Cognition assessment using the NIH Toolbox

Sandra Weintraub, PhD
Sureyya S. Dikmen, PhD
Robert K. Heaton, PhD
David S. Tulsky, PhD
Philip D. Zelazo, PhD
Patricia J. Bauer, PhD
Noelle E. Carlozzi, PhD
Jerry Slotkin, PhD
David Blitz, MA
Kathleen Wallner-Allen, PhD
Nathan A. Fox, PhD
Jennifer L. Beaumont, MS
Dan Mungas, PhD
Cindy J. Nowinski, MD, PhD
Jennifer Richler, PhD
Joanne A. Deocampo, PhD
Jacob E. Anderson, MA
Jennifer J. Manly, PhD
Beth Borosh, PhD
Richard Havlik, MD
Kevin Conway, PhD
Emmeline Edwards, PhD
Lisa Freund, PhD
Jonathan W. King, PhD
Claudia Moy, PhD
Ellen Witt, PhD
Richard C. Gershon, PhD

Correspondence to
Dr. Weintraub:
sweintraub@northwestern.edu

ABSTRACT

Cognition is 1 of 4 domains measured by the NIH Toolbox for the Assessment of Neurological and Behavioral Function (NIH-TB), and complements modules testing motor function, sensation, and emotion. On the basis of expert panels, the cognition subdomains identified as most important for health, success in school and work, and independence in daily functioning were Executive Function, Episodic Memory, Language, Processing Speed, Working Memory, and Attention. Seven measures were designed to tap constructs within these subdomains. The instruments were validated in English, in a sample of 476 participants ranging in age from 3 to 85 years, with representation from both sexes, 3 racial/ethnic categories, and 3 levels of education. This report describes the development of the Cognition Battery and presents results on test-retest reliability, age effects on performance, and convergent and discriminant construct validity. The NIH-TB Cognition Battery is intended to serve as a brief, convenient set of measures to supplement other outcome measures in epidemiologic and longitudinal research and clinical trials. With a computerized format and national standardization, this battery will provide a "common currency" among researchers for comparisons across a wide range of studies and populations. *Neurology*® 2013;80 (Suppl 3):S54–S64

GLOSSARY

CAT = computer adaptive testing; **CB** = Cognition Battery; **EF** = executive function; **NIH-TB** = NIH Toolbox for the Assessment of Neurological and Behavioral Function; **PS** = processing speed; **WM** = working memory.

Cognition is 1 of the 4 domains of behavioral and neurologic health assessed in the NIH Toolbox for the Assessment of Neurological and Behavioral Function (NIH-TB). All domain measures were intended to be freely accessible, to be usable with individuals from 3 to 85 years of age, with each domain battery not to exceed 30 minutes in duration. Expert surveys were conducted and panels of research scientists and clinicians consulted in an iterative manner to rank cognitive subdomains in order of their perceived importance for health. Information was requested from experts (N = 102) who reported sufficient familiarity with cognition to make recommendations for specific subdomains of importance. The 2 top-ranked subdomains were Executive Function (EF) (95%) and Episodic Memory (93%), followed by Language (55%), Processing Speed (52%), and Attention (50%). Many (57%) also listed a "Global Score" as important. Other cognitive subdomains were excluded because of lower priority in the rankings, coupled with the stringent time constraints on the length of the battery.

The rationale for specific cognitive constructs within subdomains and instrument selection was based on a systematic review of the literature, including evidence of the known biological associations of each. The EF subdomain was deemed to include several distinct constructs, including Switching/Set Shifting, Inhibitory Control and Attention, and Working Memory. Because of

the heavy weighting of EF by respondents, these 3 constructs were considered separate subdomains, with single instruments addressing each. Cognitive subdomains and specific constructs selected for measurement follow.

EF, also called "cognitive control," refers to the top-down cognitive modulation of goal-directed activity. Development of EF in childhood parallels the development of prefrontal, anterior cingulate, and parietal cortex, and the basal ganglia, as well as the growth of connections between these regions and others.[1] EF emerges in infancy[2] and grows rapidly between the ages of 2 and 5 years[3] with more gradual changes continuing into adolescence and early adulthood. EF is very vulnerable to aging,[4] and comparisons across the lifespan yield an inverted U-shaped pattern, with early age-related improvement followed by later age-related decline.[5] Based on factor-analytic work, there is an emerging consensus that EF can be divided into 3 partially independent subcomponents: set shifting, inhibitory control, and updating/working memory.[6] These distinctions are clearest in middle childhood and beyond, and far less distinct in children younger than age 6.[7] There is also evidence that prefrontal activation during the performance of EF tasks becomes increasingly focal and differentiated in the course of development.[8]

The set-shifting component of EF consists of the ability to shift responses based on rules or contingencies. It is measured in the Cognition Battery by a paradigm initially developed for children, the NIH-TB Dimensional Change Card Sort Test.[9] This aspect of EF is supported by a distributed neuroanatomical network involving lateral prefrontal, anterior cingulate, and inferior parietal regions.

