Published in final edited form as:
J Int Neuropsychol Soc. 2014 July ; 20(6): 567–578. doi:10.1017/S1355617714000320.

The Cognition Battery of the NIH Toolbox for Assessment of Neurological and Behavioral Function: Validation in an Adult Sample

Sandra Weintraub[1], **Sureyya S. Dikmen**[2], **Robert K. Heaton**[3], **David S. Tulsky**[4,5], **Philip David Zelazo**[6], **Jerry Slotkin**[7], **Noelle E. Carlozzi**[8], **Patricia J. Bauer**[9], **Kathleen Wallner-Allen**[10], **Nathan Fox**[11], **Richard Havlik**[10], **Jennifer L. Beaumont**[7], **Dan Mungas**[12], **Jennifer J. Manly**[13], **Claudia Moy**[14], **Kevin Conway**[15], **Emmeline Edwards**[16], **Cindy J. Nowinski**[7], and **Richard Gershon**[7]

[1]Cognitive Neurology and Alzheimer's Disease Center; Departments of Psychiatry and Neurology, Northwestern Feinberg School of Medicine, Chicago, Illinois [2]Department of Rehabilitation Medicine, University of Washington, Seattle, Washington [3]Department of Psychiatry, University of California, San Diego, California [4]Departments of Rehabilitation Medicine, Orthopedic Surgery, and General Medicine, New York University Langone Medical Center, New York, New York [5]Spinal Cord Injury and Outcomes Laboratories, Kessler Foundation, West Orange, New Jersey [6]Institute of Child Development, University of Minnesota, Minneapolis, Minnesota [7]Department of Medical Social Sciences, Northwestern University, Chicago, Illinois [8]Physical Medicine and Rehabilitation, University of Michigan, Ann Arbor, Michigan [9]Department of Psychology, Emory University, Atlanta, Georgia [10]Westat, Rockville, Maryland [11]Department of Human Development, University of Maryland, College Park, Maryland [12]Department of Neurology, University of California, Davis, California [13]Cognitive Neuroscience Division, Taub Institute for Research in Alzheimer's Disease and the Aging Brain, Columbia University, New York, New York [14]National Institute of Neurological Disorders and Stroke, Bethesda, Maryland [15]National Institute on Drug Abuse, Rockville, Maryland [16]National Center for Complementary and Alternative Medicine, Bethesda, Maryland

Abstract

This paper introduces a special series on validity studies of the Cognition Battery (CB) from the U.S. National Institutes of Health Toolbox for the Assessment of Neurological and Behavioral Function (NIHTB) (R. C. Gershon et al., 2013) in an adult sample. This first paper in the series describes the sample, each of the seven instruments in the NIHTB-CB briefly, and the general approach to data analysis. Data are provided on test-retest reliability and practice effects, and raw scores (mean, standard deviation, range) are presented for each instrument and the gold standard instruments used to measure construct validity. Accompanying papers provide details on each instrument, including information about instrument development, psychometric properties, age and education effects on performance, and convergent and discriminant construct validity. One paper in the series is devoted to a factor analysis of the NIHTB-CB in adults and another describes the psychometric properties of three composite scores derived from the individual measures representing fluid and crystallized abilities and their combination. The NIHTB-CB is designed to

provide a brief, comprehensive, common set of measures to allow comparisons among disparate studies and to improve scientific communication.

Introduction

In 1990, the “Decade of the Brain” was a joint initiative of the U.S. Library of Congress and the National Institute of Mental Health, focusing attention on human brain science and diseases of the nervous system. In 2000, the American Psychological Association adopted the moniker “The Decade of Behavior” to highlight mental diseases deserving of research support to effect changes in public health policy over the following 10 years. Both of these developments and a series of more recent initiatives supported in the U.S. by the National Institutes of Health (NIH) have highlighted the importance of brain health and have promoted an unprecedented era of research on mechanisms and treatment of central nervous system disorders. Although there have been initiatives around the globe to design common measures for research studies, to our knowledge the NIH Toolbox for the Assessment of Neurological and Behavioral Function (NIHTB) is the first initiative that is not directed at a specific disease, age group, or arena of use (e.g., school, hospital clinic). Instead, the NIHTB was conceived as a tool to measure neurological functions that would span different disciplines, apply to diverse research questions, and measure a broad range of ability across the lifespan from three to 85 years of age.

The importance of cognitive health and the impact of cognitive functioning on a wide range of behaviors and study outcomes has been made increasingly clear by growing knowledge of the effects of disease and of aging on brain health. Cognitive decline with aging, itself a looming challenge for the health care system in the U.S. (Brookmeyer, Gray, & Kawas, 1998), also could introduce a “hidden variable” into studies that are not measuring cognition as a potential modulator of outcome. For example, research results from a study of the impact of interventions to improve health literacy in older adults could be invalidated if cognition is not measured, since different aspects of health literacy are dependent on distinct components of cognition (Wolf et al., 2012).

Information about the late effects of traumatic brain injury, especially in the sports world (Erlanger, Kutner, Barth, & Barnes, 1999), has made us more aware of the potential cumulative influence of such adverse events on the brain in development and aging (McKee et al., 2009). Early lifestyle choices, such as maintaining a healthy level of physical activity, can influence the emergence and rate of cognitive decline in one's later years (Barnes & Yaffe, 2011). Health practices throughout life, such as estrogen replacement therapy in postmenopausal women, also may influence later development of cognitive dysfunction (Shao et al., 2012). Congenital or early-acquired brain disease typically has an impact on cognitive development that influences subsequent achievement in the school years and beyond (Anderson, Catroppa, Morse, Haritou, & Rosenfeld, 2005). As a result, increasing attention has been devoted to the study of clinical conditions that affect cognition and cognitive development, the effects of early and late brain injury on subsequent development, and the cognitive changes associated with normal and abnormal brain aging. Finally, there is

increasing focus on interventions that may successfully treat or reverse neurological diseases that cause cognitive impairment.

The NIHTB was designed to provide a common currency, or set of common data elements, among disparate studies using standard methodology so that differences in the outcomes of these studies would be less likely to be a result of differences in the test instruments used. It contains four modules, each addressing a different domain of neurologic/behavioral function: Cognition, Emotion, Motor and Sensory Function (see www.nihtoolbox.org). By using measures that offer a continuous scoring model from ages 3-85, the NIHTB allows for protracted longitudinal study across the life span. The development of the NIH Toolbox was conducted through the collaborative framework of the U.S. NIH Blueprint for Neuroscience Research initiative. Sixteen Institutes, centers and offices of the NIH support this initiative for neuroscience research to accelerate discoveries and reduce the burden of nervous system disorders. General methods applied to the development of measures in all four major domains are detailed in a separate series of papers introducing the full NIHTB (Coldwell et al., 2013; Cook et al., 2013; Dalton et al., 2013; Dunn et al., 2013; R. C. Gershon et al., 2013; Hodes, Insel, Landis, & Research, 2013; Nowinski, Victorson, Debb, & Gershon, 2013; Reuben et al., 2013; Rine et al., 2013; Salsman et al., 2013, 2013, in press; Salthouse, 1976; Varma, McKean-Cowdin, Vitale, Slotkin, & Hays, 2013; Victorson et al., 2013; Weintraub, Dikmen, et al., 2013; Zecker et al., 2013). The NIHTB Cognition Battery (NIHTB-CB) is the focus of the present series.

