures). More typically, constructs are measured with very different tasks at different ages. The second section of this chapter describes the steps taken to develop instruments that meet this goal of use across the lifespan. As discussed in more detail in Chapters 2–6, which are devoted to the individual instruments, some tests were borrowed from the adult literature and adapted for younger examinees, whereas others were selected from the child developmental literature and adapted for older examinees.

The final section of the chapter provides an introduction to a test for validation of the CB. The validation study included the full age range of the NIH Toolbox, ages 3–85 years, although this monograph focuses on the results of this study for children and young adolescents ages 3–15 years. The validation data from the younger and older adult populations will be published in a separate series of papers so that each population can be addressed in greater depth. To preview the conclusion: whereas the mandate to develop brief tasks to be used across the lifespan presented substantial challenges, it also afforded a significant opportunity to advance science by providing tools to further our understanding of cognitive function across the lifespan.

SUBDOMAIN SELECTION

The Cognition team was required to select the subdomains to be evaluated and then to determine the best measure of each subdomain. The selection of subdomains was based on: (1) their importance to the course of development and aging; (2) their significance for health and success in education and, in adults, for work; (3) their validation with respect to known underlying brain mechanisms; and (4) their ease of measurement and translation into brief test instruments. Evaluation of subdomains using these criteria was accomplished through widespread and reiterative input from multidisciplinary researchers and clinicians who specialize in different areas of cognitive functioning and who work with pediatric and/or adult populations. The first step was a survey of potential “end users” to determine the structure of the final NIH Toolbox and to identify subdomains to be assessed. The methods used to gather data and to establish consensus among potential NIH Toolbox end-users are detailed elsewhere (Gershon et al., 2010) and are only summarized here.

Research and clinical experts were identified via literature searches, from the Computer Retrieval of Information on Scientific Projects (CRISP) database (now known as the NIH Research Portfolio Online Reporting Tools [RePORT]), and/or by nomination by one of the

12 NIH science officers who comprised the NIH Toolbox Project Team at that time. Two Requests for Information (RFIs) were then solicited online from a total of 293 experts, and the RFIs were followed-up by telephone interviews of a subset of 44 experts. The information gathered from experts allowed us to identify cognitive domains ranked in order of their importance as judged by experts, and to determine the characteristics that would be desirable for the final instruments. The subdomains were ranked as follows: Executive function, episodic memory, language, processing speed, and attention. Table 1 shows the percentage of respondents ranking each of the sudomains among their top four subdomain rankings. Fifty-seven percent of respondents also listed a “general” or “global” cognitive score as desirable. The need for a global cognitive score was met through development of cognition composite scores, described in Chapter 8. Only 43% of respondents ranked visuospatial functions among their top four subdomains, so this subdomain was not included in the final list.

Searches of relevant databases in psychology and pediatrics were then conducted to review support for the selection of subdomains in terms of their importance for neurological and behavioral function, and to develop a test instrument library. The test instrument library was reviewed to determine whether there were existing instruments that would fulfill the needs of the NIH Toolbox. Meanwhile, large consensus meetings were held twice a year for the Steering Committee, consisting of all the NIH Toolbox major domain team leaders and NIH representatives, to evaluate the information gathered and make decisions about final choices. The Steering Committee also held monthly phone conferences to review progress. Below, each subdomain included in the final cognition battery is briefly described.

Executive Function and Attention

The subdomains of executive function (EF) and attention are described together because one of the EF measures is also a measure of selective attention. EF consists of a number of distinctive types of mental operations, subsumed by the term, “cognitive control,” that are involved in the top-down modulation of goal-directed activity. Recent factor-analytic work with adults suggests that EF can be divided into three partially independent components: cognitive flexibility, inhibitory control, and working memory (Miyake et al., 2000). EF deficits are seen in patients with acquired focal damage to prefrontal cortex who experience profound impairment in behavioral regulation despite the preservation of many basic intellectual functions (see Stuss & Knight, 2002). In children, impairments in EF or delays in its development have been linked to attention deficit hyperactivity disorder (ADHD) (Doyle, 2006), autism spectrum disorders, Conduct Disorder, and other psychiatric conditions and symptoms.

