One

NIH TOOLBOX® HISTORY AND DEVELOPMENT

Julie N. Hook
Amy K. Giella
Abigail B. Sivan

COPYRIGHTED MATERIAL

OVERVIEW

Traditional cognitive batteries, while robust, are often time-intensive and costly. Recognizing this, the National Institutes of Health (NIH), through the Blueprint for Neuroscience Research, initiated an ambitious effort to develop a common set of tools capable of measuring neurological and behavioral functions across the lifespan. This initiative led to the creation of the NIH Toolbox for the assessment of neurological and behavioral function—a milestone in digital cognitive measurement.

In 2006, the NIH Toolbox contract was awarded to develop a set of brief, psychometrically sound, and royalty-free tests that would allow for greater ease when comparing the results of neurobehavioral and neuropsychological studies (Gershon et al., 2010). After six years of development, the NIH Toolbox measurement system was released. It included a suite of tests that allowed for assessment across the lifespan (aged 3–85+) and focused on four fundamental areas of neurobehavioral functioning, including: (1) cognition, (2) motor, (3) sensation, and (4) emotion. Since its inaugural release in 2012 as a web-based system, it was reconceptualized and rereleased as an Apple iPad app in 2015. This repositioning as an app allowed for both a more flexible and an easy-to-use test administration and data collection system.

Rapid Reference 1.1

Psychometrics is a field of study focused on measuring mental processes, including theory and techniques.

In 2023, the NIH Toolbox Version 3 (V3) was released as a new iPad app with improved user interface (UI) and user experience (UX) capabilities, making it more user-friendly with streamlined test instructions and updated scoring algorithms. The NIH Toolbox V3 iPad app was initially released in English and Spanish. In addition, select Cognition tests are now available in separate language apps, which include Arabic, French, Hebrew, Italian, and Spanish, plus more translations on the horizon.

Rapid Reference 1.2

User interface (UI) focuses on the visual elements and interaction with the product's interface (e.g., buttons and layouts).

User experience (UX) is how someone interacts with the product overall (e.g., product usability or how intuitive the design is to use).

VERSION 1 (V1) DEVELOPMENT

The NIH Toolbox was launched under the NIH Blueprint for Neuroscience Research, a multiagency effort that began in 2004 to accelerate the pace of discovery in brain research.

Rapid Reference 1.3

The NIH Toolbox is a direct response to the problem of instrument heterogeneity, a key barrier to reproducible science.

The NIH Blueprint for Neuroscience Research is a collaborative effort among the NIH Office of the Director and 12 NIH institutes and centers that support research on the nervous system (National Institute of Neurological Disorders and Stroke, n.d.). The participating institutes, centers, and offices include:

- National Center for Complementary and Integrative Health (NCCIH)
- National Eye Institute (NEI)
- National Institute on Aging (NIA)
- National Institute on Alcohol Abuse and Alcoholism (NIAAA)
- National Institute of Biomedical Imaging and Bioengineering (NIBIB)
- Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD)
- National Institute on Drug Abuse (NIDA)
- National Institute of Dental and Craniofacial Research (NIDCR)

- National Institute of Environmental Health Sciences (NIEHS)
- National Institute of Mental Health (NIMH)
- National Institute of Neurological Disorders and Stroke (NINDS)
- Office of Behavioral and Social Sciences Research (OBSSR)

The NIH Blueprint is aimed at advancing research and discoveries in brain function, including disease, health, and aging, and to accelerate discoveries in neuroscience research and is focused on initiatives that were too large an undertaking for any one NIH institute. For more information about the NIH Blueprint, please refer to neuroscienceblueprint.nih.gov. Since 2004, the Blueprint has invested over $615 million to support these aims (NINDS, n.d.). An example of a supported large-scale program includes the Human Connectome Project, which has advanced our understanding of both the brain and ways to treat brain disorders. The Blueprint funds training efforts as well as research resources and tools, such as the NeuroBioBank, which includes a collection of post-mortem brain tissue that can be used to study neurological, developmental, and psychiatric disorders.