The present set of papers is the third in a series of publications that include the NIHTB-CB. The first publication introduced the Cognition Battery along with the other four modules of the NIHTB and provided an overview and summary data from the entire validation sample, children and adults (Weintraub, Dikmen, et al., 2013). The second set of publications was in the form of a monograph focusing solely on the validation study in the pediatric sample of participants from 3-15 years of age (Akshoomoff et al., 2013; P. J. Bauer & Zelazo, 2013; Carlozzi, Tulsky, Kail, & Beaumont, 2013; Fox, 2013; R.C. Gershon et al., 2013; Mungas et al., 2013; Tulsky et al., 2013; Weintraub, Bauer, et al., 2013; Zelazo et al., 2013). The present series of papers concentrates on the validation study completed in adults from 20-85 years of age. It builds on prior publications but provides more detailed description of the instruments, on the adaptations needed to make tests originally designed for children applicable to an adult sample, and on test administration, scoring procedures, and construct validity, as well as test-retest reliability. Factor structure and age and other demographic effects on performance in adults also constitute novel information. Data have not been previously reported to the degree of detail employed here.

To date, the NIHTB-CB has been validated as a research test battery and not for clinical use, nor would it substitute for a comprehensive clinical neuropsychological examination of patients with neurobehavioral symptoms and disorders. It has several potential applications in clinical research and in longitudinal, large-scale epidemiologic studies where there is the need for brief instruments that tap different cognitive constructs within a very large age range and without showing floor or ceiling effects. The NIHTB-CB can be an add-on in studies in which cognition is being tested with more specialized instruments. In that instance, it would allow comparisons with other studies also using the NIHTB-CB.

Furthermore, it can serve in studies in which cognition is not a targeted outcome, but in which a measure of cognition might be useful as a covariate, for example, to address the potentially "hidden" cognitive variables that could affect outcomes and have an impact on tailoring or personalizing treatment.

General Methods

Development of the Cognition Battery

The NIH Toolbox project team specified the following criteria for all four major domains: 1) brevity (approximately 30 minutes); 2) applicability across a broad age spectrum from 3-85 years; 3) sensitivity to the full range of normal functioning (minimal ceiling and floor effects across the adult age span); 4) comprehensiveness, covering four to six relevant subdomains; 5) state-of-the art assessment methods; and 6) absence of proprietary restrictions or costs, with limited initial equipment cost for users.

Subdomains were identified by surveying and interviewing research and clinical experts in the neurological and neuropsychological fields of cognition in adults and children [for more details about this process across all domains, see (Nowinski et al., 2013)]. Based on an initial survey of 102 cognition experts, 95% endorsed Executive Function among their top four domains to include in a battery of cognitive tests and followed by 93% for Episodic Memory, 55% for Language, 52% for Processing Speed (by 52%), and 50% for Attention. Many (57%) also listed a "Global Score" as desirable. Some cognitive subdomains (e.g., spatial cognition) were excluded due to their lower priority in the rankings and the need to limit the time for the entire battery. The selection of constructs within subdomains was based on reviews of the literature to identify those that have relevance for success in school and work, sensitivity to brain dysfunction as well as to growth in childhood and decline in aging, continuity across different age groups and well-established principles linking the construct with underlying neuroanatomical structure and function. Each accompanying paper provides the rationale for domain and construct selection.

An initial step in designing the NIHTB-CB was to collect existing instruments that tap each of the targeted constructs and to evaluate each against a list of "desirability" criteria. These criteria included: coverage of a broad age range (early childhood to late adulthood); brief administration time; availability in the public domain without proprietary restrictions or costs; availability of reliability and validity data; and representation of the domains that had been selected to test with the NIHTB-CB. After reviewing the assembled library of close to 200 instruments and batteries, however, we learned that the majority did not meet a combination of most of these criteria. As a result, the decision was made to create novel instruments and to validate them against existing "gold standard" measures for construct validity.

The need to create a "state-of-the art" instrument led to choosing a computer platform for administration of the NIHTB-CB rather than a paper-and-pencil format. Caution has been recommended in the use of computerized cognitive testing due to various sources of error, including the combination of hardware and software devices used, equipment timing issues, the operating system, and others [for a thorough review of these issues, see (Cernich,

Reeves, Sun, & Bleiberg, 2007)]. However, the advantages of greater control over stimulus presentation and response recording than is possible with human examiners, ease of data recording, and the capacity for automated scoring and simultaneous normative transformations were deemed to outweigh some of the negatives. In addition, computerized measures can be more conveniently adapted than standard paper-and-pencil measures for future modifications based on new scientific developments and needs, and on improvements in hardware and software technology.

A total of seven instruments was created for the NIHTB-CB: Flanker Inhibitory Control and Attention Test, Dimensional Change Card Sort Test, List Sorting Test, Pattern Comparison Processing Speed Test, Picture Sequence Memory Test, Picture Vocabulary Test, and Oral Reading Test. Table 1 contains brief descriptions of the NIHTB-CB tests, including test administration time, and scores derived from each. It should be noted that administration times are approximate and that the norming version has been adapted to remain within the originally intended 30-minute duration.

Since Executive Function (EF) was the most highly endorsed cognitive subdomain by the consulted experts and, because this subdomain itself contains several distinct sub-factors (Miyake et al., 2000), more than one EF test was considered justified. Thus, separate measures were designed to test inhibitory visual attention based on a flanker-type task (Fan, McCandliss, Sommer, Raz, & Posner, 2002) (the NIHTB Flanker Inhibitory Control and Attention Test) and set shifting based on a card sorting paradigm (Zelazo, 2006) (NIHTB Dimensional Change Card Sort Test.) Working memory, often considered another component of EF, was treated as a separate subdomain for the purposes of the NIHTB-CB because of its dual service in executive control and episodic memory [see (Cabeza, Dolcos, Graham, & Nyberg, 2002)]. The NIHTB List Sorting Working Memory Test was designed based on a paradigm emphasizing both *holding and manipulation* components of working memory and previously studied in English and Spanish-speaking older adults (Mungas, Reed, Marshall, & Gonzalez, 2000; Mungas, Reed, Crane, M.Z., & González, 2004).

Two language constructs were tested. The first, auditory comprehension of single word vocabulary, was based on a task requiring multiple-choice identification of items that match spoken single words (NIHTB Picture Vocabulary Test). The second, oral word reading, was based on oral letter and word pronunciation (NIHTB Oral Reading Recognition Test.) The language tests were administered according to a model of computer adaptive testing (CAT) and scored using item response theory (IRT), which allowed for a short administration time (Gershon, 2005).

Episodic memory was tested using the NIHTB Picture Sequence Memory Test. This test requires participants to observe a spatial sequence of pictures, placed one at a time on the computer screen, of individuals performing acts (e.g., planting, raking) with a related theme (e.g., gardening) but with no intrinsic temporal sequence. When the sequence is completed, the cards are “assembled” in the center of the screen and the participant must reproduce (or “imitate”) the demonstrated sequence. Finally, processing speed, a factor that has a broad influence on many types of cognitive tasks, was measured with the NIHTB Pattern

Comparison Processing Speed Test. This instrument measures speed of responses (same or different) to pairs of stimuli within a finite period of time.

Some tests were based on existing paradigms in the neuropsychological and cognitive neuroscience literature, including the NIHTB Flanker Attention Test (Fan et al., 2002) and the NIHTB Pattern Comparison Processing Speed Test, based on the work of Salthouse and colleagues (Salthouse, 1992). Another strategy employed in test design was to adapt measures created in the pediatric arena for use with adults, since few measures exist that cover the broad age spectrum for the NIHTB-CB. Thus, the Dimensional Change Card Sort (DCCS) Test (Zelazo, 2006), designed to assess set shifting in three-year olds, was adapted for use in adults. To assess episodic memory, “Elicited Imitation” of a sequence of events, also referred to as “Imitation-Based Assessment of Memory” (P.J. Bauer, 2007), a technique designed to assess learning and retention in infants (Lechuga, Marcos-Ruiz, & Bauer, 2001; Lukowski, Garcia, & Bauer, 2011), was adapted as the NIHTB Picture Sequence Memory Test for computer administration and for use with older children and adults.