Because of the importance of EF, the CB contains measures of all three components, although the measure of working memory is considered separately, below. The other measures (see Zelazo, Anderson, Richler, Wallner-Allen, & Beaumont, Chapter 2, this volume) include measures of cognitive flexibility (i.e., the ability to switch conceptual frameworks, measured by the Dimensional Change Card Sort; Zelazo, 2006), and a measure of inhibitory control (and selective attention), measured by a version of the Eriksen flanker task (Eriksen & Eriksen, 1974) that was adapted from the Attention Network Test (ANT; e.g., Rueda et al., 2004).

Episodic Memory

Episodic memory, the capacity for storing and retrieving information, is critical for the acquisition of knowledge and for building adaptive skills. This subdomain shows dramatic changes over the first two decades of life (see Bauer, Larkina, & Deocampo, 2011, for a review) and is also susceptible to a variety of diseases, including encephalitis and temporolimbic epilepsy, and, in adulthood, Alzheimer's disease (e.g., see Weintraub, Wicklund, & Salmon, 2012). Episodic memory for single object-specific actions is apparent in infants in the first year of life (e.g., Carver & Bauer, 1999). By the second year of life, infants remember temporally ordered sequences of items as well (e.g., Bauer, Wenner, Dropik, & Wewerka, 2000). In order to tap similar memory constructs in older children, adolescents, and adults for the CB, Bauer's (2007) imitation-based assessment of memory paradigm was modified to create the Toolbox Picture Sequence Memory Test (see Bauer et al., Chapter 3, this volume). This test is based on nonverbal pictorial stimuli that must be placed in a predefined sequence, with increasing numbers of pictures for older age groups.

Language

Language develops rapidly over the first 3 years of life, although further changes occur throughout childhood, and language proficiency is a fundamental skill that supports many other aspects of cognitive, social, and behavioral function. Indeed, when language development is delayed, the impact on further skill acquisition and academic progress can be profound (e.g., Dickinson, Golinkoff, & Hirsh-Pasek, 2010; Gleason & Ratner, 2009). Disorders such as dyslexia hinder otherwise talented individuals from achieving educational and career goals (e.g., Meisinger, Bloom, & Hynd, 2010; Ziegler et al., 2008).

Language consists of numerous components, including semantics, grammar, morphology, and phonology, and it is conveyed via multiple modalities including auditory comprehension, speaking, reading, and writing. Two aspects of language were selected for the CB, due in part to the ease with which they can be measured across the lifespan: auditory single word comprehension (i.e., receptive vocabulary) and single word reading aloud (oral reading recognition; see Gershon et al., Chapter 4, this volume). Auditory single word comprehension develops prior to overt speech usage in hearing individuals (Fenson et al., 1994). Vocabulary has been widely accepted as a surrogate measure for overall crystallized intelligence (Schmidt & Hunter, 2004). Oral reading proficiency also is a marker of educational opportunity in minority populations and can be used to adjust for group differences when comparing individuals of different ethnic and racial backgrounds (Manly, Byrd, Touradji, & Stern, 2004; Manly et al., 1999).

Working Memory

Although working memory is a component of executive function (e.g., Miyake et al., 2000), it is often studied on its own, or as a type of memory. From preschool age on, most working memory tasks either require retaining and reorganizing items before recalling them (e.g., backward digit span task), or completing some processing activity in between presentations of the to-be-recalled items (e.g., listening span task). Such tasks tap into both information processing and storage, and yield a measure of working memory span that corresponds to the maximal amount of accurately recalled information. Working memory shows age-related improvements across childhood (as well as age-related declines during senescence). Like executive function more broadly, working memory depends on prefrontal cortex, and is vulnerable to disruption from a wide range of cerebral insults (see Tulsky et al., Chapter 5, this volume).

Processing Speed

Processing speed (PS), which refers to the speed with which simple cognitive operations can be performed, was included in the CB because it is very sensitive to any form of cerebral insult (see DeLuca & Kalmar, 2007; Weiler, Forbes, Kirkwood, & Waber, 2003) and to changes in development (e.g., Kail, 1991). Although PS paradigms often are based on motor reaction time, Sternberg's (1966) elegant paradigm demonstrated that mental processing time can be separated from motor response time, and that it varies with the number of mental operations required by a given task. The construct represented in the CB is the simple reaction time required to make a same-different comparison between two visually presented stimuli (see Carlozzi, Tulsky, Kail, & Beaumont, Chapter 6, this volume).

