National Institutes of Health (NIH) Toolbox® V3 Technical Manual

Edited by
Erica M. LaForte, Julie N. Hook, and Amy K. Giella
Northwestern University

National Institutes of Health (NIH) Toolbox® V3 Administration Manual

Edited by
Erica M. LaForte, Julie N. Hook, and Amy K. Giella
Northwestern University

Manual Version: NIHTBV3-TM1
Applicable to App Version: 3.13.1.2 and earlier

Acknowledgements

We would like to acknowledge the efforts of the faculty and staff who have contributed countless hours to the development and analyses included in this NIH Toolbox® V3 Technical Manual (listed alphabetically):

Grace Choi, MS
Elizabeth McManus Dworak, PhD
Richard Gershon, PhD
Catherine Han, PhD
Emily H. Ho, PhD
Aaron J. Kaat, PhD
James Olsen, PhD
Zutima Tuladhar, MS

We would like to acknowledge and thank the funding sources that have helped to develop, maintain, and support the NIH Toolbox measurement system and its related products.

NIH Blueprint for Neuroscience Research (HHS-N-260-2006-00007-C)

A Cognitive Test battery for Intellectual Disabilities (NICHD 1R01HD076189-01A1)

ARMADA: Advancing Reliable Measurement in Alzheimer's Disease and cognitive Aging (NIA/U2CAG057441)

ECHO PRO Research Resource: A Developmentally-based Measurement Science Framework for Assessing Environmental Exposure and Child Health (NIH 1U24OD023319-01)

MyCog: Rapid detection of cognitive impairment in everyday clinical settings (NINDS UG3NS105562)

National Children’s Study (HHSN2752012000071)

NIH Infant and Toddler Toolbox (75N94019D00005)

The Mobile Toolbox for Monitoring Cognitive Function (NIA/ 1U2CAG060426-01)

Contents

Chapter 1: Overview of the NIH Toolbox®

The NIH Toolbox V3 app provides access to the NIH Toolbox for Assessment of Neurological and Behavioral Function, a standard set of valid, reliable, and royalty-free tools for assessing cognitive, emotional, motor, and sensory function. Designed to benefit all researchers and clinicians interested in investigating behavioral and neurological function, the NIH Toolbox app is also relevant for students and clinicians across a broad spectrum of health research. It is particularly well-suited for measuring outcomes in longitudinal epidemiological studies along with prevention or intervention trials. NIH Toolbox tests have been normed and validated across the lifespan in participants ages 3 through 85+ years.

NIH Toolbox V3 App Structure and Organization

The NIH Toolbox spans four *domains* or broad areas of health and function: Cognition, Emotion, Motor, and Sensation. A *test* is a set of *items* administered in an order determined by the app. You cannot change the content of the items or the order of item presentation within a test. Each test can be administered by itself, as part of a preset battery, or as part of a custom battery.

The NIH Toolbox app includes several *preset batteries* or sets of tests intended to be administered together in a specific order. You can change the order of the tests or add individual tests or other batteries to the assessment, before or after the preset battery. The app produces scores for each test in the battery as well as composites for the Cognition batteries and summary scores for Emotion batteries. You can also create *custom batteries* by adding tests in a certain order for use in future assessments.

Refer to the *NIH Toolbox V3 Administration Manual* (Hook & Giella 2023) for instructions on using the app.

Background and History

In 2004, 15 Institutes, Centers, and Offices at the National Institutes of Health (NIH) that support neuroscience research formed a coalition called the *Blueprint for Neuroscience Research*. The goal of this NIH Blueprint, as it is referred to, was to develop new tests, resources, and training opportunities to accelerate the pace of discovery in neuroscience research. Because the research community had long sought the development of standard tests to measure cognitive and emotional health, in 2006 the NIH Blueprint for Neuroscience Research awarded a contract to develop an innovative approach to meet this need. The outcome was the establishment of the NIH Toolbox.

The NIH Toolbox was intended to include the following domains: Cognition, Emotion, Motor, and Sensation. Initial literature and database reviews and a "Request for Information" of NIHfunded researchers identified the subdomains for inclusion in the NIH Toolbox along with the criteria affecting test selection, creation, and norming.

NIH Toolbox validation studies were conducted across the entire age range, typically including 450 to 500 participants, and statistically compared NIH Toolbox measures to existing "gold standard" measures, whenever available. For tests using item response theory (IRT) approaches to scoring, calibration samples generally included several thousand participants, ensuring robust models. In total, data were collected from more than 16,000 participants as part of field-testing, calibration, and validation activities.

The original NIH Toolbox project conducted a large national norming study in both English and Spanish languages. This study is described in detail across two journal volumes in 2013: volume 78, article 4 of the *Monographs of the Society for Research in Child Development* and in volume 80, supplement 3 of *Neurology*. A sample of 4,859 participants ages 3–85, representative of the U.S. population, was administered all the NIH Toolbox measures at sites around the United States. The normative data obtained from this study were used to generate norms for the web-based version of NIH Toolbox (V1), which was released to the general public in 2012. In 2015, the iPad app version (also known as "NIH Toolbox V2") was released.

NIH Toolbox test development has focused on the continuity of assessments throughout the lifespan. An expert team of early childhood and older adult assessment consultants was engaged to provide guidelines for administration, to offer input on test development, and to review all NIH Toolbox tests as they relate to the needs of young children and older adult participants.

NIH Toolbox tests utilize several advanced approaches in item development, test construction, and scoring. Two of these are item response theory (IRT) and computer adaptive testing (CAT), which are used in a subset of tests. IRT allows tests to be brief, yet precise and valid. Using IRT methodology, sets of items are calibrated along a continuum that covers the full range of the construct to be measured. This calibrated set of items enables the creation of measures that employ CAT, a specialized type of computer-based testing in which administration of items is based on individuals' responses, with minimal burden on participants and precise evaluation at the individual level.

The use of the NIH Toolbox grew significantly after the public release of the web-based version in 2012. In 2015, the NIH Toolbox was released in the form of an iPad app; in 2017, adjustments were made to the scoring algorithms to account for mode-of-administration differences between the desktop and iPad versions, resulting in a change in version from V1 to V2. V2 included the same tests and normative information as the original NIH Toolbox, but it also included the following user-friendly features and functions:

- relied on portable, easy-to-use technology;
- presented a complete stand-alone application that did not require internet access during test administration;
- minimized the use of custom hardware;

- included enhanced normative scores for individual tests and composites;
- allowed results to be stored locally on the iPad and exported to the iCloud, to a configurable web address, via email, and transferred directly to a computer via cable;
- provided basic reporting on an individual participant level;
- offered email support; and
- additional "experimental" tests were added in the following years.

Revision Goals

In 2018, the research team at Northwestern University began planning for the revision of the NIH Toolbox. The revision goals for the NIH Toolbox V3 were informed by many sources of information, gathered across multiple settings since the publication of the original NIH Toolbox. The user interface and workflows were redesigned and streamlined to provide a more consistent appearance and user experience across the app. Test names were shortened, and new features were added to the examiner interface (also referred to as the app shell) to facilitate an improved examiner experience. Changes to the "look and feel" of the NIH Toolbox tests included the implementation of universal fonts, background color, and overall appearance. Instructions were streamlined and are now read aloud by the audio when appropriate.

The NIH Toolbox V3 was reviewed with accessibility and usability in mind for both the examiner and the participant. Experts in design and usability were engaged to ensure the app's examiner interface and participant screens were as compliant with current accessibility standards as possible. Specific attention was given to the Section 508 amendment (updates 2008 and 2018) of the Rehabilitation Act of 1973 (updates 1998, 2008, 2028). Accessibility levels in design include A, AA, or AAA. These standards can be achieved through design elements (e.g., consistency, contrast, sizing, proportions).

The team also planned an extensive revision of the Cognition domain. Population-level cognitive abilities are susceptible to demographic and cultural shifts within the population over time, and altering the mode of administration (i.e., from the original web version to an app interface) also raises potential shifts to expected performance. According to the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014), it is the responsibility of test publishers to "renorm the test with sufficient frequency to permit continued accurate and appropriate score interpretations" (p. 104). The research team drafted the following goals to guide the revision of the Cognition Domain:

1) Gather new normative data for the Cognition domain so that the normative scores will reflect the current U.S. population for gender, race/ethnicity, and education level;
2) Implement continuous norming procedures to produce normative scores in 3-month intervals for children ages 3 through 18 and 1-year intervals for adults, to minimize "binning" associated with norms that are produced for larger age groups;
3) Implement an IRT-based Change Sensitive Score (CSS) metric;

4) Add new measures to improve the construct coverage of the domain; and
5) Revise some existing measures to address specific user concerns (e.g., reduce floor and/or ceiling effects, improve test reliability for certain ages).

The NIH Toolbox V3 app contains all NIH Toolbox V2 tests (except for some Emotion tests), as well as some new Cognition tests. The Cognition domain contains most of the changes, including streamlined workflows and new normative data. Some minor changes were made in the Emotion, Motor, and Sensation domains, though except for the Standing Balance test, no new normative data were collected for tests in these domains. For more information about Emotion, Motor, and Sensation, please refer to the *NIH Toolbox V3 Administration Manual* (Hook & Giella, 2023) and future editions of this *NIH Toolbox V3 Technical Manual*.

NIH Toolbox Cognition Domain

Cognition refers to the mental processes involved in gaining knowledge and comprehension, such as thinking, knowing, remembering, judging, and problem solving. These higher-level functions of the brain encompass language, imagination, perception, and the planning and execution of complex behaviors. The Cognition Battery includes tests measuring executive function, episodic and working memory, processing speed, language, and fluid reasoning. For each test in the Cognition domain, Table 1.1 includes the test acronym, age range, and abilities measured.

Table 1.1
NIH Toolbox V3 Cognition Tests

Test Name	*Acronym*	*Ages*	*Abilities Measured*
Dimensional Change Card Sort	DCCS	4+	Executive function
Face Name Associative Memory Exam	FNAME	18+	Memory
Face Name Associative Memory Exam Delay	FNAME Delay	18+	Memory; delayed memory
Flanker Inhibitory Control and Attention	Flanker	4+	Executive function, attention
List Sorting Working Memory	LSWM	5+	Working memory
Oral Reading Recognition	ORR	7+	Language
Oral Symbol Digit	OSD	5+	Processing speed
Pattern Comparison Processing Speed	PC	5+	Processing speed
Picture Sequence Memory	PSM	3+	Episodic memory
Picture Vocabulary	PV	3+	Language
Rey Auditory Verbal Learning	RAVLT	5+	Episodic memory
Rey Auditory Verbal Learning Delay	RAVLT Delay	5+	Episodic memory, delayed memory
Speeded Matching	SM	3–6	Processing speed
Visual Reasoning	VR	4+	Executive function

Abilities Measured in the Cognition Domain

Attention refers to the allocation of an individual's limited capacities to deal with an abundance of environmental stimulation and is the foundation for all other types of mental processes.

There are several different forms of attention, including sustained, selective, and divided. Sustained attention is closely linked to the level of wakefulness or the maintenance of an alert state. Selective attention serves to direct sensory and thought processes to a particular stimulus or sector of the visual field so action can be taken. Divided attention is the ability to attend to more than one stimulus, spatial sector, or modality simultaneously, and overlaps with executive function. In the NIH Toolbox, aspects of the Flanker Inhibitory Control and Attention (Flanker) test measure attention.

Episodic memory refers to cognitive processes involved in the acquisition (learning), storage, and retrieval of new information. It involves conscious recollection of information learned within a context and the spontaneous recollection of the information. Episodic memory can be verbal, as in remembering a conversation or a list of grocery items, or nonverbal, as in imagining a place one visited or a picture one saw a week before. In the NIH Toolbox, the Picture Sequence Memory (PSM) test is a measure of episodic memory that relies on the examinee's ability to recall a sequence of pictures from a visual and auditory presentation. The NIH Toolbox also includes the supplemental Rey Auditory Verbal Leaning (RAVLT) and Rey Auditory Verbal Leaning Delay (RAVLT Delay) tests, which can be used as an alternative measure of episodic memory for examinees who have visual or motor impairments that prevent them from completing the PSM test. The Face Name Associative Memory Exam (FNAME) and Face Name Associative Memory Exam Delay (FNAME Delay) also measure memory.

Executive function is defined as the capacity to plan, organize, and monitor the execution of behaviors that are strategically directed in a goal-oriented manner. The NIH Toolbox focuses on the following components of executive function: (1) set shifting, or the capacity for switching among multiple aspects of a strategy or task, as measured by the Dimensional Change Card Sort (DCCS) test; (2) inhibition of automatic response tendencies that may interfere with achieving a goal, as measured by the Flanker test; and (3) nonverbal and visual reasoning, as measured by the Visual Reasoning (VR) test.

Language refers to a set of mental processes that serve to translate thought into symbols (words, gestures) that can be shared among individuals for purposes of communication. The NIH Toolbox focuses on two aspects of language: vocabulary and reading. The Picture Vocabulary (PV) test is a measure of receptive vocabulary knowledge that is fundamental to learning and that also has a very high association with overall intelligence (or what has been called the "g- factor"). The Oral Reading Recognition (ORR) test measures oral reading skill that reflects level and quality of prior educational experiences. This measure provides a robust indication of verbal intelligence that is undisturbed by many medical conditions affecting the brain.

Processing speed is defined as either the amount of time it takes to process a set amount of information, or, conversely, the amount of information that can be processed within a certain amount of time. It is a measure of mental efficiency. Processing speed is central for many cognitive functions and domains and is sensitive to change and/or disease. In the NIH Toolbox, the Pattern Comparison Processing Speed (PC) test measures processing speed. The NIH Toolbox includes the Speeded Matching (SM) test and the Oral Symbol Digit (OSD) test as supplemental measures. SM is for children 3 to 6 years old and can replace PC when testing young children. OSD can be used as an alternative measure of processing speed for individuals with, for example, motor impairments.

Working memory refers to the ability to maintain and manipulate information in active attention (Schneider & McGrew, 2018). It requires an individual 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 in the same short-term buffer. This concept updates the traditional construct of "short-term memory," which refers to a passive storage buffer, to include the notion of an active computational workspace. Working memory overlaps with constructs of attention and executive function. In the NIH Toolbox, the List Sorting Working Memory (LSWM) test is a measure of working memory.

Cognition Batteries

The NIH Toolbox app contains two preset batteries in the Cognition domain. You may select a battery to administer, or you may create a custom battery from among the tests provided.

Cognition Battery: The Cognition Battery is recommended for participants ages 7+, but it can also be administered to participants as young as age 6 and will produce “extended” normative data. This battery includes tests that assess the following constructs: attention, episodic memory, executive function, language, processing speed, and working memory. Tests in this battery include Picture Vocabulary, Flanker Inhibitory Control and Attention, List Sorting Working Memory, Dimensional Change Card Sort, Pattern Comparison, Picture Sequence Memory, and Oral Reading Recognition. Administering all tests in the Cognition Battery will produce a Fluid Cognition score, a Crystallized Cognition score, and a Total Cognition Composite score.

Early Childhood Cognition Battery: The Early Childhood Cognition Battery was designed to be developmentally appropriate for participants ages 4 to 6, but it can also be administered to participants ages 7 and 8 and will produce “extended” normative data. This battery includes tests that assess the following constructs: attention, episodic memory, executive function, processing speed, and language. Tests in this battery include Picture Vocabulary, Flanker Inhibitory Control and Attention, Dimensional Change Card Sort, Picture Sequence Memory, and Speeded Matching.

Uses of the NIH Toolbox Cognition Domain

Since its release in 2012, the NIH Toolbox has been used in hundreds of studies. In fact, its use has expanded from its original conceptualization of a research tool to also include use in clinical, school, and clinical/pharmaceutical trials. For example, a recent scoping review (Fox et al., 2022) found that there were 281 articles that used the NIH Toolbox in clinical samples. Most of these studies, 80%, used the Cognition domain tests, compared to tests in Emotion (17%), Motor (10%), or Sensation (5%). The most represented clinical areas included neurologic disorders which comprised 40% of the studies, followed by psychological disorders and cancer, 14% and 11% of the studies respectively. Notably, one area of specialized use of these cognitive tools is with the assessment of individuals with intellectual disability. The brevity of these tests makes them well suited for use in this population; there have been several articles published in this area (e.g., Hessl et al., 2016) and the *NIH Toolbox® V3 Cognition Battery Supplemental Administration Manual for Intellectual and Developmental Disabilities* (McKenzie et al., 2023) is also available in the User Resources at nihtoolbox.org.

Organization of this Technical Manual

The following chapters of this manual include important technical information about the NIH Toolbox V3 Cognition domain. Chapter 2 contains information about the development and revision of the NIH Toolbox V3 Cognition tests, as well as detailed information about the scaling, administration, and scoring of the revised tests in the Cognition domain. Chapter 3 describes the NIH Toolbox V3 norming study. Chapter 4 contains information about the reliability and validity of the tests and composites in the NIH Toolbox. Future updates to this manual will include details about the Emotion, Motor, and Sensation domains.

Chapter 2: Measure Development and Revision

This chapter describes the goals and objectives for the revision of tests in the NIH Toolbox® V3 update. For measures that are new to the NIH Toolbox in V3, the development of the measures is described.

NIH Toolbox App Design Updates

The entire NIH Toolbox user interface was redesigned and streamlined in V3 to provide a more consistent appearance and user experience across the app. The redesign of the app focused on the user interface, or the means through which the user interacts with the device; the user experience, or the user's overall experience when using the app; accessibility; consistency in look and feel; and flexibility for various use cases.

In response to user feedback, several updates were made to the NIH Toolbox to make it more intuitive and contemporary. First, a uniform font and color design was applied to the app. Button sizes and shapes were standardized across the app. Title screens for each test were updated to include color-coding by domain and icons were added to indicate if additional materials are needed for that test. A consistent and intuitive iconography palette was introduced.

In addition to the updated look and feel of the app, several V3 updates improve the overall user experience for both the administrator and the participant. Many of the instructions that were read aloud by the administrator in V2 are audio-recorded in V3. The use of audio-recorded instructions not only simplifies test administration, but it also improves the standardization of the test administration and reduces the impact of construct-irrelevant variance in participant performance resulting from variations in administrator reading speed, pronunciation, and oral reading ability. For most tests, the instructions still appear on the iPad screen and the administrator has the option to mute the audio recording and read the instructions orally.
The "Touch and hold" gesture from the NIH Toolbox V2 app was replaced with a more intuitive and user-friendly, iPad-native "Slide to continue" gesture. The "Slide to continue" icon appears on screens with important instructional information and/or on transition screens prior to practice items and live items. A "Back" button was added to many instruction screens to give the examiner flexibility to go back to a previous screen to replay instructions if the participant was distracted or did not fully understand the first time.

User feedback about the "administration gesture" that was used in V2 to pause, stop, or skip a test was that the gesture was difficult to perform intentionally but was prone to being implemented accidentally by the participant during test administration. In V3, the administration gesture was replaced with a new gesture that requires users to swipe to the left with three fingers anywhere on the screen. This updated administration gesture is native to the iPad and easier for users to perform.

Two of the speeded Cognition tests in the V2 app utilized a “Home Base” feature, which required participants to place their finger over a printed dot on a piece of paper between items. While the intention of the Home Base was to standardize the test administration across participants, NIH Toolbox user feedback suggested that it was used inconsistently. In response to this feedback, Home Base was removed from the V3 app; participants are now allowed to assume whatever hand position is most comfortable during speeded test administration. Relatedly, the Pattern Comparison test no longer requires the participant to use their dominant hand to respond to items.

Cognition Test Updates

Among the four NIH Toolbox domains, the Cognition domain underwent the most extensive revision in the V3 update. Several existing tests were revised by updating test items, workflows and test logic, and, in some cases, recalibrating the item pools. New scoring models were developed for several existing tests. Five new tests were added to the Cognition domain, and the entire battery was renormed to reflect the current demographic characteristics of the United States. Detailed descriptions of the development processes for new tests, and revisions to existing tests, is included in the sections below. Where relevant, changes or updates to the test administration procedures scoring models are described.

Dimensional Change Card Sort

The Dimensional Change Card Sort (DCCS) test is a measure of executive functioning, specifically cognitive flexibility. Two target pictures are presented that vary on two dimensions (i.e., shape and color). Participants are asked to match a series of bivalent test pictures (i.e., yellow balls and blue trucks) to the target pictures, first according to one dimension (e.g., color) and then, after a number of trials, according to the other dimension (i.e., shape). “Switch” trials are also employed, in which the participant must change the dimension being matched. For example, after four straight trials matching on shape, the participant may be asked to match on color on the next trial and then go back to shape, thus requiring the cognitive flexibility to quickly choose the response option that is consistent with that dimension. This test is recommended for ages 4 to 85+.

Test Development and Revisions

Feedback from NIH Toolbox users informed several challenges in the V2 DCCS workflow. There were different versions of the test for ages 3 to 7, ages 8 to 11, and ages 12+, and these versions differed in several ways. All examinees took an identical set of five “Shape” Practice items and five “Color” Practice items in V2, but examinees ages 8 and up were routed directly from Color practice to live mixed trials, whereas examinees ages 3 to 7 took an additional set of five Color items and five Shape items, with feedback for correct and incorrect responses, before proceeding to the live mixed trials items. The accuracy scores for these additional items, called “Preswitch” and “Postswitch” for color and shape, respectively, were included in the total test score for participants ages 3 to 7. Participants ages 8 and older who were not administered the Preswitch and Postswitch items were given 10 points “free” credit in their accuracy scores. Item presentation also differed by age in the V2 test. Children ages 11 and under heard the word “shape” or “color” along with the visual presentation of the word on the screen for each item, whereas the version of the test for ages 12 and older did not contain audio prompts. Users also noted that the test felt quite long for young children who were administered Shape practice, Color practice, Preswitch, and Postswitch prior to the live items.

In addition to the workflow challenges, data from several studies employing the V2 test suggested that the test was relatively easy for most participants ages 12 and older; in general, these participants completed the test items very quickly and with 95% or greater accuracy. This resulted in a lack of score differentiation among examinees in adolescence and young adulthood.

Prior to the V3 norming study, two small pilot studies were conducted to investigate the impact of several proposed changes to the DCCS test. The results of the pilot studies supported the following changes to the test:

1) The lower age range of the test was changed from 3 years to 4 years. Data collected during the pilot phase showed that many 3-year-olds did not respond with higher-than-chance accuracy on the DCCS test, even when they were provided with scaffolding for learning the test task.
2) Data from very young children and children with disabilities suggested that the transition in test task from Shape practice to Color practice in the V2 DCCS test mimicked the transition from Preswitch to Postswitch, resulting in unnecessary redundancy in these phases of the test. In the V3 test, the Shape practice was eliminated so that all examinees begin with Color practice, and then advance to Preswitch (Color) and Postswitch (Shape). This change shortens the test for young children and removes the set-switch redundancy that existed between Shape and Color practice in the V2 test.
3) Instructions, practice items, live items, and scoring were standardized so that all participants ages 4 to 85+ receive the same version of the test.
4) Some changes were made to the V3 item stimulus presentation format to “speed up” the items and minimize examinee habituation and anticipation, to make the test more difficult for older children and adults. First, the fixation star was removed before the presentation of the “SHAPE” or “COLOR” prompt for all items. Second, the delay between the appearance of the response buttons and the appearance of the prompt word “SHAPE” or “COLOR” on the screen was reduced from 1.0 seconds to 0.8 seconds.

Administration and Scoring

After the app presents a short audio-recorded introduction and demonstration of the test task, the participant completes five Color practice items. The app provides feedback for both correct and incorrect responses. Next, the participant receives 5 additional Color (Preswitch) items, followed by five Shape (Postswitch) items. Finally, the participant proceeds to the set of 30 live items. For the live items, the app captures the participant’s reaction time to touch down, item response, and item score.

The RCS is computed as the number of correct responses (out of 30) divided by the sum of the response times across all items (in seconds). The RCS represents the number of correct responses per second and considers accuracy and response time on all 30 items. The RCS is converted to a Change Sensitive Score (CSS) for reporting purposes. Age-adjusted normative scores are available for participants ages 4 to 85+ years, and age-and-education–adjusted normative scores are available for individuals ages 22 to 85+ years.

The DCCS test contributes to the Fluid Composite, Total Cognition Composite, and Early Childhood Composite.

Face Name Associative Memory Exam

The Face Name Associative Memory Exam (FNAME) test is an assessment of delayed associative memory—a fluid cognitive ability. FNAME is administered in two parts. The first part is a learning phase (i.e., the "FNAME" test in the NIHTB V3 app) that shows 12 faces, one at a time, paired with a name. The second part is a testing phase (i.e., the "FNAME Delay" test in the app) that has three sections. In the first section of FNAME Delay, Face Seen Before (FSB), the examinee is presented with pictures of three faces and asked which face was presented before. In the second section, First Name Letter (FNL), the examinee is shown a person's face and must tap the first letter of the person's name on an on-screen keyboard. In the final section, Face-Name Matching (FNM), the examinee is shown a picture of a face and must select (from among three options) the name that goes with the face.

Test Development

FNAME and FNAME Delay are new supplemental measures in the NIH Toolbox V3 app. An experimental version was originally tested in the V2 app, specifically for older adults aged 60 years and older. The V3 test content is largely the same as the V2 experimental measure, though normative scores are now available, and it covers a wider age range. Both versions are based on an original laboratory paradigm developed to assess associative memory and impairments related to dementia and cognitive decline. As part of the V3 norming, scoring procedures were updated and revised. Supplemental process scores are available for each of the three subcomponents: Face Seen Before (FSB), First Name Letter (FNL), and Face-Name Matching (FNM). The sum of these three subcomponents is converted to a Change Sensitive Score (CSS) using a linear transformation to the underlying item response theory (IRT) logit scale. Given the complexity of the test (three subcomponents and repeated faces across the components), the IRT model used for scaling was a bifactor model with one general factor representing associative memory, one specific factor representing the entire subcomponent of FSB, and twelve additional specific factors modeling the shared residual dependency for the same face across the FNL and FNM subcomponents. The IRT score is derived using the Lord and Wingersky algorithm (Cai, 2014; Huang & Cai, 2021) for expected *a posteriori* scoring from sum scores of bifactor models.