Gold standard measures were identified from standardized published neuropsychological tests and matched to the extent possible to the constructs measured to the NIHTB-CB tests on the basis of consensus from the cognition domain team. For example, the Picture Sequence Memory Test assesses verbally mediated and visual episodic memory across learning trials. Thus, the gold standard selected for comparison consisted of the average score from two episodic memory tests with learning trials, one nonverbal and the other verbal, namely, the Brief Visuospatial Memory Test-Revised (Benedict, 1997) and the Rey Auditory Verbal Learning Test (RAVLT) (Rey, 1958), respectively. Table 2 lists the gold standard tests identified for each NIHTB-CB instrument along with the scores used in analyses. The rationale for the selection of each is described in greater detail in each of the accompanying papers.

Early on, it was decided to require an examiner to administer the tests to assure compliance, especially in the youngest and oldest subjects, and whenever the NIHTB-CB is used to assess individuals or groups who may require monitoring and/or assistance in understanding and following standard instructions. A test manual was constructed with instructions for administration. An examiner training module is available on the NIH Toolbox website(http://www.nihtoolbox.org/HowDoI/HowToAdministerTheToolbox/Training%20Manuals/NIH%20Toolbox%20Training%20Manual-English%209-25-12.pdf).

Test development was completed in stages. For each measure, a prototype instrument was designed and piloted and a Beta-1 version was subsequently created. The Beta-1 version was piloted in ten 3-year-olds and 11 young adults to identify any significant flaws and was then revised (Beta-2). The Beta-2 version went through three additional adjustments, each based on testing with similarly small groups, to adjust factors such as size and clarity of stimuli and number of trials to be administered in each subtest to assure brevity. The resulting Beta-3 version was then piloted on 123 individuals to determine if the measures were broadly sensitive to age. Based on that experience, further adjustments were made and Beta-4 was piloted on 146 individuals, who also were administered a number of well-validated measures of the same construct in an initial attempt to gauge construct validity.

The participants in all four Beta versions of the instruments came largely from convenience samples at each participating site and did not participate in the present validation study. Based on the results of the Beta-4 test, a final revision (Validation NIHTB-CB) was used in the study reported here.

Validation Study

Participants—Adult participants were recruited from 4 testing sites: 25 at NorthShore University Health System in Evanston, IL, 84 at the Northwestern Cognitive Neurology and Alzheimer's Disease Center (CNADC) in Chicago, IL, 92 at New Jersey's Kessler Foundation Research Center in West Orange, NJ, and 67 at the University of Washington in Seattle, WA. The younger participants in the sample (ages 20-60) were recruited with the use of flyers in the communities of each contributing institution. Although advertisements indicated the need for healthy individuals, participants were not screened prior to recruitment. Of the 109 participants 65 and older, the group most at risk for cognitive decline/dementia, 62 were recruited from among a pool of known cognitively healthy volunteers participating in the Clinical Core registry of the NIA-funded CNADC and the rest from the community via flyers. The lack of objective cognitive screening may have resulted in inclusion of individuals, particularly those from the community, with some cognitive impairment. However, the NIHTB-CB was intended to cover the full normal distribution of ability and a subsequent examination of floor and ceiling effects (see Results) did not suggest skewing of the older sample with respect to cognitive impairment.

It should be noted that there are gaps in the ages sampled for the validation study. Thus, results showing test scores by age in each accompanying paper are graphed for age bands that differ in the number of years encompassed by each. We had previously determined that a total sample size of 400-500 participants (children and adults) would be required for the validation study, and decided to focus on age bands where there was evidence for significant developmental differences from childhood through old age. Therefore, for the validation study, we oversampled on both ends of the age spectrum. For the adult sample, this resulted in oversampling the age range from 65 to 85 years. We did not recruit participants aged 36 to 39 and 61 to 64 years. In the Results, below, Figure 2 shows the distribution of the sample across different age bands. For the normative study, to be reported in future publications, the full age range was covered.

Self-report questionnaires were collected from participants to provide information on current health status, family income, and employment status.

A subset of 89 participants (33% of the sample) was retested 7 to 21 days later (Mean= 15.5 days, SD=4.8) to assess test-retest reliability and practice effects. Informed consent was obtained from all participants via a protocol approved by the institutional review boards at the respective institutions.

Equipment—The validation study was conducted with the use of a Windows 7 laptop, facing the examiner, connected to a 19” touch-screen external monitor with1440 × 900 resolution, facing the participant. It is planned to continue upgrading software to run on current versions of Windows and Internet Explorer into the future (including Windows 8+).

Extensive user directions have been provided to ensure that the computer is set up correctly. The following website links can be accessed for hardware requirements and technical details: (http://www.nihtoolbox.org/WhatAndWhy/Technology%20Support%20Documents/Intro%20to%20Computer%20and%20Special%20Equipment-revisions%208-5-13.pdf) (http://www.nihtoolbox.org/HowDoI/TechnicalManual/CognitionTechnicalManuals/Pages/default.aspx). The tests were designed to minimize the likelihood that the use of computers could introduce unwanted variance. For example, for the few tests where exact item level timing is important to assess a given trait, we utilized the hardwired keyboard itself as an entry device in order to not be subject to the same delays often encountered when utilizing differing types of mice or other connected peripherals. Variability in item display timing (which is often subject to differences in hardware quality or background software programs such as virus checkers) was removed as an element in test level timing—the software turns off the test timer during the period of time required to display a test item, and it is only turned back on when the display is complete. A new feature to check for browser compatibility will be introduced later this year.

The participant and examiner sat perpendicular to one another at a table, with the examiner facing the laptop (Figure 1). The examiner controlled the initiation of each test via the laptop. The examiner's laptop also served to display correct responses for the NIHTB Oral Reading Test and a space to record if the oral reading responses were correct or not. Examiners had been previously trained on the correct pronunciation of the reading items with the use of audio training CDs. The examiner also entered the responses for the NIHTB-CB List Sorting Test. Responses to all other NIHTB-CB subtests were entered by the examinee and recorded automatically by the computer.

Data and Analysis

Analyses used unadjusted scaled scores for both the NIHTB and “gold standard” tests. Scaled scores were created by first ranking the raw scores, next applying a normative transformation to the ranks to create a standard normal distribution, and finally rescaling the distribution to have a mean of 10 and a standard deviation of 3. These scaled scores were not age-adjusted.

In the remaining papers a variety of data analysis methods and statistics are used to report results. Pearson correlation coefficients between age and test performance were calculated to assess the ability of the NIHTB-CB tests to detect age-related cognitive decline during adulthood. Intraclass correlation coefficients (ICC) with 95% confidence intervals were calculated to evaluate test-retest reliability. Across measures, ICC less than 0.40 was considered poor test-retest reliability, 0.40 - 0.75 adequate, and 0.75 or greater was good to very good. Practice effects were evaluated using paired t-tests and effect sizes (mean change from time 1 to time 2 / SD of Time 1) were calculated as a standardized estimate of the mean change. This method for deriving Cohen's d statistic (Cohen, 1992) has been used in studies of test-retest reliability in standardized neuropsychological batteries (Dikmen, Heaton, Grant, & Temkin, 1999; Duff et al., 2005). Convergent validity was assessed with Pearson correlation coefficients between the NIHTB-CB measure and a well-established “gold standard” measure of the same construct. Convergent and discriminant validity results

are reported in the accompanying papers for each measure and not contained in the present paper. Across measures, correlations less than 0.3 were considered poor, 0.3 – 0.6 adequate, and 0.6 or greater were good to very good evidence of convergent validity, based on recommendations made by Andresen (Andresen, 2000). Evidence of discriminant validity consisted of lower correlations with selected "gold standard" measures of a *different* cognitive construct.

Analyses of variance (ANOVA) were performed to examine other demographic associations with performance, adjusted for age and other relevant covariates. Group comparisons were then performed using general linear models to examine other demographic associations with performance, adjusted for age, gender, and education, where appropriate.

Floor and ceiling effects represented by the percent of participants scoring at the minimum or maximum possible score are also reported.