INSTRUMENT SELECTION

Following identification of the cognitive subdomains, additional experts were recruited to help develop tests. These experts held weekly conference calls to review decisions, update progress, and assure consistency of methods across all subdomains. Individual subdomain teams were also formed and these teams convened as needed to work on their specific tests. In addition to the Steering Committee meetings and conference calls, the entire CB team held a day-long meeting in July, 2007, to determine the particular instruments to be subjected to validation testing. During that meeting, the CB team derived criteria for validation studies, including acceptable levels of test–retest reliability and convergent and discriminant validity. In 2008, a public conference was held in Bethesda, Maryland, to present the NIH Toolbox to an expert advisory panel and obtain feedback. Written critiques of the subdomains and instruments were reviewed by the CB team and addressed. In 2010, we conducted several conference calls with 16 expert consultants to present the version of the CB created for validation testing and to invite feedback prior to initiating the validation study.

The general principles that guided decision-making regarding the NIH Toolbox instruments were:

1. *Versatility*: Measures should be capable of monitoring neurological and behavioral health status and function over time (as in longitudinal epidemiological studies), and capable of evaluating effectiveness of interventions and treatments (as in clinical trials). Instruments should be readily portable from one type of study design to another and have minimal ceiling and floor effects.
2. *Brevity*: Measures should be brief, to ensure low respondent burden. The targeted total time for the CB was 30 min (ages 7–85 years), and 20 min for children from 3 to 6 years of age.
3. *Methodological Soundness*: Measures should demonstrate validity and reliability.
4. *Dynamic*: Measures should be internally flexible (e.g., adaptive testing), and instruments should demonstrate sensitivity to change over time.
5. *State-of-the-art design*: Measures should employ modern psychometric approaches to the measurement of latent dimensions (e.g., item response theory models and computer-adaptive testing, to the extent relevant).
6. *Diversity*: Measures should have known properties across cultures and age ranges. English and Spanish versions should be developed and validated in culturally and geographically diverse groups.

In addition to these features, the instruments were submitted to scrutiny for their adaptability to the key populations that were to be assessed using the NIH Toolbox. Four special working groups were convened to examine the constructs selected and the instruments and procedures being developed with consideration of people from different ethnic, racial, and cultural backgrounds; the needs of older adults; the needs of people with disabilities; and the needs of children. These working groups were made up of project scientists and external consultants. Each group reviewed the instruments and procedures being developed from its particular perspective and identified areas of concern and proposed ways to address those concerns. The working groups held meetings and conference calls, discussed issues within groups, and sometimes partnered with each other when similar issues arose. Each group provided recommendations to the Cognition team, as well as to the Emotional, Sensation, and Motor teams, on ways to enhance the usability and relevance of the NIH Toolbox for diverse populations from ages 3 to 85 years.

The Cultural Working Group strived to ensure that the measures were culturally sensitive and conceptually appropriate across different cultures and languages. For example, they made recommendations for wording changes, picture changes, and suggested guidelines for determining language proficiency. The Geriatric Working Group and the Accessibility Working Group addressed issues relating to the suitability of instruments for those with motor and sensory impairments, often seen in the elderly, and for those with other impairments related to disability. Issues such as font size, image size, type of motoric response required, and color of stimuli (with respect to color blindness) were considered within the context of working to increase the accessibility of tasks for those in the general US population with a disability. The Pediatric Working Group addressed the difficulties of designing instruments suitable for use with young children. Because the work of this group was most relevant to the pediatric data reported here, we discuss it more fully.

Developing instruments for use across the broad age range of 3–85 years presented significant challenges. Children differ from adults in many ways, including social, emotional, and selfregulatory ways that may affect performance on tasks designed to measure cognition. A major challenge was to structure tasks and the testing environment so that differences in task performance across different ages would more likely reflect differences in competence for the construct of interest rather than differences in other performance factors. For example, children are able to demonstrate competence for constructs at younger ages when simple, easy-to-follow instructions are used and when task materials are engaging, concrete, and familiar. In addition, compared to healthy young adults, most children and even adolescents have shorter attention spans, are more easily distracted by external stimuli, and are less proficient at regulating their attention, behavior, and level of motivation to the task at hand.