Administration and Scoring

The FNAME test is administered in two portions. The first portion (i.e., “FNAME”) is an initial learning phase where a face-name pair is presented and the examinee is required to indicate whether it will be easy or hard to remember the face-name pair. Regardless of the selection (and even if a participant does not make a selection), after a fixed interval exposure window the test moves on to the next face-name pair. After all pairs have been presented, the learning phase ends. The second phase of FNAME is the testing portion (i.e., “FNAME Delay”), which has three subcomponents. In the first FNAME Delay subcomponent, the examinee is presented with three faces and required to indicate which one they saw before (visual recognition). In the second subcomponent, examinees are shown the faces they saw during the learning component, one at a time, and are asked to indicate the first letter of the individual’s name (associative recall). In the final subcomponent, each face is presented with three names and the examinee must match the face and name pairing (associative recognition).

In operational administration, the delay between learning and testing (i.e., FNAME and FNAME Delay) is intended to be approximately 15 minutes, but any delay from 5 to 25 minutes is within the acceptable administration window. Outside of this window, the app will generate a notification to the examiner to say that insufficient or excessive time has elapsed since the learning component. The examiner may choose to override this warning; in this case, an administration irregularity is noted in the export and score reports.

The learning portion of the test does not produce any scores. The delayed portion of the test produces three process scores—one for each subcomponent of the test: Face Seen Before (FSB), First Name Letter (FNL), and Face-Name Matching (FNM). Supplemental process scores, ranging from 0 to 12, are available for each of these three subcomponents. The raw score is the sum of these three process scores. This raw score is converted to a theta score and associated standard error, which is based on the sum score expected *a posteriori* IRT score using the Lord and Wingersky algorithm (Cai, 2014; Huang & Cai, 2021) for bifactor models. The theta score is then converted to a CSS for reporting purposes. Age-adjusted normative scores are available for examinees from 18 to 85+ years old, and age-and-education-adjusted normative scores are available for individuals from 22 to 85+ years old. The FNAME and FNAME Delay tests are supplementary measures and do not contribute to any composites within the NIH Toolbox V3 Cognition Battery.

Flanker Inhibitory Control and Attention

The Flanker Inhibitory Control and Attention (Flanker) test measures a participant’s attention and inhibitory control, which are considered executive functions. Executive functioning (EF) is a set of neurocognitive skills required for goal-directed problem-solving; it taps into aspects of working memory, inhibitory control, and cognitive flexibility (Carlson, Zelazo, & Faja, 2013). The Flanker test is considered a measure of *fluid ability*, or the capacity for new learning and information processing in novel situations. The test requires the participant to focus on the middle stimulus while inhibiting attention to flanking fish stimuli. This test is recommended for ages 4 to 85+.

Test Development and Revisions

Feedback from NIH Toolbox users suggested several drawbacks of the V2 Flanker test workflow. First, there were different versions of the test for ages 3 to 7, ages 8 to 11, and ages 12+, and these versions varied with respect to the inclusion (or not) of a practice item set, the particular live items presented (fish, arrows, or both), and the scoring model employed (participants ages 8+ received scores based on accuracy and speed, while some participants ages 3 to 7 received scores based only on accuracy). These differences in the age-specific versions of the test resulted in scores that were not always comparable across the versions of the test. Additionally, data from several studies employing the V2 test suggested that the test was relatively easy for most participants ages 12 and older; most participants completed the test items very quickly and with 95% accuracy. This resulted in a lack of score differentiation among examinees in adolescence and young adulthood.

Prior to the V3 norming study, two small pilot studies were conducted to investigate the impact of several proposed changes to the Flanker test to address these shortcomings. The results of the pilot studies supported the following changes to the Flanker test:

1) The lower age range of the test was changed from 3 years to 4 years. Data collected during the pilot phase showed that many 3-year-olds were not capable of better-than-chance responding on the Flanker test, even when they were provided with scaffolding for learning the task.
2) Instructions, practice items, live items, and scoring were standardized so that all participants ages 4 to 85+ receive the same version of the test.
3) Although the fish item trials were only administered to children ages 3 to 7 in V2, pilot research revealed that most adults worked slower and with lower accuracy on the fish trials than on the arrow trials. Because one revision goal was to decrease speed and accuracy to better differentiate among adolescent and adult participants, the “arrow” item set was dropped from the test. The V3 test includes only one set of live “fish with arrows” item trials.
4) The number of live items was increased from 20 to 30. The first 20 “fish with arrows” items remain identical to the V2 implementation. Ten additional fish with arrows items, including eight congruent items and two incongruent items, were added to the end of the test for a new total of 10 incongruent and 20 congruent items. With this change, the percentage of incongruent items on the test was reduced from 40% to 33%. Because participant reaction time tends to get faster on the later incongruent items, the inclusion of fewer incongruent items in subsequent trials can mitigate this increase in vigilance.
5) Several changes were made to the V3 item stimulus presentation format to “speed up” the items and make the test more difficult for older children and adults. First, the delay between the appearance of item and the fixation star was made variable, alternating pseudo-randomly between 0.3 second and 1.0 seconds. Second, the auditory “middle” prompt was removed from between the fixation star and the item stimulus presentation for all items. Third, the time between the presentation of the fixation star and the item presentation was reduced from 1.0 seconds to 0.8 seconds for all items.
6) Finally, a 10-second response time limit was implemented for all live items. If the participant does not respond by tapping a response within 10 seconds of the

appearance of the fish on the screen, the item will be counted as incorrect, and the test will advance automatically to the next item.

Administration and Scoring

After the app presents a short audio-recorded introduction and demonstration of the test task, the participant completes five practice items. The app provides feedback for both correct and incorrect responses. The participant then proceeds to the set of 30 live test items. For the live items, the app captures the participant's reaction time to touch down, item response, and item score. The test takes approximately 3 minutes to administer.

To account for the interaction of accuracy and speed in examinee performance, a new rate correct score (RCS) model is employed in the V3 Flanker test. The RCS is computed as the number of correct responses (out of 30) divided by the sum of the response times across all items (in seconds). The RCS represents the number of correct responses per second and considers accuracy and response time on all 30 items. The RCS is converted to a Change Sensitive Score (CSS) for reporting purposes. Age-adjusted normative scores are available for participants ages 4 to 85+ years, and age-and education adjusted normative scores are available for individuals ages 22 to 85+ years.

The Flanker test contributes to the Fluid Composite and the Total Cognition Composite and the Early Childhood Composite.

List Sorting Working Memory

The List Sorting Working Memory (LSWM) test measures a participant's working memory, or the ability to hold information in a short-term buffer and then manipulate the information. Pictures and written names of different foods and animals are displayed with accompanying audio-recorded and visual prompts (e.g., "elephant," "banana"), and the participant is asked to say the items back in size order from smallest to largest, first within a single dimension (either animals or foods, called "1-List") and then on two dimensions (foods, then animals, called "2-List"). Each dimension contains two unique trials at each sequence length, with sequence lengths ranging from two pictures to seven pictures. Each item is scored either correct or incorrect by the test administrator, using an external keyboard. This test is recommended for ages 5 to 85+.

Test Development and Revisions

There were different versions of the V2 test for ages 3 to 6 and for ages 7+. The version for young children included four practice items with 10 questions each in 1-List (e.g., "Which animal is the smaller animal?"), and four practice items with 12 questions each in 2-List (e.g., "Which one is a food?"). The V2 test for ages 7+ had two practice items each for 1-List and 2-List, and each practice item had three possible trials if the response was incorrect.

User feedback from the V2 LSWM test and pre-norming piloting informed some changes to simplify the test and improve the participant experience. These include:

1) The minimum age for the test was increased from 3 years to 5 years. Despite the extensive scaffolding in the V2 ages 3 to 6 version of the test, children as young as 3 or 4 years often had trouble understanding the test task, resulting in chance-level responding to even the easiest 1-List items.
2) In V3 LSWM, all participants ages 5+ see the same practice items—two practice items for 1-List and two practice items for 2-List, each with two possible trials if the response was incorrect.
3) In the V2 test, participants who responded correctly to the first item at any sequence length were automatically assigned credit for the second item of that sequence length. Calibration of the item sequences following the V3 norming study revealed that the two items at each sequence length often varied in difficulty, suggesting that administration of the second item at each sequence for all participants could increase the measurement precision of the test. In response to this finding, in V3 LSWM participants are administered both items at each sequence length, regardless of their performance on the first item in the length.
4) In the V2 test, each participant began testing with the two-picture 1-List sequence and tested forward through successively longer 1-List sequences until both items at a sequence length were failed or until the participant completed the 7-picture sequence(s); then, the participant advanced to the two-picture 2-List sequence and continued testing until both items at a sequence length were failed or until the participant completed the 7-picture sequences. The V3 norming data suggested that the 2- and 3-picture sequences in the 1-List section were much too easy for most examinees younger than age 9, and the 2- and 3-picture sequences in the 2-List section were much too easy for most examinees younger than age 15; those items were not providing useful information for older children and adults. To maximize information and minimize testing time, the V3 workflow utilizes differential starting points and basal/ceiling rules in both the 1-List and 2-List sections of the test. Although this test is not a pure CAT, the differential starting points together with the basal and ceiling rules mimic the rules of an adaptive test administration, thereby minimizing testing time and maximizing the efficiency of test administration.
5) A chime sound was added after the presentation of each item in V3 to alert the participant that the item presentation has ended. This update was in response to V2 user feedback suggesting that participants often began responding orally to item prompts before the stimulus presentation was complete or waited too long after the item ended because it was unclear when the last stimulus was presented.
6) Finally, V2 user feedback suggested that two pictures were not familiar to participants across all cultures. To address this feedback, the blueberry and pumpkin in the V2.1 test were replaced with a bean and a cake, respectively, on the V3 test.

Administration and Scoring

In each of the two sections (1-List and 2-List), the app first presents a short audio-recorded introduction to the test task, followed by two practice items of 2-picture and 3-picture sequence lengths, respectively. The practice items provide feedback for both correct and incorrect responses. After the practice items, the participant is routed to an age-appropriate starting sequence length and testing proceeds via basal and ceiling rules. Within each section, a *basal* is obtained when the examinee responds correctly to both sequences in the shortest set of administered sequences, OR when the examinee has taken both 2-picture sequences. A *ceiling* is obtained when the examinee responds incorrectly to both sequences in the longest set of administered sequences, OR when the examinee has taken both 7-picture sequences. For each sequence administered, the app captures the item response as entered by the test administrator (1 for correct responses; 0 for incorrect responses). Response time is *not* captured for LSWM because the test responses are scored, then entered by the administrator. The test takes approximately 6 minutes to administer.

The total raw score is the number of administered 1-List sequences that were scored correct, plus the number of administered 2-List sequences that were scored correct, plus one point for every unadministered item sequence below the basal in each part, for a maximum of 24 points. The raw score is converted to a Change Sensitive Score (CSS) for reporting purposes. Age-adjusted normative scores are available for participants ages 5 to 90 years, and age-and education adjusted normative scores are available for individuals ages 22 to 90 years.

The List Sorting Working Memory test contributes to the Fluid Composite and the Total Cognition Composite.

Oral Reading Recognition

The Oral Reading Recognition (ORR) test is an assessment of reading decoding skills and crystallized abilities. It is administered in a computer adaptive test (CAT) format. Participants ages 7 and up see a word on the iPad screen and read it aloud, and the examiner scores each item as correct or incorrect on an external (Bluetooth) keyboard.

Test Development and Revisions

The V3 ORR test maintains the item content from the V2 version; no words were changed. The V2 items were recalibrated during the V3 norming study.

Although the ORR test is CAT-administered in the NIHTB V3 app, a fixed-form calibration design was utilized during the V3 norming study to ensure that adequate item-level N-counts were obtained to produce stable item difficulty estimates. The goal of the calibration study design was to maximize statistical information by presenting items to each examinee that were well-targeted to the examinee's ability level. Historical item difficulty estimates were used to assemble forms and estimate average form difficulty.

Data from all 263 items and 3,516 participants were administered ORR in the V3 norming study were concurrently calibrated using the Rasch model. Item response N-counts ranged from 24 to 3,153; four items in the pool had low response N-counts (<50) due to the limited number of young examinees in the study sample. As a check on the item stability estimates from the V3 calibration, the V3 examinees were scored separately on both the V3 and V2 ORR item difficulty parameter estimates; the scores correlated at .998, providing evidence for the stability of the

V3 item difficulty estimates. Item statistics were reviewed to assess model fit. All items showed adequate fit to the Rasch model; the final ORR item pool retains all 263 calibrated items.

Administration and Scoring

The ORR test begins with an instructional screen to introduce the examinee to the test task. In V3, this instructional text was modified slightly, and audio recorded to standardize the presentation. Following the instructional screen, test items are presented on the screen, one item at a time, and read aloud by the examinee. The examiner scores the responses by typing 1 (for correct) or 0 (for incorrect) on a Bluetooth keyboard.

The ORR CAT algorithm administers a minimum of 20 items and a maximum of 35 items with a target probability of correct response of 0.675. Once the CAT has reached the minimum test length, the test will end if the standard error of the ability estimate falls below 0.44 logits. The CAT algorithm produces an interval-scaled score on the logit metric, which is converted to a Change Sensitive Score (CSS) for reporting purposes. Age-adjusted normative scores are available for examinees ages 7 to 85+ years, and age-and-education-adjusted normative scores are available for individuals ages 22 to 85+ years.

The Oral Reading Recognition test contributes to the Crystallized Composite and the Total Cognition Composite.

Oral Symbol Digit

The Oral Symbol Digit (OSD) test is a measure of speed of information processing. In OSD, a coding key with nine abstract symbols is presented – each paired with a number between 1 and 9. Participants are asked to orally indicate which numbers go with symbols that are presented in a long string on the paper ‘Oral Symbol Digit Examination Sheet’. The participant is given 120 seconds to call out as many numbers that go with the corresponding symbols as they can, without skipping any. The examiner marks the corresponding items as correct or incorrect on the iPad. This test is administered to ages 5-85+. The OSD test can be administered as an accommodation in place of the Pattern Comparison Processing Speed (PC) test for those with significant motor limitations in the upper extremities.

Test Development and Revisions

There were some OSD updates from V2 to V3. Response time was not reported for live items. A Back button and Next button were added to where appropriate to facilitate examiners moving between instruction screens. The on-screen instructions, which are to be read aloud by the examiner, were split across separate rows so the examiner can more easily identify what they should demonstrate and what they should say aloud.

Administration and Scoring

Examinees are presented with nine practice items to teach them how to respond to this test. The examiner should provide corrective feedback on these items. Regardless of performance in practice, all examinees move forward to live test items. Examinees have 120 seconds to match as many symbols with numerical digits as possible. Items continue to be presented until the cumulative item time reaches 120 seconds (up to a maximum of 143 items). No new items are presented after that, but an examinee can respond to the last item on the screen. If the participant responds to all 143 items prior to the maximum 120 second time limit, the test ends. Including practice items, OSD takes approximately three minutes to complete.

The raw score for OSD is calculated as the sum of the items for which the examinee correctly responded to within 120 seconds. Items that were not responded to are considered incorrect. Item-level scores are reported (1 = correct, 0 = incorrect or non-response). Items were calibrated to the Rasch model, where all items are equally related to the underlying processing speed trait, but each item ranged in difficulty due to positioning in the test (and how thus how likely a person was to reach that item). Examinees were scored using maximum likelihood scoring to get an interval-scaled score on the logit metric, which is converted to the change sensitive score for interpretability.

Age-adjusted normative scores are available for examinees aged 5 to 85+ years old, and Age-and-Education-adjusted normative scores are available for individuals aged 22 to 85+ years old. The OSD test is a supplementary measure and does not contribute to any composites within the NIH Toolbox V3 Cognition Battery.

Pattern Comparison Processing Speed

Pattern Comparison (PC) measures speed of information processing. Participants see a pair of simple pictures on the iPad screen and must quickly discern whether the pictures are the same or different. For each pair of pictures, the examinee taps the *Yes* button if the pictures are the same, or the *No* button if the pictures are different. This test is recommended for ages 5 to 85+.

Test Development and Revisions

The V3 Pattern Comparison test retains the item-pair content from the original NIH Toolbox test, although the pictures have been redrawn for higher-resolution display. For item pairs that are not the same, the differences are based on color discrimination, adding or taking something away from one image, or a one-versus-many discrimination. The pictures were intentionally designed to be simple to measure pure processing speed that is not confounded by other cognitive abilities.

There were two versions of the Pattern Comparison test in the V2 app: one version for 3- to 6-year-old examinees, and one version for examinees age 7+. Although both versions contained the same 130 live items, the two versions differed in the instructional language, practice items and feedback, and response button format. Pilot research conducted prior to the V3 norming study suggested that examinees ages 3 to 4 may not consistently perform better than chance responding on this test; therefore, the lower age for the V3 test was increased to 5 years.

With this change, the separate version of the test for the youngest examinees was deemed unnecessary. The V3 pre-norming pilot study confirmed that examinees as young as 5 years were able to understand the shortened test instructions and were able to respond to test items with high accuracy using the *Yes* and *No* buttons rather than the “happy face” and “sad face” buttons that were used in the V2 3- to 6-year-old version of the test. Additionally, the pilot study results suggested that a reduction in the number of practice items from seven to three shortened the administration time but did not negatively impact examinee performance on the live items. As such, the V3 version of the test contains only three practice items, with one practice item to represent each way that a pair of items could be different (color, or completeness, or one-versus-many). Three practice items that were removed from V2 were added to the live item set in V3, bringing the total number of live items up to 133.

For older children and young adults, the V2 Pattern Comparison test showed a relative ceiling effect; most participants in this age range completed all the items in fewer than 90 seconds with high accuracy. For this reason, the V3 test was shortened to 80 seconds. In the analysis of the V3 norming data, all responses made after the 80-second mark were treated as "not reached" (i.e., incorrect). Item difficulties were then estimated with the Rasch dichotomous model.

Administration and Scoring

Examinees are first presented with three practice items to teach the test task and to familiarize the youngest examinees, especially non-readers, with the *Yes* and *No* response buttons. Corrective feedback is provided on the practice items. Regardless of performance on the practice items, all examinees proceed to the live test items. Examinees have 80 seconds of actual presentation time (excluding transitions between items) to respond to as many items as possible (up to a maximum of 133). Prior to each item, there is a 300-millisecond delay, which is not included in the overall 80-second test time. Item timing is calculated from when an item presentation is complete until the examinee touches down on either the "Yes" or "No" response button. Items continue to be presented until the cumulative item time reaches 80 seconds. No new items are presented after that, but an examinee can respond to the item presented on the screen at that time. If all 133 items are presented prior to the 80-second time limit, the test ends. The test takes less than three minutes to administer.

The total raw score is the number of items answered correctly during the 80 seconds of presentation time. The raw score is converted to a Change Sensitive Score (CSS) for reporting purposes. Age-adjusted normative scores are available for participants ages 5 to 85+ years, and age-and education adjusted normative scores are available for individuals ages 22 to 85+ years. The Pattern Comparison test contributes to the Fluid Composite and the Total Cognition Composite.

Picture Vocabulary

The Picture Vocabulary test is an assessment of receptive vocabulary—a crystallized ability—that is administered via a computer adaptive test (CAT) format. Participants ages 3 and older listen to a word presented by audio recording, then select one of the four pictures on the screen that means the same as the word.

Test Development and Revisions

The V3 Picture Vocabulary test maintains the original item content; with the exception of one item that was removed after analysis of the calibration data (see below), no recorded words or pictures were changed. The items were recalibrated during the V3 norming study.

Although Picture Vocabulary is CAT-administered in the NIHTB V3 app, a fixed-form calibration design was utilized during the V3 norming study to ensure that adequate item-level *N*-counts were obtained to produce stable item difficulty estimates. The goal of the calibration study design was to maximize statistical information by presenting items to each examinee that were well-targeted to the examinee's ability level. Historical item difficulty estimates were used to assemble forms and estimate average form difficulty.

Data from all 373 items and 3,605 study participants were concurrently calibrated using the Rasch model. Item response *N*-counts ranged from 5 to 2,633; 65 of the easiest items in the pool had low response *N*-counts (<50) due to the limited number of young examinees in the study sample. To improve the stability of these item difficulty estimates, calibrations were re-run with historical item-level data from the original NIHTB norming study included. Items with response-level *N*-counts greater than 50 were anchored to their V3 difficulty estimates, and all other items were allowed to calibrate freely. Item-level *N*s for the 65 items increased (range *N* = 31–643); item-level difficulty estimates for the unanchored items and person-level ability estimates for the V3 study participants did not change significantly with the addition of the historical data.

Item statistics were reviewed to assess model fit. Two items were flagged for misfit; of these two items, one was retained because the misfit appeared to be the result of a few unexpected outlying correct responses from low-ability examinees. The other misfitting item was determined to have a second plausible correct response and was subsequently removed from the pool. Difficulty estimates obtained from the calibration of the remaining 372 items were included in the V3 app.

Several changes were made to the Picture Vocabulary test workflow in V3. First, the instructions and practice items were standardized across all ages; there are no longer separate versions of the test for participants of different ages. In the V2 test, the recorded item prompts for ages 3 through 6 were preceded by, "Touch the picture of …." This was removed from the V3 recordings so that all examinees hear only the stimulus word.

Administration and Scoring

After hearing a short audio-recorded introduction, the participant is administered two practice items. The practice items provide feedback for both correct and incorrect responses. The participant then proceeds to the live test items.

In operational administration, the CAT administers a minimum of 20 items and a maximum of 35 items with a target probability of correct response of 0.675. Once the CAT has reached the minimum test length, the test will end if the standard error of the ability estimate falls below 0.44 logits.

The CAT algorithm produces an interval-scaled score on the logit metric, which is converted to a Change Sensitive Score (CSS) for reporting purposes. Age-adjusted normative scores are available for examinees ages 3 to 85+ years, and age-and-education-adjusted normative scores are available for individuals ages 22 to 85+ years.

The Picture Vocabulary test contributes to the Crystallized Composite and the Total Cognition Composite and to the Early Childhood Composite.

Picture Sequence Memory

Picture Sequence Memory (PSM) measures episodic memory, defined as the acquisition (i.e., learning), storage, and retrieval of new information. Participants are presented with a series of pictures on the iPad screen, each accompanied by an audio recorded phrase. Then, the pictures are shuffled on the screen and the examinee must drag them, one at a time, back into the presented order. Points are awarded for placing the pictures in correct adjacent-pairs order; in other words, examinees receive 1 point for each pair of pictures that are placed in the correct adjacent order, regardless of whether the two pictures are in their respective correct boxes on the screen. The pictures in each sequence share a common theme (either “going to the fair” or “playing in the park,” depending on the test form), but do not follow a logical chronological ordering; in other words, the examinee must learn and remember the picture order that is presented in the test and will not be advantaged by having attended a fair or played in a park. This test is recommended for ages 3 to 85+.

Test Development and Revisions

The V3 PSM test underwent several changes from V2. In V2, there were three forms of the test (Park, Fair, and Farm), each with separate versions for ages 3 to 4, ages 5 to 6, age 7, and ages 8+. The V3 test has been streamlined to only two forms (A and B), each with an item presentation that includes either a Park sequence, or a Fair sequence, or both (depending on the examinee’s performance). The V3 test automatically selects items based on the examinee’s age, eliminating the need for separate versions of the test for examinees of different ages. The instructional screens, which teach the examinee how to drag pictures into boxes on the screen and were administered by the examiner in V2, were replaced by an interactive, animated training module in V3.

To ensure that Forms A and B of the test are parallel in difficulty, the test was normed using a multi-form, counterbalanced design. In this design, a subset of the examinees was administered sequences of varying lengths and themes. The items (sequences) were then calibrated using the Rasch partial-credit model to place the 3-, 6-, 9-, and 15-picture, Park- and Fair-themed sequences onto a common underlying scale.

In the V2 version of the PSM test, examinees ages 7 and older were administered two sequences of the same theme, but of varying lengths. To better assess the acquisition aspect of episodic memory in V3, the administration of the PSM test was changed so that each examinee is administered one pair of *identical* sequences[1], allowing the examinee to learn from the second administration of the sequence. Also, the V3 test employs a multi-stage routing design to ensure each examinee is administered the test sequence (3, 6, 9, or 15 pictures) best targeted to their specific ability.

Administration and Scoring

PSM is administered in three phases: Instructions, Practice, and Test Items. In the Instruction phase, all examinees are administered a brief animated tutorial that teaches the task of dragging pictures into boxes on the screen.

The Practice phase contains age-appropriate practice sequences as follows:

- Examinees ages 3 to 5 are administered a 2-picture "Ice Cream" sequence and a 3-picture "Birthday Cake" sequence. Examinees are allowed up to two trials of each practice sequence. If an examinee does not place all the pictures in the correct boxes in at least one trial of either sequence, the test ends and no live items are administered. All other examinees proceed to the live items.
- Examinees ages 6+ are administered a 4-picture "Camping" sequence. Examinees are allowed up to two trials; if an examinee does not place all the pictures in the correct boxes in at least one trial, the test ends and no live items are administered. All other examinees proceed to the live items.

Examinees who continue into the Test Items phase will be routed to a first sequence according to their age; 3-year-old examinees begin with the Fair 3-picture sequence in both Form A and Form B of the test. Examinees ages 4 to 5 begin with either the Park (Form A) or Fair (Form B) 6-picture sequence. Examinees ages 6 to 7 *or* age 65+ begin with either the Park (Form A) or Fair (Form B) 9-picture sequence. Examinees ages 8 to 64 begin with the Park (Form A) or Fair (Form B) 15-picture sequence. Examinees receive one point for each pair of pictures that is correctly placed in order, such that each sequence is worth L–1 points, where L is the number of pictures in the sequence.

After the first test picture sequence is administered, routing proceeds as follows:

- Three-year-old examinees are all administered a second trial of the 3-picture Fair sequence. Then, examinees are further routed based on their total adjacent pairs scores from the two trials of the 3-picture Fair sequence. Examinees with a score of 0–2 points are routed to the 3-picture Park sequence, and examinees with a total score of 3–4 points are routed to the 6-picture Park sequence. The total test score is the sum of adjacent-pairs scores from all four administered sequences.
- Examinees ages 4 to 5 are routed based on the score from the first (6-picture) sequence administered. Examinees who score 0 points are administered two trials of the 3-picture sequence of the opposite theme (Park or Fair); examinees who score 1–3 points are administered the same 6-picture sequence a second time; and examinees who score 4–5 points are administered two trials of the 9-picture sequence of the opposite theme

[1] Three-year old examinees are administered two pairs of identical sequences.