One issue with traditional neurobehavioral assessment tools is that they frequently vary in methodology. Take, for example, two tests of processing speed; they could both use different items, test designs, scoring methods, and normative data. With all these possible differences, it can make it difficult to compare scores when different tests are used. This can hinder the comparability of findings across studies, as well as limit meta-analytic power. Further, these issues can become more pronounced in longitudinal and multisite research when different tests are used.

VERSION I DEVELOPMENT

In 2006, the development of the NIH Toolbox began when the Blueprint awarded Dr. Richard Gershon, principal investigator, a contract to develop a suite of assessment tools to address the need of developing a standardized set of instruments to assess neurobehavioral functioning. This project lasted six years and included input from over 250 scientists and about 80 institutions.

The framework for the NIH Toolbox was created by performing an extensive literature review and interviewing or surveying around 100 experts (Nowinski et al., 2013). From this work, the following domains and subdomains were identified (Table 1.1).

There were several other factors that drove the selection of tests for inclusion in the NIH Toolbox, and these factors included:

- appropriate for use across the age span (ages 3–85 years),
- ability to use in diverse settings and groups,
- no fees for intellectual property use,
- strong reliability and validity findings (i.e., psychometric soundness), and
- brevity in assessment.

Table 1.1 NIH Toolbox Version 1 Domains and Subdomains

Domain	Subdomain	References
Cognition	Attention, episodic memory, working memory, expressive and receptive language, executive function, and processing speed	Weintraub et al. (2013)
Emotion	Psychological well-being, stress and self-efficacy, social relationships, and negative affect	Salsman et al. (2013)
Motor	Balance, dexterity, strength, endurance, and locomotion	Rine et al. (2013), Reuben et al. (2013)
Sensation	Audition, vision, vestibular, olfaction, pain, and gustation	Coldwell et al. (2013), Cook et al. (2013), Dalton et al. (2013), Dunn et al. (2013), Varma et al. (2013), and Zecker et al. (2013)

Administration of the core set of tests in each domain needed to be 30 minutes or less, so all four core test domains (Cognition, Emotion, Motor, and Sensation) could be administered in 2 hours or less. Test instructions needed to be understandable to examinees with low literacy, and tests needed to allow for the assessment of people from diverse backgrounds. Further, tests needed to assess typical functioning rather than being specific to functioning associated with a specific disorder or disease. For example, a test specific to the assessment of one particular disorder, like autism, was not considered. However, a test of cognitive function that could assess typically developing cognition, in addition to being sensitive to cognitive issues associated with autism or other disorders, was selected.

Across all four domains, 1,391 measures were reviewed. After the downward selection based on the criteria listed herewith, 48 tests remained. This smaller group was then reviewed by expert working groups covering: (1) culture, (2) language, (3) accessibility, (4) use in pediatrics, and (5) use in geriatrics. Tests then underwent further evaluation when needed and were examined in validation studies of 450–500 participants using criterion measures in their respective domains and subdomains (see Akshoomoff et al., 2013; Weintraub et al., 2014; Zelazo et al., 2014). For instance, validity studies for the Cognition tests included the *Wechsler Adult Intelligence Scale* IV (WAIS; Wechsler, 2008), *Delis Kaplan Executive Function System* (D-KEFS; Delis et al., 2001), and *Peabody Picture Vocabulary Test, 4th Edition* (PPVT-4; Dunn & Dunn, 2007). Next, tests were identified to see if they were suitable for administration and scoring via computerized adaptive testing (CAT) and item response theory (IRT). Because these techniques require more participants, this resulted in 16,000 participants being assessed as part of validation, calibration, and field test studies.

After test selection and refinement, there was a large national standardization study conducted in English and Spanish with 4,859 participants. Ages ranged from 3 to 85 years and were representative of the 2010 US population based on geographic location, socioeconomic status, ethnicity/race, and sex assigned at birth (see Casaletto et al., 2015 for more details). Normative scores were collected and calculated for each year from 3 through 17 years and then in the following age bands: 18–29, 30–39, 40–49, 50–59, 60–69, 70–79, and 80–85 years (Beaumont et al., 2013).

In 2012, version 1 of the NIH Toolbox was launched as a validated test suite for online use. To use version 1, examiners frequently needed a laptop with two auxiliary keyboards and mice and monitors. Other specialized equipment was also required for non-cognition domain tests in sensation and motor.