In consideration of the challenges of measurement in children in particular, the Pediatric Working Group developed a set of pediatric assessment principles to inform instrument design and to standardize assessment procedures across all NIH Toolbox instruments. These principles addressed instrument design characteristics, the testing environment, the psychological and physical needs of the child, and the nature and extent of the interactions among the test administrator, the child, and the parent (where appropriate). For example, it was considered important to use simple instructions, have practice trials to ensure understanding, establish stop rules to minimize failure experiences, and have an examiner present during testing. The Pediatric Working group reviewed all NIH Toolbox instruments and considered whether each was age appropriate, assessed an appropriate construct, was appropriately sized, and was nonthreatening. The Pediatric Working Group advocated for building flexibility into the computer interface (e.g., the ability to repeat instructions) and made recommendations for an appropriate response mechanism (e.g., touchscreen, mouse). Recommendations were also made on whether task instructions should be “live” or “prerecorded” and provided by computer to standardize presentation and, when it was decided to use a prerecorded voice, the group consulted on what the quality of the voice and the gender of the speaker should be. The guidance provided by the Pediatric Working Group, as well as the other working groups, significantly improved the NIH Toolbox overall and strengthened its ability to obtain valid assessments from diverse populations—not just from children.

VALIDATION STUDY

To determine the reliability and validity of the instruments as measures of the target subdomains, the Cognition team conducted a validation study involving a total of 476 participants recruited from multiple sites (Chicago's NorthShore University HealthSystems, Emory University in Atlanta, New Jersey's Kessler Institute for Rehabilitation, and the University of Minnesota). Eligible participants were 3–85 years of age and sample recruitment was distributed across age, gender, race, and education strata. Table 2 illustrates the pediatric validation sample, including the 208 three- to fifteen year olds whose data are featured in this monograph. There were a total of one hundred twenty 120 three- to six year olds and 88 eight- to fifteen year olds. As Table 3 indicates, not all ages were sampled in this study; also, education levels indicated in the table are defined as highest parental education. A subset of 66 child participants (approximately 32%) completed a retest 7–21 days later to assess test–retest reliability and practice effects.

Validation Measures

Validation measures were selected by reviewing published tests commonly used in neuropsychological practice to assess the constructs being tapped by the CB tests. Table 3 shows each CB measure and its associated validation measures. Table 4 shows CB measures and the validation measures by age group to which each was administered. Table 5 shows a sample of the criterion grid established for judging validity, using the measure of working memory (The NIH Toolbox List Sorting Working Memory Test) as an example.

Pearson correlation coefficients between age and test performance were calculated separately for children and adults to describe the developmental and aging-related associations for each measure. Intraclass correlation coefficients (*ICC*) were calculated to evaluate test–retest reliability. Across measures, $ICC < .4$ were considered poor, .4 to .74 were considered adequate and ≥ .75 were considered excellent. Convergent validity was assessed with correlations between each Toolbox measure and a well-established validation measure of the same construct. Across measures, $r < .3$ were considered poor, .3 to .59 were considered adequate, and ≥ .6 were considered excellent. Evidence of discriminant validity consisted of lower correlations with selected validation measures of a *different* cognitive construct. The rationale for selection of each validation instrument is discussed in the context of the individual chapters of the monograph.

PLAN FOR THE REST OF THE MONOGRAPH

Chapters 2–6 each addresses a single subdomain. In each chapter, we review the rationale for inclusion of that subdomain in the battery, and the importance of that subdomain for health. We also review the literature on developmental changes in the subdomain throughout childhood and into adolescence, and the evidence linking the subdomain or construct to brain functioning. The test instruments are described in greater detail, including the adaptations to enable testing across the 3–85 years age range. We present results of a validation study and describe the psychometric properties of the new CB measures. Chapter 7 of the monograph reports the results of a confirmatory factor analysis of the CB validation study. Chapter 8 reports the creation of CB composite scores and the relations of demographic variables to these scores. The final chapter provides brief summaries of the rationale for development of the CB and the major findings from the validation study, followed by discussion of the implications of the NIH Toolbox CB for the study of cognitive development, the limitations of the battery, and directions for further development of the instrument.