The ability to focus, sustain, and shift attention is a prerequisite for performing most conscious cognitive operations frequently tested experimentally or clinically.[10] The developmental syndrome of attention deficit hyperactivity disorder has been associated with poor outcomes in academic achievement and adult life adaptation, including increased risk of accidents.[11] Visual spatial attention, critical at many developmental time points and important for safety in a variety of environments, is mediated by a well-studied, distributed large-scale neuroanatomical network composed of the frontal eye fields, the posterior parietal cortex, and the anterior cingulate area and their interconnections with one another and with subcortical structures in the thalamus and basal ganglia.[12,13] A measure of visuospatial inhibitory attention, the NIH-TB Flanker Inhibitory Control and Attention Test, was chosen for the Cognition Battery.

Working memory (WM) refers to a limited-capacity storage buffer that becomes overloaded when the amount of information exceeds that capacity. Conceptually, WM refers to the ability to 1) process information across a series of tasks and modalities, 2) hold the information in a short-term buffer, 3) manipulate the information, and 4) hold the products of that manipulation in the same short-term buffer. Cortical networks associated with spatial and nonspatial WM include prefrontal and posterior parietal regions.[14] WM has been studied extensively across the lifespan.[15,16] Its integrity has been linked to scholastic development[17] and letter knowledge,[18] and its impairment to reading disabilities.[19] WM improves significantly as children develop and WM span is thought to double in capacity between the ages of 5 and 10.[20] WM is relatively stable throughout adulthood. A reduction in performance in older adults may be attributable to a reduction in processing speed, rather than to changes in WM per se.[21] The test chosen to measure this construct is the NIH-TB List Sorting Working Memory Test.

Memory is composed of different systems of information storage and retrieval. The memory construct selected for the NIH-TB was episodic memory, a system involved in storage of unique events or experiences encoded in a time-specific manner. Episodic memory is fragile, sensitive to decay and interference and to both "normal" aging and many brain diseases. Episodic memory provides the building blocks for cognitive growth during development and is the system we rely on to update reality. Its absence, as in the historic case of the patient H.M., results in an existence in which there is only the present.[22] Episodic memory has a protracted course of development, with pronounced changes throughout the first 2 decades of life.[23,24] It is one of the first cognitive functions to show age-related decline and the most

susceptible to developmental disorders,[25] brain trauma, and neurodegenerative diseases such as Alzheimer disease.[26] The large-scale neuroanatomical network that supports episodic memory in addition to the hippocampus includes the hypothalamus, thalamus, medial temporal regions, cingulate cortex, and prefrontal cortex.[22,27] The NIH-TB Picture Sequence Memory Test is the measure of episodic memory in the Cognition Battery.

Language is a system of conventional symbols for communication, linked to a large-scale neuroanatomical network in the left cerebral hemisphere.[28] Developmental disorders of language and communication (e.g., autism, dyslexia) and limited opportunities to acquire literacy have a significant impact on academic achievement and life adaptation. Language scores can predict occupational attainment and performance.[29] Many acquired conditions can impair language in adulthood, including aphasia due to stroke and neurodegenerative brain disease. After much deliberation considering the various language components that the NIH-TB could test, 2 measures were designed: a single-word oral reading test, the NIH-TB Oral Reading Recognition Test, and a single-word vocabulary comprehension test, the NIH-TB Picture Vocabulary Test.

Reading was selected because it is a proxy for a broad range of cognitive, educational, and socioeconomic factors. The ability to pronounce low-frequency words with irregular orthography has been used as an estimate of overall intelligence.[30] Single-word reading recognition tasks are strong predictors of health and cognitive outcomes across the lifespan, and performance on these tasks is also an estimate of the quality of education, accounting for some of the racial/ethnic differences on neuropsychological test performance seen in older adults.[31,32]

Vocabulary represents the lexical component of language and is highly associated with general measures of "crystallized intelligence," or "g_c,"[33] overall cognitive functioning, and success in school and work.[29,34] Single-word auditory comprehension is a fundamental language skill that children learn very early, even before they are able to speak. Infants may have a repertoire of as many as 50 words they can understand before age 1.[35] Syntactic proficiency is equally important

for development,[36,37] but is more challenging to measure and to translate across different languages than single-word processing.

The final subdomain, Processing Speed (PS), is defined as either the amount of time it takes to process a set amount of information, or the amount of information that can be processed within a certain unit of time.[38] Simple PS tasks require a simple motor response to a target stimulus. Measures of complex PS, in contrast, require more concentration, as well as some mental manipulation.

The greatest growth in PS is observed relatively early and becomes more attenuated during childhood and adolescence.[39] Performance declines in young adulthood and steadily as people age.[40] PS measures are among the most sensitive indicators of cerebral dysfunction,[41] and slowed PS has been demonstrated in traumatic brain injury, multiple sclerosis, Parkinson disease, symptomatic HIV, chronic fatigue syndrome, dementia, and schizophrenia.[42] Slowed PS has been associated with changes in neurotransmitter activity (e.g., reduced cholinergic function, reduced numbers of D_2 dopamine receptors, and altered glutamate activity), white matter integrity, glucose metabolism, and nerve conduction velocities (e.g., as measured by evoked potentials, event-related potentials, and EEG).[42] For the Cognition Battery, the NIH-TB Pattern Comparison Processing Speed Test was chosen to measure PS.

The data reported in this article are derived from the validation study of the Cognition Battery. Results are reported for test-retest reliability, the effects of age on performance, and convergent and discriminant construct validity. More extensive details of test design and administration and scoring are available for the pediatric portion of the sample (ages 3–15),[43] and similar details will be presented for the adult sample (ages 20–85) in future publications.