Results

The main results for the validation study are divided among the remaining papers in this series in detail. In this section, we describe demographics of the sample, test-retest reliability, practice effects and floor and ceiling effects across the entire adult sample, for each instrument.

A total of 268 adults, ranging in age from 20 to 85 years, were recruited: 149 females and 119 males (Table 3). Race/ethnicity composition of the sample was 148 Caucasian [non-Hispanic White], 75 African American, 38 Hispanic, and seven multiracial (excluded from subsequent ethnicity comparisons). Mean age (*SD*) was 52.3 *(21.0)* years, and mean education (*SD*) was 13.4 *(2.9)* years. Education was categorized as less than high school graduate (25% of the sample), high school graduate or some college (37%), and Bachelor's degree or higher (38%).

The following indicates the percentage of individuals falling into each of five levels of family income: < $20,000 [18%], $20,000 to $39,999 [24%], $40,000 to $74,999 [29%], $75,000 to $99,999 [12%], and ≥ $100,000 [13%]; 4% "don't know" or refused. Current health status was self-reported by participants as Excellent (24% of participants), Very Good (41%), Good (26%), or Fair to Poor (9%). Current employment status categories were designated "Employed for wages or Self-employed" (44% of participants), "Retired" (31%), "Out of work" (12%), or "Other" (e.g., homemaker or student) (13%).

Figure 2 illustrates the number of individuals in the adult sample, at each age band that participated in the validation study and for whom data are reported in each of the accompanying papers. It should be noted that in the normative study, the age gaps are fully covered.

Test-retest reliability was comparable to published results obtained for the gold standard measures. Table 4 shows the ICC's for test-retest reliability for the NIHTB-CB tests. Values ranged from 0.73 to 0.90. Table 4 also shows effect sizes for the practice effects for each NIHTB-CB test and for the gold standard measures administered. Effect sizes ranged from

0.08 on the NIHTB-CB Vocabulary test to 0.42 on the Picture Sequence Memory Test. These values are quite comparable to the effect sizes obtained for practice effects in each of the gold standard measures. The language measures are administered via CAT methods and thus participants may be exposed to a different set of items from one administration to another, significantly reducing the practice effect.

Table 5 shows the mean raw scores for each NIHTB-CB instrument and gold standard measure and unadjusted scale scores for composite measures derived from one or more subtests. The medians and ranges also are provided to assist in evaluating the range of ability covered in this adult sample. A small ceiling effect was observed for NIHTB Picture Sequence Memory Test with 2.6% achieving the maximum score possible (see Table 5). All were in their 20's or 30's with the exception of one 59-year-old. Two people, both in their early 30's, obtained the maximum possible score for NIHTB Flanker Attention and List Sorting Tests. Two participants, both age 65 or older, scored the lowest possible score for the NIHTB Picture Sequence Memory Test.

Summary

The results reported in this paper for the NIHTB-CB validation study in adults from 20-65 years of age shows that the instruments have good test-retest reliability over a relatively short interval of time; that practice effects are consistent with those reported in the literature for similar instruments; and that there are minimal floor and ceiling effects for the age range studied. These properties are encouraging for its use in research studies, particularly those that will require measurement over multiple time points and longitudinal follow up from young to advanced adulthood.

Series Outline

Each accompanying paper in this series is dedicated to a different aspect of the validation project. One paper reports the results of the confirmatory factor analysis of the validation study in adults (Mungas, et al., this issue). Another describes the derivation of NIHTB-CB “Fluid”, “Crystallized” and “Total” composite scores for adults, their psychometric properties, including the effects on these scores of reported health status, associations with prior school difficulties and current employment status, and the demographic variables of sex, education and age (Heaton et al., this issue). The remaining papers each address a single subdomain and review in detail the rationale for its selection; the specific construct identified for testing within the subdomain; the evidence linking the domain/construct to brain functioning; the importance of that domain/construct for health, and everyday functioning; and the design of the instruments, including adaptations to enable testing across the age spectrum from three to 85 years (Tulsky et al., this issue; Carlozzi et al., this issue; Gershon et al., this issue; Zelazo et al., this issue; Dikmen et al., this issue.)

Future Directions

The validation study led to further refinements of the NIHTB-CB instruments, including shifting from a computer touch screen to a keyboard button press mode of response. Although initially attractive for its transparency to computer-naïve examinees, the touch screen introduced an undesirable variable for reaction time tests, namely the added amount

of time to move the entire hand to the screen. The final normative study used the button press version of the NIHTB-CB on a large national census-matched sample (N = 4,700), and a Spanish version was created (Beaumont et al., 2013) and normed on 750 individuals. Results from the normative studies are being evaluated and will appear in future publications.

A number of studies have already utilized the NIHTB-CB Validation Version. The feasibility and validity of the NIHTB-CB have been evaluated in a cohort of patients with Parkinson's disease with and without depression (PI: Mustafa M. Husain), in an acute neuro-rehabilitation setting (PI: Victor Mark), and in patients with traumatic brain injury, spinal cord injury, stroke (PI's: David Tulsky and Allen Heinemann), and HIV infection (PI: Robert Heaton). Preliminary results suggest that it is feasible to use the NIHTB-CB with all of these populations and that it is sensitive to brain dysfunction. The children's battery has also been used to collect phenotypic information on children ages 3-21 who are enrolled in the Pediatric Imaging, Neurocognition, and Genetics (PING) Study (PI: Terry Jernigan) (Akshoomoff et al., 2014) and is also being used in the National Children's Study “Vanguard Study” protocol for children ages 36 and 60 months and their parents.

The NIH has supported many multi-institute initiatives in the United States to facilitate communication among researchers and comparisons among different studies focusing on similar questions. The NIH Toolbox for Assessment of Neurological and Behavioral Function represents one of these accomplishments, and is designed to serve as a common currency for comparing and enriching broad types of research supported by the NIH. The NIH Toolbox CB is a research tool to facilitate this goal.

The use of common instruments that cover the lifespan allows for information to be collected efficiently on large numbers of research participants across the lifespan and to leverage the research investment by permitting comparisons among disparate studies. Detailed information on the NIHTB and how to obtain the cognitive, sensory, emotional and motor modules is available on: www.nihtoolbox.org.

Disclosures

This study is funded in whole or in part with Federal funds from the Blueprint for Neuroscience Research, National Institutes of Health, under Contract No. HHS-N-260-2006-00007-C.

Dr. Weintraub is funded by NIH grants # R01DC008552, P30AG013854, and the Ken and Ruth Davee Foundation and conducts clinical neuropsychological evaluations (35% effort) for which her academic-based practice clinic bills. She serves on the editorial board of *Dementia & Neuropsychologia* and advisory boards of the *Turkish Journal of Neurology* and *Alzheimer's and Dementia*.

Dr. Dikmen receives research grant funding from NIH R01 NS058302 and R01HD061400, NIDRR H133A080035, NIDRR H133G090022, and NIDRR, H133A980023, and DoD W81XWH-0802-0159

Dr. Heaton is funded by NIH grants # P30MH062512, HHSN271201000036C, R01MH92225, R01MH094160, and P50DA026306. He is on the editorial board of the Journal of the International Neuropsychological Society and The Clinical Neuropsychologist.

Dr. Tulsky is funded by NIH contracts H133B090024, H133N060022, H133G070138, B6237R, cooperative agreement U01AR057929, and grant, R01HD054659. He has received consultant fees from the Institute for Rehabilitation and Research, Frazier Rehabilitation Institute/Jewish Hospital, Craig Hospital, and Casa Colina Centers for Rehabilitation.

Dr. Zelazo serves on the editorial boards of Child Development, Development and Psychopathology, Frontiers in Human Neuroscience, Cognitive Development, Emotion, Developmental Cognitive Neuroscience, and Monographs of the Society for Research in Child Development. He is a Senior Fellow of the Mind and Life Institute and President of the Jean Piaget Society. He receives research funding from the Canadian Institute for Health Research (Grant # 201963), Institute of Education Science (R305A110528), National Institutes of Health (P20MH085987, R41 TR 000367), and the Character Lab.