Neuropsychological Test Design

The NIH Toolbox introduced several groundbreaking features in its test battery. It stands among the first fully digital collections of norm-referenced neurobehavioral assessments, with the added distinction of being co-normed across various tests and functional domains. These measures were crafted using innovative statistical methodologies, enabling administration in under five minutes for most tests while preserving strong psychometric integrity. Importantly, the tests were designed to evaluate a broad spectrum of neurobehavioral functioning, rather than focusing on specific clinical conditions.

Digital Norm-Referenced and Co-normed Testing

Norm-referenced tests evaluate an individual's performance by comparing it to that of a defined reference group. For the NIH Toolbox, this group consists of a representative sample of community-dwelling adults and children from across the United States (as described herewith). While the value of co-norming, especially in fixed batteries, has been supported by some experts for enhancing cross-test comparability (e.g., Russell, 2004), others have questioned its necessity (Rohling et al., 2015). To that end, the NIH Toolbox accommodates both fixed and flexible testing formats. Examiners can administer the standard core battery in each domain or customize assessments by selecting individual tests across domains, with all options benefiting from co-norming.

Item Response Theory and Computerized Adaptive Testing in the NIH Toolbox

A key strength of the NIH Toolbox lies in its efficiency, with most tests completed in five minutes or less. This is largely due to the integration of IRT and CAT (Cella et al., 2007). In fact, the NIH Toolbox was among the first neurobehavioral assessment tools to utilize these advanced psychometric approaches.

To use IRT, the trait being measured must vary along a continuum (i.e., one can possess more or less of it), and test items must vary in difficulty along that continuum. This setup allows individuals to be accurately placed based on their level of the trait. Developing a CAT requires assembling a large pool of potential items. These items undergo rigorous evaluation and calibration to determine their measurement properties. In the NIH

Toolbox's development, over 16,000 individuals completed early versions of items assessing cognition (e.g., vocabulary and reading), emotion, and vision. Only items that met strict quality standards, such as showing no gender-based bias, were retained in the final item banks.

When a CAT test is administered to an examinee, the test algorithms select items from the item bank to estimate the individual's ability level. The CAT algorithm uses preset rules, such as starting difficulty, test length, precision thresholds, and how much item difficulty can change between questions (Gershon, 2005; Gershon et al., 2003). For example, when administering the NIH Toolbox Picture Vocabulary test to an adult, the first item is chosen based on the person's education level. If they respond correctly, the next item is harder; if not, an easier one is given. The test continues until the result reaches the desired level of precision. On average, about 22 items are needed to achieve a standard error of 0.75 or lower.

VERSION 2 (V2) DEVELOPMENT

To improve mobility and simplify test administration, in 2015, the NIH Toolbox was relaunched as an iPad app. As part of this relaunch, an equivalency study was conducted, and scoring algorithms in the V2 iPad app were updated for several timed cognitive assessments to account for device-specific differences. The normative data provided in Casaletto et al. (2015) became the gold standard. Users who wanted to continue using the V1 (the web version) were informed that the normative data was updated but not integrated on the web platform; V1 web users interested in the updated norms needed to apply formulas manually or use the Python Norming Program. Then-current NIH Toolbox V2 users and future users were encouraged to begin new studies or data collection on the iPad only (i.e., not the web version), as the V2 iPad app was updated with the new normative formulas, and old data collected in the V2 app was automatically rescored with the new formulas when that data was reexported. Those interested in using both the V1 web version and the V2 iPad app version were alerted to some score discrepancies that were identified for some tests, and formulas were provided so users could utilize the Python Norming Program to rescore the data and align the scores with the V2 iPad app version.

Rapid Reference 1.4

When the decision was being made to select a more portable device, one of the reasons Apple's iPad was selected was that Apple had a more tightly controlled "ecosystem." For example, developers had fewer device variations and operating system versions to account for, making it easier to ensure compatibility and consistent performance across Apple devices.

Cognition Domain Access, Use, and Structure

While the NIH Toolbox assessments are examiner-administered (versus self-administered), the degree of examiner involvement varies. For instance, in the Cognition domain, the List Sorting Working Memory test relies on the examiner to judge the accuracy of an examinee's spoken response. The examiner, using an external keyboard, scores and records the examinee's response. In contrast, the Picture Vocabulary test is automated, administered, and scored by the app with minimal examiner input.