(Park or Fair). The total test score is the sum of adjacent-pairs scores from the last two sequences administered.

- Examinees ages 6 to 7 *and* ages 65+ are routed based on their score from the first (9-picture) sequence administered. Examinees who score 0–1 point are administered two trials of the 6-picture sequence of the opposite theme (Park or Fair); examinees who score 2–5 points are administered the same 9-picture sequence a second time; and examinees who score 6–8 points are administered two trials of the 15-picture sequence of the opposite theme (Park or Fair). The total test score is the sum of adjacent-pairs scores from the last two sequences administered.
- Examinees ages 8 to 64 years are routed based on their score from the first (15-picture) sequence administered. Examinees who score 0–2 points are administered two trials of the 9-picture sequence of the opposite theme (Park or Fair); examinees who score 3–14 points are administered the same 15-picture sequence a second time. The total test score is the sum of adjacent-pairs scores from the last two sequences administered.

The Picture Sequence Memory test contributes to the Fluid Composite, Total Cognition Composite, and Early Childhood Composite.

Rey Auditory Verbal Learning

The Rey Auditory Verbal Learning Test (RAVLT; Rey, 1941) was one of the first standardized methods to evaluate verbal learning and memory using a list of 15 unrelated words presented to the participant over repeat trial. The test requires the examinee to recall as many words as possible in any order after each presentation of the list. During this immediate recall phase of the test (i.e., “RAVLT” in the V3 app), the participant is learning the list. After a 5 to 25-minute delay, the examinee is asked to freely recall as many words as possible from this list, which is the delayed recall phase of the test (i.e., “RAVLT Delay” in the V3 app). In V3, this test is recommended for ages 5 to 85+.

Test Development and Revisions

The V3 RAVLT underwent several changes from V2. In the immediate recall phase of the test, a chime was added at the end of the presentation of the wordlist to give the participant an indication that the list had finished, and they could start responding. That is, in NIH Toolbox version of RAVLT, the wordlist is read aloud by a recorded voice on the app. This has had the benefit of offering greater standardization in how the list is presented but in V2 had the drawback of participants being unsure that the list presentation was complete. Typically, when an examiner (and not a recording) presents a word list, the examiner looks up at the participant to give the non-verbal cue the list was complete and now it is their turn to respond. Thus, in V3, the chime acts as this cue to participants that it is time to respond.

Before being added to the NIH Toolbox, the RAVLT was well-known throughout the neuropsychology and other psychology testing communities as a verbal learning and delayed memory test. In V2, however, the delay portion of the test was not part of the app. Users who wanted this would need to record this outside of the app. With the V3 release, however, this delay was added as a separate “RAVLT Delay” test.

Another notable change from V2 is that in V3, the RAVLT immediate and delayed portions of the test are normed. Please note that during the V3 normative data collection the V3 Visual Reasoning and Oral Symbol Digit tests were administered during the delay.

Administration and Scoring

Part 1: RAVLT has three wordlist trials. The instructions are presented on the screen and read aloud via audio recording. The participant does not interact with the app during this test, rather the examiner records the participant's responses after each list presentation. Examiners score words as correct if the word is the same or closely related to target word (like a plural versus a singular, other closely related form of the word like garden or gardening). The immediate raw score is based on the number recalled correctly at the end of each wordlist presentation.

Part 2: RAVLT Delay is administered to the participant 5 to 25 minutes after the immediate portion of the test. During the delay, it is recommended that no tasks or activities are done that could interfere in the recalling of the target words. Again, the instructions are presented on the screen and read aloud via audio recording. The participant does not interact with the screen, but the examiner records the participant's response. The delayed raw score is based on the total number of words recalled correctly**.**

RAVLT and RAVLT Delay are supplementary measures and does not contribute to any composites within the NIH Toolbox V3 Cognition Battery.

Speeded Matching

Speeded Matching (SM) measures speed of information processing in young children ages 3 to 6. The test uses the well-established match-to-sample paradigm to assess processing speed (Kaat, McKenzie, Shields, LaForte, Coleman, Michalak, & Hessl, 2021). A target image is presented at the top of the screen and a field of four images are presented below it. Examinees are required to identify and tap the stimulus in the field that exactly matches the target stimulus. This test was designed for younger examinees or those with lower ability that may impede performance on Pattern Comparison; identifying an exact match is cognitively less demanding than judging whether two stimuli are identical or not. The images are simple, child-friendly line drawings of animal faces in varying colors. Overall, the test takes less than three minutes to administer.

Test Development

The Speeded Matching test was added to NIHTB V3 to address a need for a measure of processing speed for young children for whom higher-order executive functioning skills may not have emerged. An experimental version was originally tested in the V2 app, and the same items are used in V3. The instructions were slightly tweaked, audio was added so that all instructions are played as audio, and some aspects of the test feedback were updated (e.g., slower blinking lights around the boxes to draw attention to them). The fail rule was also updated from 4 to 8 incorrect responses to the practice items.

Administration and Scoring

Speeded Matching has three phases of administration. In the Demonstration phase, the examinee watches the examiner correctly respond to an item, and then the examinee must respond to the same item. Next, in the Practice phase, the examinee completes up to two trials each of four additional items, with corrective feedback. If the examinee does not correctly answer at least one trial of one practice item, the test terminates. Examinees who proceed to the Live item phase have 90 seconds of actual presentation time (excluding transitions between items) to respond to as many items as possible (up to a maximum of 130). Prior to each item, there is a short delay, which was not included in the overall test time. Item timing is calculated from when an item presentation is complete until the examinee touches down on any response button in the bottom array. Items continue to be presented until the cumulative item time reaches 90 seconds. No new items are presented after that, but an examinee can respond to the item presented on the screen at that time. If all 130 items are presented prior to the 90-second time limit, the test ends.

The raw score for Speeded Matching is calculated as the number of items the examinee answered correctly within 90 seconds of testing time. The raw score value ranges from 0 to 130. Items were calibrated to the Rasch model, where all items are equally related to the underlying processing speed trait, but they ranged in difficulty due to positioning in the test (i.e., later items are reached by only the most able examinees, and therefore appear more difficult). Examinees are scored using maximum likelihood scoring to obtain an interval-scaled score on the logit metric, which is converted to a change-sensitive score for reporting purposes.

Speeded Matching contributes to the Early Childhood Cognition battery (ages 4-6 years), with supplemental norms available from ages 3 to 8-years and 9-months old. Consistent with all NIHTB V3 Cognition tests, change sensitive scores are available for all ages, even those outside of the recommended age ranges. Age-adjusted normative scores are available for examinees throughout the core and supplemental age ranges (i.e., ages 3 to 8-years and 9-months old).

Visual Reasoning

Visual Reasoning is a new test in the NIHTB V3 that measures the executive functions of nonverbal and visual reasoning. Examinees are presented with a series of pictures at the top of the screen, in varying formations (e.g., a horizontal pattern or a matrix), with one picture missing from the series. The examinee must select, from among four options, the picture that best completes the series. Test items require the examinee to use analogic reasoning, serial reasoning, spatial visualization, and mental rotation. Visual Reasoning is a CAT-administered test. It is recommended for ages 4+.

Test Development

The NIHTB Visual Reasoning (VR) test was developed in response to an identified need for a measure of fluid reasoning that does not rely on verbal or language skills, and that can be administered in 10 minutes or less. The VR test was developed *de novo* to fulfill this need in the NIHTB.

Several goals guided the development of the test:

- Minimize verbal instructions by utilizing animations, visual graphics, and examiner gestures for examinee task training and feedback;
- Limit expressive language demands on the examinee by using a multiple-choice, touch-response format;
- Limit processing speed demands; and
- Minimize testing time through a computer-adaptive testing (CAT) administration format.

A total of 240 items representing the following types were initially developed: simple comparison, sequences, pattern comparison, analogy, serial reasoning, spatial visualization, and multi-rule reasoning. All initial items included between four and six response options. Two pilot studies were conducted. In the first study, a convenience sample of 95 individuals (37 children and 58 adults) were administered between 40 and 60 items each on an iPad. Cognitive interviews were conducted with a subset of study participants to gauge test-taking strategies. Data were analyzed to assess item difficulty, examinee performance, and reaction time. Results from the initial pilot study suggested that while the simple comparison items were appropriate for 4-year-old examinees, these young children struggled with the more complex item types. In contrast, children ages 5 and older, and adults, responded successfully to all item types. Generally, children responded more quickly than adults, and for adults, slower response times were associated with higher accuracy.

Information gleaned from the first pilot study informed the design of the second study. Items were preliminarily ordered by difficulty and separated into forms that contained varied item types targeted to specific age groups. The second pilot study included 779 participants (489 children and 290 adults). Participants each completed between 30 and 45 items; just under half of the participants (379) took the test on an iPad, and (to maximize efficiency in data collection) the remaining examinees took the test remotely via a web-based administration platform. Analysis of the data from the second study revealed age-group performance patterns similar to those in the first study. Data from the second study were Rasch-analyzed to evaluate item and distractor functioning. Based on these analyses, seven items were removed from the pool, two new medium-difficulty items were written, and all items were standardized to include only four answer options. After removal of several items to use as practice items, 193 items remained in the pool after the second pilot study.

The newly developed VR test was administered in the NIHTB V3 norming study. Multiple fixed forms of 25 to 30 items each were assembled from the item pool, and the study design included both horizontal and vertical form linking. To optimize item difficulty targeting and maximize statistical information, forms were targeted to specific ages (either ages 3 to 4, ages 5 to 10, or ages 11+). Study participants were randomly assigned to an age-group-appropriate form, and the items on each form were presented in randomized order. Testing time was capped at 10 minutes for each participant, although no time limit was introduced at the item level.

Data from the V3 norming study were calibrated with a 2-parameter IRT model. Item statistics were reviewed for fit and discrimination. Thirteen items were removed from the item pool due to poor item statistics. Simulation analyses were conducted using the item difficulty and response time data from the V3 norming study to determine the minimum and maximum administration times that would be required for CAT algorithm convergence for each age group.

Administration and Scoring

Visual Reasoning is administered in three phases: Instructions, Practice, and Test items. Instructions are presented on the iPad screen with accompanying audio. The Practice phase includes three items of varying types, with feedback for both correct and incorrect responses. The examinee is allowed up to two attempts on each practice item. All examinees proceed to the live test items, regardless of performance on the practice items.

The live test items are administered via a computer-adaptive testing (CAT) algorithm. An appropriate starting item is selected based on the examinee's age (for participants younger than 19 years old) or education (for participants 20 years and older). The CAT administers a minimum of 20 items and a maximum of 35 items with a target probability of correct response of 0.675. Once the CAT has reached the minimum test length, the test will end if the standard error of the ability estimate falls below 0.44 logits.

The CAT algorithm produces an interval-scaled score on the logit metric, which is converted to a Change Sensitive Score (CSS) for reporting purposes. Age-adjusted normative scores are available for examinees ages 4 to 85+ years, and age-and-education-adjusted normative scores are available for individuals ages 22 to 85+ years.

The Visual Reasoning test is a supplementary measure and does not contribute to any composites within the NIH Toolbox V3 Cognition Battery.

Chapter 3: Norming

The NIH Toolbox® V3 Cognition Battery norming study was conducted from June through September of 2021. During this period, the Cognition tests were administered to 3,956 individuals, including 2,248 children ages 3 to 17 and 1,708 adults ages 18 to 90+ years. The goals of the V3 norming study were to:

1) add new tests to the Cognition battery to improve the construct coverage of the battery;
2) update existing tests to address specific user concerns regarding floor effects, ceiling effects, and reliability;
3) update the Cognition Battery norms to reflect the current U.S. population for sex assigned at birth, race, ethnicity, and education level;
4) implement continuous norming procedures to minimize the “binning” phenomenon associated with discrete norming procedures; and
5) recalibrate as many tests as possible to a common IRT-based “Change-Sensitive Score” metric.

Description of the NIH Toolbox V3 Norming Study

The V3 norming study data collection was conducted by a private market research firm under the direction of researchers at Northwestern University. The private market research firm hired examiners, recruited participants, and conducted the assessments at their offices. Norming data were collected at 12 sites across the four regions of the United States, with at least one site in each U.S. census division:

Midwest Region:
- Appleton, WI (East North Central)
- Chicago, IL (East North Central)
- Columbus, OH (East North Central)
- St Louis, MO (West North Central)

Northeast Region:
- Boston, MA (New England)
- Iselin, NJ (Middle Atlantic)

South Region:
- Atlanta, GA (South Atlantic)
- Baltimore, MD (South Atlantic)
- Dallas, TX (West South Central)
- Nashville, TN (East South Central)

West Region:
- Los Angeles, CA (Pacific)
- Phoenix, AZ (Mountain)

Northwestern University and the private market research firm employed a “train the trainer” model to prepare examiners for data collection. Ten of the private market research firm’s trainers and several project staff were trained on NIHTB administration during a 5-day, in-person session led by the Northwestern University team, which consisted of a faculty member and a project manager. The trainers then returned to their respective sites, where they practiced administration with participants of various ages. The trainers videorecorded their final practice cases and submitted the video files to the Northwestern University training team for review and certification. The Northwestern University training team provided feedback on any administration errors that were made during the certification case.

The respective site trainers then implemented the same training regimen with the 10-12 site-specific examiners at each location, culminating in the observation of a final practice case for certification.

Examiners who administered external validity measures were trained and certified in NIHTB administration in the same way as the regular examiners. The validity study examiners were also required to administer several external validity measures, including several batteries that are available on Pearson's Q-Interactive platform. Although experience with the Q-Interactive platform was a requirement for examiners who gathered data for the external validity studies, these examiners were given a half-day training/refresher course on these measures by Northwestern University project staff.

Norming study participants were recruited from the private market research firm's national database through phone calls and emails. Those who met the quotas provided by Northwestern University for census region, sex assigned at birth, age, race, ethnicity, and education level (or parent education level, for children) were enrolled for participation. All testing took place in the private market research firm's offices. Sampling targets were monitored daily to ensure the study demographic targets were met.

Characteristics of the V3 Norming Sample

Target demographics for the NIHTB V3 norming study were developed in consultation with an epidemiologist familiar with the US Census Bureau catalog of surveys and products. The sampling cells were based on the 1-year estimates from the 2017 American Community Survey (ACS) and Current Population Survey (CPS). As the primary V3 norming objective was to create age-adjusted reference values, the norming study sample was not selected to represent the US population for age. Child participants were oversampled to allow precise estimation of reference values for the ages where most cognitive abilities show rapid growth and large variation. One-year sampling cells were utilized for children (ages 3-17 years); 2-year sampling cells were used for young adults (ages 18-21 years); a single sampling cell for young adults (ages 22-29 years); and 10-year sampling cells were utilized for most of adulthood (30-79 years). For older adults (age 80+), the sampling cell was halved, with one half representing ages 80-84 and the second half representing anyone over ages 85 years who was able to complete the NIHTB battery.

Additional demographics were sampled to ensure adequate representation of the U.S. population. These characteristics were nested within geographic region (West, South, Northeast, and Midwest) and broader age categories. The broader age categories chosen were children (ages 3-17 years), young adults (18-21 years), middle-aged adults (22-59 years), and older adults (60 years or older). Target demographic characteristics that were nested within these broader age categories and geographic regions included sex assigned at birth (male or female), race and ethnicity (Hispanic ethnicity regardless of race, Non-Hispanic White, Non-Hispanic Black, Non-Hispanic Asian, or Non-Hispanic Other), and education [for children, parental education categories included (a) less than a high school education, (b) a high school education or GED, (c) some college, including technical or trade schools and 2-year Associate degrees, or (d) a Bachelor’s degree or higher; for adults, obtained education levels included (a) less than a high school diploma, (b) a high school diploma or GED, (c) some college, (d) a 4-year Bachelor’s degree, or (e) a graduate or professional degree, including Master’s or Doctoral degree, Medical Degree, or other professional certification].

A 10% margin around the 2017 ACS and CPS proportions were considered acceptable for the purpose of sampling during the data collection phase of the study. After data collection was complete, however, the sampling cells were updated from the 2017 1-year ACS and CPS estimates to the 2020 Decennial Census augmented with the 2019 ACS as needed. These updated proportions were then used to compute the examinee weights for the V3 norm calculations.

Figure 3.1 provides a density plot for the sample weights by broad age group. By definition, the mean sample weight was 1.0, but the median ranged from 0.70 to 0.93 depending on the age group. The largest sample weights were assigned to individuals from demographic groups less-represented in the sample—particularly individuals with less than a high school education (or children of parents with less than a high school education).

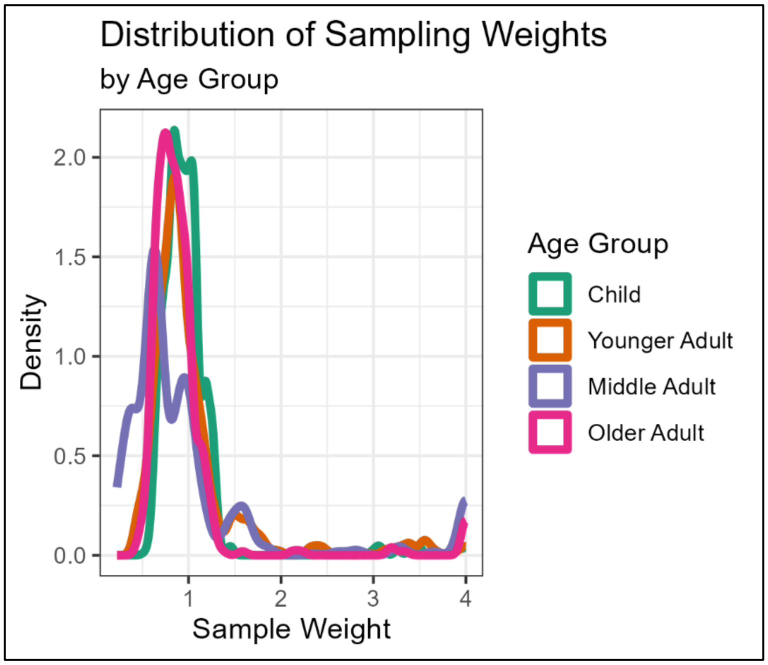


Figure 3.1: Distribution of sampling weights in the NIHTB V3 norming study

The sample demographics for child, young adult, middle-aged adult, and older adult age groups, and the total sample, are reported in Table 3.1. As appropriate, the demographic categories were aggregated and combined for weighting purposes; the demographic unweighted and weighted distributions for the variables of interest are reported in Tables 3.2a through 3.2d. The raw count of participants is provided, as well as the population proportion based on the 2020 Decennial Census and/or 2019 ACS. The unweighted and weighted proportions for the obtained sample are then reported, with comparisons to the target population proportions.

Table 3.1
Demographic Characteristics of the NIHTB V3 Norming Sample, Total and by Age Group

Demographic Characteristic		*Child*	*Younger Adult*	*Middle Adult*	*Older Adult*	*Total Sample*
Sex Assigned at Birth	Male	1122	200	323	237	1882
	Female	1126	210	362	375	2073
	Not Reported	0	0	0	1	1
Gender Identity	Male	1092	191	321	227	1831
	Female	1108	205	354	349	2016
	Not Reported	48	14	10	37	109
Hispanic Ethnicity	Non-Hispanic	1682	317	563	587	3149
	Hispanic – Mexican, Mexican American, or Chicano	340	59	71	14	484
	Hispanic – Puerto Rican	58	7	15	1	81
	Hispanic – Cuban	11	2	0	0	13
	Hispanic – Other Hispanic, Latinx, or Spanish Origin	155	25	36	6	222
	Hispanic – Origin Not Reported	2	0	0	5	7
Race	Black	522	73	130	60	785
	American Indian or Alaskan Native	48	3	7	3	61
	Hawaiian	1	1	1	0	3
	Other Pacific Islander	20	10	4	0	34
	Asian – Chinese	52	10	12	7	81
	Asian – Indian	79	6	25	2	112
	Asian – Filipino	30	7	9	5	51
	Asian – Vietnamese	15	7	4	1	27
	Asian – Korean	15	3	4	1	23
	Asian – Japanese	14	3	3	3	23
	Asian – Other Specified Ethnicity	27	2	5	0	34
	White	1633	282	483	531	2929
	Middle Eastern or North African	13	10	3	3	29
Educational Attainment (or Parental Educational Attainment for Children)	Less than High School Graduate	617	200	215	213	1245
	High School Graduate or GED	561	191	200	206	1158
	Some College, including Technical School, Trade School, or 2-Year Associate's Degree	721	205	210	162	1298
	Bachelor's Degree	570	3	175	136	884
	Graduate or Professional Degree	337	2	85	102	526
	Educational Attainment Not Reported	3	0	0	0	3

Note. Sex assigned at birth and gender identity were asked separately. Individuals were able to choose more than one racial or ethnic group. Consistent with U.S. Census standards, and for the purposes of sample weighting, Hawaiian and Other Pacific Islander were combined, all Asian ethnicities were combined, and Middle Eastern or North African was classified with White race. For children, Bachelor's Degree and Graduate or Advanced Degrees were combined given the demographic proportions available.

Table 3.2a
Census Representation of the NIHTB V3 Norming Study Sample and Norming Weights, Child Sample

Region	*Characteristic*	*Target*	*Census Proportion*	*Obtained Sample*	*Unweighted Proportion*	*Weighted Proportion*	*Difference*
Northeast	Race / Ethnicity	Non-Hispanic Asian American	0.011	19	0.008	0.011	0.000
		Non-Hispanic Black	0.020	52	0.023	0.020	0.000
		Hispanic – Any Race	0.033	69	0.031	0.033	0.000
		Non-Hispanic Other	0.007	20	0.009	0.007	0.000
		Non-Hispanic White	0.089	189	0.084	0.089	0.000
South		Non-Hispanic Asian American	0.008	14	0.006	0.008	0.000
		Non-Hispanic Black	0.026	68	0.030	0.026	0.000
		Hispanic – Any Race	0.027	52	0.023	0.027	0.000
		Non-Hispanic Other	0.011	25	0.011	0.011	0.000
		Non-Hispanic White	0.139	348	0.155	0.139	0.000
Midwest		Non-Hispanic Asian American	0.014	33	0.015	0.014	0.000
		Non-Hispanic Black	0.080	207	0.092	0.080	0.000
		Hispanic – Any Race	0.096	194	0.086	0.096	0.000
		Non-Hispanic Other	0.018	53	0.024	0.018	0.000
		White	0.179	359	0.160	0.179	0.000
West		Non-Hispanic Asian American	0.021	36	0.016	0.021	0.000
		Non-Hispanic Black	0.011	27	0.012	0.011	0.000
		Hispanic – Any Race	0.099	249	0.111	0.099	0.000
		Non-Hispanic Other	0.019	38	0.017	0.019	0.000
		Non-Hispanic White	0.092	196	0.087	0.092	0.000
Northeast	Sex Assigned at Birth	Female	0.078	171	0.076	0.078	0.000
		Male	0.082	178	0.079	0.082	0.000
South		Female	0.103	240	0.107	0.103	0.000
		Male	0.108	267	0.119	0.108	0.000
Midwest		Female	0.190	440	0.196	0.190	0.000
		Male	0.198	406	0.181	0.198	0.000
West		Female	0.118	275	0.122	0.118	0.000
		Male	0.124	271	0.121	0.123	0.000
Any Region	Parental Educational Attainment	Less than High School Graduate	0.090	58	0.026	0.090	0.000
		High School Graduate or GED	0.242	561	0.250	0.242	0.000
		Some College, including Technical School, Trade School, or 2-Year Associate's Degree	0.251	721	0.321	0.251	0.000
		Bachelor's Degree or Higher	0.418	908	0.404	0.418	0.000

Table 3.2b
Census Representation of the NIHTB V3 Norming Study Sample and Norming Weights, Young Adult Sample

Region	Characteristic	Target	Census Proportion	Obtained Sample	Unweighted Proportion	Weighted Proportion	Difference
Northeast	Race / Ethnicity	Non-Hispanic Asian American	0.012	9	0.019	0.012	0.000
		Non-Hispanic Black	0.022	13	0.027	0.022	0.000
		Hispanic – Any Race	0.031	20	0.042	0.031	0.000
		Non-Hispanic Other	0.005	1	0.002	0.005	0.000
		Non-Hispanic White	0.102	54	0.113	0.102	0.000
South		Non-Hispanic Asian American	0.009	2	0.004	0.009	0.000
		Non-Hispanic Black	0.025	12	0.025	0.025	0.000
		Hispanic – Any Race	0.023	9	0.019	0.023	0.000
		Non-Hispanic Other	0.008	3	0.006	0.008	0.000
		Non-Hispanic White	0.146	75	0.157	0.146	0.000
Midwest		Non-Hispanic Asian American	0.013	8	0.017	0.013	0.000
		Non-Hispanic Black	0.082	37	0.078	0.082	0.000
		Hispanic – Any Race	0.084	38	0.080	0.084	0.000
		Non-Hispanic Other	0.014	6	0.013	0.014	0.000
		White	0.186	76	0.159	0.186	0.000
West		Non-Hispanic Asian American	0.022	9	0.019	0.022	0.000
		Non-Hispanic Black	0.012	8	0.017	0.012	0.000
		Hispanic – Any Race	0.092	44	0.092	0.092	0.000
		Non-Hispanic Other	0.016	7	0.015	0.016	0.000
		Non-Hispanic White	0.097	46	0.096	0.097	0.000
Northeast	Sex Assigned at Birth	Female	0.085	46	0.096	0.085	0.000
		Male	0.086	51	0.107	0.086	0.000
South		Female	0.104	50	0.105	0.104	0.000
		Male	0.108	51	0.107	0.108	0.000
Midwest		Female	0.186	82	0.172	0.186	0.000
		Male	0.194	83	0.174	0.194	0.000
West		Female	0.115	57	0.119	0.115	0.000
		Male	0.122	57	0.119	0.122	0.000
Any Region	Educational Attainment	Less than High School Graduate	0.149	9	0.019	0.075	-0.073
		High School Graduate or GED	0.319	221	0.463	0.346	0.027
		Some College, including Technical School, Trade School, or 2-Year Associate's Degree	0.404	224	0.470	0.439	0.035
		Bachelor's Degree or Higher	0.120	19	0.040	0.130	0.010
		Advanced Degree	0.009	4	0.008	0.010	0.001