References

Bauer, PJ. Remembering the times of our lives: Memory in infancy and beyond. Mahwah, NJ: Erlbaum; 2007.

Bauer, PJ.; Larkina, M.; Deocampo, J. Early memory development. In: Goswami, U., editor. The Wiley-Blackwell handbook of childhood cognitive development. 2. Oxford, UK: Wiley-Blackwell; 2011. p. 153-179.

Bauer PJ, Wenner JA, Dropik PL, Wewerka SS. Parameters of remembering and forgetting in the transition from infancy to early childhood. Monographs of the Society for Research in Child Development. 2000; 65(4):i–vi. 1–204. [PubMed: 12467092]

Brocki K, Fan J, Fossella J. Placing neuroanatomical models of executive function in a developmental context: Imaging and imaging–genetic strategies. Annals of the New York Academy of Sciences. 2008; 1129:246–255. [PubMed: 18591485]

Carver LJ, Bauer PJ. When the event is more than the sum of its parts: Nine-month-olds' long-term ordered recall. Memory. 1999; 7:147–174. [PubMed: 10645377]

DeLuca, J.; Kalmar, JH. Information processing speed in clinical populations. London: Psychology Press; 2007.

Dickinson DK, Golinkoff RM, Hirsh-Pasek K. Speaking out for language: Why language is central to reading development. Educational Researcher. 2010; 39(4):305–310.

Doyle AE. Executive functions in attention-deficit/hyperactivity disorder. Journal of Clinical Psychiatry. 2006; 67 (Suppl 8):21–26. [PubMed: 16961426]

Eriksen BA, Eriksen CW. Effects of noise letters upon the identification of a target letter in a nonsearch task. Perception and Psychophysics. 1974; 16:143–149.

Fenson L, Dale PS, Reznick JS, Bates E, Thal DJ, Pethick SJ. Variability in early communicative development. Monographs of the Society for Research in Child Development. 1994; 59:1–173. [PubMed: 7845413]

Gershon RC, Cella D, Fox NA, Havlik RJ, Hendrie HC, Wagster MV. Assessment of neurological and behavioural function: The NIH Toolbox. Lancet Neurology. 2010; 9(2):138–139. [PubMed: 20129161]

Gleason, JB.; Ratner, NB., editors. The development of language. 7. Boston: Pearson/Allyn & Bacon; 2009.

Kail R. Processing time declines exponentially during childhood and adolescence. Developmental Psychology. 1991; 27:259–266.

Manly JJ, Byrd DA, Touradji P, Stern Y. Acculturation, reading level, and neuropsychological test performance among African American elders. Applied Neuropsychology. 2004; 11(1):37–46. [PubMed: 15471745]

Manly JJ, Jacobs DM, Sano M, Bell K, Merchant CA, Small SA, et al. Effect of literacy on neuropsychological test performance in nondemented, education-matched elders. Journal of the International Neuropsychological Society. 1999; 5(3):191–202. [PubMed: 10217919]

Meisinger EB, Bloom JS, Hynd GW. Reading fluency: Implications for the assessment of children with reading disabilities. Annals of Dyslexia. 2010; 60(1):1–17. [PubMed: 20033795]

Miyake A, Friedman NP, Emerson MJ, Witzki AH, Howerter A, Wager TD. The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. Cognitive Psychology. 2000; 41:49–100. [PubMed: 10945922]

NICHD Early Child Care Research Network. Child care and child development: Results from the NICHD Study of Early Child Care and Youth Development. New York: Guilford Publications; 2005.