Dr. Bauer serves as a member of the editorial board for the journal Journal of Experimental Child Psychology, as Associate Editor for the journals Developmental Review and Memory, and as Editor of the Monographs of the Society for Research in Child Development, for which she receives a stipend. She has received royalties from the publication of Memory in Infancy and Beyond (2007, Erlbaum), and Advances in Child Development and Behavior (Volumes 37 and 38, 2009 and 2010, respectively; Elsevier); and is funded by NIH grants HD067359, HD074724, and HD071845.

Dr. Carlozzi is funded by NIH grants R03NS065194, R01NR013658, R01NS077946, U01NS056975. She was previously funded by contracts H133B090024, B6237R, H133G070138, H133A070037-08A and a grant from the NJ Department of Health and Senior Services.

Dr. Slotkin reports no disclosures.

Dr. Wallner-Allen reports no disclosures.

Dr. Fox is funded by NIH grants R37HD017899, MH074454, U01MH080759, R01MH091363, P50MH078105, P01HD064653. He is Associate Editor of the *International Journal of Behavioral Development* and serves on the scientific board of the National Scientific Council for the Developing Child.

Ms. Beaumont served as a consultant for NorthShore University HealthSystem, FACIT.org, and Georgia Gastroenterology Group PC. She received funding for travel as an invited speaker at the North American Neuroendocrine Tumor Symposium.

Dr. Mungas is funded by research grants from the National Institute on Aging and a grant from the California Department of Public Health California Alzheimer's Disease Centers program.

Dr. Nowinski receives or has received research support from the National Institutes of Health (contracts HHSN265200423601C, HHSN260200600007C and HHSN267200700027C), the Department of Veteran's Affairs, the Analysis Group, Novartis and Teva Pharmaceuticals. She has also received honoraria for writing and updating an article for Medlink.

Dr. Manly is funded by NIH grants R01AG028786, R01AG037212; she had received funding previously from NIH grant R01AG016206 and a grant from the Alzheimer's Association (IIRG 05-14236). She is a consulting editor for the Journal of the International Neuropsychological Society. She serves on the Medical and Scientific Advisory Board of the Alzheimer's Association, and as a member of the Advisory Council on Alzheimer's Research, Care, and Services.

Dr. Havlik reports no disclosures.

Dr. Conway reports no disclosures.

Dr. Moy reports no disclosures.

Dr. Edwards reports no disclosures.

Dr. Gershon has received personal compensation for activities as a speaker and consultant with Sylvan Learning and the American Board of Podiatric Surgery. He is currently funded by several grants awarded by the NIH: N01-AG-6-0007, HHSN260200600007, 1U01DK082342-01, HD05469, 1RC2AG036498-01; NIDRR: H133B090024.

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

Akshoomoff N, Beaumont JL, Bauer PJ, Dikmen SS, Gershon RC, Mungas D, Heaton RK. Nih toolbox cognition battery (cb): composite scores of crystallized, fluid, and overall cognition. Monographs of the Society for Research in Child Development. 2013; 78(4):119–132.10.1111/mono.12038 [PubMed: 23952206]

Akshoomoff N, Newman E, Thompson WK, McCabe C, Bloss CS, Chang L, Jernigan TL. The NIH Toolbox Cognition Battery: Results from a large normative developmental sample (PING). Neuropsychology. 2014; 28(1):1–10.10.1037/neu0000001 [PubMed: 24219608]

Anderson V, Catroppa C, Morse S, Haritou F, Rosenfeld J. Functional plasticity or vulnerability after early brain injury? Pediatrics. 2005; 116(6):1374–1382.10.1542/peds.2004-1728 [PubMed: 16322161]

Andresen EM. Criteria for assessing the tools of disability outcomes research. Arch Phys Med Rehabil. 2000; 81(12 Suppl 2):S15–20. [PubMed: 11128900]

Barnes DE, Yaffe K. The projected effect of risk factor reduction on Alzheimer's disease prevalence. Lancet Neurol. 2011; 10(9):819–828.10.1016/S1474-4422(11)70072-2 [PubMed: 21775213]

Bauer PJ, Zelazo PD. Ix Nih toolbox cognition battery (cb): summary, conclusions, and implications for cognitive development. Monographs of the Society for Research in Child Development. 2013; 78(4):133–146.10.1111/mono.12039 [PubMed: 23952207]

Bauer, PJ. Remembering the times of our lives: memory in infancy and beyond. Mahwah, N.J: Lawrence Erlbaum Associates; 2007.

Beaumont JL, Havlik R, Cook KF, Hays RD, Wallner-Allen K, Korper SP, Gershon R. Norming plans for the NIH Toolbox. Neurology. 2013; 80(11 Suppl 3):S87–92.10.1212/WNL.0b013e3182872e70 [PubMed: 23479550]

Benedict, R. Brief Visuospatial Memory Test-Revised. Odessa, FL: Psychological Assessment Resources, Inc; 1997.

Brookmeyer R, Gray S, Kawas C. Projections of Alzheimer's disease in the United States and the public health impact of delaying disease onset. American Journal of Public Health. 1998; 88(9): 1337–1342. [PubMed: 9736873]

Cabeza R, Dolcos F, Graham R, Nyberg L. Similarities and differences in the neural correlates of episodic memory retrieval and working memory. Neuroimage. 2002; 16(2):317–330.10.1006/nimg.2002.1063 [PubMed: 12030819]

Carlozzi NE, Tulsky DS, Kail RV, Beaumont JL. Vi Nih toolbox cognition battery (cb): measuring processing speed. Monographs of the Society for Research in Child Development. 2013; 78(4):88–102.10.1111/mono.12036 [PubMed: 23952204]

Cernich A, Reeves D, Sun W, Bleiberg J. Automated Neuropsychological Assessment Metrics sports medicine battery. Archives of Clinical Neuropsychology. 2007; 22(Suppl 1):S101–114. [PubMed: 17118625]

Cohen J. A power primer. Psychological Bulletin. 1992; 112:155–159. [PubMed: 19565683]

Coldwell SE, Mennella JA, Duffy VB, Pelchat ML, Griffith JW, Smutzer G, Hoffman HJ. Gustation assessment using the NIH Toolbox. Neurology. 2013; 80(11 Suppl 3):S20–24. [PubMed: 23479539]

Cook KF, Dunn W, Griffith JW, Morrison MT, Tanquary J, Sabata D, Gershon RC. Pain assessment using the NIH Toolbox. Neurology. 2013; 80(11 Suppl 3):S49–53. [PubMed: 23479545]

Dalton P, Doty RL, Murphy C, Frank R, Hoffman HJ, Maute C, Slotkin J. Olfactory assessment using the NIH Toolbox. Neurology. 2013; 80(11 Suppl 3):S32–36. [PubMed: 23479541]

Dikmen SS, Heaton RK, Grant I, Temkin NR. Test-retest reliability and practice effects of expanded Halstead-Reitan Neuropsychological Test Battery. Journal of the International Neuropsychological Society. 1999; 5(4):346–356. [PubMed: 10349297]

Duff K, Beglinger LJ, Schoenberg MR, Patton DE, Mold J, Scott JG, Adams RL. Test-retest stability and practice effects of the RBANS in a community dwelling elderly sample. Journal of Clinical and Experimental Neuropsychology. 2005; 27:265–575. JCEN 1005 27 565-575.