METHODS Although the entire battery is computerized and includes automated scoring, it is necessary for an examiner to present task instructions, monitor compliance, and ensure valid results. For accessibility, all instructions are administered visually on the screen and also presented orally.

NIH-TB Cognition Battery tests. *NIH-TB Flanker Inhibitory Control and Attention Test (Executive/Attention).* This test is a version of the Eriksen flanker task derived from the Attention Network Test.[44] It tests the ability to inhibit visual attention to irrelevant task dimensions. On each trial, a central directional

target (fish for children younger than 8, arrows for ages 8 and older) is flanked by similar stimuli on the left and right. The task is to indicate the direction of the central stimulus. On congruent trials, the flankers face the same direction as the target. On incongruent trials, they face the opposite direction. A scoring algorithm integrates accuracy, a suitable measure in early childhood, and reaction time, a more relevant measure of adult performance on this task, yielding scores from 0 to 10. There are 40 trials and the average time to complete the task is 4 minutes.

NIH-TB Dimensional Change Card Sort Test (Executive/Shifting). The NIH-TB Dimensional Change Card Sort Test,[9] originally designed for children, was adapted for adults to assess the set-shifting component of EF. A target visual stimulus must be matched to 1 of 2 choice stimuli according to shape or color. Participants younger than 8 years receive a block of trials in which only 1 dimension is relevant and then a second block (switch) in which the other dimension is critical. Those who succeed following the switch also receive a mixed block, in which color is relevant on the majority of trials with occasional, unpredictable shifts to shape. Participants 8 years and older receive only the mixed block. The relevant criterion word, "color" or "shape," appears on the screen and for young children is also delivered orally via the computer. Scoring is similar to that for the flanker task, with an algorithm that weights accuracy for children and reaction time for adults. A total of 40 trials require 4 minutes.

NIH-TB List Sorting Working Memory Test (Working Memory). This task is an adaptation of Mungas' List Sorting task from the Spanish and English Neuropsychological Assessment Scales.[45,46] A series of stimuli is presented on the computer screen visually (object) and orally (spoken name), 1 at a time. Participants are instructed to repeat the stimuli to the examiner in order of size, from smallest to largest. In 1 condition, all stimuli come from 1 category. In the second, stimuli are presented from 2 categories, following which the participant must report first all stimuli from 1 category, then from the other, in order of size within each. The number of items in each series increases from one trial to the next and the test is discontinued when 2 trials of the same length are failed. The prototype task has been previously validated in an elderly sample.[45,47] The List Sorting task takes approximately 7 minutes to administer. Test scores consist of total items correct across all trials.

NIH-TB Picture Sequence Memory Test (Episodic Memory). The NIH-TB Picture Sequence Memory Test is a new measure derived from imitation-based tasks (elicited and deferred imitation) used in research with infants and young children.[48–50] The original stimuli were 3-dimensional props used to produce action sequences that the infant or child imitates. For the NIH-TB, the stimuli are pictured objects and activities, thematically related but with no inherent order. For each trial, pictures appear in the center of the computer screen and then are moved 1 at a time into a fixed spatial order, as an audio file simultaneously describes the content of each (e.g., "Plant the tomatoes"), until the entire sequence is displayed on the screen. Then the pictures return to the center of the screen in a random display and the participant must move them into the sequence demonstrated. The score is derived from the cumulative number of adjacent pairs of pictures remembered correctly over 3 learning trials. Based on pilot testing, level of task difficulty was adjusted for the various age groups. Thus, for ages 3 to 4 years, 6 pictures were administered; 5 to 6 years, 9 pictures; 8 years, 12 pictures; 9 to 60 years, 15 pictures; and 65 to 85 years, 9 pictures. Administration time is approximately 10 minutes.

NIH-TB Oral Reading Recognition Test (Language). This test measures the ability to pronounce single printed words and/or to recognize letters. An English item bank, controlled for frequency of word use, complexity of letter-sound relationships, and orthographic typicality, was developed with an initial set of item response theory

calibrations. Letters and other "prereading" items are included. Items are presented on the computer screen one by one and the participant is asked to read them aloud. Items are administered by computer adaptive testing (CAT) and participant responses are entered by the examiner. The CAT item bank in final form will contain approximately 250 items, although only 30 to 40 will be presented, depending on performance. Average administration time is 4 minutes.

NIH-TB Picture Vocabulary Test (Language). Single words are presented via an audio file, paired simultaneously with 4 screen images of objects, actions, and/or depictions of concepts. The task is to pick the picture that matches the spoken word. The test is CAT administered, which reduces the amount of time to identify performance level. The test does not require speaking and can be performed by individuals who are preliterate and illiterate. Items were recalibrated and final parameter estimates were obtained after norming for optimal CAT administration. Total administration time is approximately 5 minutes.

NIH-TB Pattern Comparison Processing Speed Test (Processing Speed). This test is modeled after Salthouse's Pattern Comparison Task,[51] an extensively researched assessment of choice reaction time, easily adapted for computerized administration. Participants are asked to identify whether 2 visual patterns are the "same" ("Yes" button) or "not the same" ("No" button). Children younger than 8 years indicate these choices with a "smiley" or "frowny" face button. Type, complexity, and number of stimuli are varied to ensure adequate variability of performance across the age spectrum from 3 to 85 years. The NIH-TB Pattern Comparison Processing Speed Test requires 3 minutes to administer and the score is the number of correct items (of a possible 130) completed in 90 seconds.