Rueda MR, Fan J, McCandliss BD, Halparin JD, Gruber DB, Lercari LP, et al. Development of attentional networks in childhood. Neuropsychologia. 2004; 42:1029–1040. [PubMed: 15093142]

Schmidt FL, Hunter J. General mental ability in the world of work: Occupational attainment and job performance. Journal of Personality and Social Psychology. 2004; 86(1):162–173. [PubMed: 14717634]

Shaywitz SE, Shaywitz BA, Fulbright RK, Skudlarski P, Mencl WE, Constable RT, et al. Neural systems for compensation and persistence: Young adult outcome of childhood reading disability. Biological Psychiatry. 2003; 54(1):25–33. [PubMed: 12842305]

Sternberg S. High-speed scanning in human memory. Science. 1966; 153:652–654. [PubMed: 5939936]

Weiler MD, Forbes P, Kirkwood M, Waber D. The developmental course of processing speed in children with and without learning disabilities. Journal of Experimental Child Psychology. 2003; 85(2):178–194. [PubMed: 12799167]

Weintraub, S.; Wicklund, AH.; Salmon, DP. The neuropsychological profile of Alzheimer disease. In: Selkoe, D.; Holzman, D.; Mandelkow, E., editors. The biology of Alzheimer's Disease. Woodbury NY: Cold Spring Harbor Laboratory Press; 2012. p. 25-52.

Zelazo PD. The Dimensional Change Card Sort (DCCS): A method of assessing executive function in children. Nature Protocols. 2006; 1:297–301.

Ziegler JC, Castel C, Pech-Georgel C, George F, Alario FX, Perry C. Developmental dyslexia and the dual route model of reading: Simulating individual differences and subtypes. Cognition. 2008; 107(1):151–178. [PubMed: 17959161]

TABLE 1

Percentage of Expert Raters (*N* = 147) Ranking Subdomain in Top 4

Subdomain	**%**
Executive function	95
Memory	93
General/global score	57
Language	55
Processing speed	52
Attention	50
Visuospatial function	43
Other 1	7
Other 2	3

TABLE 2

Validation Pediatric Sample Demographics (*N* = 208)

		Gender		**Race/Ethnicity**		
Age Groups	**Level of Parents' Education**	**Male**	**Female**	**White**	**Black**	**Hispanic/Other/Multiple Races**
3–6 years	<High School	6	5	5	5	1
N = 120	High School Graduate	29	27	23	19	14
	College Graduate+	29	24	26	16	11
Total		64	56	54	40	26
8–15 years	<High School	4	6	4	5	1
N = 88	High School Graduate	22	23	18	13	14
	College Graduate+	14	19	16	7	10
Total		40	48	38	25	25

TABLE 3

Cognition Battery (CB) and Corresponding Convergent Validation Measures for Children

Cognition Subdomain	NIH Toolbox CB Measures	Validation Measure
Executive Function	Flanker Inhibitory Control and Attention Test Dimensional Change Card Sort Test	WPPSI-III Block Design (3–6 years) D-KEFS Color Word Interference (8–15 years)
Episodic Memory	Picture Sequence Memory Test	NEPSY-II Sentence Repetition (3–6 years) Rey Auditory-Verbal Learning Test (RAVLT; 3-trial version; 8–15 years)[a] Brief Visuospatial Memory Test-Revised (BVMT-R; 8–15 years)[a]
Language	Picture Vocabulary Test Oral Reading Recognition Test	Peabody Picture Vocabulary Test 4th Edition (PPVT-IV) Wide Range Achievement Test 4th Edition (WRAT-IV) Reading Subtest
Working Memory	List Sorting Working Memory Test	NEPSY-II Sentence Repetition (3–6 years) WISC-IV Letter Number Sequencing (8–15 years)
Processing Speed	Pattern Comparison Processing Speed Test	WPPSI-III or WISC-IV Processing Speed Composite, as appropriate Paced Auditory Serial Addition Test (PASAT; 8–15 years)

Note. WPPSI-III, Wechsler Preschool and Primary Scale of Intelligence, 3rd Edition; D-KEFS, Delis–Kaplan Executive Function Scales; NEPSY-II, Developmental Neuropsychological Assessment Battery, 2nd Edition; WISC-IV, Wechsler Intelligence Scale for Children, 4th Edition.

[a] Two validation measures were used in order to capture both verbal and visuospatial memory.