Dunn W, Griffith JW, Morrison MT, Tanquary J, Sabata D, Victorson D, Gershon RC. Somatosensation assessment using the NIH Toolbox. Neurology. 2013; 80(11 Suppl 3):S41–44. [PubMed: 23479543]

Erlanger DM, Kutner KC, Barth JT, Barnes R. Neuropsychology of sports-related head injury: Dementia Pugilistica to Post Concussion Syndrome. Clin Neuropsychol. 1999; 13(2):193–209.10.1076/clin.13.2.193.1963 [PubMed: 10949160]

Fan J, McCandliss BD, Sommer T, Raz A, Posner MI. Testing the efficiency and independence of attentional networks. Journal of Cognitive Neuroscience. 2002; 14(3):340–347. [PubMed: 11970796]

Fox NA. Commentary on zelazo and bauer (editors), national institutes of health toolbox cognition battery (cb): validation for children between 3 and 15 years. Monographs of the Society for Research in Child Development. 2013; 78(4):150–155.10.1111/mono.12044 [PubMed: 23952209]

Gershon RC. Computer adaptive testing. J Appl Meas. 2005; 6(1):109–127. [PubMed: 15701948]

Gershon RC, Wagster MV, Hendrie HC, Fox NA, Cook KF, Nowinski CJ. Responding to the needs of the neurology research community: introduction to the NIH Toolbox for the Assessment of Neurological and Behavioral Function. Neurology. 2013; 80(11 Suppl 3):S2–6. [PubMed: 23479538]

Gershon RC, Slotkin J, Manly J, Blitz D, Beaumont J, Schnipke D, Weintraub S. NIH Toolbox Cognitive Battery (CB): Measuring Language (Vocabulary Comprehension and Reading Decoding). Monographs of the Society for Research in Child Development. 2013; 78(9):49–69. [PubMed: 23952202]

Hodes RJ, Insel TR, Landis SC, Research NIH Blueprint for Neuroscience. The NIH toolbox: setting a standard for biomedical research. Neurology. 2013; 80(11 Suppl 3):S1.10.1212/WNL.0b013e3182872e90 [PubMed: 23479536]

Lechuga MT, Marcos-Ruiz R, Bauer PJ. Episodic recall of specifics and generalisation coexist in 25-month-old children. Memory. 2001; 9(2):117–132. [PubMed: 11338937]

Lukowski AF, Garcia MT, Bauer PJ. Memory for events and locations obtained in the context of elicited imitation: evidence for differential retention in the second year of life. Infant Behav Dev. 2011; 34(1):55–62. doi: S0163-6383(10)00104-9 [pii]. 10.1016/j.infbeh.2010.09.006. [PubMed: 21047688]

McKee AC, Cantu RC, Nowinski CJ, Hedley-Whyte ET, Gavett BE, Budson AE, Stern RA. Chronic traumatic encephalopathy in athletes: progressive tauopathy after repetitive head injury. J Neuropathol Exp Neurol. 2009; 68(7):709–735.10.1097/NEN.0b013e3181a9d503 [PubMed: 19535999]

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(1):49–100.10.1006/cogp.1999.0734 [PubMed: 10945922]

Mungas D, Reed BR, Marshall SC, Gonzalez HM. Development of psychometrically matched English and Spanish language neuropsychological tests for older persons. Neuropsychology. 2000; 14(2): 209–223. [PubMed: 10791861]

Mungas D, Reed BR, Crane PK, Haan MZ, González G. Spanish and English Neuropsychological Assessment Scales (SENAS): further development and psychometric characteristics. Psychol Assessment. 2004; 16:347–359.

Mungas D, Widaman K, Zelazo PD, Tulsky D, Heaton R, Slotkin J, Gershon RC. NIH Toolbox Cognitive Battery (CB): Factor Structure for 3- to 15-year-olds. Monographs of the Society for Research in Child Development. 2013; 78:103–118. [PubMed: 23952205]

Nowinski CJ, Victorson D, Debb SM, Gershon RC. Input on NIH Toolbox inclusion criteria: surveying the end-user community. Neurology. 2013; 80(11 Suppl 3):S7–12. [PubMed: 23479548]

Reuben DB, Magasi S, McCreath HE, Bohannon RW, Wang YC, Bubela DJ, Gershon RC. Motor assessment using the NIH Toolbox. Neurology. 2013; 80(11 Suppl 3):S65–75. [PubMed: 23479547]

Rey, A. L'examen clinique en psychologie. Paris: Presses Universitaires de France; 1958.

Rine RM, Schubert MC, Whitney SL, Roberts D, Redfern MS, Musolino MC, Slotkin J. Vestibular function assessment using the NIH Toolbox. Neurology. 2013; 80(11 Suppl 3):S25–31.10.1212/WNL.0b013e3182872c6a [PubMed: 23479540]

Salsman JM, Butt Z, Pilkonis PA, Cyranowski JM, Zill N, Hendrie HC, Cella D. Emotion assessment using the NIH Toolbox. Neurology. 2013; 80(11 Supplement 3):S76–86. [PubMed: 23479549]

Salsman JM, Butt Z, Pilkonis PA, Cyranowski JM, Zill N, Hendrie HC, Cella D. Emotion assessment using the NIH Toolbox. 2013 in press.

Salthouse TA. Speed and age: multiple rates of age decline. Experimental Aging Research. 1976; 2(4): 349–359. [PubMed: 1017454]

Salthouse TA. Influence of processing speed on adult age differences in working memory. Acta Psychologica. 1992; 79(2):155–170. [PubMed: 1598844]

Shao H, Breitner JC, Whitmer RA, Wang J, Hayden K, Wengreen H, For the Cache County Investigators. Hormone therapy and Alzheimer disease dementia: New findings from the Cache County Study. Neurology. 2012; 79(18):1846–1852.10.1212/WNL.0b013e318271f823 [PubMed: 23100399]

Tulsky DS, Carlozzi NE, Chevalier N, Espy KA, Beaumont JL, Mungas D. NIH toolbox cognition battery (cb): measuring working memory. Monographs of the Society for Research in Child Development. 2013; 78(4):70–87.10.1111/mono.12035 [PubMed: 23952203]

Varma R, McKean-Cowdin R, Vitale S, Slotkin J, Hays RD. Vision assessment using the NIH Toolbox. Neurology. 2013; 80(11 Suppl 3):S37–40.10.1212/WNL.0b013e3182876e0a [PubMed: 23479542]

Victorson D, Manly J, Wallner-Allen K, Fox N, Purnell C, Hendrie H, Gershon R. Using the NIH Toolbox in special populations: considerations for assessment of pediatric, geriatric, culturally

diverse, non-English-speaking, and disabled individuals. Neurology. 2013; 80(11 Suppl 3):S13–19.10.1212/WNL.0b013e3182872e26 [PubMed: 23479537]

Weintraub S, Bauer PJ, Zelazo PD, Wallner-Allen K, Dikmen SS, Heaton RK, Gershon RC. I. NIH Toolbox Cognition Battery (CB): introduction and pediatric data. Monographs of the Society for Research in Child Development. 2013; 78(4):1–149.10.1111/mono.12031 [PubMed: 23952199]

Weintraub S, Dikmen SS, Heaton RK, Tulsky DS, Zelazo PD, Bauer PJ, Gershon RC. Cognition assessment using the NIH Toolbox. Neurology. 2013; 80(11 Suppl 3):S54–64.10.1212/WNL.0b013e3182872ded [PubMed: 23479546]

Wolf MS, Curtis LM, Wilson EA, Revelle W, Waite KR, Smith SG, Baker DW. Literacy, cognitive function, and health: results of the LitCog study. Journal of General Internal Medicine. 2012; 27(10):1300–1307.10.1007/s11606-012-2079-4 [PubMed: 22566171]

Zecker SG, Hoffman HJ, Frisina R, Dubno JR, Dhar S, Wallhagen M, Wilson RH. Audition assessment using the NIH Toolbox. Neurology. 2013; 80(11 Suppl 3):S45–48.10.1212/WNL.0b013e3182872dd2 [PubMed: 23479544]

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

Zelazo PD, Anderson JE, Richler J, Wallner-Allen K, Beaumont JL, Weintraub S. NIH Toolbox Cognitive Battery (CB): Measuring Executive Function and Attention. Monographs of the Society for Research in Child Development. 2013:16–33. [PubMed: 23952200]