Table 3.2c
Census Representation of the NIHTB V3 Norming Study Sample and Norming Weights, Middle Adult Sample

Region	*Characteristic*	*Target*	*Census Proportion*	*Obtained Sample*	*Unweighted Proportion*	*Weighted Proportion*	*Difference*
All Regions	Race / Ethnicity	Non-Hispanic Other	0.043	13	0.021	0.043	0.000
Northeast		Non-Hispanic Asian American	0.012	8	0.013	0.012	0.000
		Non-Hispanic Black	0.022	16	0.026	0.022	0.000
		Hispanic – Any Race	0.031	12	0.019	0.031	0.000
		Non-Hispanic White	0.102	71	0.115	0.102	0.000
South		Non-Hispanic Asian American	0.009	4	0.006	0.009	0.000
		Non-Hispanic Black	0.025	16	0.026	0.025	0.000
		Hispanic – Any Race	0.023	13	0.021	0.023	0.000
		Non-Hispanic White	0.146	95	0.154	0.146	0.000
Midwest		Non-Hispanic Asian American	0.013	13	0.021	0.013	0.000
		Non-Hispanic Black	0.082	58	0.094	0.082	0.000
		Hispanic – Any Race	0.084	36	0.058	0.084	0.000
		White	0.186	117	0.189	0.186	0.000
West		Non-Hispanic Asian American	0.022	20	0.032	0.022	0.000
		Non-Hispanic Black	0.012	10	0.016	0.012	0.000
		Hispanic – Any Race	0.092	43	0.070	0.092	0.000
		Non-Hispanic White	0.097	73	0.118	0.097	0.000
Northeast	Sex Assigned at Birth	Female	0.085	58	0.094	0.083	-0.003
		Male	0.086	49	0.079	0.084	-0.003
South		Female	0.104	73	0.118	0.104	0.001
		Male	0.108	58	0.094	0.108	0.001
Midwest		Female	0.186	125	0.202	0.187	0.001
		Male	0.194	105	0.170	0.195	0.001
West		Female	0.115	81	0.131	0.116	0.001
		Male	0.122	69	0.112	0.123	0.001
Any Region	Educational Attainment	Less than High School Graduate	0.082	15	0.024	0.082	0.000
		High School Graduate or GED	0.261	170	0.275	0.261	0.000
		Some College, including Technical School, Trade School, or 2-Year Associate's Degree	0.253	191	0.309	0.253	0.000
		Bachelor's Degree or Higher	0.258	159	0.257	0.258	0.000
		Advanced Degree	0.147	83	0.134	0.147	0.000

Table 3.2d
Census Representation of the NIHTB V3 Norming Study Sample and Norming Weights, Older Adult Sample

Region	Characteristic	Target	Census Proportion	Obtained Sample	Unweighted Proportion	Weighted Proportion	Difference
All Regions	Race / Ethnicity	Non-Hispanic Other	0.043	3	0.005	0.021	-0.022
Northeast		Non-Hispanic Asian American	0.012	1	0.002	0.007	-0.005
		Non-Hispanic Black	0.022	8	0.014	0.025	0.003
		Hispanic – Any Race	0.031	2	0.004	0.014	-0.016
		Non-Hispanic White	0.102	98	0.175	0.117	0.016
South		Non-Hispanic Asian American	0.009	1	0.002	0.007	-0.002
		Non-Hispanic Black	0.025	14	0.025	0.029	0.004
		Hispanic – Any Race	0.023	2	0.004	0.014	-0.009
		Non-Hispanic White	0.146	110	0.196	0.168	0.023
Midwest		Non-Hispanic Asian American	0.013	4	0.007	0.015	0.002
		Non-Hispanic Black	0.082	27	0.048	0.095	0.013
		Hispanic – Any Race	0.084	4	0.007	0.029	-0.056
		White	0.186	158	0.282	0.215	0.029
West		Non-Hispanic Asian American	0.022	12	0.021	0.025	0.003
		Non-Hispanic Black	0.012	6	0.011	0.013	0.002
		Hispanic – Any Race	0.092	13	0.023	0.093	0.001
		Non-Hispanic White	0.097	98	0.175	0.111	0.015
Northeast	Sex Assigned at Birth	Female	0.085	65	0.116	0.081	-0.004
		Male	0.086	44	0.078	0.082	-0.004
South		Female	0.104	72	0.128	0.107	0.003
		Male	0.108	55	0.098	0.111	0.004
Midwest		Female	0.186	125	0.223	0.180	-0.006
		Male	0.194	70	0.125	0.188	-0.006
West		Female	0.115	74	0.132	0.121	0.006
		Male	0.122	56	0.100	0.129	0.006
Any Region	Educational Attainment	Less than High School Graduate	0.104	7	0.012	0.050	-0.054
		High School Graduate or GED	0.312	182	0.324	0.349	0.037
		Some College, including Technical School, Trade School, or 2-Year Associate's Degree	0.254	147	0.262	0.273	0.019
		Bachelor's Degree or Higher	0.192	129	0.230	0.193	0.001
		Advanced Degree	0.137	96	0.171	0.135	-0.002

NIH Toolbox V3 Norming Procedures

Following procedures outlined by DeBell and Krosnick (2009), iterative proportional fitting was used to derive probability weights for each individual, and then the weights were trimmed to a maximum value of 4. Timmerman et al.'s (2021) regression-based norming method using generalized additive models (GAM), which models the change-sensitive score (CSS) distribution of each test as a function of age as a continuous variable, was employed. Figure 3.2 provides a visual description of the norming processes for both tests and composites; these procedures are described in detail below.

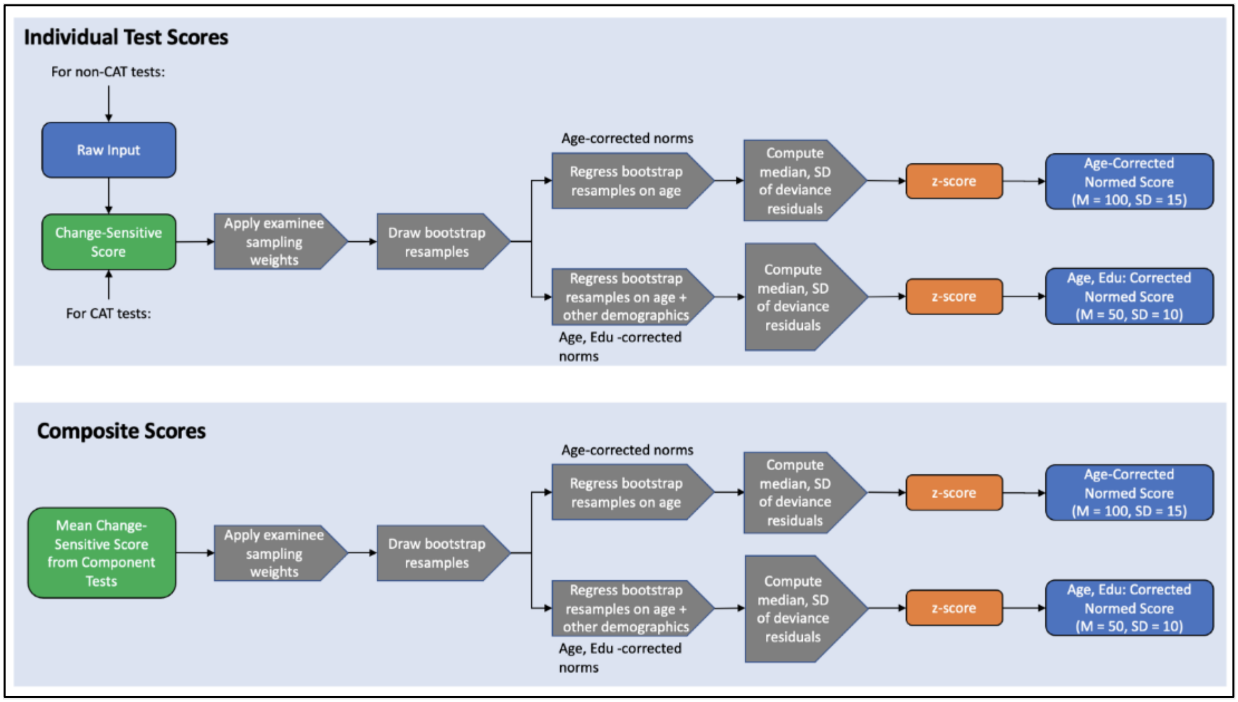

Individual Test Scores
For non-CAT tests:
Raw Input
Change-Sensitive Score
For CAT tests:
Apply examinee sampling weights
Draw bootstrap resamples
Age-corrected norms
Regress bootstrap resamples on age
Compute median, SD of deviance residuals
z-score
Age-Corrected Normed Score (M = 100, SD = 15)
Regress bootstrap resamples on age + other demographics
Age, Edu -corrected norms
Compute median, SD of deviance residuals
z-score
Age, Edu: Corrected Normed Score (M = 50, SD = 10)
Composite Scores
Mean Change-Sensitive Score from Component Tests
Apply examinee sampling weights
Draw bootstrap resamples
Age-corrected norms
Regress bootstrap resamples on age
Compute median, SD of deviance residuals
z-score
Age-Corrected Normed Score (M = 100, SD = 15)
Regress bootstrap resamples on age + other demographics
Age, Edu -corrected norms
Compute median, SD of deviance residuals
z-score
Age, Edu: Corrected Normed Score (M = 50, SD = 10)

Figure 3.2: Steps to produce continuous norms for NIHTB V3 tests and composites

Norming of Test Scores

For tests scored using the Rasch model or another Item Response Theory (IRT) model, the final thetas and corresponding standard errors were used to generate the CSSs; for other measures, the raw item level input (e.g., raw scores or rate-corrected scores) were used to generate the CSSs[2]. CSSs are centered at 500, where 500 represents the median ability of 10-year-old participants in the V3 norming sample.

Individual sampling weights were applied to every case in the sample. Bootstrap resamples were drawn and each CSS was utilized for the selected cases. Within each resample, plausible value imputation was used to sample an expected score for each examinee based off of their obtained CSS and its associated standard error. The bootstrap resamples were regressed on age to develop age-adjusted norms for each measure. The process for developing age-and-education-adjusted norms was similar, except that the bootstrap resamples were regressed on

[2] See Chapter 2 for more information about the scoring model used to derive CSSs for each Cognition test in NIHTB V3.

age-within-education strata. From these models, the median and *SD* of the deviance residuals were computed, and a *z*-score transformation was applied, such that age-adjusted normed scores were distributed at *M* = 100, *SD* = 15, and age-and-education-adjusted normed scores were distributed at *M* = 50, *SD* = 10.

Appendix A contains summary statistics, by age group, for all test- and composite-level CSS scores for the V3 norming study participants.

V3 Norming of Composite Scores

The composite CSS scores were created by averaging the CSSs from the requisite tests (i.e., Picture Vocabulary and Oral Reading for the Crystallized Composite; Dimensional Change Card Sorting, Flanker Inhibitory Control and Attention, List Sorting Working Memory, Pattern Comparison Processing Speed, and Picture Sequence Memory for the Fluid Composite; and Picture Vocabulary, Flanker Inhibitory Control and Attention, Dimensional Change Card Sort, Picture Sequence Memory, and Speeded Matching for the Early Childhood Composite). The standard error for the average of the CSS scores was calculated as the square root of the variance of a mixture distribution composed of Gaussian random variables, whereby the variance is the sum of the variance of the requisite components plus a correction factor for dispersion of the means. From there, the procedures to produce normed scores followed that of the V3 test norming procedures, including drawing a plausible value for each individual within the bootstrap resampling. To obtain the Total Cognition Composite score, the average of the Crystalized and Fluid composites was calculated and regressed on age (or age-within-education strata, for the age-and-education–corrected norms).

Chapter 4: Reliability and Validity

Reliability

Reliability refers, generally, to the consistency of scores across replications of a test (AERA, APA, & NCME, 2014). High reliability indices imply that changes in the score reflect actual changes in the underlying measured variable (i.e., latent trait).

Reliability Coefficients for Tests and Composites

For the NIH Toolbox® tests, empirical reliability was computed using the Change-Sensitive Scores (CSSs) as

$$\text{Variance}_{\text{CSS}} / (\text{Variance}_{\text{CSS}} + SEM^{2}_{\text{CSS}}),$$

for all norming participants, and separately for children and adults. Table 4.1 contains sample *n*s, reliability coefficients, and SEMs for all tests and composites. Reliability indexes can be interpreted as the average proportion of observed variance (within each age group) in the test's scores that is due to true differences in the latent trait, and not to random measurement error.

Table 4.1
Sample *ns*, Empirical Reliability Indices, and *SEM*s for NIHTB Tests and Composite Change-Sensitive Scores, for Child, Adult, and Total Norming Samples

	Ages 3 to 17			Ages 18 and Older			Total Sample		
Test / Composite	*n*	CSS R_{11}	*SEM* (CSS)	*n*	CSS R_{11}	*SEM* (CSS)	*n*	CSS R_{11}	*SEM* (CSS)
DCCS	2163	0.91	8.54	1590	0.88	11.01	3753	0.90	9.75
Flanker	1970	0.91	7.32	1466	0.86	7.31	3436	0.90	7.37
FNAME	NA	NA	NA	1594	0.78	6.56	1595	0.78	6.56
LSWM	1898	0.92	7.23	1595	0.87	6.88	3493	0.91	7.11
ORR	2070	0.99	5.39	1596	0.90	4.49	3666	0.99	5.10
OSD	1831	0.99	3.71	1582	0.98	3.81	3413	0.99	3.76
PC	1938	0.99	3.26	1594	0.99	3.58	3532	0.99	3.41
PSM	2151	0.97	6.42	1573	0.95	6.05	3724	0.97	6.27
PV	2200	0.94	4.37	1600	0.87	4.39	3800	0.95	4.48
RAVLT	1916	0.96	3.43	1586	0.95	3.35	3502	0.95	3.40
RAVLT Delay	1264	0.72	7.09	1466	0.77	7.44	2730	0.75	7.31
SM	838	0.98	3.82	NA	NA	NA	838	0.98	3.82
VR	2171	0.88	4.57	1590	0.76	4.47	3761	0.86	4.60
Early Childhood Composite	728	0.98	-	NA	NA	-	728	0.98	-
Fluid Composite	1647	0.98	-	1595	0.97	-	3242	0.98	-
Crystallized Composite	1650	0.98	-	1601	0.93	-	3251	0.98	-
Total Cognition Composite	1646	0.99	-	1595	0.95	-	3241	0.99	-

Note. DCCS = Dimensional Change Card Sort, Flanker = Flanker Inhibitory Control and Attention, FNAME = Face Name Associative Memory Exam, LSWM = List Sorting Working Memory, ORR = Oral Reading

Recognition, OSD = Oral Symbol Digit, PC = Pattern Comparison Processing Speed, PSM = Picture Sequence Memory, PV = Picture Vocabulary, RAVLT = Rey Auditory Verbal Learning, RAVLT Delay = Rey Auditory Verbal Learning Delay, SM = Speeded Matching, VR = Visual Reasoning.

Test-Retest Reliability

Test-retest reliability describes how well a test score remains consistent in measuring an individual's performance across multiple administrations. To assess test-retest reliability for the NIHTB V3 tests, a sample of 190 V3 norming participants were administered each test between 1 and 14 days following the first administration. Table 4.2 shows the Pearson correlations, absolute agreement ICCs, and mean-score differences and associated Cohen's *D* statistics for the Change-Sensitive Scores (CSSs) from the first and second administrations of each NIHTB test, for children (under 18 years old), adults (18 years and older), and for all participants in the sample. Pearson correlations are generally moderate to high, ranging from 0.63 to 0.98. Not unexpectedly, tests that include the repeated presentation of a set of stimuli that the examinee must remember—such as the Rey Auditory Verbal Learning tests, Picture Sequence Memory, and Oral Symbol Digit—show larger mean CSS changes and higher associated Cohen's D values than do tests that are CAT-administered such as Picture Vocabulary and Oral Reading Recognition (where examinees will encounter different items on each administration). A notably large practice effect is present for Pattern Comparison Processing Speed (differences of 20.8 and 18.4 for children and adults, respectively), suggesting that examinees do benefit from having performed the test task earlier; however, the relatively high correlations between Time 1 and Time 2 (0.88 and 0.80, respectively) suggest that all examinees benefit from this practice effect in a similar way.

Table 4.2
Test-Retest Reliability for V3 Cognition Tests

	Under 18 years old				*18 years old and older*				*All participants*			
Measure	*Pearson r*	*ICC*	*CSS mean difference*	*Cohens D*	*Pearson r*	*ICC*	*CSS mean difference*	*Cohens D*	*Pearson r*	*ICC*	*CSS mean difference*	*Cohens D*
DCCS	0.79	0.79	2.76	0.17	0.71	0.66	7.37	0.39	0.79	0.77	5.09	0.29
Flanker	0.88	0.88	1.84	0.17	0.77	0.71	5.7	0.5	0.86	0.84	3.79	0.33
FNAME	-	-	-	-	0.7	0.60	-4.67	0.54	0.7	0.6	-4.67	0.54
LSWM	0.75	0.76	5.84	0.34	0.81	0.77	5.46	0.47	0.79	0.78	5.64	0.39
ORR	0.98	0.98	0.01	0	0.73	0.71	3.47	0.32	0.97	0.97	1.76	0.17
OSD	0.8	0.7	18.09	0.7	0.66	0.57	16.14	0.58	0.76	0.68	17.04	0.64
PC	0.88	0.68	20.8	1.17	0.8	0.66	18.4	0.89	0.83	0.68	19.55	1.02
PSM	0.78	0.64	17.87	0.88	0.76	0.58	18.95	0.94	0.77	0.62	18.41	0.91
PV	0.86	0.86	0.31	0.04	0.86	0.86	0.28	0.05	0.93	0.93	0.29	0.04
RAVLT	0.7	0.39	13.94	1.14	0.77	0.47	13.45	1.25	0.74	0.44	13.68	1.19
RAVLT Delay	0.63	0.45	14.39	0.81	0.64	0.5	13.63	0.68	0.65	0.48	13.99	0.74
SM	0.83	0.8	5.37	0.32	-	-	-	-	0.83	0.8	5.37	0.32
VR	0.79	0.79	-0.31	0.04	0.73	0.71	0.65	0.08	0.78	0.77	0.17	0.02

Note. DCCS = Dimensional Change Card Sort, Flanker = Flanker Inhibitory Control and Attention, FNAME = Face Name Associative Memory Exam, LSWM = List Sorting Working Memory, ORR = Oral Reading

Recognition, OSD = Oral Symbol Digit, PC = Pattern Comparison Processing Speed, PSM = Picture Sequence Memory, PV = Picture Vocabulary, RAVLT = Rey Auditory Verbal Learning, RAVLT Delay = Rey Auditory Verbal Learning Delay, SM = Speeded Matching, VR = Visual Reasoning.

Validity

The *Standards for Educational and Psychological Assessment* (AERA, APA, & NCME, 2014) guided all aspects of the NIH Toolbox V3 Cognition battery development. The descriptions of the processes for the planning, piloting, and V3 norming phases of the project presented elsewhere in this document, together with the results of the analyses presented in this section, serve as evidence to support the validity of the NIH Toolbox Cognition test and composite scores for measuring specific aspects of crystallized intelligence, fluid reasoning, memory, processing speed, and overall cognitive functioning. The evidence presented in this section follows the framework presented in the *Standards* (2014).

Evidence Relevant to Test Content and Construct Coverage

Validity evidence relevant to test content and construct coverage evaluates how well a test score (or composite score) describes an individual's performance on the construct it was intended to measure (AERA, APA, & NCME, 2014; Cizek, 2020). Each NIH Toolbox Cognition test was designed to measure a specific aspect of cognition while avoiding the introduction of construct-irrelevant variance that might confound the interpretation of the test score. Evidence supporting the content and construct coverage aspects of validity, including descriptions of the cognitive abilities measured by the tests, descriptions of the test tasks, and descriptions of the scoring models and interpretation, can be found in Chapters 1 and 2 of this manual. Empirical sources of evidence relating to test content, including cross-sectional growth curves and test and composite intercorrelations, are included in this section.

Cross-Sectional Growth Curves

For tests such as the NIH Toolbox Cognition tests, which are intended to measure ability from childhood through late adulthood, additional evidence to support the test content aspect of validity can be provided by the cross-sectional growth curves for the tests and composite scores. These curves show how test and composite scores change, on average, for examinees from childhood through adolescence and into adulthood. Divergent cross-sectional growth curves among the tests and composites provide evidence that the tests and composites measure unique abilities (Carroll, 1993). Additionally, curve trajectories that conform to theoretical expectations about the growth and decline of human abilities over the lifespan provide additional evidence to support the content aspect of validity.

Figure 4.1 shows the cross-sectional growth curves for the eight tests that are included in the Total Cognition Composite score. To place the curves in the same frame of reference, the origin of each curve represents the median CSS score for 6-year-old examinees in the V3 norming sample. All other points are plotted as the difference between the median score at each age and the median score for 6-year-olds. The scores for the two crystallized tests (Picture Vocabulary and Oral Reading), represented respectively by the light orange and dark orange lines, steadily increase through childhood and adolescence, becoming relatively flat—but still increasing—into late adulthood. Although these two curves share similar shapes, the Oral Reading scores tend show a larger absolute change from age 6 than do the Picture Vocabulary scores. This is not unexpected, given the rapid growth of reading skills in middle childhood. The fluid reasoning scores (represented by the various green and blue lines) tend to show similar growth trajectories throughout childhood and adolescence, but are characterized by distinctive peaks around age 20 followed by relatively rapid decline for the remainder of the adult years. Scores from the two memory tests (Picture Sequence Memory and List Sorting Working Memory, represented by the dark and light green lines, respectively) show a rapid increase from age 6 through about age 10, a relatively slower rate of increase between age 10 and 20, and then a gradual decline through the rest of adulthood. Although these two tests peak at about the same level in late-adolescence, Picture Sequence Memory shows a much larger absolute decline in late adulthood than List Sorting Working Memory does; in fact, the median Picture Sequence Memory score among 90-year-olds in the V3 norming sample is the same as the median score for 6-year-olds. In contrast to the memory tests, the processing speed tests (DCCS, shown in medium blue; Flanker, shown in light blue; and Pattern Comparison, shown in dark blue) show relatively rapid growth from age 6 until their peaks in the early 20s, then drop off through adulthood. Like the memory tests, the processing speed test scores for 90-year-olds in the norming sample are about the same as the scores obtained by 6- to 8-year-olds.

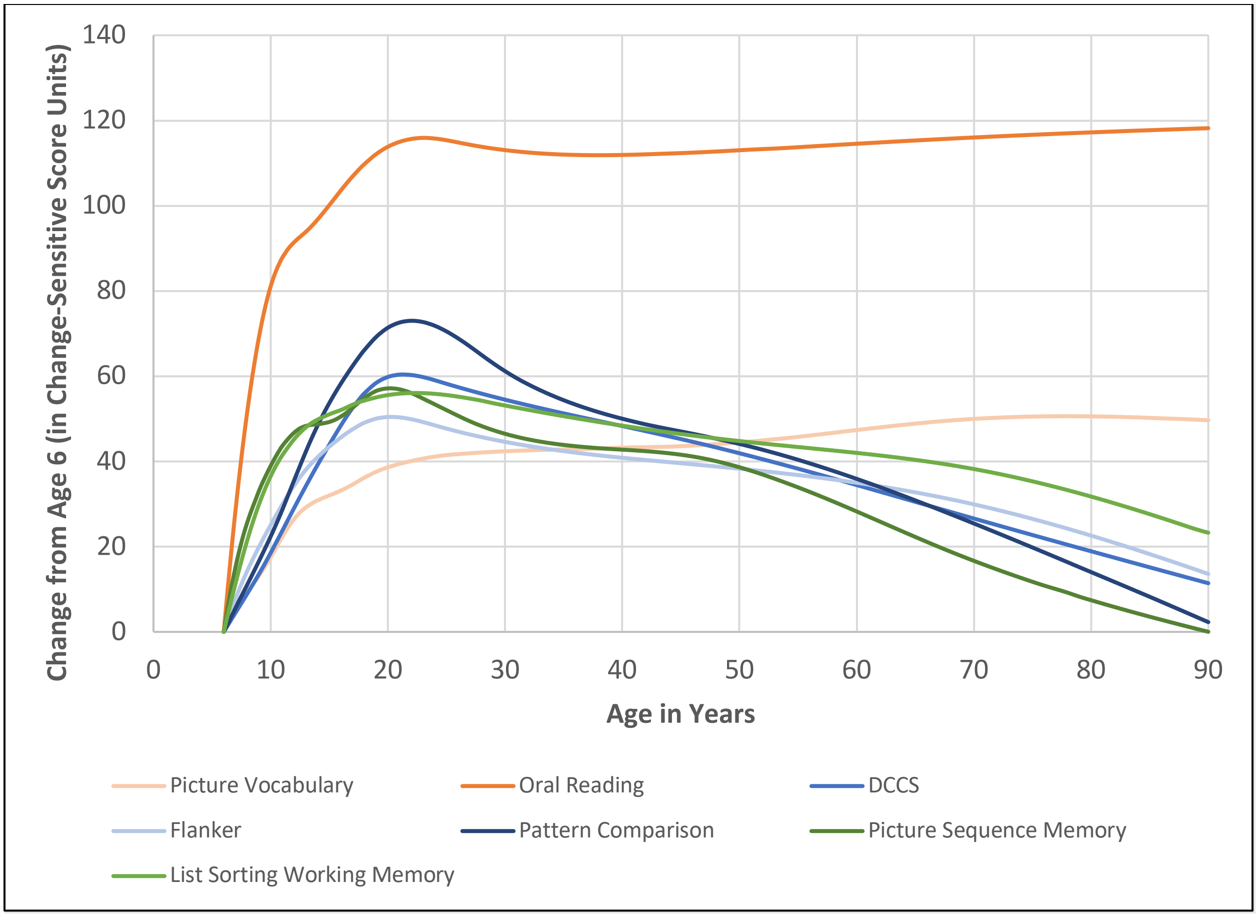


Figure 4.1: Cross-sectional growth curves for the eight core NIH Toolbox Cognition tests

Figure 4.2 shows the cross-sectional growth curves for the Fluid (blue) and Crystallized (orange) Composite scores. As in Figure 4.1, the curves represent the change, in CSS units, from age 6 among examinees in the NIH Toolbox V3 norming sample. Consistent with the behavior of the individual test score growth curves, the Crystallized and Fluid Composite scores show distinctive and predictable trajectories. While both composite scores increase rapidly in childhood and adolescence, the Crystallized Composite scores tend to remain relatively high, continuing to increase—albeit at a slower rate—over the adult years. In contrast, the Fluid Composite scores show a sharp and rapid decline across the adult years, after reaching a peak around age 20 to 22.