TABLE 4

NIH Toolbox Cognition Battery (CB) and Validation Measures by Age Cohort

	3–4	**5–6**	**8–15**	**20–85**
NIH Toolbox CB Measures				
Dimensional Change Card Sort Test	Yes	Yes	Yes	Yes
Flanker Inhibitory Control and Attention Test	Yes	Yes	Yes	Yes
Picture Sequence Memory Test	Yes	Yes	Yes	Yes
List Sorting Working Memory Test	Yes	Yes	Yes	Yes
Pattern Comparison Processing Speed Test	Yes	Yes	Yes	Yes
Oral Reading Recognition Test	Yes	Yes	Yes	Yes
Picture Vocabulary Test	Yes	Yes	Yes	Yes
Validation Measures				
Wisconsin Card Sort Test-64 cards			Yes	Yes
Paced Auditory Serial Addition Test			Yes	Yes
EXAMINER Dot Counting Test			Yes	Yes
Delis–Kaplan Executive Function: Color/Word Interference			Yes	Yes
Rey Auditory Verbal Learning Test			Yes	Yes
Brief Visual Memory Test-Revised			Yes	Yes
Wide Range Achievement Test-IV Reading	Yes	Yes	Yes	Yes
Peabody Picture Vocabulary Test-IV	Yes	Yes	Yes	Yes
NEPSY-II				
Sentence Repetition	Yes	Yes		
Speeded Naming	Yes	Yes		
WPPSI–III				
Block Design	Yes	Yes		
Coding		Yes		
Symbol Search		Yes		
WISC-IV				
Coding			Yes	
Letter-Number Sequencing			Yes	
Symbol Search			Yes	
WAIS-IV				
Coding				Yes
Letter-Number Sequencing				Yes
Symbol Search				Yes
Questionnaires				
Child Behavior Questionnaire	Yes	Yes		
Sociodemographics Form-Parent	Yes	Yes	Yes	
Sociodemographics Form-Subject				Yes
Cognitive Information Form	Yes	Yes	Yes	Yes

Note. NINDS EXAMINER: National Institute of Neurological Diseases and Stroke battery of “Domain Specific Test of Executive Function,” http://examiner.ucsf.edu/index.htm; NEPSY-II, Developmental Neuropsychological Assessment Battery, 2nd Edition; WPPSI-III, Wechsler

Preschool and Primary Scale of Intelligence, 3rd Edition; WISC-IV, Wechsler Intelligence Scale for Children, 4th Edition; WAIS-IV, Wechsler Adult Scale of Intelligence, 4th Edition.

TABLE 5

Sample Criteria for Validation of Toolbox List Sorting Working Memory Test; One List, Two List, and Total Scores

Analysis	**Criterion**
Measure-level selection criteria	
Convergent validity (NEPSY-II Sentence Repetition; WISC-IV Letter Number Sequencing) by age band and overall	Minimum correlation of .5
Divergent validity: PPVT-IV, D-KEFS Color Word Interference, WCST-64	.1 *less than* correlation with others
Test–Retest Reliability (*ICC*)	.5
Correlation between One List and Two List	.75
Internal Consistency Reliability (Split-half)	.65
Internal Consistency Reliability (alpha)	.55
Age effects	*r*-Squared approx. .2; linear effect through childhood; ages 20–35 stable (highest functioning); >age 35 declining; from middle age on, *r*-squared again approx. .2
Demographic effects (education, ethnicity, & gender) by age	Education: Linear relation for adults (*r*-squared approx. .1); for children, age and education will be highly correlated (just look at age)
Floor/ceiling effects	No evidence of floor or ceiling effects
Test timing by age band and overall	5 min
Primary test score means stratified separately on demographic variables (education, ethnicity, and gender)	No differential functioning (significant differences) between referent and focal groups
Percent of respondents making it to an item	Same as frequency pattern for total score
Item-level selection criteria	
Item-total correlations	Should range .15–.40 (should covary with *p*-value)
Item *p*-values	*p*-Values should range .15–.95

Note. PPVT-IV, Peabody Picture Vocabulary Test-4th Edition; WISC-IV, Wechsler Intelligence Scale for Children, 4th Edition; NEPSY-II, Developmental Neuropsychological Assessment, 2nd Edition; D-KEFS, Delis–Kaplan Executive Function Scales; WCST-64, Wisconsin Card Sorting Test-64 Card version.