Subjects. The sample (N = 476) was based on a stratification plan to include adequate numbers of individuals within age bands, level of education, and racial/ethnic backgrounds. A marketing research firm assisted in recruitment of community-dwelling individuals. Testing was completed at Northwestern University (Chicago) and 5 additional sites: Emory University (Atlanta), the University of Minnesota (Minneapolis), the University of Washington (Seattle), NorthShore University HealthSystem (Evanston, IL), and Kessler Foundation Research Center (West Orange, NJ). Eligible participants were 3 to 85 years of age living in the community. See table 1 for age, sex, race, and education strata. Not all ages were sampled in this study. Education levels in the table are defined as actual years of education completed by adult participants (ages 20+ years) and highest parental education for children (ages 3–15 years). One-third of the sample was randomly selected to repeat testing 7 to 21 days later to assess test-retest reliability.

Analyses. This initial report includes results from analyses of test-retest reliability, associations of test scores with age, and convergent and discriminant construct validity. Age associations reflect the validity of the tests for measuring cognitive development during childhood and age-related cognitive decline during adulthood. Convergent and discriminant validity results provide evidence that the Cognition Battery is measuring the intended constructs.

Pearson correlation coefficients between age and test performance were calculated separately for children and adults. Intraclass correlation coefficients were calculated to evaluate test-retest reliability. Convergent validity was assessed by correlations between each NIH-TB measure and a well-established "gold standard" measure of the same construct; evidence of discriminant validity was assessed by correlations with gold standards of a *different* cognitive construct. Gold standard measures for each NIH-TB instrument are listed in table 2.

Gold standard measures were scarce for 3- to 6-year-olds, and we were unable to identify well-established measures to test convergent or discriminant validity in this age group for the measures of

Table 1   Initial validation sample demographics (N = 476)[a]

Age groups	Education (self/parent)	Sex		Race/ethnicity		
		Male	Female	White	Black	Hispanic/other/multiple
3–6 y, n = 120	<High school	6	5	5	5	1
	High school graduate	29	27	23	19	14
	College+	29	24	26	16	11
Total		64	56	54	40	25
8–15 y, n = 88	<High school	4	6	4	5	1
	High school graduate	22	23	18	13	14
	College+	14	19	16	7	10
Total		40	48	38	25	26
20–60 y, n = 159	<High school	22	26	21	15	12
	High school graduate	29	31	26	19	15
	College+	24	27	24	15	12
Total		75	84	71	49	39
65–85 y, n = 109	<High school	9	11	9	10	1
	High school graduate	12	27	26	11	2
	College+	23	27	42	5	3
Total		44	65	77	26	6

[a] Parental education was used for participants ages 3 to 15 years and participant education was used for adults (ages 20+).

attention, episodic memory, EF, and PS. Thus, in this age group, only convergent validity was measured between the Cognition Battery measures and a measure of general cognitive ability (i.e., "*g*") obtained by averaging *z* scores of the Wechsler Preschool and Primary Scale of Intelligence–3rd edition Block Design subtest[52] and the Peabody Picture Vocabulary Test–4th edition.[53] More detailed psychometric information on individual measures and on challenges related to testing for construct validity in very young children is detailed in Zelazo et al. (in press).[9]

RESULTS **Test-retest reliability.** Test-retest reliability was strong for the entire sample and separately for

children (ages 3–15 years) and adults (ages 20–85 years; table 3). Intraclass correlation coefficients for the entire sample on the NIH Toolbox measures ranged from 0.78 for the Picture Sequence Memory Test to 0.99 on the Oral Reading Recognition Test, with most other values falling above 0.90.

Age effects. All cognitive abilities are expected to improve during childhood, and most are expected to show some age-related decline during adulthood, with the exception of language skills and other

Table 2   Convergent and discriminant validity ("gold standard") measures used for ages 8 to 85 years

NIH Toolbox measure	Convergent validity measure	Discriminant validity measure
Flanker (attention)	WISC-IV/WAIS-IV Letter-Number Sequencing/Coding/Symbol Search average[a]	PPVT-4
Flanker (EF)	D-KEFS Inhibition	PPVT-4
DCCS	D-KEFS Inhibition	PPVT-4
PSMT	BVMT-R/RAVLT average	PPVT-4
List Sorting	WISC-IV/WAIS-IV Letter-Number Sequencing[a]/PASAT average	PPVT-4
Pattern Comparison	WISC-IV/WAIS-IV Coding/Symbol Search average[a]	PPVT-4
Vocabulary	PPVT-4	BVMT-R/RAVLT average
Reading	WRAT-4	BVMT-R/RAVLT average

Abbreviations: BVMT-R = Brief Visuospatial Memory Test–revised; DCCS = Dimensional Change Card Sort; D-KEFS = Delis-Kaplan Executive Function System; EF = executive function; PASAT = Paced Auditory Serial Addition Test; PPVT-4 = Peabody Picture Vocabulary Test-4th edition; PSMT = Picture Sequence Memory Test; RAVLT = Rey Auditory Verbal Learning Test; WAIS-IV = Wechsler Adult Intelligence Scale-4th edition; WISC-IV = Wechsler Intelligence Scale for Children-4th edition; WRAT-4 = Wide Range Achievement Test-4th edition.
[a] Depending on subject's age.