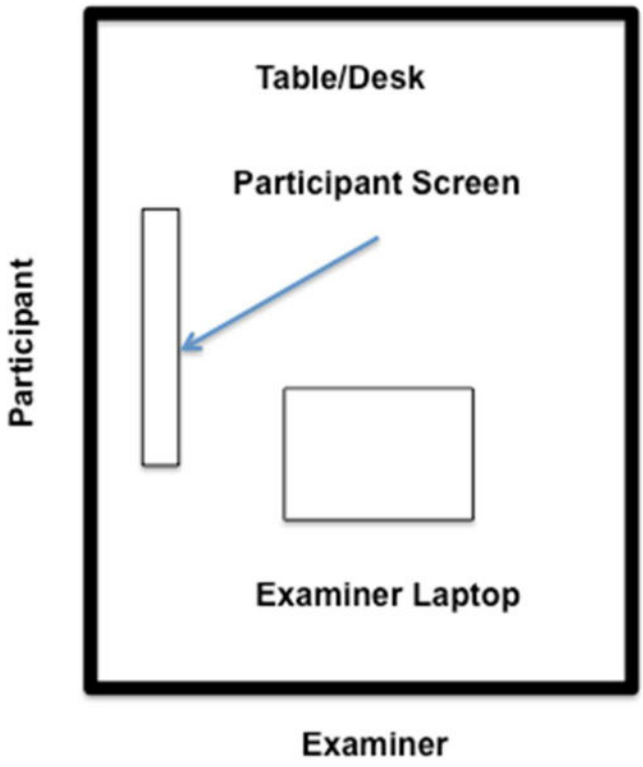

Figure 1. Testing Arrangement

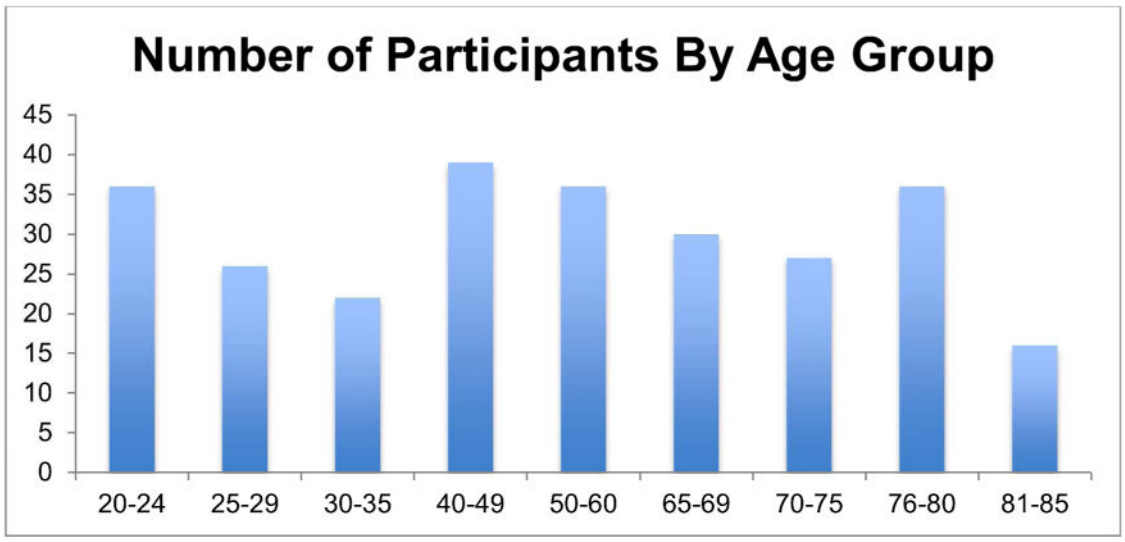

Figure 2. Distribution of Adult Participants In the Validation Study by Age Band Sampled

Table 1

NIH Toolbox Cognition Battery Tests

Test Name	TB Subdomain/Construct Administration time*	Stimulus and Task Description	Scores Computed For Validation Study
NIHTB Flanker Inhibitory Control And Attention Test	Executive Attention 6 minutes	Visual display of central arrow pointing left or right, flanked by arrows in the same or opposite direction as the central arrow. Task: Indicate direction of central stimulus (leftward or rightward pointing) when flankers are in the same (congruous) or in opposite (incongruous) directions from the central	Total Trials=40 Total Score=10 Score is based on an algorithm derived from both accuracy and reaction time if the former is >80%. If less than 80%, score is accuracy.
NIHTB Dimensional Change Card Sort Test	Executive Category Switching 7 minutes	Visual display of two different stimuli side-by-side, each in a different color. Test stimulus matching one of the two display stimuli in either color or shape appears at the bottom of the screen. Task: On some trials, shape is the sorting criterion, on others, color. There are 5 pre-switch trials (one category), 5 post-switch trials (after shift to second category) and 30 mixed category trials (shifting)	Total Trials=40 Total Score=10 Score is based on an algorithm derived from both accuracy and reaction time if the former is >80%. If less than 80%, score is accuracy.
NIHTB List Sorting Working Memory Test	Working Memory: information holding and manipulation 7 minutes	Stimuli from a single category (animals) or two categories (fruits and animals) are presented sequentially visually and aurally in series ranging from 2 to 8 items. Task: Orally repeat the sequence of items in order of size. For two-category items, order by size from one specified category first and then from the second.	Total Items correctly sequenced on the one- and two-category trials, of a possible 28.
NIHTB Pattern Comparison Processing Speed Test	Processing speed: number of items completed in a finite amount of time Maximum score=130 2 minutes	Pairs of stimuli appear side by side. Task: Indicate if the stimuli are the “same” or “not the same.”	Total number of correct responses within 90 seconds.
NIHTB Picture Sequence Memory Test	Episodic Memory for a sequence of pictured events 9 minutes (depending on length of series)	A series of pictures of people performing acts related to a single theme, but not in any intrinsic order, is presented one at a time on the computer screen. After the last item, the pictures are all “collected” in random array in the center of the screen. Task: The respondent must place the pictures in the same demonstrated sequence. There are three trials of learning with presentation of the sequence followed by replication of the sequence in each.	Total number of correct placements across three learning trials (total possible= 48)
NIHTB Oral Reading Test	Language: Written word pronunciation 6 minutes	Single printed words are presented in the center of the screen to be read aloud by the respondent. Examiner enters if response is correct or not. Items are presented via CAT method based on the participant's responses.	Theta score based on IRT
NIHTB Vocabulary Test	Language: Auditory word-visual picture matching 4 minutes	Four pictures are presented in a two-by-two array on the screen. A single recorded word is presented aurally and the participant must indicate which of the pictures matches the word.	Theta score based on IRT

*Administration times are approximate. The norming version has been shortened to remain within the desired 30 minutes originally planned. IRT=item response theory; NIHTB=NIH Toolbox

Table 2
Convergent and Discriminant Validity ("Gold Standard") Measures For Ages 20-85

NIHTB-CB SUBDOMAIN	NIH Toolbox Measure	Convergent Validity Measure	Discriminant Validity Measure
EXECUTIVE FUNCTION	NIHTB Flanker Inhibitory Control and Attention Test	WAIS-IV[1] Letter-Number Sequencing/Coding/Symbol Search * D-KEFS[2] Inhibition **	PPVT-4[3] **
	NIHTB Dimensional Change Card Sort Test	D-KEFS Inhibition **	PPVT-4 **
EPISODIC MEMORY	NIHTB Picture Sequence Memory Test	BVMT-R[4]/RAVLT[5] *	PPVT-4 **
WORKING MEMORY	NIHTB List Sorting Test	WAIS-IV Letter-Number Sequencing*/ PASAT[6] *	PPVT-4 **
PROCESSING SPEED	NIHTB Pattern Comparison Processing Speed Test	WAIS-IV Coding/Symbol Search *	PPVT-4 **
LANGUAGE	NIHTB Picture Vocabulary Test	PPVT-4 **	BVMT-R/RAVLT *
	NIHTB Oral Reading Recognition Test	WRAT-4[7] Reading Test **	BVMT-R/RAVLT *