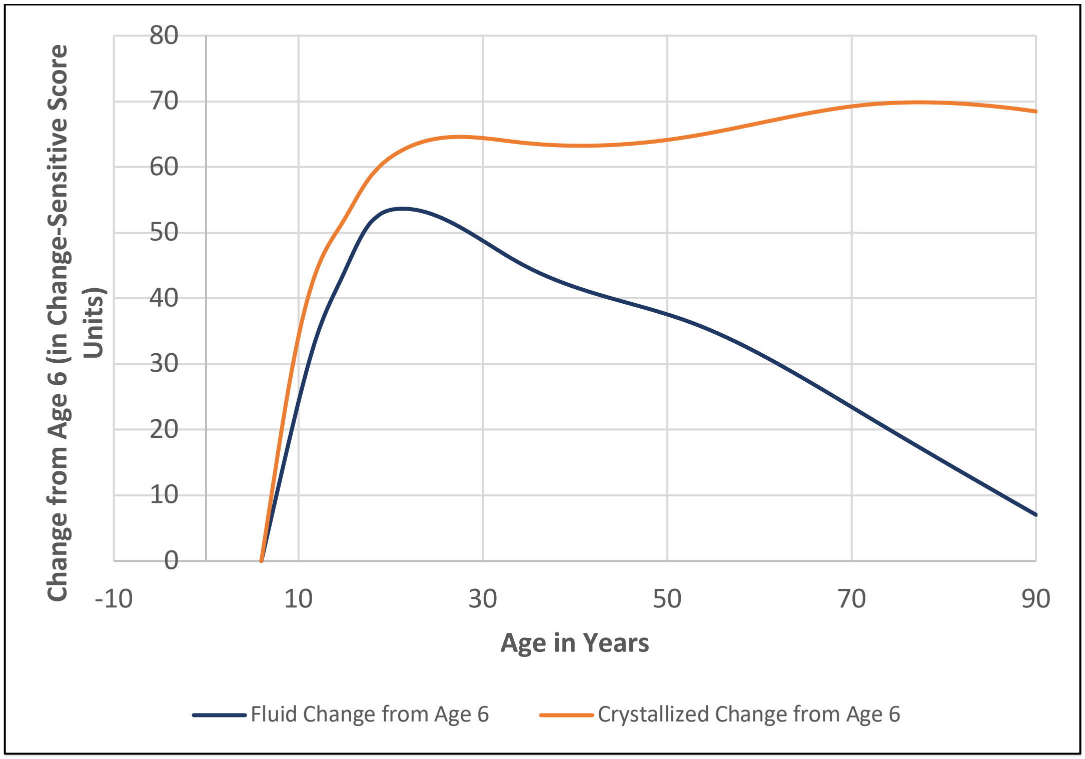


Figure 4.2: Cross-sectional growth curves for NIH Toolbox Fluid and Crystallized Composite scores

The growth and decline of the NIH Toolbox Fluid and Crystallized ability scores, both tests and composites, conform to theoretical expectations (Baltes, Staudinger, & Lindenberger, 1999).

Test and Composite Intercorrelations

An examination of the relationship between and among the NIHTB tests and composites can provide additional evidence relevant to the test content; namely, that the strength of the relationships among these scores varies in expected ways, given the understanding of the latent trait(s) underlying each score.

Tables 4.3, 4.4, and 4.5 contain correlations between the NIHTB test and composite scores for the total sample, and for children and adults separately.

Table 4.3
Intercorrelation Matrix for NIHTB Tests and Composites, Total Sample

NIHTB Composite / Test	*Crystallized Composite*	*Fluid Composite*	*Total Cognition Composite*	*PV*	*ORR*	*Flanker*	*DCCS*	*LSWM*	*PSM*	*PC*	*OSD*	*RAVLT*	*RAVLT Delay*	*VR*	*FNAME*
Crystallized Composite	1.00														
	(3236)														
Fluid Composite	0.35	1.00													
	(3227)	(3228)													
Total Cognition Composite	0.78	0.86	1.00												
	(3227)	(3227)	(3227)												
PV	0.78	0.30	0.62	1.00											
	(3235)	(3226)	(3226)	(3785)											
ORR	0.92	0.31	0.70	0.49	1.00										
	(3223)	(3220)	(3220)	(3363)	(3364)										
Flanker	0.27	0.73	0.62	0.22	0.25	1.00									
	(2860)	(2858)	(2857)	(3241)	(2981)	(3244)									
DCCS	0.21	0.77	0.63	0.18	0.18	0.55	1.00								
	(3049)	(3049)	(3048)	(3455)	(3175)	(3227)	(3458)								
LSWM	0.43	0.54	0.59	0.36	0.39	0.25	0.22	1.00							
	(3213)	(3209)	(3209)	(3475)	(3338)	(3078)	(3280)	(3477)							
PSM	0.24	0.54	0.48	0.22	0.20	0.19	0.18	0.29	1.00						
	(3201)	(3201)	(3200)	(3712)	(3326)	(3203)	(3415)	(3435)	(3715)						
PC	0.11	0.71	0.53	0.11	0.11	0.45	0.51	0.17	0.13	1.00					
	(3228)	(3226)	(3225)	(3514)	(3362)	(3113)	(3318)	(3472)	(3476)	(3517)					
OSD	0.26	0.48	0.46	0.20	0.25	0.31	0.35	0.34	0.33	0.29	1.00				
	(3190)	(3187)	(3187)	(3400)	(3299)	(3014)	(3209)	(3369)	(3368)	(3399)	(3401)				
RAVLT	0.31	0.33	0.39	0.26	0.27	0.14	0.19	0.36	0.34	0.11	0.35	1.00			
	(3204)	(3201)	(3201)	(3486)	(3337)	(3085)	(3289)	(3443)	(3448)	(3482)	(3375)	(3487)			
RAVLT Delay	0.25	0.28	0.31	0.22	0.21	0.11	0.14	0.27	0.37	0.07	0.33	0.70	1.00		
	(2323)	(2322)	(2322)	(2323)	(2322)	(2031)	(2191)	(2315)	(2303)	(2321)	(2310)	(2323)	(2323)		
VR	0.41	0.28	0.41	0.36	0.35	0.20	0.17	0.34	0.29	0.04	0.28	0.30	0.26	1.00	
	(3214)	(3211)	(3210)	(3641)	(3348)	(3219)	(3434)	(3457)	(3601)	(3497)	(3390)	(3471)	(2315)	(3644)	
FNAME	0.21	0.28	0.30	0.20	0.18	0.15	0.16	0.26	0.30	0.12	0.31	0.31	0.34	0.18	1.00
	(1582)	(1580)	(1580)	(1581)	(1581)	(1362)	(1480)	(1579)	(1563)	(1580)	(1570)	(1573)	(1461)	(1577)	(1582)

Note. DCCS = Dimensional Change Card Sort, Flanker = Flanker Inhibitory Control and Attention, FNAME = Face Name Associative Memory Exam, LSWM = List Sorting Working Memory, ORR = Oral Reading Recognition, OSD = Oral Symbol Digit, PC = Pattern Comparison Processing Speed, PSM = Picture Sequence Memory, PV = Picture Vocabulary, RAVLT = Rey Auditory Verbal Learning, RAVLT Delay = Rey Auditory Verbal Learning Delay, VR = Visual Reasoning.

Table 4.4
Intercorrelation Matrix for NIHTB Tests and Composites, Child (Ages 3 to 17) Sample

NIHTB Composite / Test	*Crystallized Composite*	*Fluid Composite*	*Total Cognition Composite*	*Early Childhood Composite*	*PV*	*ORR*	*Flanker*	*DCCS*	*LSWM*	*PSM*	*PC*	*OSD*	*RAVLT*	*RAVLT Delay*	*VR*	*SM*
Crystallized Composite	1.00															
	(1747)															
Fluid Composite	0.40	1.00														
	(1742)	(1743)														
Total Cognition Composite	0.79	0.86	1.00													
	(1742)	(1742)	(1742)													
Early Childhood Composite	0.41	0.87	0.70	1.00												
	(262)	(263)	(262)	(700)												
PV	0.71	0.34	0.59	0.43	1.00											
	(1747)	(1742)	(1742)	(698)	(2297)											
ORR	0.93	0.34	0.71	0.40	0.42	1.00										
	(1737)	(1736)	(1736)	(401)	(1878)	(1878)										
Flanker	0.28	0.72	0.62	0.69	0.22	0.26	1.00									
	(1590)	(1589)	(1588)	(624)	(1972)	(1713)	(1974)									
DCCS	0.22	0.76	0.61	0.53	0.18	0.18	0.52	1.00								
	(1664)	(1664)	(1663)	(657)	(2071)	(1791)	(1961)	(2073)								
LSWM	0.44	0.57	0.60	0.44	0.36	0.40	0.27	0.25	1.00							
	(1728)	(1727)	(1727)	(523)	(1991)	(1855)	(1812)	(1898)	(1992)							
PSM	0.27	0.55	0.49	0.69	0.25	0.22	0.19	0.17	0.29	1.00						
	(1733)	(1733)	(1732)	(681)	(2245)	(1859)	(1949)	(2047)	(1970)	(2247)						
PC	0.13	0.68	0.51	0.54	0.12	0.12	0.41	0.49	0.15	0.13	1.00					
	(1743)	(1743)	(1742)	(551)	(2030)	(1878)	(1846)	(1935)	(1990)	(2010)	(2032)					
OSD	0.32	0.48	0.48	0.48	0.23	0.30	0.29	0.33	0.35	0.32	0.27	1.00				
	(1715)	(1714)	(1714)	(459)	(1925)	(1825)	(1756)	(1836)	(1896)	(1909)	(1926)	(1926)				
RAVLT	0.33	0.32	0.38	0.24	0.27	0.28	0.11	0.19	0.34	0.31	0.10	0.35	1.00			
	(1727)	(1726)	(1726)	(540)	(2009)	(1861)	(1825)	(1913)	(1969)	(1989)	(2007)	(1909)	(2010)			
RAVLT Delay	0.32	0.30	0.36	--	0.30	0.26	0.09	0.17	0.29	0.40	0.05	0.39	0.72	1.00		
	(958)	(958)	(958)	(--)	(958)	(957)	(869)	(921)	(953)	(953)	(958)	(953)	(958)	(958)		
VR	0.43	0.31	0.43	0.37	0.35	0.37	0.22	0.18	0.35	0.29	0.06	0.33	0.29	0.34	1.00	
	(1732)	(1731)	(1730)	(687)	(2160)	(1867)	(1955)	(2054)	(1978)	(2138)	(2017)	(1919)	(1999)	(954)	(2162)	
SM	0.30	0.55	0.47	0.71	0.22	0.30	0.52	0.43	0.31	0.16	0.53	0.41	0.15	--	0.26	1.00
	(260)	(261)	(260)	(698)	(804)	(399)	(622)	(655)	(521)	(769)	(549)	(457)	(538)	(--)	(685)	(807)

Note. DCCS = Dimensional Change Card Sort, Flanker = Flanker Inhibitory Control and Attention, LSWM = List Sorting Working Memory, ORR = Oral Reading Recognition, OSD = Oral Symbol Digit, PC = Pattern Comparison Processing Speed, PSM = Picture Sequence Memory, PV = Picture Vocabulary, RAVLT = Rey Auditory Verbal Learning, RAVLT Delay = Rey Auditory Verbal Learning Delay, SM = Speeded Matching, VR = Visual Reasoning.

Table 4.5
Intercorrelation Matrix for NIHTB Tests and Composites, Adult (Ages 18+) Sample

NIHTB Composite / Test	*Crystallized Composite*	*Fluid Composite*	*Total Cognition Composite*	*PV*	*ORR*	*Flanker*	*DCCS*	*LSWM*	*PSM*	*PC*	*OSD*	*RAVLT*	*RAVLT Delay*	*VR*	*FNAME*
Crystallized Composite	--														
	(1489)														
Fluid Composite	0.30	--													
	(1485)	(1485)													
Total Cognition Composite	0.76	0.85	--												
	(1485)	(1485)	(1485)												
PV	0.86	0.27	0.66	--											
	(1488)	(1484)	(1484)	(1488)											
ORR	0.92	0.27	0.69	0.59	--										
	(1486)	(1484)	(1484)	(1485)	(1486)										
Flanker	0.25	0.74	0.63	0.21	0.23	--									
	(1270)	(1269)	(1269)	(1269)	(1268)	(1270)									
DCCS	0.20	0.79	0.65	0.18	0.18	0.61	--								
	(1385)	(1385)	(1385)	(1384)	(1384)	(1266)	(1385)								
LSWM	0.41	0.51	0.56	0.36	0.37	0.21	0.19	--							
	(1485)	(1482)	(1482)	(1484)	(1483)	(1266)	(1382)	(1485)							
PSM	0.20	0.54	0.47	0.17	0.17	0.19	0.20	0.30	--						
	(1468)	(1468)	(1468)	(1467)	(1467)	(1254)	(1368)	(1465)	(1468)						
PC	0.09	0.75	0.56	0.08	0.09	0.52	0.53	0.19	0.14	--					
	(1485)	(1483)	(1483)	(1484)	(1484)	(1267)	(1383)	(1482)	(1466)	(1485)					
OSD	0.19	0.48	0.43	0.16	0.18	0.33	0.38	0.31	0.33	0.32	--				
	(1475)	(1473)	(1473)	(1475)	(1474)	(1258)	(1373)	(1473)	(1459)	(1473)	(1475)				
RAVLT	0.28	0.34	0.39	0.25	0.26	0.19	0.17	0.38	0.38	0.12	0.33	--			
	(1477)	(1475)	(1475)	(1477)	(1476)	(1260)	(1376)	(1474)	(1459)	(1475)	(1466)	(1477)			
RAVLT Delay	0.20	0.27	0.28	0.17	0.18	0.12	0.13	0.26	0.35	0.09	0.29	0.69	--		
	(1365)	(1364)	(1364)	(1365)	(1365)	(1162)	(1270)	(1362)	(1350)	(1363)	(1357)	(1365)	(1365)		
VR	0.40	0.25	0.39	0.37	0.34	0.17	0.16	0.34	0.28	0.01	0.23	0.30	0.22	--	
	(1482)	(1480)	(1480)	(1481)	(1481)	(1264)	(1380)	(1479)	(1463)	(1480)	(1471)	(1472)	(1361)	(1482)	
FNAME	0.21	0.28	0.30	0.20	0.18	0.15	0.16	0.26	0.30	0.13	0.31	0.30	0.33	0.18	--
	(1486)	(1484)	(1484)	(1485)	(1485)	(1268)	(1384)	(1483)	(1467)	(1484)	(1474)	(1477)	(1365)	(1481)	(1486)

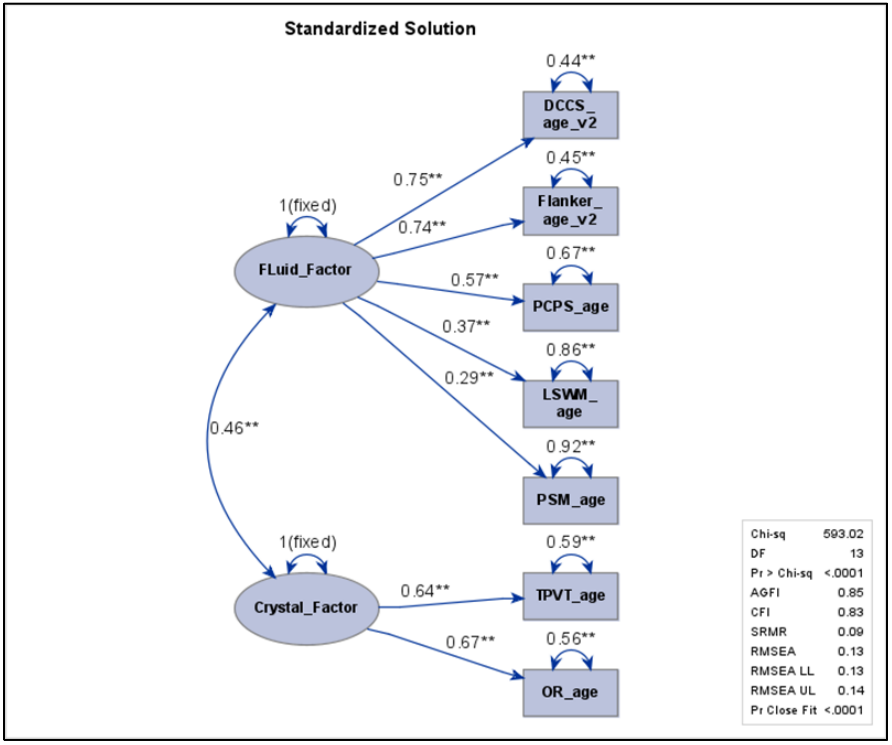
Note. DCCS = Dimensional Change Card Sort, Flanker = Flanker Inhibitory Control and Attention, FNAME = Face Name Associative Memory Exam, LSWM = List Sorting Working Memory, ORR = Oral Reading Recognition, OSD = Oral Symbol Digit, PC = Pattern Comparison Processing Speed, PSM = Picture Sequence Memory, PV = Picture Vocabulary, RAVLT = Rey Auditory Verbal Learning, RAVLT Delay = Rey Auditory Verbal Learning Delay, VR = Visual Reasoning.

Evidence Relevant to the Internal Battery Structure

Based on the intercorrelations in Tables 4.3 through 4.5, evidence for a hypothesized structure emerges. The overall relatively low to moderate intercorrelations imply that each measure represents a distinct ability but also represents a positive manifold of ability constructs. Additionally, moderate to strong relationships between the measures that share hypothesized constructs also become evident. For example, scores on comprehension-based measures tend to show moderate correlations. For example, Picture Vocabulary and Oral Reading Recognition, the two crystallized tests in the NIHTB, show moderate correlations (*r* = 0.49, 0.42, and 0.59 for the total, child, and adult samples, respectively). Similarly, executive function and processing speed measures tend to show moderate correlations; for example, Flanker, Dimensional Change Card Sort, and Pattern Comparison Processing Speed correlations range from 0.41 to 0.61 across the total, child, and adult samples). Based on these relationships and early exploratory analyses, we propose a two-factor solution that yields Fluid and Crystallized composites. This section contains details about the confirmatory analyses of this proposed structure.

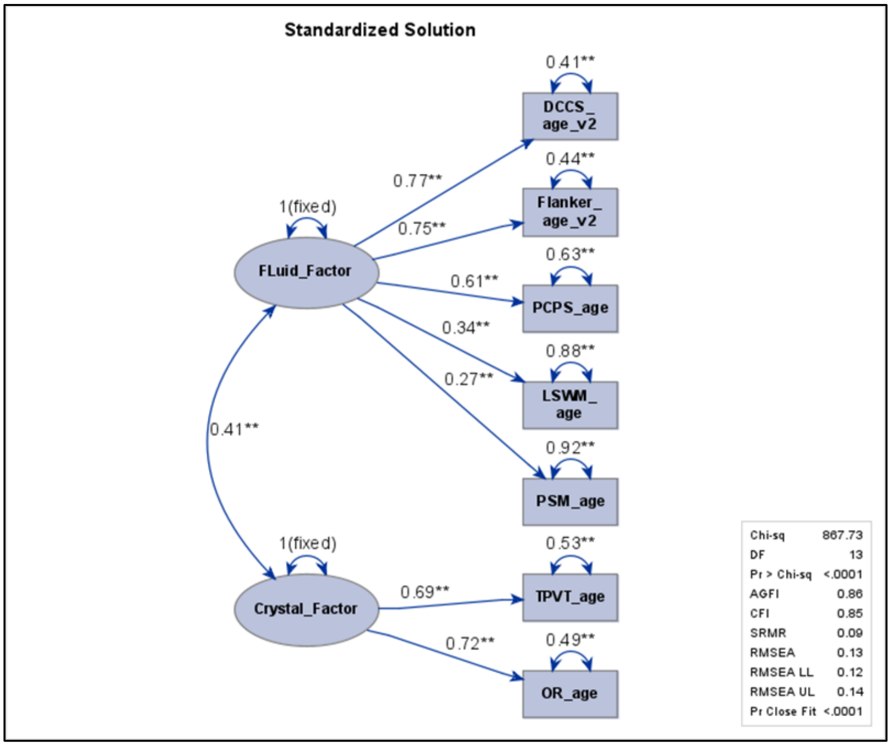
Factor Structure

The conformity of the V3 measures to the proposed structure of Fluid and Crystallized cognitive abilities can be evaluated via a traditional factor analysis that assumes latent variables. Confirmatory factor analyses were conducted assuming a two-factor model (one Fluid factor and one Crystallized factor) in youth (ages 4-20) and the total sample (ages 4-88) for age-adjusted scores (see Figures 4.3 and 4.4). All analyses used maximum likelihood or full information maximum likelihood estimation procedures.



Chi-sq	593.02
DF	13
Pr > Chi-sq	<.0001
AGFI	0.85
CFI	0.83
SRMR	0.09
RMSEA	0.13
RMSEA LL	0.13
RMSEA UL	0.14
Pr Close Fit	<.0001

Figure 4.3: Correlated two-factor model for Fluid and Crystallized composites using age-adjusted scores; child (3 to 20 years) sample



Chi-sq	867.73
DF	13
Pr > Chi-sq	<.0001
AGFI	0.86
CFI	0.85
SRMR	0.09
RMSEA	0.13
RMSEA LL	0.12
RMSEA UL	0.14
Pr Close Fit	<.0001

Figure 4.4: Correlated two-factor model for Fluid and Crystallized composites using age-adjusted scores; total (ages 4 to 85+) sample

Statistical criteria for goodness-of-fit values recommend that the Root Mean Square Error of Approximation (RMSEA) should have value of <.06 for a close fit and <.08 for a reasonable fit (Browne & Cudek, 1993; Joreskog & Sorbom, 1993; Hu & Bentler, 1999; Hooper, Coughlan, & Mullen, 2008). The RMSEA for the children and adolescent EFA and CFA showed a close fit (less than 0.06). The adult sample RMSEA showed a reasonable fit for the EFA and CFA with values less than 0.08.

The Standardized Root Mean Residual (SRMR) should have a value <0.05 for a good fit (Byrne, 1998; Hooper, Coughlan, & Mullen, 2008) and value <.08 for a reasonable fit (Hu & Bentler, 1999). The SRMR for children was 0.028 for the EFA and 0.044 for the CFA. The SRMR for adults was 0.026 for the EFA and 0.051 for the CFA. According to the SRMR, a reasonable fit was found for youth and adults.

The Comparative Fit Index (CFI) and the Tucker Lewis Index (TLI) values >0.95 show a good fit (Hu & Bentler, 1999; Hooper et al., 2008) and values < 0.90 to 0.95 show a reasonable fit (Kline, 2023).

The CLI index shows a good fit for the EFAs for youth and adults and a reasonable fit for the CFAs for youth and adults. The TLI index shows a good fit for reasonable fit for the CFA across all samples.

Due to the large sample sizes the Chi Square tests were not interpreted, as these tests are often statistically significant for large samples.

The Fluid and Crystallized factors showed significant correlations of 0.46 for the child sample (ages 4 to 20) and 0.41 for the total sample (ages 4 to 85+). Factor loadings for the child and adult samples are reported in Table 4.6. For both the child and the total samples, Picture Vocabulary and Oral Reading Recognition significantly loaded onto the Crystallized composite factor (ts > 30, ps < .0001), whereas Dimensional Change Card Sort, Flanker, Pattern Comparison, List Sorting Working Memory, and Picture Sequence Memory significantly loaded on the Fluid composite factor (see Table 4.6).

Table 4.6
Factor Loadings for the Child and Total Samples for a Correlated Two-Factor Model

		Child Sample (4 to 20 years)				*Total Sample (4 to 85+ years)*			
Factor	*Test*	*Factor Loading*	*SE*	*t*	*p*	*Factor Loading*	*SE*	*t*	*p*
Fluid	DCCS	0.75	0.02	45.7	<.0001	0.77	0.01	69.5	<.0001
	Flanker	0.75	0.02	45.4	<.0001	0.75	0.01	66.7	<.0001
	PC	0.57	0.02	30.0	<.0001	0.61	0.02	47.7	<.0001
	LSWM	0.37	0.02	16.3	<.0001	0.34	0.02	20.8	<.0001
	PSM	0.29	0.02	12.0	<.0001	0.27	0.02	16.2	<.0001
Crystallized	PV	0.64	0.03	20.2	<.0001	0.69	0.03	30.3	<.0001
	ORR	0.67	0.03	20.5	<.0001	0.72	0.02	30.7	<.0001

Note. DCCS = Dimensional Change Card Sort, Flanker = Flanker Inhibitory Control and Attention, LSWM = List Sorting Working Memory, ORR = Oral Reading Recognition, PC = Pattern Comparison Processing Speed, PSM = Picture Sequence Memory, PV = Picture Vocabulary.