Table 3	Test-retest reliability of instruments in the NIH-TB Cognition Battery						
	All		**Children aged 3–15 y**		**Adults aged 20–85 y**		
NIH-TB cognition subdomain/instrument	**No.**	**ICC (95% CI)**	**No.**	**ICC (95% CI)**	**No.**	**ICC (95% CI)**	
Executive Function/Flanker	125	0.96 (0.94, 0.97)	52	0.95 (0.92, 0.97)	73	0.80 (0.70, 0.87)	
Executive Function/DCCS	123	0.94 (0.92, 0.96)	49	0.92 (0.86, 0.95)	74	0.88 (0.82, 0.92)	
Attention/Flanker	130	0.95 (0.93, 0.96)	57	0.95 (0.92, 0.97)	73	0.94 (0.91, 0.96)	
Episodic Memory/PSMT	155	0.78 (0.71, 0.83)	66	0.76 (0.64, 0.84)	89	0.77 (0.67, 0.84)	
Working Memory/List Sorting	155	0.89 (0.85, 0.92)	66	0.87 (0.80, 0.92)	89	0.77 (0.67, 0.84)	
Processing Speed/Pattern Comparison	148	0.82 (0.76, 0.87)	59	0.84 (0.75, 0.90)	89	0.72 (0.60, 0.81)	
Language/Vocabulary	155	0.94 (0.92, 0.96)	66	0.84 (0.75, 0.90)	89	0.81 (0.73, 0.87)	
Language/Reading	154	0.99 (0.99, 0.99)	65	0.99 (0.98, 0.99)	89	0.91 (0.87, 0.94)	

Abbreviations: CI = confidence interval; DCCS = Dimensional Change Card Sort; ICC = intraclass correlation coefficient; NIH-TB = NIH Toolbox for the Assessment of Neurological and Behavioral Function; PSMT = Picture Sequence Memory Test.

aspects of "crystallized" intelligence. Therefore, correlations between age and NIH-TB test performance were conducted separately for children (ages 3–15 years) and adults (ages 20–85 years). Table 4 presents these results.

All NIH-TB Cognition Battery measures showed robust associations between test performance and age in the child group ($r = 0.58–0.87$), where scores improved with age. With the exception of the language measures (Vocabulary and Reading, $r = 0.15$ and -0.02, respectively), age and test scores ($r = -0.46$ to -0.65) on the remaining NIH-TB Cognition Battery measures were negatively associated, with lower scores at higher age levels. Thus, on the Picture Sequence Memory Test, performance improved during childhood and early adolescence, with gradual decline in scores across adult age ranges beginning in the 30s (figure, A). In contrast, the NIH-TB Picture Vocabulary Test showed gradual, linear improvement with age until the mid 50s and then stabilized, whereas Oral Reading Recognition showed a much sharper increase until early grade-school years (age 7–8) and then followed the same pattern of more gradual improvement and then stability in the older age groups (figure, B).

Convergent and discriminant validity. In children from 3 to 6 years of age, all NIH-TB Cognition Battery measures were significantly correlated (ranging from $r = 0.54$ to $r = 0.74$) with our measure of general cognitive ability ("g"), indicating that they are sensitive to a range of different cognitive ability levels within this age cohort. Table 5 shows results for convergent and discriminant validity for ages 8 to 85 years. For all NIH-TB CB instruments, correlations for convergent validity measures ranged from $r = 0.48$ to $r = 0.93$ (all $p < 0.0001$), suggesting that the NIH-TB measures are tapping the desired constructs. Correlations for discriminant validity measures ranged from $r = 0.05$ to $r = 0.30$, indicating

lack of, or relatively weak, relationship with measures that tap different constructs.

DISCUSSION This article introduces the NIH Toolbox Cognition Battery, a brief series of cognitive tests for the purpose of supplementing measures in epidemiologic and longitudinal studies to constitute a "common currency" among researchers. The Cognition Battery has 7 computerized instruments that measure 6 ability subdomains important for cognitive health from the ages of 3 to 85 years. Data are presented for 208 normal children (age 3–15 years) and 268 normal adults (age 20–85 years) on 3 important psychometric characteristics: test-retest reliability, sensitivity to cognitive growth during childhood and age-related decline during adulthood, and construct validity.
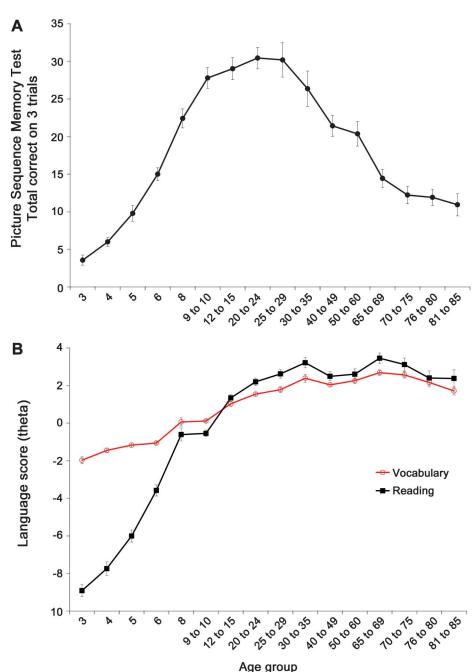
Table 4	Correlations between NIH-TB cognition subdomain scores and age			
	Ages 3–15 y		**Ages 20–85 y**	
Cognition subdomain/test	**No.**	**r**	**No.**	**r[a]**
Executive Function/Flanker	176	0.67[b]	235	−0.50[b]
Executive Function/DCCS	171	0.72[b]	236	−0.53[b]
Attention/Flanker	190	0.58[b]	237	−0.54[b]
Episodic Memory	204	0.80[b]	266	−0.65[b]
Working Memory	201	0.79[b]	265	−0.46[b]
Processing Speed	189	0.77[b]	265	−0.53[b]
Language/Vocabulary	200	0.82[b]	265	0.15
Language/Reading	204	0.87[b]	266	−0.02