* average of rescaled raw scores

** raw score rescaled

[1] Wechsler Adult Intelligence Scale – 4th Edition

[2] Delis-Kaplan Executive Function System

[3] Peabody Picture Vocabulary Test – 4th Edition

[4] Brief Visuospatial Memory Test – Revised

[5] Rey Auditory Verbal Learning Test

[6] Paced Auditory Serial Addition Test

[7] Wide Range Achievement Test – 4th Edition

Table 3

Adult validation sample demographics

Age Groups	Education	Gender		Race/Ethnicity		
		Male	Female	White	Black	Hispanic/ Other
20-60 Yrs. N=159	< High School	22	26	21	15	12
	High School Gradate	29	31	26	19	15
	College +	24	27	24	15	12
65-85 Yrs. N=109	< High School	9	11	9	10	1
	High School Gradate	12	27	26	11	2
	College +	23	27	42	5	3
TOTAL N=268		119	149	148	75	45

Table 4

Test re-test reliability for N=89 participants (unless otherwise indicated) on NIHTB-CB tests and practice effects on NIHTB-CB tests and gold standard measures. All mean scores are unadjusted scaled scores.

NIHTB-CB Test	ICC for test-retest reliability	Time 1 Mean (SD)	Time 2 Mean (SD)	t-test	p-value	Effect Size
Pattern Comparison Processing Speed	0.73	10.3 (2.9)	10.8 (2.8)	2.30	0.024	0.18
Vocabulary	0.80	9.8 (3.0)	10.1 (3.0)	1.20	0.232	0.08
Reading	0.90	9.9 (3.0)	10.0 (3.3)	0.67	0.505	0.03
Picture Sequence Memory Test	0.84	10.2 (2.9)	11.4 (3.0)	6.96	<0.001	0.42
List Sorting	0.77	10.0 (2.9)	10.8 (3.1)	3.95	<0.001	0.27
DCCS (n=78)	0.81	9.8 (2.9)	10.7 (3.0)	5.24	<0.001	0.33
Flanker (n=73)	0.83	10.0 (3.0)	10.8 (3.3)	3.90	<0.001	0.27

Gold Standard Tests For Construct Validity	Time 1 Mean (SD)	Time 2 Mean (SD)	t-test	p-value	Effect Size
WAIS-IV Coding (n=88)	10.0 (2.8)	11.0 (3.1)	6.14	<0.001	0.35
WAIS-IV Symbol Search	10.0 (2.9)	11.0 (2.9)	5.55	<0.001	0.35
Average of Coding and Symbol Search	10.0 (2.8)	11.1 (3.1)	6.65	<0.001	0.39
PPVT	9.8 (3.0)	10.0 (2.9)	1.30	0.196	0.06
WRAT-4 Reading	9.8 (3.0)	10.1 (3.0)	1.81	0.073	0.11
BVMT-R (n=88)	10.3 (3.2)	11.6 (3.3)	6.19	<0.001	0.41
RAVLT (n=87)	10.1 (3.2)	11.7 (3.8)	7.61	<0.001	0.51
Average BVMT and RAVLT	10.2 (3.2)	11.8 (3.6)	8.28	<0.001	0.49
WAIS-IV Letter-Number Seq (n=88)	9.9 (2.7)	10.4 (3.0)	2.59	0.011	0.18
PASAT (n=85)	9.9 (2.7)	10.7 (2.8)	4.49	<0.001	0.28
D-KEFS Inhibition (n=88)	10.0 (3.0)	10.8 (3.2)	5.40	<0.001	0.27

BVMT-R: Brief Visuospatial Memory Test-Revised; DCCS: Dimensional Change Card Sort Test;D-KEFS: Delis-Kaplan Executive Function System; PASAT: Paced Auditory Serial Addition Test; PPVT: Peabody Picture Vocabulary Test; RAVLT: Rey Auditory Verbal Learning Test; Seq: Sequencing; WAIS-IV: Wechsler Adult Intelligence Scale, 4th edition; WRAT-4: Wide Range Achievement Test, fourth edition

Table 5

Raw test scores for NIH Toolbox Cognition Battery Instruments and Gold Standard Measures Across Entire Adult Sample (20-85 years of age) and unadjusted Scaled scores for selected composite measures

	N	Mean (SD)	Median (Range)	Percent at floor (NIHTB-CB)	Percent at ceiling (NIHTB-CB)
SAMPLE VARIABLES					
Age (Years)	268	52.3 (21.0)	53 (20 – 85)		
Numeric Years of Education	266	13.4 (2.9)	12 (4 – 20)		
NIH TOOLBOX MEASURES					
Reading Theta	265	2.7 (1.6)	2.8 (-3.2 – 7.0)	0%	0%
Vocabulary Theta	263	2.1 (0.9)	2.1 (-1.4 – 4.4)	0%	0%
Picture Sequence Memory Test Number Correct for 3 trials	265	20.2 (11.0)	19 (0 – 48)	0.8%	2.6%
List Sorting, Total Correct, 1 and 2 category lists	264	18.3 (3.0)	18 (8 – 28)	0%	0.4%
Flanker Score, 0-10	237	8.2 (1.0)	8.4 (0.5 – 9.7)	0%	0%
DCCS Score, 0-10	244	7.8 (1.6)	8.1 (0.6 – 9.6)	0%	0%
Pattern Comparison, Total Correct in 90 seconds	264	41.7 (9.4)	41 (15 – 73)	0%	0%
Mean Reaction Time Flanker, DCCS Composite, Scaled Score	257	10.0 (3.0)	10.0 (1.6 – 18.4)	0%	0%
Toolox Processing Speed Composite Scaled Score[1]	267	10.0 (3.0)	10.0 (1.5 – 18.5)	0%	0%
GOLD STANDARD MEASURES					
WRAT-4 Reading RAW	264	57.0 (7.2)	59 (26 – 70)		
PPVT-4 Raw	263	204.0 (19.8)	208 (22 – 227)		
RAVLT Sum of Scores for 3 Trials	261	23.3 (6.1)	24 (8 – 39)		
BVMT-R Sum of Scores for 3 Trials	262	18.9 (7.6)	19 (0 – 36)		
Average of BVMT-R and RAVLT Scaled Score	264	10.0 (3.0)	10.0 (1.5 – 18.5)		
PASAT Total Score	256	30.8 (12.0)	31 (3 – 49)		
Wechsler (WISC-IV/ WAIS-IV) Letter-Number RAW	262	18.8 (3.6)	19 (5 – 27)		
Wechsler (WISC-IV/WAIS-IV) Coding RAW	263	60.5 (17.2)	60 (12 – 104)		
Wechsler (WISC IV / WAIS IV) Symbol Search RAW	264	28.8 (9.2)	29 (3 – 54)		
Average of Coding and Symbol Search Scaled Score	264	10.0 (3.0)	10.0 (1.5 – 18.5)		
D-KEFS Inhibition Total Score	257	59.5 (17.6)	55 (29 – 140)		
Wisconsin Card Sort Total Errors	260	20.1 (11.0)	16 (6 – 51)		

[1] average of Flanker & DCCS reaction time and Pattern Comparison scaled scores

BVMT-R: Brief Visuospatial Memory Test-Revised; DCCS: Dimensional Change Card Sort; DKEFS: Delis-Kaplan Executive Function System; PASAT: Paced Auditory Serial Addition Test; PPVT-4: Peabody Picture Vocabulary Test, fourth edition; RAVLT: Rey Auditory Verbal Learning Test; StDev: standard deviation of raw scores; WISC: Wechsler Intelligence Scale for Children; WAIS: Wechsler Adult Intelligence Scale; WRAT-4: Wide Range Achievement Test, fourth edition