Evidence Relevant to the Relationship of the NIHTB to Other Variables

Strong correlations with “gold standard” external measures that are widely used, highly reliable, and well-researched can provide evidence for the utility of test and composite score interpretations (AERA, APA, & NCME, 2014; Cizek, 2020). Due to their shared task demands with the NIHTB Cognition tests and/or similar theoretical bases, there are several commercially available test batteries that are appropriate gold-standard measures for evaluating both convergent and divergent validity for the NIHTB tests. It is especially important that convergent and divergent validity evidence be amassed for the new tests in the NIH Toolbox V3 Cognition domain (Visual Reasoning, Face Name Associative Memory Exam, Rey Auditory Verbal Learning, Rey Auditory Verbal Learning Delay, and Speeded Matching), as this evidence contributes to the understanding of how the interpretation of these new tests’ scores is similar to—and different from—the familiar interpretations of test scores from other batteries.

Relationship of NIH Toolbox Cognition Tests and Composites to Other Cognition Measures

During the NIH Toolbox V3 norming study, convergent validity studies were conducted with the following gold-standard batteries: the *Wechsler Adult Intelligence Scale, 4th Edition* (WAIS-IV; Wechsler, 2008); the *Wechsler Intelligence Scale for Children, 5th Edition* (WISC-V; Wechsler, 2014); the *Wechsler Preschool and Primary Scale of Intelligence, 4th Edition* (WPPSI-IV;

Wechsler, 2012); the *California Verbal Learning Test, 3rd Edition* (CVLT3; Delis, Kramer, Kaplan, & Ober, 2017); and the *Wechsler Memory Scale, 4th Edition* (WMS-IV; Wechsler, 2009). The results of these studies are described below.

Wechsler Adult Intelligence Scales, 4th Edition (WAIS-IV)

The *Wechsler Adult Intelligence Scale, 4th Edition* (WAIS-IV; Wechsler, 2008) is used to assess cognitive abilities in adolescents and adults ages 16 to 90. It comprises ten core subtests that yield a Full Scale IQ (FSIQ) score and four factor-based index scores: Verbal Comprehension, Perceptual Reasoning, Working Memory, and Processing Speed (Abdelhamid, Bassiouni, & Gómez-Benito, 2021). The WAIS-IV was standardized in the United States on 2,200 individuals between the ages of 16 to 90 (Wechsler, 2008).

Subtests that contribute to the WAIS-IV Full Scale IQ were administered to a randomly selected subsample of 180 adult NIH Toolbox V3 norming study participants (mean age = 44 years, *SD* = 18 years). The study sample was approximately half (53%) male, 56% white, and 25% Hispanic; approximately 67% of the examinees had attended at least some college. Table 4.7 contains the mean scores and standard deviations for the tests and composites from each battery, as well as the correlations between the scores. An examination of the means and standard deviations for each battery suggests that the sample was of average ability. All correlations are positive and range from 0.03 to 0.75.

Not unexpectedly, the NIH Toolbox Total Cognition Composite and the WAIS-IV FSIQ are highly correlated (0.74). Although the theoretical nomenclature and relative weighting of the component constructs varies slightly between the two batteries, both of these *g*-composite scores contain measures that generally fall into the categories of fluid reasoning, crystallized intelligence, memory, and processing speed. The correlation between the NIH Toolbox Crystallized Composite and the WAIS-IV Verbal Comprehension Index (VCI) is a moderate 0.67; the NIHTB Crystallized Composite is also moderately correlated with the WAIS-IV FSIQ (0.64), Perceptual Reasoning Index (PRI; 0.53), and Working Memory Index (WMI; 0.60) but less strongly correlated with the WAIS-IV Processing Speed Index (PSI; 0.30). This provides evidence for the interpretation of the NIHTB Crystallized Composite as an overall *g* score that does not rely on the contribution of fluid reasoning, working memory, or processing speed, which is useful for neuropsychological assessment of cognitive abilities that typically do not decline with age or brain insult or injury. This type of assessment is called a ‘hold’ test by some professionals when they want an estimate of a person’s function prior to a brain injury, such as a traumatic brain injury (Hook & Kuentzel, 2023). The collection of tests that comprise the NIHTB Fluid Composite is fairly heterogenous, including tests measuring nonverbal reasoning, processing speed, and working memory. Thus, it is not surprising that the NIHTB Fluid Composite is moderately correlated with the WAIS-IV FSIQ (0.56), WMI (0.52), and PSI (0.57). Interestingly, the NIHTB Fluid Composite has a weaker correlation (0.48) with the WAIS-IV PRI; however, the PRI contains three subtests that all measure aspects of visual and/or nonverbal reasoning, whereas the NIHTB Fluid Composite contains only one such test (Visual Reasoning).

Table 4.7
Correlations Between the NIH Toolbox V3 Tests and Composites and the WAIS-IV Subtests and Composites

			WAIS-IV														
			Composites					*Subtests*									
NIHTB Composites / Tests	*M*	*SD*	*FSIQ*	*VCI*	*PRI*	*WMI*	*PSI*	*Block Design*	*Similarities*	*Digit Span*	*Matrix Reasoning*	*Vocabulary*	*Arithmetic*	*Symbol Search*	*Visual Puzzles*	*Information*	*Coding*
Total Cognition Composite	98.9	14.7	0.74	0.60	0.62	0.69	0.54	0.49	0.45	0.54	0.51	0.62	0.56	0.43	0.55	0.46	0.53
Crystallized Composite	99	14.5	0.64	0.67	0.53	0.60	0.30	0.38	0.53	0.46	0.48	0.67	0.56	0.22	0.45	0.52	0.33
Fluid Composite	98.8	14.5	0.56	0.32	0.48	0.52	0.57	0.42	0.22	0.45	0.35	0.34	0.37	0.48	0.44	0.24	0.54
PV	99.1	12.5	0.67	0.73	0.56	0.57	0.37	0.41	0.59	0.36	0.50	0.70	0.55	0.30	0.49	0.61	0.37
ORR	99.4	14.4	0.60	0.60	0.49	0.61	0.26	0.35	0.48	0.51	0.46	0.62	0.55	0.18	0.41	0.46	0.29
Flanker	98.6	15.3	0.31	0.15	0.31	0.21	0.37	0.27	0.15	0.13	0.22	0.14	0.15	0.35	0.29	0.10	0.31
DCCS	98.6	13.2	0.43	0.32	0.31	0.35	0.50	0.34	0.29	0.30	0.15	0.32	0.25	0.42	0.31	0.22	0.48
LSWM	99.9	13.7	0.61	0.39	0.61	0.66	0.33	0.47	0.29	0.66	0.49	0.35	0.52	0.25	0.57	0.35	0.35
PSM	101	14.2	0.32	0.15	0.30	0.30	0.33	0.30	0.04	0.32	0.29	0.18	0.17	0.23	0.16	0.15	0.37
PC	98.2	14.1	0.31	0.11	0.20	0.31	0.45	0.13	0.08	0.20	0.12	0.20	0.21	0.42	0.26	0.03	0.38
OSD	99.5	12.0	0.61	0.28	0.56	0.45	0.75	0.51	0.24	0.41	0.42	0.33	0.29	0.63	0.49	0.16	0.75
RAVLT	100	15.2	0.37	0.24	0.30	0.41	0.33	0.19	0.06	0.35	0.32	0.28	0.36	0.28	0.21	0.24	0.30
RAVLT Delay	100	13.9	0.32	0.18	0.30	0.30	0.27	0.17	0.11	0.21	0.30	0.20	0.26	0.23	0.26	0.16	0.23
FNAME	99.8	13.9	0.29	0.20	0.28	0.24	0.31	0.11	0.14	0.17	0.33	0.22	0.27	0.25	0.24	0.14	0.31
VR	98.9	13.5	0.48	0.39	0.47	0.38	0.33	0.32	0.33	0.38	0.50	0.31	0.34	0.22	0.35	0.36	0.38
M			101	101	99.2	101	105	9.92	10.1	10.2	10.2	10.4	10.1	11	9.6	10.1	10.8
SD			13.6	12.2	13.7	14.4	13.4	2.65	2.17	2.69	2.94	2.43	2.88	2.89	2.96	3.12	2.61

Note. DCCS = Dimensional Change Card Sort, Flanker = Flanker Inhibitory Control and Attention, FNAME = Face Name Associative Memory Exam, LSWM = List Sorting Working Memory, ORR = Oral Reading Recognition, OSD = Oral Symbol Digit, PC = Pattern Comparison Processing Speed, PSM = Picture Sequence Memory, PV = Picture Vocabulary, RAVLT = Rey Auditory Verbal Learning, RAVLT Delay = Rey Auditory Verbal Learning Delay, VR = Visual Reasoning, FSIQ = Full Scale IQ, VCI = Verbal Comprehension Index, PRI = Perceptual Reasoning Index, WMI = Working Memory Index, PSI = Processing Speed Index. Due to incomplete test administrations for some participants, correlation *N*s ranged from 85 to 180. Means and standard deviations are shown in standard score units for all NIH Toolbox tests and composites and for WAIS-IV composites, and in scaled score units for WAIS-IV subtests. All correlations have been corrected for range restriction using Thorndike's Case 2 correction (Thorndike, 1947).

At the test level, there is a notably high correlation between the NIHTB Picture Vocabulary test and the WAIS-IV VCI (0.73), suggesting that the Picture Vocabulary test on its own is a strong measure of verbal or crystallized intelligence. There is also a high correlation between the NIHTB Oral Symbol Digit test and the WAIS-IV Coding subtest. This is not unexpected given that these two tests share very similar task demands. The correlation between NIHTB Oral Symbol

Digit and the WAIS-IV PSI is also high (0.75), which suggests that Oral Symbol Digit may be tapping into perceptual-verbal abilities in addition to processing speed. Other test pairs that share similar task demands and/or content also show moderate to high correlations; for example, NIHTB List Sorting Working Memory with WAIS-IV Digit Span (0.66), and NIHTB Picture Vocabulary with WAIS-IV Vocabulary (0.70).

Among the new tests in the NIH Toolbox, three of the four (Rey Auditory Verbal Learning, Rey Auditory Verbal Learning Delay, and Face Name Associative Memory Exam) showed low or negligible correlations with all WAIS-IV subtests and indices. This is not unexpected, as these new NIHTB tests measure constructs (verbal learning, recall, and associative memory, respectively) not purported to be measured by the WAIS-IV subtests included in this study. The new NIHTB Visual Reasoning test correlated moderately with the similar WAIS-IV Matrix Reasoning subtest (0.50) and the WAIS-IV PRI (0.47), providing support for the Visual Reasoning test as a measure of fluid and perceptual reasoning.

Wechsler Intelligence Scales for Children, 5th Edition (WISC-V)

The *Wechsler Intelligence Scale for Children, Fifth Edition* (WISC-V; Wechsler, 2014) is used to assess cognitive abilities in children ages 6 to 16. The battery contains 10 primary subtests, 6 secondary subtests, and 5 complementary subtests, which can be administered in different combinations to obtain a Full Scale IQ (FSIQ) score, five primary index scores (Verbal Comprehension Index, Visual Spatial Index, Fluid Reasoning Index, Working Memory Index, and Processing Speed Index). Several ancillary index scores are also available but are outside the scope of this discussion. The WISC-V was normed on a sample of 2,200 children between the ages of 6 and 16 in the United States.

The seven subtests that contribute to the WISC-V FSIQ were administered to a randomly selected subsample of 50 children who were NIH Toolbox V3 norming study participants (mean age = 11.2 years, *SD* = 3.19 years). The sample was 40% male, 51% White, and 29% Hispanic; the maternal education level of the sample was fairly high (67% some college or higher). Table 4.8 contains the mean scores and standard deviations for the tests and composites from each battery, as well as the correlations between the scores. An examination of the means and standard deviations for each battery suggests that the sample was of average ability. Correlations range from negative (-0.07) to moderate (0.65).

Among the highest correlations in Table 4.8 are those between the NIHTB Crystallized and Total Composites and the three WISC-IV composite scores. The NIHTB Total Composite score correlated in the 0.57 to 0.64 range with all WISC-IV composite scores. The NIHTB Fluid Composite correlations with all WISC-V composites were noticeably weaker (0.37 to 0.48); however, this is not unexpected given the relative heterogeneity of the NIHTB Fluid Composite, which includes tests of memory, processing speed, and executive functioning, in addition to the more traditional fluid reasoning Visual Reasoning test. The NIHTB Early Childhood composite is positive but relatively weakly correlated with all WISC-V composites; however, these correlations should be interpreted with caution due to the low *n*-counts for those cells.

Among the test-level correlations presented in Table 4.8, the NIHTB Picture Vocabulary test shows moderate correlations with all three WISC-V composites as well as the WISC-IV Similarities (0.63), Matrix Reasoning (0.50), and Vocabulary (0.60) tests. This is consistent with the results of the WAIS-IV study (Table 4.7) and suggests that the NIHTB Picture Vocabulary test is a strong overall general indicator of general intelligence in this age range. Other notable and relatively strong correlations exist between the NIHTB List Sorting Working Memory test and the WISC-V FSIQ (0.54) and VCI (0.63), and the Similarities (0.59), Vocabulary (0.58), and Figure Weights (0.51) subtests. As in the WAIS-IV study results with adults, List Sorting Working Memory in this age range appears to be a relatively mixed measure of working memory, verbal skills, and fluid reasoning.

Table 4.8
Correlations Between the NIH Toolbox V3 Tests and Composites and the WISC-V Subtests and Composites

			WISC-V Measures									
			Composites			*Subtests*						
NIHTB Composites / Tests	*M*	*SD*	*FSIQ*	*VCI*	*FRI*	*Block Design*	*Similarities*	*Matrix Reasoning*	*Digit Span*	*Coding*	*Vocabulary*	*Figure Weights*
Crystallized Composite	94.4	16.0	0.60	0.55	0.56	0.29	0.53	0.46	0.56	0.27	0.50	0.56
Fluid Composite	98.7	16.1	0.48	0.48	0.37	0.39	0.34	0.35	0.23	0.22	0.55	0.32
Total Composite	95.5	15.5	0.64	0.59	0.57	0.41	0.50	0.48	0.48	0.31	0.60	0.54
Early Childhood Composite	102.0	14.8	0.40	0.34	0.24	0.18	-0.07	0.35	0.22	0.42	0.58	0.04
PV	99.0	14.3	0.62	0.65	0.53	0.17	0.63	0.50	0.43	0.42	0.60	0.45
ORR	95.2	16.4	0.49	0.49	0.49	0.22	0.46	0.38	0.53	0.13	0.45	0.50
Flanker	100.0	16.0	0.32	0.20	0.25	0.36	0.05	0.26	0.08	0.23	0.29	0.19
DCCS	100.6	17.0	0.25	0.22	0.25	0.42	0.10	0.27	0.03	0.05	0.29	0.19
LSWM	98.3	16.5	0.54	0.63	0.41	0.23	0.59	0.23	0.38	0.20	0.58	0.51
PSM	96.8	13.5	0.37	0.37	0.18	0.14	0.29	0.24	0.19	0.45	0.40	0.08
SM	103.5	18.8	0.28	0.08	0.27	-0.05	-0.04	0.25	0.17	0.58	0.15	0.19
PC	100.7	13.6	0.24	0.28	0.25	0.26	0.15	0.29	0.08	0.01	0.35	0.18
OSD	97.7	11.4	0.55	0.42	0.55	0.44	0.38	0.50	0.56	0.46	0.40	0.49
RAVLT	100.4	16.1	0.27	0.29	0.22	0.09	0.23	0.16	0.34	0.28	0.30	0.22
VR	100.8	14.4	0.65	0.42	0.56	0.38	0.41	0.55	0.49	0.34	0.38	0.45
M			96.4	98.8	95.6	9.9	9.6	9.3	9.7	8.3	9.9	9.2
SD			15.2	14.3	16.3	3.0	2.7	3.1	3.0	2.8	2.9	3.2

Note. DCCS = Dimensional Change Card Sort, Flanker = Flanker Inhibitory Control and Attention, LSWM = List Sorting Working Memory, ORR = Oral Reading Recognition, OSD = Oral Symbol Digit, PC = Pattern Comparison Processing Speed, PSM = Picture Sequence Memory, PV = Picture Vocabulary, RAVLT = Rey Auditory Verbal Learning, SM = Speeded Matching, VR = Visual Reasoning, FSIQ = Full Scale IQ, VCI = Verbal Comprehension Index, FRI = Fluid Reasoning Index. Due to the limited age range for the NIHTB

Speeded Matching test and the Early Childhood Composite, *N*s for those two measures ranged from 12 to 13. All other correlation *N*s ranged from 45 to 50. Means and standard deviations are shown in standard score units for all NIH Toolbox tests and composites and for WISC-V composites, and in scaled score units for WISC-V subtests. All correlations have been corrected for range restriction using Thorndike's Case 2 correction (Thorndike, 1947).

Among the new NIHTB tests, Oral Symbol Digit has an expectedly moderate correlation with Digit Span (0.56) but a lower-than-expected correlation with Coding (0.46), given the shared task demands of these two tests. However, Oral Symbol Digit correlates moderately with both the WISC-V FSIQ (0.55) and FRI (0.55), providing evidence to support its use as a fluid reasoning measure in the NIH Toolbox battery. The NIHTB Rey Auditory Verbal Learning Test shows low correlations (.09 to 0.34) with all WISC-V subtests, with the highest of these being the 0.34 correlation with Digit Span, a working memory subtest. These relatively low correlations are not unexpected, given that the Rey Auditory Verbal Learning Test has different task demands than any of the WISC-V subtests. The new NIHTB Visual Reasoning test is highly correlated with the WISC-V FSIQ (0.65), FRI (0.56) and Matrix Reasoning subtest (0.55), providing evidence to support its use as a measure of fluid reasoning among 7- to 16-year-old individuals.

Wechsler Preschool & Primary Scales of Intelligence, 4th Edition (WPPSI-IV)

The *Wechsler Preschool and Primary Scale of Intelligence, Fourth Edition* (WPPSI-IV; Wechsler, 2012) is an individually administered early childhood intelligence test. The battery[3] contains 10 primary subtests and 5 secondary subtests, which can be administered in different combinations to obtain a Full Scale IQ score, five primary index scores (Verbal Comprehension Index, Visual Spatial Index, Fluid Reasoning Index, Working Memory Index, and Processing Speed Index). Several ancillary index scores are also available but are outside the scope of this discussion. The WPPSI-IV was normed on a sample of 1,700 children between the ages of 3 and 7 in the United States.

Subtests that contribute to the WPPSI-IV FSIQ were administered to a randomly selected subsample of 43 children who were NIH Toolbox V3 norming study participants (mean age = 5.2 years, *SD* = 0.81 years). The study sample was 54% male, 40% White, and 37% Hispanic. The maternal education level of the sample was fairly high; 63% had attended at least some college. Table 4.9 contains the mean scores and standard deviations for the tests and composites from each battery, as well as the correlations between the scores. An examination of the means and standard deviations for each battery suggests that the sample was of average ability. Correlations range from negative (-0.06) to high (0.84).

The NIHTB Early Childhood composite shows relatively low correlations with the WPPSI-IV FSIQ (0.44) and VCI (0.39); however, this is not unexpected, as the compositions of these composite

[3] The WPPSI-IV is divided into two distinct age bands (2 years, 6 months to 3 years, 11 months and 4 years, 0 months to 7 years, 7 months) corresponding to different subtest batteries due to significant cognitive ability and developmental changes during the age range covered. This study employed only the Ages 4:0 to 7:7 battery; therefore, the 2:6 to 3:11 battery features are not discussed here.

scores are quite distinct. Whereas the WPPSI-IV FSIQ score contains two subtests measuring verbal comprehension and one subtest each measuring visual spatial abilities, fluid reasoning, working memory, and processing speed, the NIHTB contains one verbal test (Picture Vocabulary), two tests measuring aspects of executive functioning (Flanker Inhibitory Control and Attention and Dimensional Change Card Sort), a processing speed test (Speeded Matching), and one episodic memory test (Picture Sequence Memory).

Table 4.9
Correlations Between the NIH Toolbox V3 Tests and Composites and the WPPSI-IV Subtests and Composites

			WPPSI-IV Measures							
			Composites		*Subtests*					
NIHTB Composites / Tests	*M*	*SD*	*FSIQ*	*VCI*	*Information*	*Similarities*	*Block Design*	*Matrix Reasoning*	*Picture Memory*	*Bug Search*
Early Childhood Composite	95.8	17.8	0.44	0.39	0.35	0.40	0.18	0.38	0.32	0.47
PV	98.7	14.8	0.58	0.71	0.68	0.66	0.14	0.42	0.44	0.45
ORR	96.5	18.5	0.71	0.61	0.56	0.56	0.00	0.57	0.66	0.55
Flanker	96.0	15.9	0.26	0.37	0.31	0.38	0.08	0.20	0.12	0.37
DCCS	101.3	20.9	0.03	0.08	-0.02	0.14	-0.04	-0.01	0.01	0.20
LSWM	97.9	11.9	0.62	0.60	0.62	0.54	0.12	0.62	0.63	0.63
PSM	98.5	14.5	0.41	0.32	0.26	0.32	0.42	0.44	0.29	0.26
PC	98.5	17.8	0.39	0.43	0.33	0.37	0.12	0.12	0.24	0.32
OSD	99.6	9.9	0.53	0.43	0.66	0.39	-0.06	0.70	0.84	0.65
RAVLT	99.0	14.6	0.63	0.41	0.36	0.42	0.28	0.36	0.46	0.45
VR	100.9	14.4	0.53	0.50	0.36	0.51	0.19	0.45	0.28	0.25
SM	101.1	14.6	0.35	0.27	0.33	0.23	0.32	0.04	0.17	0.60
M			93.1	92.0	8.8	8.1	8.7	9.7	8.6	9.1
SD			11.2	15.4	3.0	3.6	2.3	3.2	2.5	2.7

Note. DCCS = Dimensional Change Card Sort, Flanker = Flanker Inhibitory Control and Attention, FNAME = Face Name Associative Memory Exam, LSWM = List Sorting Working Memory, ORR = Oral Reading Recognition, OSD = Oral Symbol Digit, PC = Pattern Comparison Processing Speed, PSM = Picture Sequence Memory, PV = Picture Vocabulary, RAVLT = Rey Auditory Verbal Learning, RAVLT Delay = Rey Auditory Verbal Learning Delay, SM = Speeded Matching, VR = Visual Reasoning, FSIQ = Full Scale IQ, VCI = Verbal Comprehension Index. Because the NIHTB Oral Reading and Oral Symbol digit tests were not appropriate for the youngest children in this study, Ns for those correlations are 18 or 19 (Oral Reading), or between 20 and 23 (Oral Symbol Digit). All other correlation *N*s ranged from 32 to 50. Means and standard deviations are shown in standard score units for all NIH Toolbox tests and composites and for WPPSI-IV composites, and in scaled score units for WPPSI-V subtests. All correlations have been corrected for range restriction using Thorndike's Case 2 correction (Thorndike, 1947).

Although the NIHTB Early Childhood Composite is not highly correlated with the WPPSI-FSIQ in this age range, several of the NIHTB tests do show moderate correlations with the FSIQ score. These include Picture Vocabulary (0.58), List Sorting Working Memory (0.62), Rey Auditory Verbal Learning (0.63), and Visual Reasoning (0.53). Among these, only Picture Vocabulary contributes to the NIHTB Early Childhood composite score, which suggests that even the NIHTB measures that do not contribute to that composite are strong indicators of general cognitive ability in this age range. Oral Reading Recognition also shows a high correlation with FSIQ (0.71); however, as noted in the table footnote this correlation is based on a relatively low *n* for the NIHTB Oral Reading Recognition test, should therefore be interpreted with caution.

Among the new NIHTB tests, Speeded Matching has a moderate correlation with WPPSI-IV Bug Search (0.60), a processing speed subtest, and relatively low correlations (0.04 to 0.33) with all other WPPSI-IV subtests. This provides evidence for the use of the NIHTB Speeded Matching test as a measure of processing speed in this age range. The NIHTB Oral Symbol Digit test has a similarly moderate (0.65) correlation with Bug Search, but also shows moderate to high correlations (0.39 to 0.84) with several other WPPSI-IV subtests and a negative correlation (-0.06) with Block Design. The Oral Symbol Digit correlations should be interpreted with caution, however, due to the low *n*s for those cells. Finally, the new NIHTB Visual Reasoning test is moderately correlated with the WPPSI-IV FSIQ (0.53), VCI (0.50), and Similarities subtest (0.51). However, it shows a lower correlation with the WPPSI-IV Matrix Reasoning subtest (0.45). This may be related to the higher verbal reasoning requirements of the NIHTB Visual Reasoning test at this age range.

California Verbal Learning Test, 3rd Edition **(CVLT-3)**

The *California Verbal Learning Test, Third Edition* (CVLT3; Delis, Kramer, Kaplan, & Ober, 2017) is a clinical and research battery designed to assess the strategies and processes involved in learning, recalling, and recognizing verbal information in adolescents and adults ages 16 to 90 years. In the CVLT-3, the examinee is first asked to recall a list of 16 words immediately after presentation on five trials; the list includes four words in each of four semantic categories. An interference list of 16 different words is then presented for one trial. The interference trial is followed by Short Delay Free and cued Recall trials of the first word list. After a 20-minute delay, another free recall and cued recall trial is administered, followed by a Yes/No Recognition trial (Holdnack, Drozdick, & Courville, 2017). The CVLT-3 was normed on 700 individuals ages 16 to 90 in the United States. The CVLT-3 was administered to a randomly selected subsample of 51 adults (mean age = 46 years, *SD* = 19.3 years) who were participants in the NIH Toolbox V3 norming study. The sample was 47% male, 55% White, and 24% Hispanic. Sixty-one percent of the sample had attended at least some college. The goal of this study was to investigate the relationship of the scores from the NIHTB memory tests—and primarily the new NIH Toolbox Rey Verbal Auditory Learning and Rey Verbal Auditory Learning Delay tests—with the CVLT-3. Because they were most relevant to the test tasks from the NIHTB Rey Verbal Auditory Learning tests, only the following scores were included in this study: Trials 1 to 5 total score, Interference Trial Free Recall score, and Short-Delay Free Recall Response Total.