Abbreviations: DCCS = Dimensional Change Card Sort; NIH-TB = NIH Toolbox for the Assessment of Neurological and Behavioral Function; r = Pearson correlation coefficient.
[a] Correlations adjusted for education in adult group.
[b] $p < 0.001$.

(A) NIH Toolbox Picture Sequence Memory Test scores show improvement into early adulthood and then decline from the 50s on. Administration set size varied by age group as follows: ages 3 to 4 years, 6 pictures; 5 to 6 years, 9 pictures; 8 years, 12 pictures; 9 to 60 years, 15 pictures, and 65 to 85 years, 9 pictures. (B) Results for NIH Toolbox Reading and Vocabulary scores (reported as a "theta," or individual ability score, based on item response theory analyses) show improvement sustained into adulthood. The data points in both A and B represent the mean score ± standard error.

The subject sample for this study deliberately emphasized representation of ethnic minorities (almost 50%) and the oldest and youngest groups (3–6 years and 65–85 years; together, 48%) to ensure that the tests would perform as needed in these important segments of the population. Thus, the participants in this study (or their parents, in the case of children) tended to be rather highly educated, particularly in the youngest and oldest groups (see table 1). A more representative population-based sampling strategy was implemented for the NIH-TB norming study.

Adequate test-retest reliability was considered essential for the NIH-TB Cognition Battery, particularly because of its anticipated use in longitudinal studies. The results suggest that test-retest reliability of all NIH-TB measures is good to excellent across a large age range. Composite scores, which have higher reliability than individual test scores, are being developed to

| | Ages 8-85 y | | Ages 8-85 y | | Ages 3-6 y | |
| | Convergent validity[a] | | Discriminant validity[a] | | Correlation with average of WPPSI Block Design and PPVT-4 z scores | |
	No. of subjects	r	No. of subjects	r	No. of subjects	r
Flanker/Executive Attention	312	−0.48[b]	317	0.15[c]	89	0.70[b]
DCCS/Executive Switching	311	−0.51[b]	317	0.14[d]	85	0.74[b]
PSMT/Episodic Memory	351	0.69[b]	350	−0.08	115	0.59[b]
List Sorting/Working Memory	350	0.58[b]	350	0.30[b]	112	0.64[b]
Processing Speed	349	0.49[b]	349	0.12[d]	99	0.54[b]
Vocabulary/Language	350	0.78[b]	351	0.08	111	0.74[b]
Reading/Language	351	0.93[b]	351	0.19[b]	115	0.65[b]

Table 5 Convergent and discriminant validity of NIH-TB cognitive function battery measures

Abbreviations: DCCS = Dimensional Change Card Sort; NIH-TB = NIH Toolbox for the Assessment of Neurological and Behavioral Function; PPVT-4 = Peabody Picture Vocabulary Test–4th edition; PSMT = Picture Sequence Memory Test; r = Pearson correlation coefficient; WPPSI = Wechsler Preschool and Primary Scale of Intelligence–3rd edition.
[a] Gold standard measures used to test convergent and discriminant validity for ages 8 to 85 y appear in table 2.
[b] $p < 0.001$.
[c] $p < 0.01$.
[d] $p < 0.05$.

increase the potential for use of the battery in clinical trials and other longitudinal research.

Evidence of test validity can take many forms, and derives from both clinical and nonclinical subject samples. We presented the relationship of Cognition Battery performance with age in cognitively normal children and adults. The Reading and Vocabulary scores showed the expected associations with age, growing through adolescence and stabilizing in older adulthood. Not surprisingly, Reading showed an especially steep improvement from age 3 to the early school years, when both formal and informal educational experiences ideally promote such development. The Language subdomain tests, as expected, are experience-based and peaked somewhat later than measures of other cognitive subdomains, and then remained relatively stable even into the ninth decade of life. The nonlanguage subdomain tests in the battery, in contrast, conformed to the pattern expected with cognitive ability measures in that they peaked in early adulthood and then declined in later adulthood at different rates, depending on the measure.

Another validity measure we included in this report expresses how well the tests in the battery measure the intended constructs (convergent validity) as opposed to different cognitive constructs (discriminant validity). The "gold standard" tests related to the Cognition Battery instruments in the expected ways for participants across a wide age band (ages 8–85 years), demonstrating both convergent and discriminant validity. Evaluating construct validity in young children (ages 3–6 years) was challenging because of the absence of specific gold standard measures of targeted constructs appropriate for these ages. (See Zelazo et al., in press,[54] for discussion.) The lack of such measures may reflect the fact

that different subdomains of cognition become more differentiated with experience and development.[55] The correlations between the NIH-TB measures and general cognitive ability, our index of convergent validity in young children (ages 3–6 years), were high (0.54–0.74), possibly supporting such a notion.