Table 4.10 contains the mean scores and standard deviations for the tests and composites from the NIH Toolbox and the relevant CVLT-3 trial scores, as well as the correlations between the measures. An examination of the means and standard deviations for each battery suggests that the sample was of average ability. Correlations range from negligible (0.03) to moderate (0.65). The Rey Auditory Verbal Learning and Rey Auditory Verbal Learning Delay tests are moderately correlated with the CVLT-3 Trials 1 to 5 standard score (0.54 and 0.65, respectively), followed by the Short-Delay Free Recall Scaled Score (0.50 and 0.59, respectively) and the Interference Trial Recall Scaled Score (0.50 and 0.54, respectively). Overall, the NIHTB RAVLT Delay shows slightly higher correlations with the CVLT-3 than does the RVALT. These correlations with the CVLT-3 Trials 1 to 5 standard score are similar to those between several other NIHTB tests and the CVLT-3 Trials 1 to 5 standard score. While one might expect the RAVLT-CVLT-3 correlations to be stronger, these results could be explained by the differences in task presentation between the NIHTB RAVLT and the CVLT-3 item format. The NIHTB RAVLT word list is not divided into semantic categories; therefore, it may preclude the use of some strategies that examinees can utilize on the CVLT-3 trials. Not surprisingly, NIHTB Picture Sequence Memory—a test that has similar task demands to the RVALT but uses visual stimuli—is also moderately correlated with the CVLT-3 Trials 1 to 5 standard score (0.55) and Short-Delay Free Recall Scaled Score (0.64). Notably, there are low correlations between the NIHTB crystallized tests and composites and executive functioning tests and the CVLT-3 scores; this provides divergent validity evidence and supports the utility of the NIHTB RAVLT scores as measures of verbal learning, immediate recall, and delayed recall.

Table 4.10
Correlations Between the NIH Toolbox V3 Tests and Composites and the CVLT-3

NIHTB Composites / Tests	*M*	*SD*	*CVLT-3 Scores*		
			Trials 1–5 Standard Score	*Interference Trial Free Recall Scaled Score*	*Short-Delay Free Recall Scaled Score*
Crystallized Composite	101	12.4	0.62	0.26	0.07
Fluid Composite	97.3	14.9	0.41	0.43	0.50
Total Composite	98.4	14	0.33	0.44	0.41
PV	101	13.3	0.11	0.18	0.03
ORR	99.9	10.4	0.04	0.35	0.15
Flanker	98.5	12.6	0.06	0.25	0.14
DCCS	96.2	11.8	0.19	0.31	0.30
LSWM	101	13.5	0.29	0.49	0.29
PSM	102	14.6	0.55	0.48	0.64
PC	96.3	11.9	0.50	0.24	0.56
OSD	99.3	12.7	0.55	0.50	0.55
RAVLT	100	16.2	0.54	0.50	0.50
RAVLT Delay	99.3	13.5	0.65	0.54	0.59
FNAME	101	11.2	0.43	0.32	0.45
VR	101	12.6	0.38	0.35	0.44
M			97.39	9.65	9.57
SD			16.24	2.49	3.39

Note. DCCS = Dimensional Change Card Sort, Flanker = Flanker Inhibitory Control and Attention, FNAME = Face Name Associative Memory Exam, LSWM = List Sorting Working Memory, ORR = Oral Reading Recognition, OSD = Oral Symbol Digit, PC = Pattern Comparison Processing Speed, PSM = Picture Sequence Memory, PV = Picture Vocabulary, RAVLT = Rey Auditory Verbal Learning, RAVLT Delay = Rey Auditory Verbal Learning Delay, VR = Visual Reasoning. Correlation *N*s ranged from 47 to 51. Means and standard deviations are shown in standard score units for all NIH Toolbox tests and composites. All correlations have been corrected for range restriction using Thorndike's Case 2 correction (Thorndike, 1947).

Wechsler Memory Scale, 4th Edition (WMS-IV)

The *Wechsler Memory Scale, Fourth Edition* (WMS-IV; Wechsler, 2009) is an individually administered assessment of memory functioning for adolescents and adults ages 16 to 90. The WMS-IV assesses the learning and memory constructs of encoding, storage, and retrieval (Drozdick, Raiford, Wahlstrom, & Weiss, 2018). The WMS-IV battery includes six primary subtests and one optional subtest. Two WMS-IV batteries are available: the Adult battery, designed for adolescents and adults ages 16 to 69; and the Older Adult battery, designed for use with adults ages 65 to 90. For NIHTB V3 validation, separate studies were conducted that included the WMS-IV Verbal Paired Associates subtest from their respective batteries. In Verbal Paired Associates I, the examinee is read a series of word pairs, and the asked to provide the second word when each first word is read aloud. This is repeated three times, for a total of four learning trials. In Verbal Paired Associates II, the examinee completes three delayed memory tasks. In the delayed recall task, the examinee recalls the second word in each pair. In the recognition task, the examinee recalls the second word pair and is asked whether it is a pair from the list. In the word recall task, the examinee is asked to recall as many of the individual words as possible.

WMS-IV Adult Battery Verbal Paired Associates I and II Subtests

The Verbal Paired Associates I and II subtests from the WMS-IV Adult Battery were administered to a randomly selected subsample of 102 adults (mean age = 43 years, *SD* = 14.6 years) who participated in the NIH Toolbox V3 norming study. The sample was 39% male, 46% White, and 12% Hispanic. Fifty-three percent of the participants had attended at least some college.

Table 4.11 contains the mean scores and standard deviations for the tests and composites from the NIH Toolbox and the WMS-IV Verbal Paired Associates I and Verbal Paired Associates II scaled scores, as well as the correlations between the NIHTB and WMS-IV scores. An examination of the means and standard deviations for each battery suggests that the sample was of average ability. Correlations range from negligible (0.03) to moderate (0.65). Among the NIHTB tests, the Picture Sequence Memory (0.51 and 0.51), RAVLT (0.54 and 0.60), FNAME (0.57 and 0.53), and RAVLT Delay (0.63 and 0.65) tests are the most highly correlated with the WMS-IV VPA I and II tests, respectively. The List Sorting Working Memory test also showed a moderate correlation (0.60) with the WMS-IV VPA II subtest, but a slightly weaker correlation (0.47) with the VPA I subtest.

Table 4.11

Correlations Between the NIH Toolbox V3 Tests and Composites and the WMS-IV Adult Battery Verbal Paired Associates Subtests

NIHTB Composites / Tests	*M*	*SD*	*WMS-IV Verbal Paired Associates I*	*WMS-IV Verbal Paired Associates II*
Crystallized Composite	100	14.2	0.45	0.39
Fluid Composite	98.6	14.3	0.36	0.39
Total Composite	99.2	14.6	0.47	0.45

PV	99.8	13.6	0.46	0.39
ORR	100	13.4	0.38	0.35
Flanker	98.7	15.3	0.08	0.06
DCCS	97.8	11.9	0.19	0.25
LSWM	101	12.4	0.47	0.60
PSM	102	15.1	0.51	0.51
PC	97.5	13.1	0.06	0.03
OSD	100	12.5	0.44	0.39
RAVLT	102	16.9	0.54	0.60
RAVLT Delay	101	15	0.63	0.65
FNAME	100	13.3	0.57	0.53
VR	101	12.9	0.42	0.41
M			10.7	11.11
SD			2.78	2.91

Note. DCCS = Dimensional Change Card Sort, Flanker = Flanker Inhibitory Control and Attention, FNAME = Face Name Associative Memory Exam, LSWM = List Sorting Working Memory, ORR = Oral Reading Recognition, OSD = Oral Symbol Digit, PC = Pattern Comparison Processing Speed, PSM = Picture Sequence Memory, PV = Picture Vocabulary, RAVLT = Rey Auditory Verbal Learning, RAVLT Delay = Rey Auditory Verbal Learning Delay, VR = Visual Reasoning.

WMS-IV Older Adult Battery Verbal Paired Associates I and II Subtests

The Verbal Paired Associates I and II subtests from the WMS-IV Older Adult Battery were administered to a randomly selected subsample of 26 adults (mean age = 72 years, *SD* = 5.9 years) who were participants in the NIH Toolbox V3 norming study. The sample was 35% male, and 77% White. There were no Hispanic individuals in the sample. Thirty-one percent of the participants had attended at least some college.

Table 4.12 contains the mean scores and standard deviations for the tests and composites from the NIH Toolbox and the WMS-IV Older Adult Battery Verbal Paired Associates I and Verbal Paired Associates II scaled scores, as well as the correlations between the NIHTB and WMS-IV scores. An examination of the means and standard deviations for each battery suggests that the sample was of low-average to average ability. Correlations range from negative (-0.46) to moderate (0.57). The correlations between the NIHTB memory tests and the WMS-IV Older Adult VPA I and II subtest scores were similar to those in Table 4.11 for the WMS-IV Adult Battery, except that the FNAME test showed a weaker correlation (0.28) with the VPA I score among the older adults.

Table 4.12
Correlations Between the NIH Toolbox V3 Tests and Composites and the WMS-IV Older Adult Battery Verbal Paired Associates Subtests

NIHTB Composites / Tests	*M*	*SD*	*WMS-IV Verbal Paired Associates I*	*WMS-IV Verbal Paired Associates II*
Crystallized Composite	94.3	11.8	0.24	0.18
Fluid Composite	90.4	15.4	0.44	0.22
Total Composite	90.7	14.2	0.40	0.23
PV	96	11.6	0.22	0.17

ORR	95.2	11.6	0.21	0.15
Flanker	93.7	13.7	0.06	-0.13
DCCS	92.4	13	0.33	0.19
LSWM	94.6	17.3	0.49	0.37
PSM	97.1	14.7	0.56	0.40
PC	92.5	11.1	-0.17	-0.46
OSD	97.2	8.9	0.22	0.13
RAVLT	101	14.3	0.56	0.54
FNAME	97.4	10.9	0.28	0.51
VR	96.1	11.6	0.40	0.44
RAVLT Delay	99.6	11.8	0.57	0.51
M			11.38	12.6
SD			2.62	2.35

Note. DCCS = Dimensional Change Card Sort, Flanker = Flanker Inhibitory Control and Attention, FNAME = Face Name Associative Memory Exam, LSWM = List Sorting Working Memory, ORR = Oral Reading Recognition, OSD = Oral Symbol Digit, PC = Pattern Comparison Processing Speed, PSM = Picture Sequence Memory, PV = Picture Vocabulary, RAVLT = Rey Auditory Verbal Learning, RAVLT Delay = Rey Auditory Verbal Learning Delay, VR = Visual Reasoning.

Relationship Between NIH Toolbox V3 and V2

As described in Chapter 1, the NIH Toolbox battery has evolved significantly since it was normed as a stand-alone desktop application and first released as a web-based desktop assessment system in 2012. Shortly after its initial release, the battery was adapted for iPad administration. iPad administration presented several advantages for users, including increased portability, offline administration, and minimized reliance on custom hardware. However, it also introduced several substantive differences in the user experience, some of which had the potential to impact examinee performance. For example, there are differences in screen size and response modes (e.g., tapping responses on the iPad versus using directional keys on the keyboard in the desktop version), and in the presentation of instructions and items. Additionally, there are inherent differences in the way that the two types of devices handle the capture of response times that are integral to the speeded tests in the NIHTB Cognition battery. To assess the impact of the mode-of-administration differences between a desktop browser to an iPad, an equivalency study was conducted in 2016. While the results of this study informed updates to the scoring algorithms for several of the Cognition tests (Northwestern University, 2017); the underlying test norms were not updated at that time.

The goal of examining the relationship between the NIHTB Version 2 (V2) and NIHTB Version 3 (V3) in the current study was to understand how changes to the test items, instructions, scoring, and norming demographics may affect evidence of concordance across the test and composite scores from the two versions. Data were collected in parallel with the larger NIHTB V3 norming study in June through September of 2021.

Sample

Participants were recruited for the V3/V2 study from the larger NIHTB V3 norming study. Specifically, participants from the norming study were asked if they were interested in taking part in supplemental research studies. Participants who answered in the affirmative and who met the demographic requirements for the studies were eligible for assignment into supplemental studies. From this pool of individuals, a sample of 150 participants was randomly selected to participate in the NIHTB V2 study.

Table 4.13 contains a detailed breakdown of participant's demographics for the overall study sample and separately for children and adult participants. The participants were between the ages of 6 and 79 (*M* = 30 years; *SD* = 22.46 years), and about half of them (51.3%) were male. The largest proportion (44.0%) of these participants reported that their highest level of education (or maternal education for those younger than 18 years) was a high school diploma or a GED. A majority of participants (86.7%) reported their racial identity as White and, regardless of race, most participants (87.3%) reported being not Hispanic or Latino.

Table 4.13
NIHTB V3/V2 Study Sample Demographics

	All Participants (N = 150)	*Children (n = 74)*	*Adults (n = 76)*
Age			
Mean (SD)	30.38 (22.46)	11.11 (3.10)	49.14 (16.42)
Range	[6, 79]	[6, 16]	[21, 79]
	% (*n*)	**% (*n*)**	**% (*n*)**
Sex Assigned at Birth			
Female	48.70 (73)	47.30 (35)	50.00 (38)
Male	51.30 (77)	52.70 (39)	50.00 (38)
Gender			
Female	48.70 (73)	47.30 (35)	50.00 (38)
Male	51.30 (77)	52.70 (39)	50.00 (38)
Racial Identity			
White or Caucasian	86.70 (130)	81.10 (60)	92.10 (70)
Black or African American	7.33 (11)	10.80 (8)	3.95 (3)
Asian	4.00 (6)	5.41 (4)	2.63 (2)
Multiracial or More Than One Race	1.33 (2)	2.70 (2)	0.00 (0)
Other	0.67 (1)	0.00 (0)	0.67 (1)
Ethnic Identity			
Hispanic / Latino (Any Race)	12.70 (19)	20.30 (15)	5.26 (4)
Not Hispanic / Latino (Any Race)	87.30 (131)	79.70 (59)	94.70 (72)
Highest Level of Education (or Mother's Highest Level of Education for child participants)			
Less than HS	1.33 (2)	1.35 (1)	1.32 (1)
HS Diploma or GED	44.00 (66)	36.50 (27)	51.30 (39)
Some College	26.00 (39)	25.70 (19)	26.3 (20)
College or Bachelor's Degree (4-year degree)	15.30 (23)	16.20 (12)	14.50 (11)
Graduate or Professional Degree (Any Level)	13.30 (20)	20.30 (15)	6.58 (5)

Measures and Procedure

Participants were administered the English versions of the NIHTB V3 Cognition Battery and supplementary Cognition tests, followed by NIHTB V2 Cognition Battery and Standing Balance test. Of these measures, this study focused on the seven measures used to create the V3 Total

Composite scores. These included: Dimensional Change Card Sort, Flanker Inhibitory Control and Attention, List Sorting Working Memory, Oral Reading Recognition, Pattern Comparison Processing Speed, Picture Sequence Memory, and Picture Vocabulary.

Analyses

Concurrence between the NIHTB V3 and NIHTB V2 Cognition Batteries was examined through Spearman Rho correlations. These analyses were completed using the full sample and separately for the child and adult samples across the seven measures and the age-relevant composite scores. For the NIHTB V2, unadjusted (uncorrected) scores and age-adjusted standard scores were used for analyses. For the NIHTB V3, change-sensitive scores (CSSs) and age-adjusted scores were used for analyses. Both uncorrected scores and CSSs are scores that operate on different scales.

In addition to examining the correlations between NIHTB V2 and V3 measures, interclass correlations were examined for age-adjusted composite scores.

Results

Table 4.14 presents the Spearman Rho correlations between the NIHTB V3 and V2 test and composite scores. The V3 CSSs are strongly correlated (0.76 to 0.92) with V2 uncorrected scores for all composites, suggesting that the raw-scoring algorithms for the two editions tend to rank-order examinees in the same manner based on overall Fluid, Crystallized, and Total Cognition scores. The Oral Reading Recognition, Picture Vocabulary, and Pattern Comparison Processing Speed V3 CSS scores also show high to very high positive correlations with their V2 uncorrected counterpart scores. This is not unexpected, as these three tests underwent very few substantive changes in the V3 revision. Correlations are weaker, but still in the moderate to high range, for the Dimensional Change Card Sort, Flanker Inhibitory Control and Attention, and List Sorting Working Memory tests. These tests underwent more significant workflow changes in the V3 revision[4], including changes to test instructions, practice items, item timing, and scoring algorithms. It is not unexpected that these changes may have had a slightly differential impact on examinee performance, resulting in slightly lower correlations than the tests that underwent fewer changes. For DCCS and Flanker, examinees taking the V2 and V3 versions of these tests will encounter the same live items in the same order, so these tests may be slightly more susceptible to practice effects on repeat administrations than, for example, Oral Reading Recognition and Picture Vocabulary, which are administered via a CAT algorithm. If practice effects were differential, this could help explain the slightly weaker correlations between V3 CSS and V2 uncorrected scores. The correlations between the Picture Sequence Memory V3 CSSs and V2 uncorrected scores for both children (0.41 and 0.58, respectively) are noticeably lower than for the other tests. Again, this is not unexpected given the more extensive changes the Picture Sequence Memory test underwent in the V3 revision. Additionally, due to the administration routing rules in the V3 version, it is possible that some examinees in this study may have seen similar pictures presented in slightly different orders than in the V2 and V3 versions of the test. This may have resulted in a *negative* practice effect for some examinees,

[4] For details on the V3 changes to individual tests, see Chapter 2.

whereby those who encountered certain combinations of V3 and V2 items may have actually done *worse* on the V2 items if they relied on (different, and therefore incorrect) knowledge of item content from the V3 item presentation. Further analysis is required to determine the extent to which a possible negative practice effect may have resulted in lower scores for a subset of examinees who were administered specific item combinations; if this is true, then those practice effects (and resulting impacts on V3/V2 correlations) may be less pronounced for examinees who are administered those versions with a longer delay between administrations.

Correlations for the V3 and V2 age-adjusted standard scores in Table 4.14 are generally similar but slightly lower than the CSS correlations for the overall sample and for adults. For children, however, several correlations are noticeably lower. This is also not unexpected, as the continuous norming procedures utilized in the V3 norming study had a much larger impact on the way that normative scores were derived for children than for adults, compared with the procedures used for the V2 norming.[5]

Table 4.14
Spearman Rho Correlations Between the NIHTB V3 and V2 Test and Composite Scores

	V3 Change Sensitive Scores & V2 Uncorrected Scores			***V3 Age Adjusted Scores & V2 Age Adjusted Scores***		
Composites / Tests	***All Participants***	***Children***	***Adults***	***All Participants***	***Children***	***Adults***
Total Composite	0.92	0.92	0.88	0.70	0.66	0.78
Early Childhood Composite	0.79	0.79	–	0.63	0.63	–
Fluid Composite	0.85	0.84	0.87	0.62	0.54	0.73
Crystallized Composite	0.92	0.90	0.76	0.73	0.74	0.74
DCCS	0.68	0.66	0.62	0.47	0.41	0.60
Flanker	0.66	0.71	0.52	0.53	0.57	0.49
LSWM	0.70	0.68	0.70	0.57	0.55	0.58
ORR	0.90	0.91	0.71	0.72	0.78	0.69
PSM	0.50	0.41	0.58	0.40	0.29	0.47
PV	0.88	0.86	0.64	0.65	0.61	0.74
PC	0.83	0.87	0.80	0.66	0.62	0.71

Note. DCCS = Dimensional Change Card Sort, Flanker = Flanker Inhibitory Control and Attention, LSWM = List Sorting Working Memory, ORR = Oral Reading Recognition, PC = Pattern Comparison Processing Speed, PSM = Picture Sequence Memory, PV = Picture Vocabulary.

Table 4.15 presents the intraclass correlations between the V3 and V2 age-adjusted composite scores. For the total sample and for the adult group, the Crystallized Composite scores show the highest ICCs (0.80 and 0.80, respectively). As discussed above, this is not unexpected given that the two tests contributing to that composite score—Picture Vocabulary and Oral Reading Recognition—both underwent minimal changes in the V3 revision. The Crystallized Composite correlation for children (0.72) is slightly lower than for adults. The Fluid Composite score shows weaker, but still moderate, V3/V2 correlations for both children (0.53) and adults (0.64). The Early Childhood Composite correlation (0.81) is strong. The relatively low correlations for the Fluid Composite scores for children (0.53) and adults

[5] The V3 norming procedures, including how they differed from the V2 norming procedures, are discussed in Chapter 3.

(0.64) are likely due to the more significant V3 revisions to the tests that contribute to the Fluid Composite (as discussed above).

Table 4.15
Intraclass Correlations Between the V3 and V2 Age-Adjusted Composite Sores

Score	*Participant Group*	*V3 Age-Adjusted Scores & V2 Age-Adjusted Scores ICC*
Early Childhood Composite	All	–
	Adult	–
	Child	0.81
Fluid Composite	All	0.59
	Adult	0.64
	Child	0.53
Crystallized Composite	All	0.80
	Adult	0.80
	Child	0.72
Total Composite	All	0.72
	Adult	0.79
	Child	0.69

Implications for Use of the NIHTB in Longitudinal or Pooled Research Studies

The NIHTB was originally developed to address the need for a common metric, or "common currency" for comparing neuropsychological constructs between studies or within longitudinal studies (Gershon, Wagster, Hendrie, Fox, Cook, & Nowinski, 2013). In the time since its initial V1 publication, the NIHTB has been cited in over 450 published papers, and has been used to assess neurologic and behavioral functioning among participants in clinical samples in over 200 published studies (Fox et al., 2022). While continuity and comparability of measured variables is crucial for the analysis of data in large studies, the reality of changing population demographics is a very real challenge faced by test developers and test users. Flynn (1984) famously noted that mean IQ scores of Americans have been rising steadily over the past several generations, based on the norms for major intelligence batteries; this underscores the need for norm-referenced tests to undergo regular updating. The *Standards for Educational and Psychological Testing* (2014) clearly prescribe that test developers must "renorm [tests] with sufficient frequency to permit continued accurate and appropriate score interpretations" (p. 104). For norm-referenced score interpretations, this implies that the norms must reflect the ability of the population from which individual research participants are drawn. Proportional changes in race, ethnicity, education level, and other demographic variables can impact the interpretation of norm-referenced scores, which are in essence population-derived rank-order metrics. Indeed, an inspection of the 2020 and 2010 census statistics reveal an increase in the proportion of the U.S. population that is non-White and non-Hispanic from 36.3% to 42.2%, and an increase in the Hispanic (any race) proportion from 16.3% to 18.7%. Although the overall level of education among adults did not increase since the original NIHTB norming study, the changes in education levels between racial and ethnic subgroups has changed significantly. For instance, from 2012 to 2022 the percentage of adults aged 25 years and older in the U.S. with a college degree or higher increased from 34.5% to 41.8% for the White, non-Hispanic group, from 21.2% to 27.6% for the Black group, from 51.0% to 59.3% for the Asian group, and from 14.5% to 20.9% for the Hispanic group (US Census Bureau, 2023).

As these population changes have occurred gradually in the time since the 2012 NIHTB V1 publication, scores from research studies collected with the NIHTB Cognition prior to the V3 release necessarily contain differential amounts of construct-irrelevant variance related to sampling error; in other words, more recent studies will contain a greater mismatch between the study participants and their respective normative (2010) reference groups, due to population shifts from 2012 to 2022.

In addition to changing population demographics, another challenge facing test developers and test users in the age of digital assessments is the impact of changing technology. As a case in point, between the original 2010 norming study and the public release of V1 in 2012, the NIHTB had already undergone a significant technology update from a standalone desktop program to a browser-based desktop program. The 2015 NIHTB V2 release as an iPad app represented another generation of removal from the delivery format represented by the test scoring algorithms and norms; indeed, as described above, this transition resulted in large score differences on some tests that required adjustment formulas (Northwestern University, 2017). Similar to the effect of shifts on normative scores, changes to the test delivery mode, administration procedures, and scoring algorithms can introduce differences between scores administered via different modes.

The V3 renorming of the NIHTB Cognition battery essentially amounts to a “reset” in the concordance between the current V3 test format and reference population, and the scores derived from the NIHTB Cognition battery. For long-term longitudinal studies spanning NIHTB versions, the individual-level score error associated with the aging norm tables and shifting administration modes (i.e., from stand-alone desktop, to browser, to iPad) across time will likely cancel itself out in large samples. In other words, at the group level, we expect fewer noticeable overall differences in scores collected before and after the renorming than we would see at the individual participant level. However, each researcher needs to review their study protocol against the information contained in this manual to determine the extent to which their future study data may be impacted by the V3 update. As Table 4.14 shows, uncorrected (for age variance) scores from V3 and V2 are highly correlated at the composite level. Researchers who rely on these scores for their unit of analysis can feel comfortable that the overall Fluid, Crystallized, and Total Composite scores from different versions of NIHTB are adequately parallel in their interpretations. Researchers who rely on individual test scores, or who rely solely on age-corrected test scores for child samples (which are less highly correlated across editions V3 and V2 due to differences in norming procedures), will need to consider very carefully the V3 updates described in this manual to determine the extent to which those changes may impact the interpretations of their participant scores over time. In some cases, it may be necessary for researchers to conduct comparability analyses using their own study data to determine whether and to what extent their V2 NIHTB scores need to be adjusted to harmonize with V3 scores for group-level data analysis.

References

Abdelhamid, G. S. M., Bassiouni, M. G. A., & Gómez-Benito, J. (2021). Assessing Cognitive Abilities Using the WAIS-IV: An Item Response Theory Approach. *International journal of environmental research and public health*, *18*(13), 6835. https://doi.org/10.3390/ijerph18136835

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Baltes, P. B., Staudinger, U. M., & Lindenberger, U. (1999). LIFESPAN PSYCHOLOGY: Theory and Application to Intellectual Functioning. *Annual Review of Psychology*, *50*(1), 471–507. https://doi.org/10.1146/annurev.psych.50.1.471

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen and J. S. Long (Eds.), Testing structural equation models (pp. 136-162). Newbury Park, CA: Sage.

Byrne, B. M. (1998). *Structural Equation Modeling With Lisrel, Prelis, and Simplis Basic Concepts, Applications, and Programming*. Routledge.