The NIH-TB Cognition Battery was designed as a brief, diverse, accessible, and psychometrically sound set of instruments that will be broadly applicable in research studies of normal and abnormal groups across a wide age range. The current results regarding age effects, test-retest reliability, and construct validity are promising. The next phase of development established normative standards for the NIH-TB measures using a large, demographically diverse sample, including a Spanish-language version of the measures. More detailed information will also be available about associations of test performances with various aspects of everyday functioning (e.g., school performance in the child sample) and relationships with additional demographic characteristics (educational level/socioeconomic status, sex, and ethnicity). The NIH-TB Cognition Battery was not developed as a clinical measure to either screen for cognitive impairment or to substitute for a full, competent neuropsychological evaluation. However, future studies with clinical populations are expected to generate another source of validation of the Cognition Battery as a sound set of measures of a broad range of normal and abnormal cognitive functioning, with implications for brain health in large-scale research studies.

AUTHOR CONTRIBUTIONS

Sandra Weintraub: drafting/revising the manuscript, study concept or design, analysis or interpretation of data, acquisition of data. Sureyya

ACKNOWLEDGMENT

STUDY FUNDING

DISCLOSURE

DISCLAIMER

The views and opinions expressed in this report are those of the authors and should not be construed to represent the views of NIH or any of the sponsoring organizations, agencies, or the U.S. government.

REFERENCES

1. Diamond A. Normal development of prefrontal cortex from birth to young adulthood: cognitive functions anatomy and biochemistry. In: Stuss D, Knight B, editors. Principles of Frontal Lobe Function. New York: Oxford University Press; 2002:466–503.

2. Diamond A. Frontal lobe involvement in cognitive changes during the first year of life. In: Gibson KR, Petersen AC; Council SSR, editors. Brain Maturation and Cognitive Development: Comparative and Cross-cultural Perspectives. New York: De Gruyter; 1991:127–180.

3. Zelazo PD, Carlson SM, Kesek A. Development of executive function in children. In: Nelson CA, Luciana M, editors. Handbook of Developmental Cognitive Neuroscience, 2nd ed. Cambridge, MA: MIT Press; 2008:553–574.

4. Daniels K, Toth J, Jacoby L. The aging of executive functions. In: Bialystok E, Craik FIM, editors. Lifespan Cognition: Mechanisms of Change. New York: Oxford University Press; 2006:96–111.

5. Zelazo PD, Craik FI, Booth L. Executive function across the life span. Acta Psychol 2004;115:167–183.

6. 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. Cogn Psychol 2000; 41:49–100.

7. Wiebe SA, Sheffield T, Nelson JM, Clark CAC, Chevalier N, Espy K. The structure of executive function in 3-year-olds. J Exp Child Psychol 2011;108:436–452.

8. Durston S, Davidson MC, Tottenham N, et al. A shift from diffuse to focal cortical activity with development. Dev Sci 2006;9:1–8.

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

10. Weintraub S. Neuropsychological assessment of mental state. In: Mesulam MM, editor. Principles of Cognitive and Behavioral Neurology. New York: Oxford University Press; 2000:121–173.

11. Swensen A, Birnbaum HG, Ben Hamadi R, Greenberg P, Cremieux PY, Secnik K. Incidence and costs of accidents among attention-deficit/hyperactivity disorder patients. J Adolesc Health 2004;35:346.e341–346.e349.

12. Mesulam M. A cortical network for directed attention and unilateral neglect. Ann Neurol 1981;10:309–325.

13. Posner MI, Petersen SE. The attention system of the human brain. Annu Rev Neurosci 1990;13:25–42.

14. D'Esposito M, Aguirre GK, Zarahn E, Ballard D, Shin RK, Lease J. Functional MRI studies of spatial and nonspatial working memory. Cogn Brain Res 1998;7:1–13.

15. Conlin JA, Gathercole SE, Adams JW. Children's working memory: investigating performance limitations in complex span tasks. J Exp Child Psychol 2005;90:303–317.

16. Salthouse TA, Meinz EJ. Aging, inhibition, working memory, and speed. J Gerontol B Psychol Sci Soc Sci 1995;50: P297–P306.

17. Hitch GJ, Towse JN, Hutton U. What limits children's working memory span? Theoretical accounts and applications for scholastic development. J Exp Psychol Gen 2001; 130:184–198.

18. de Jong PF, Olson RK. Early predictors of letter knowledge. J Exp Child Psychol 2004;88:254–273.

19. de Jong PF. Working memory deficits of reading disabled children. J Exp Child Psychol 1998;70:75–96.

20. Riggs KJ, McTaggart J, Simpson A, Freeman RP. Changes in the capacity of visual working memory in 5- to 10-year-olds. J Exp Child Psychol 2006;95:18–26.

21. Salthouse TA, Coon VE. Interpretation of differential deficits: the case of aging and mental arithmetic. J Exp Psychol Learn Mem Cogn 1994;20:1172–1182.

22. Corkin S, Amaral DG, Gonzalez RG, Johnson KA, Hyman BT. H. M.'s medial temporal lobe lesion: findings from magnetic resonance imaging. J Neurosci 1997;17: 3964–3979.

23. Bauer PJ. Remembering the Times of Our Lives: Memory in Infancy and Beyond. Mahwah, NJ: Lawrence Erlbaum Associates; 2007.

24. Perner J, Ruffman T. Episodic memory and autonoetic consciousness: developmental evidence and a theory of childhood amnesia. J Exp Child Psychol 1995;59:516–548.