Cai, L. (2014). Lord–Wingersky Algorithm Version 2.0 for Hierarchical Item Factor Models with Applications in Test Scoring, Scale Alignment, and Model Fit Testing. *Psychometrika*, *80*(2), 535–559. https://doi.org/10.1007/s11336-014-9411-3

Carlson, S. M., Zelazo, P. D., & Faja, S. (2013). Executive Function. In Zelazo, P.D.(Ed.), *The Oxford Handbook of Developmental Psychology, Vol. 1: Body and Mind.* New York: Oxford University Press.

Carroll, J. B. (1993). *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. Cambridge University Press.

Cizek, G. J. (2020). *Validity*. Routledge.

DeBell, M., & Krosnick, J. A. (2009). Computing Weights for American National Election Study Survey Data. ANES Technical Report Series. No. NES012427

Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (2017). *California Verbal Learning Test* (3rd Ed.). San Antonio, TX: Pearson.

Drozdick, L. W., Raiford, S. E., Wahlstrom, D., & Weiss, L. G. (2018). The Wechsler Adult Intelligence Scale—Fourth Edition and the Wechsler Memory Scale—Fourth Edition. In D.P. Flanagan & McDonough, E.M. (Eds.) *Contemporary Intellectual Assessment: Theories, Tests, and Issues.* New York: Guilford.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin*, *95*(1), 29–51. https://doi.org/10.1037/0033-2909.95.1.29

Fox, R. S., Zhang, M., Amagai, S., Bassard, A., Dworak, E. M., Han, Y. C., Kassanits, J., Miller, C. H., Nowinski, C. J., Giella, A. K., Stoeger, J. N., Swantek, K., Hook, J. N., & Gershon, R. C. (2022). Uses of the NIH Toolbox® in Clinical Samples: A Scoping Review. *Neurology: Clinical Practice*, *12*(4), 307–319. https://doi.org/10.1212/CPJ.0000000000200060

Gershon, R. C., Wagster, M. V., Hendrie, H. C., Fox, N. A., Cook, K. F., & Nowinski, C. J. (2013). NIH Toolbox for Assessment of Neurological and Behavioral Function. *Neurology*, *80*(Issue 11, Supplement 3), S2–S6. https://doi.org/10.1212/wnl.0b013e3182872e5f

Hessl, D., Sansone, S. M., Berry-Kravis, E., Riley, K., Widaman, K. F., Abbeduto, L., Schneider, A., Coleman, J., Oaklander, D., Rhodes, K. C., & Gershon, R. C. (2016). The NIH Toolbox Cognitive Battery for intellectual disabilities: three preliminary studies and future directions. *Journal of Neurodevelopmental Disorders*, *8*(1). https://doi.org/10.1186/s11689-016-9167-4

Hodes, R. J., Insel, T. R., & Landis, S. C. (2013). The NIH Toolbox: Setting a standard for biomedical research. *Neurology*, *80*(Issue 11, Supplement 3), S1–S1. https://doi.org/10.1212/wnl.0b013e3182872e90

Holdnack, J. A., Drozdick, L. W., & Courville, T. G. (2017). Substitution of CVLT–3 Scores for WMS–IV Verbal Paired Associates Scores. San Antonio, TX: Pearson.

Hook & Giella. (2023). *NIH Toolbox V3 Administration Manual.*

Hook, J. N., & Kuentzel, J. G. (2023). *Pediatric Neuropsychology.* In E. K Hodges, J.G. Kuentzel, & J.N. Hook (Eds.). (pages 2-16) New York, NY: Routledge Taylor Francis Group.

Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural Equation Modelling: Guidelines for Determining Model Fit Structural equation modelling: guidelines for determining model fit. Electron J Bus Res Methods. 2008; 6: 53–60

Hu, L., & Bentler, P. M. (1999). Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria versus New Alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Huang, S., & Cai, L. (2021). Lord–Wingersky Algorithm Version 2.5 with Applications. *Psychometrika*, *86*(4), 973–993. https://doi.org/10.1007/s11336-021-09785-y

Joreskog & Sorbom (2022). LISREL 12.

Kaat, A. J., McKenzie, F. J., Shields, R. H., LaForte, E., Coleman, J., Michalak, C., & Hessl, D. R. (2021). Assessing processing speed among individuals with intellectual and developmental disabilities: A match-to-sample paradigm. *Child Neuropsychology*, *28*(1), 1–13. https://doi.org/10.1080/09297049.2021.1938987

Kline, R. B. (2023). *Principles and Practice of Structural Equation Modeling*. Guilford Publications.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT True-Score and Equipercentile Observed-Score "Equatings." *Applied Psychological Measurement*, *8*(4), 453–461. https://doi.org/10.1177/014662168400800409

McKenzie, F., Drayton, A., Shields, R., Dakopolos, A., Glassman, D., Kaat, A., Coleman, J., Thompson, T., Sansone, S., Riley, K., Berry-Kravis, E., Gershon, R., Hessl, D. (2023). *National Institutes of Health Toolbox® V3 Cognition Battery Supplemental Administration Manual for Intellectual and Developmental Disabilities. UC Davis MIND Institute Translational Psychophysiology and Assessment Laboratory.*

Northwestern University. (2017). *NIH Toolbox V2 Cognition Analysis Guide.*

Rehabilitation Act of 1973, Pub. L. No. 93-112, § 508, 29 U.S.C. § 794d (1998), as amended by Pub. L. No. 110-325, § 102(a), 122 Stat. 3553 (2008); Pub. L. No. 115-233, § 501, 132 Stat. 1918 (2018).

Rey, A. (1941) L'examen psychologique dans les cas d'encephopathie traumatique. *Archives de Psychologie*, 28, 286-340. Corwin, J. and Bylsma, F.W., Translated (1993) *The Clinical Neuropsychologist*, 7, 4-9.

Schneider, W. J., & McGrew, K. S. (2018). The Cattell–Horn–Carroll theory of cognitive abilities. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (4th ed., pp. 73–163). The Guilford Press.

Thorndike, R. L. (1947). *Research problems and techniques.* (Rep. No. 3 AAF Aviation Psychology Program Research Reports). Washington, DC: U.S. Government Printing Office.

Timmerman, M. E., Voncken, L., & Albers, C. J. (2020). A tutorial on regression-based norming of psychological tests with GAMLSS. *Psychological Methods*. https://doi.org/10.1037/met0000348

US Census Bureau. (2023, February 16). *Census Bureau Releases New Educational Attainment Data*. Census.gov. https://www.census.gov/newsroom/press-releases/2023/educational-attainment-data.html

Wechsler, D. (2008). WAIS-IV administration and scoring manual (4th ed.). San Antonio, TX: Pearson.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale* (4th ed.). San Antonio, TX: Pearson.

Wechsler, D. (2009). *Wechsler Memory Scale* (4th ed.). San Antonio, TX: Pearson.

Wechsler, D. (2012). *Wechsler Preschool and Primary Scale of Intelligence* (4th ed.). San Antonio, TX: Pearson.

Wechsler, D. (2014). *Wechsler Intelligence Scale for Children* (5th ed.). San Antonio, TX: Pearson.

Weintraub, S., Bauer, P. J., Zelazo, P. D., Wallner-Allen, K., Dikmen, S. S., Heaton, R. K., Tulsky, D. S., Slotkin, J., Blitz, D. L., Carlozzi, N. E., Havlik, R. J., Beaumont, J. L., Mungas, D., Manly, J. J., Borosh, B. G., Nowinski, C. J., & Gershon, R. C. (2013). I. NIH TOOLBOX COGNITION BATTERY (CB): INTRODUCTION AND PEDIATRIC DATA. *Monographs of the Society for Research in Child Development*, *78*(4), 1–15. https://doi.org/10.1111/mono.12031

Appendix A: Change-Sensitive Score Summary Statistics for the NIHTB V3 Norming Sample

Appendix Table A1 contains sample *n*s and Change-Sensitive Score (CSS) means, *SD*s, and *SEM*s for age groups from the V3 norming sample for all NIHTB tests. Appendix Table A2 contains the summary statistics for composite CSS scores. For ages 3 to 19, where growth of the underlying abilities is relatively rapid, the summary statistics are provided for single years of age. The reported adult age groups are 20 to 21 years, 22 to 29 years, followed by 10-year groups from ages 30 to age 79. Ages 80 to 84 are reported together, and all examinees older than 85 are combined into a single 85+ group.

Table A.1
CSS Summary Statistics for NIHTB Tests, by Age Group

Age Group	*Statistic*	*DCCS*	*FNAME*	*Flanker*	*LSWM*	*ORR*	*OSD*	*PC*	*PSM*	*PV*	*RAVLT*	*RAVLT Delay*	*SM*	*VR*
Age 3	*n*	92	—	—	—	—	—	—	92	112	—	—	108	105
	M_{CSS}	468.9	—	—	—	—	—	—	402.8	463.9	—	—	442.2	470.7
	SD_{CSS}	5.83	—	—	—	—	—	—	20.33	10.17	—	—	8.24	6.85
	SEM_{CSS}	NA	—	—	—	—	—	—	18.08	9.13	—	—	7.68	5.91
Age 4	*n*	134	—	122	—	141	—	—	142	147	—	—	146	142
	M_{CSS}	474.4	—	458.5	—	372.1	—	—	427.5	472.6	—	—	451	475.8
	SD_{CSS}	8.92	—	10.52	—	27.47	—	—	30.66	9.5	—	—	11.4	9.34
	SEM_{CSS}	8.21	—	9.51	—	26.4	—	—	29.18	8.48	—	—	10.95	8.54
Age 5	*n*	137	—	128	133	146	95	144	145	147	142	—	148	144
	M_{CSS}	478.6	—	466.8	457.2	392.2	453.2	474.9	447.1	478.3	474.8	—	458.7	480.6
	SD_{CSS}	7.36	—	10.44	16.1	33.8	14.96	10.46	27.32	9.92	12.1	—	12.47	8.05
	SEM_{CSS}	6.75	—	9.41	14.27	33.04	14.26	10.04	26.34	9.05	11.54	—	11.97	6.94
Age 6	*n*	137	—	133	131	141	116	144	134	144	141	—	144	143
	M_{CSS}	484.2	—	476.8	468.9	428.5	467.9	481.4	467.3	485.5	479.5	—	473.5	486.9
	SD_{CSS}	9.23	—	13.17	19.11	37.25	13.34	12.58	18.79	7.67	12.64	—	18.25	8.25
	SEM_{CSS}	8.51	—	11.9	17.68	36.78	12.91	12.21	17.74	6.68	12.13	—	17.81	7.26
Age 7	*n*	137	—	131	144	143	136	145	144	145	140	—	144	143
	M_{CSS}	488	—	482.9	481.5	463.9	474.9	486.6	473.2	490.4	487.1	—	484.3	490.8
	SD_{CSS}	10.6	—	12.89	18.96	29.47	16.17	12.47	16.91	7.6	13.11	—	20.44	8.98
	SEM_{CSS}	9.76	—	11.69	17.62	28.8	15.83	12.09	16.1	6.59	12.67	—	20.01	8
Age 8	*n*	139	—	136	144	146	141	147	145	147	146	—	145	144
	M_{CSS}	492.2	—	490.3	489.8	475.7	487.1	492.7	488.3	494.3	491.2	—	497.7	495.5
	SD_{CSS}	10.76	—	15.31	17.57	26.03	19.11	15.7	20.91	7.79	11.98	—	21.8	8.5
	SEM_{CSS}	9.91	—	13.86	16.2	25.56	18.82	15.38	20.29	6.77	11.5	—	21.37	7.39
Age 9	*n*	140	—	136	146	149	145	149	148	149	147	125	—	148
	M_{CSS}	496.5	—	492.3	498.6	492.5	496.5	495.6	493.6	498.1	497.8	499	—	499.7
	SD_{CSS}	13.6	—	13.91	19.83	17.2	22.96	18.14	19.99	8.03	13.9	12.16	—	8.77
	SEM_{CSS}	12.52	—	12.61	18.57	16.54	22.7	17.87	19.29	6.99	13.49	10.21	—	7.49
Age 10	*n*	145	—	143	151	152	151	153	152	153	152	137	—	151
	M_{CSS}	499.4	—	499.2	501.4	498	502.2	501.6	498.2	500.2	499.8	500	—	499.8
	SD_{CSS}	15.79	—	15.87	19.22	17.19	21.47	18.81	20.25	9.3	13.39	12.72	—	8.69
	SEM_{CSS}	14.54	—	14.38	17.95	16.56	21.19	18.55	19.6	8.31	12.96	10.76	—	7.46
Age 11	*n*	144	—	139	150	149	150	151	151	152	148	140	—	151
	M_{CSS}	508.3	—	504.3	507.6	504.3	515.4	510.6	502.1	508.6	502.9	501.6	—	500.6
	SD_{CSS}	17.92	—	15.42	18.74	16.89	22.31	23.13	22.14	7.72	12.7	12.08	—	7.45
	SEM_{CSS}	16.47	—	13.95	17.44	16.23	22.01	22.91	21.47	6.71	12.26	10.06	—	6.19
Age 12	*n*	145	—	139	151	151	149	151	149	152	151	143	—	152
	M_{CSS}	513.1	—	508.1	512.6	508.9	520.9	517.7	506.5	510.5	504.4	503.6	—	501.7
	SD_{CSS}	22.57	—	17.08	16.58	17.12	25.23	25.73	20.04	9.09	12.32	12.47	—	8.21
	SEM_{CSS}	20.79	—	15.49	15.14	16.52	24.96	25.53	19.25	8.17	11.87	10.4	—	7.03
Age 13	*n*	142	—	133	146	149	146	149	148	149	147	136	—	148
	M_{CSS}	519.1	—	510.6	511.7	510.4	526.5	522.1	501.3	512	504.6	502.7	—	501.7

	SD_{CSS}	23.68	—	18.48	17.87	13.91	24.86	27.05	22.52	9.18	12.9	14.34	—	8.14
	SEM_{CSS}	21.83	—	16.73	16.52	13.22	24.58	26.85	21.86	8.28	12.46	12.33	—	6.91
Age 14	*n*	141	—	135	145	146	145	147	146	147	146	141	—	146
	M_{CSS}	523.3	—	516.2	513.3	515.3	531.9	530.8	507.6	513.9	505.3	503.3	—	503.6
	SD_{CSS}	26.45	—	18.64	20.64	15.04	26.19	29.65	22.66	8.13	13.69	13.58	—	8.35
	SEM_{CSS}	24.19	—	16.75	19.37	14.36	25.92	29.46	21.9	7.14	13.26	11.54	—	7.15
Age 15	*n*	156	—	146	160	161	161	161	161	161	160	154	—	161
	M_{CSS}	526	—	515.3	515.9	518.6	541.1	534.8	508.1	515.6	508.1	505.7	—	503.5
	SD_{CSS}	26.63	—	19.75	17.56	13.39	25.65	30.2	21.18	8.97	13.71	15.01	—	8.09
	SEM_{CSS}	24.65	—	17.96	16.16	12.66	25.37	30	20.33	8.05	13.29	12.91	—	6.9
Age 16	*n*	139	—	128	145	145	145	145	144	145	145	143	—	145
	M_{CSS}	527.5	—	516.6	515.2	519.6	544	534.7	512.2	515.5	507.1	503.5	—	504.4
	SD_{CSS}	28.45	—	20.28	17.59	14.54	26.07	29.19	23.21	9.31	12.53	13.16	—	9.2
	SEM_{CSS}	26.19	—	18.35	16.2	13.88	25.79	28.99	22.38	8.4	12.07	11.1	—	8.03
Age 17	*n*	140	—	129	149	149	149	149	149	149	148	144	—	147
	M_{CSS}	535.3	—	519.7	516.5	523.8	544.9	537	508	518.3	506.2	504.4	—	506
	SD_{CSS}	39.53	—	21.07	17.32	14.49	27.29	30.03	22.81	9.74	12.43	14.05	—	7.61
	SEM_{CSS}	36.33	—	19.08	15.9	13.81	27.01	29.83	22.08	8.85	11.98	12.01	—	6.34
Age 18	*n*	96	95	95	96	96	96	96	96	97	96	96	—	96
	M_{CSS}	536.4	500.1	519.7	519.3	525	549.5	547.5	507.7	518.2	509.3	506	—	506.2
	SD_{CSS}	27.6	11.54	19.38	20.96	15.22	27.19	30.74	24.84	10.2	14.9	16.07	—	10.34
	SEM_{CSS}	23.92	9.68	17.43	19.7	14.54	26.9	30.5	24.11	9.23	14.48	14	—	9.16
Age 19	*n*	144	145	132	144	145	145	145	145	145	145	141	—	144
	M_{CSS}	546.9	501.2	525	521.4	529.7	555.4	551.6	513.5	520.3	510.8	507	—	505.4
	SD_{CSS}	42.39	12.55	20.18	16.37	11.84	26.18	31.46	21.28	8.92	13.4	15.77	—	7.8
	SEM_{CSS}	39.69	10.74	18.29	14.86	11	25.9	31.19	20.37	7.97	12.97	13.61	—	6.57
Age 20–21	*n*	147	157	139	157	157	157	157	157	157	156	150	—	155
	M_{CSS}	541.5	502.3	523.5	519.4	531.9	558.5	555.1	516.7	523.3	510.3	508.1	—	506.6
	SD_{CSS}	35.76	12.26	18.98	15.62	12.7	25.32	28.75	21.43	8.7	13.85	15.55	—	8.67
	SEM_{CSS}	32.91	10.41	17.2	14.09	11.93	25.04	28.44	20.51	7.68	13.44	13.34	—	7.45
Age 22–29	*n*	152	162	138	161	162	162	162	161	162	160	158	—	162
	M_{CSS}	534	503	514.6	516.7	529.4	549.3	540.7	507.4	522.7	506.2	503.3	—	504.6
	SD_{CSS}	31.96	13.1	17.82	16.98	13.43	25.48	32.88	24.55	9.42	14.64	15.76	—	9.15
	SEM_{CSS}	29.49	11.31	16.11	15.55	12.69	25.19	32.65	23.7	8.47	14.24	13.76	—	7.99
Age 30–39	*n*	158	169	141	168	169	168	168	167	169	169	163	—	169
	M_{CSS}	534.8	499.5	516.9	515	527.2	549.9	534.2	502.7	526.2	506.6	502.7	—	505.1
	SD_{CSS}	33.48	12.68	17.33	17.46	15.91	28.4	31.88	23.61	11.14	12.87	15.3	—	8.94
	SEM_{CSS}	30.94	10.91	15.75	16.07	15.26	28.13	31.67	22.81	10.29	12.43	13.33	—	7.76
Age 40–49	*n*	159	173	152	173	173	172	173	173	173	172	164	—	173
	M_{CSS}	525.1	497	510.9	512	529.9	543.1	524.9	497.3	526.9	502.8	497.8	—	503.2
	SD_{CSS}	27.02	10.82	16.66	15.64	15.32	23.8	30.32	23.1	11.55	13.08	13.7	—	9.44
	SEM_{CSS}	24.77	8.93	15.11	14.15	14.65	23.48	30.12	22.41	10.69	12.65	11.82	—	8.29
Age 50–59	*n*	157	172	144	172	172	172	172	170	172	172	159	—	172
	M_{CSS}	518.5	492.1	508	506.4	528	532.2	518.3	492	528.1	500.2	494.5	—	500.3
	SD_{CSS}	23.37	11.52	16.17	15.34	14.88	21.61	26.54	22.04	11.44	12.96	12.18	—	8.62
	SEM_{CSS}	21.56	9.72	14.7	13.85	14.18	21.28	26.34	21.35	10.59	12.54	10.24	—	7.43
	n	146	168	139	168	168	166	168	166	168	167	151	—	168

Age 60–69	M_{CSS}	512.4	489.9	504.2	503.2	529.9	524.1	507.2	480.6	530.9	497	492.5	—	500.4
	SD_{CSS}	17.44	11.71	13.62	16.69	12.5	26.31	19.95	19.18	11.08	12.33	11.39	—	7.7
	SEM_{CSS}	16.08	9.94	12.43	15.29	11.71	26.01	19.7	18.55	10.16	11.88	9.42	—	6.45
Age 70–79	*n*	140	157	124	157	156	157	157	156	157	155	134	—	157
	M_{CSS}	504.4	486.9	499.5	499.2	535	515.7	500.9	467.2	533.5	495.8	492.2	—	497.9
	SD_{CSS}	12.93	10.52	12.17	15.26	13.46	20.16	16.3	15.86	12.02	12.05	12.77	—	7.33
	SEM_{CSS}	11.88	8.66	10.94	13.77	12.65	19.81	15.99	14.81	11.16	11.59	10.82	—	6.04
Age 80–84	*n*	132	133	118	135	134	129	133	128	136	132	107	—	133
	M_{CSS}	498.1	481.9	492.3	494	531.4	501.3	488.3	464.5	532.9	490	487	—	496.8
	SD_{CSS}	12.19	11.4	15.14	14.43	16.62	19.25	16.98	15.74	14.08	11.83	11.02	—	7.24
	SEM_{CSS}	11.1	9.59	13.71	12.91	15.99	18.94	16.69	14.42	13.3	11.37	8.86	—	5.91
Age 85+	*n*	64	64	54	65	65	59	64	55	65	63	44	—	62
	M_{CSS}	493.9	480	486.9	490.2	535.4	496.5	482.8	456.2	532.2	486.7	485	—	495.6
	SD_{CSS}	15.41	12.84	18.76	18.24	15.53	20.6	16.66	17.57	14.46	9.59	8.27	—	7.87
	SEM_{CSS}	14.42	11.13	17.09	16.93	14.8	20.32	16.38	16.17	13.65	9.03	5.95	—	6.54

Note. DCCS = Dimensional Change Card Sort, Flanker = Flanker Inhibitory Control and Attention, FNAME = Face Name Associative Memory Exam, LSWM = List Sorting Working Memory, ORR = Oral Reading Recognition, OSD = Oral Symbol Digit, PC = Pattern Comparison Processing Speed, PSM = Picture Sequence Memory, PV = Picture Vocabulary, RAVLT = Rey Auditory Verbal Learning, RAVLT Delay = Rey Auditory Verbal Learning Delay, SM = Speeded Matching, VR = Visual Reasoning.

Table A.2
CSS Summary Statistics for NIHTB Composites, by Age Group

Age Group	*Statistic*	*Early Childhood*	*Crystallized*	*Fluid*	*Total Cognition*
Age 3	n	—	—	—	—
	M_{CSS}	—	—	—	—
	SD_{CSS}	—	—	—	—
Age 4	n	145	—	—	—
	M_{CSS}	455.02	—	—	—
	SD_{CSS}	10.65	—	—	—
Age 5	n	148	—	—	—
	M_{CSS}	464.87	—	—	—
	SD_{CSS}	9.61	—	—	—
Age 6	n	144	—	—	—
	M_{CSS}	476.96	—	—	—
	SD_{CSS}	10.07	—	—	—
Age 7	n	145	145	145	144
	M_{CSS}	483.72	477.31	482.44	479.87
	SD_{CSS}	9.25	15.94	9.78	11.31
Age 8	n	146	147	147	147
	M_{CSS}	492.57	485.06	490.57	487.81
	SD_{CSS}	10.18	14.73	9.77	10.15
Age 9	n	—	149	149	149
	M_{CSS}	—	495.28	495.21	495.24
	SD_{CSS}	—	10.59	11.76	9.61
Age 10	n	—	153	153	153
	M_{CSS}	—	499.06	500.03	499.54
	SD_{CSS}	—	11.6	12.16	10.05
Age 11	n	—	152	151	151
	M_{CSS}	—	506.54	506.62	506.59
	SD_{CSS}	—	10.44	12.94	9.28
Age 12	n	—	152	151	151
	M_{CSS}	—	509.54	511.56	510.74
	SD_{CSS}	—	11.65	12.79	9.96
Age 13	n	—	149	148	148
	M_{CSS}	—	511.19	513.03	512.15
	SD_{CSS}	—	10.01	14.46	10.56
Age 14	n	—	147	147	147
	M_{CSS}	—	514.52	518.01	516.26
	SD_{CSS}	—	9.92	16.29	11.31
Age 15	n	—	161	161	161
	M_{CSS}	—	517.11	519.83	518.47
	SD_{CSS}	—	9.77	15.84	10.94
Age 16	n	—	145	145	145
	M_{CSS}	—	517.53	521.24	519.39
	SD_{CSS}	—	10.61	15.38	11.24
Age 17	n	—	149	149	149
	M_{CSS}	—	521.09	522.87	521.98
	SD_{CSS}	—	10.51	18.34	11.68
Age 18	n	—	97	96	96

	M_{CSS}	—	521.47	526.2	523.94
	SD_{CSS}	—	10.98	15.14	10.55
Age 19	n	—	145	145	145
	M_{CSS}	—	525.01	531.77	528.39
	SD_{CSS}	—	8.82	18.57	11.26
Age 20–21	n	—	157	157	157
	M_{CSS}	—	527.61	531.09	529.35
	SD_{CSS}	—	9.58	16.43	10.36
Age 22–29	n	—	162	162	162
	M_{CSS}	—	526.04	522.92	524.48
	SD_{CSS}	—	10.15	16.9	11.19
Age 30–39	n	—	169	169	169
	M_{CSS}	—	526.7	520.46	523.58
	SD_{CSS}	—	12.25	17.31	12.76
Age 40–49	n	—	173	173	173
	M_{CSS}	—	528.39	513.87	521.13
	SD_{CSS}	—	11.98	14.87	10.68
Age 50–59	n	—	172	172	172
	M_{CSS}	—	528.01	508.24	518.12
	SD_{CSS}	—	11.71	14.64	11.07
Age 60–69	n	—	168	167	167
	M_{CSS}	—	530.38	501	515.73
	SD_{CSS}	—	10.56	11.83	8.67
Age 70–79	n	—	157	157	157
	M_{CSS}	—	534.12	493.76	513.94
	SD_{CSS}	—	11.75	9.57	8.2
Age 80–84	n	—	136	133	133
	M_{CSS}	—	532.11	487.43	509.81
	SD_{CSS}	—	14.11	10.08	10.11
Age 85+	n	—	66	65	65
	M_{CSS}	—	533.22	482.52	508.2
	SD_{CSS}	—	14.13	11.71	9.81