25. Gadian DG, Aicardi J, Watkins KE, Porter DA, Mishkin M, Vargha-Khadem F. Developmental amnesia associated with early hypoxic-ischaemic injury. Brain 2000;123(pt 3):499–507.

26. Albert MS. Cognitive and neurobiologic markers of early Alzheimer disease. Proc Natl Acad Sci USA 1996;93: 13547–13551.

27. Mitchell AS, Dalrymple-Alford JC. Lateral and anterior thalamic lesions impair independent memory systems. Learn Mem 2006;13:388–396.

28. Price CJ. The anatomy of language: contributions from functional neuroimaging. J Anat 2000;197(pt 3):335–359.

29. Schmidt FL, Hunter J. General mental ability in the world of work: occupational attainment and job performance. J Pers Soc Psychol 2004;86:162–173.

30. Grober E, Sliwinski M. Development and validation of a model for estimating premorbid verbal intelligence in the elderly. J Clin Exp Neuropsychol 1991;13:933–949.

31. Manly JJ, Byrd DA, Touradji P, Stern Y. Acculturation, reading level, and neuropsychological test performance among African American elders. Appl Neuropsychol 2004;11:37–46.

32. Manly JJ, Jacobs DM, Sano M, et al. Effect of literacy on neuropsychological test performance in nondemented, education-matched elders. J Int Neuropsychol Soc 1999;5:191–202.

33. Cattell RB. Intelligence: Its Structure, Growth and Action. Amsterdam: Elsevier; 1987.

34. Kastner JW, May W, Hildman L. Relationship between language skills and academic achievement in first grade. Percept Mot Skills 2001;92:381–390.

35. Fenson L, Dale PS, Reznick JS, Bates E, Thal DJ, Pethick SJ. Variability in early communicative development. Monogr Soc Res Child Dev 1994;59:1–173.

36. Gleason JB, Ratner NB, editors. The Development of Language, 7th ed. Boston: Pearson/Allyn & Bacon; 2009.

37. Hirsh-Pasek K, Golinkoff RM. The Origins of Grammar: Evidence from Early Language Comprehension. Cambridge, MA: M.I.T. Press; 1996.

38. Kalmar JH. Information processing speed in multiple sclerosis: a primary deficit? In: DeLuca J, Kalmar JH, editors. Information Processing Speed in Clinical Populations. New York: Taylor & Francis; 2007:153–172.

39. Fry AF, Hale S. Relationships among processing speed, working memory and fluid intelligence in children. Biol Psychol 2000;54:1–34.

40. Salthouse TA. Aging and measures of processing speed. Biol Psychol 2000;54:35–54.

41. Hawkins KA. Indicators of brain dysfunction derived from graphic representations of the WAIS-III/WMS-III technical manual clinical samples data: a preliminary approach to clinical utility. Clin Neuropsychologist 1998;12:535–551.

42. DeLuca J, Kalmar JH. Information Processing Speed in Clinical Populations. London: Psychology Press; 2007.

43. Zelazo PD, Bauer PJ, Editors. National Institutes of Health Toolbox—Cognitive Function Battery (NIH Toolbox CFB): Validation for Children Between 3 and 15 Years. Monogr Soc Res Child Dev (in press).

44. Rueda MR, Fan J, McCandliss BD, et al. Development of attentional networks in childhood. Neuropsychologia 2004;42:1029–1040.

45. Mungas D, Reed BR, Tomaszewski Farias S, DeCarli C. Criterion-referenced validity of a neuropsychological test battery: equivalent performance in elderly Hispanics and non-Hispanic Whites. J Int Neuropsychol Soc 2005;11:620–630.

46. Crane PK, Narasimhalu K, Gibbons LE, et al. Item response theory facilitated cocalibrating cognitive tests and reduced bias in estimated rates of decline. J Clin Epidemiol 2008;61:1018–1027.e9.

47. Mungas D, Widaman KF, Reed BR, Tomaszewski Farias S. Measurement invariance of neuropsychological tests in diverse older persons. Neuropsychology 2011;25:260–269.

48. Bauer PJ. Long-term recall memory: behavioral and neuro-developmental changes in the first two years of life. Curr Dir Psychol Sci 2002;11:137–141.

49. Bauer PJ. Developments in declarative memory. Psychol Sci 2005;16:41–47.

50. Bauer PJ. Constructing a past in infancy: a neuro-developmental account. Trends Cogn Sci 2006;10:175–181.

51. Salthouse TA, Babcock RL, Shaw RJ. Effects of adult age on structural and operational capacities in working memory. Psychol Aging 1991;6:118–127.

52. Wechsler D. WPPSI-III Wechsler Preschool Primary Scale of Intelligence, 3rd ed. San Antonio: Psychological Corporation; 2002.

53. Dunn LM, Dunn LM. Peabody Picture Vocabulary Test Fourth Edition (PPVT-IV). Bloomington, MN: NCS Pearson; 2007.

54. Zelazo PD, Anderson JE, Richler J, Wallner-Allen K, Beaumont JL, Weintraub S. NIH Toolbox cognitive function battery (CFB): measuring executive function and attention. Monogr Soc Res Child Dev (in press).

55. Johnson MH. Interactive specialization: a domain-general framework for human functional brain development? Developmental Cognitive Neuroscience 2011;1:7–21.