The Cognition domain offers recommended core batteries tailored to the participant's age, recognizing that not all assessments are suitable for young children. Examiners have the flexibility to administer individual tests, design custom batteries across domains, or use the preset/preconfigured core Cognition battery. All test data is scored and saved directly on the iPad used for administration. This means examinee data does not need to be uploaded or transferred elsewhere, as it is scored locally on the iPad. Examiners can optionally export score reports or raw data files (i.e., CSV spreadsheets) and can even print score reports directly from the app, if desired.

Access to the NIH Toolbox Cognition Battery requires completion of a credential verification process, in alignment with professional guidelines outlined in the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014). These standards emphasize that test users must demonstrate appropriate levels of training, knowledge, and experience to ensure ethical and accurate use of psychological assessments.

In accordance with these principles, NIH Toolbox Cognition tests are classified as *Level C* measures, which is the highest classification level used in test security protocols (APA, 2002). As a comparison, this is the same level of credential requirements that the Wechsler Intelligence Scales (Wechsler, 2008) would require. Level C assessments require users to possess a doctoral degree in psychology, education, or a closely related field, or to hold current state licensure or national certification to practice independently in a relevant professional capacity. Furthermore, users must have formal training in the administration, scoring, and interpretation of standardized assessments. Individuals who do not meet these qualifications must be supervised by a qualified professional to ensure test integrity and ethical standards are upheld.

Restricting access to these assessments also serves to protect the validity and reliability of the measures by preventing overexposure of test content. Exposure to specific test items or procedures may lead examinees to respond in a way that undermines the purpose of the test, such as exaggerating impairments to appear impaired or rehearsing responses to inflate apparent ability (Ben-Porath, 2012). This concern is especially relevant in neuropsychological and high-stakes assessments, where test preparation or prior knowledge can significantly alter outcomes (Bush et al., 2005).

In V2, the Cognition Battery included seven core tests, which contribute to the calculation of three composite scores for individuals aged 7–85: *Total Cognition Composite*, *Fluid Cognition Composite*, and *Crystallized Cognition Composite*.

For younger participants aged 3–6 years, the *Early Childhood Cognition Composite* is provided as a developmentally appropriate alternative to the Total Cognition Composite. All individual tests and composite scores are norm-referenced.

Rapid Reference 1.5

A composite score comprises two or more tests. It is a summary score.

A complete listing of the V2 Cognition tests and composites is provided in Table 1.2.

Version 2 Scoring and Normative Considerations

The NIH Toolbox Cognition Battery V2 generated a comprehensive set of scores at both the individual test and composite level scores, which included raw scores (or ability estimates, such as theta scores for CAT-administered tests), uncorrected standardized scores, age-corrected standardized scores, and fully demographically adjusted T-scores.

Table 1.2 Version 2 NIH Toolbox Cognition Tests by Age and Subdomain

			Normed Composite Scores			
Test Name	Subdomain	Ages	Early Childhood (3–6 years)	Fluid (7+ years)	Crystallized (7+ years)	Total (7+ years)
Dimensional Change Card Sort	Executive function	3–85	X	X		X
Flanker Inhibitory Control and Attention	Attention and executive function	3–85	X	X		X
List Sorting Working Memory	Working memory	7–85		X		X
Pattern Comparison	Processing speed	7–85		X		X
Picture Sequence Memory	Episodic memory	3–85	X	X		X
Picture Vocabulary	Receptive language	3–85	X		X	X
Oral Reading Recognition	Expressive language	3–85			X	X
Auditory Verbal Learning (Rey)	Immediate recall	8–85	Supplemental tests Only raw scores available			
Oral Symbol Digit	Processing speed	8–85				

The demographic correction model accounts for age, sex assigned at birth, educational attainment (or maternal education for participants under 18), and self-identified race/ethnicity (Gershon et al., 2013). To produce valid norm-referenced composite scores, all contributing subtests for a given composite must be administered in their entirety.

At the test level and the composite level, examiners receive raw scores (or thetas for CAT-administered tests), uncorrected standardized scores, age-corrected standardized scores, and a fully demographically corrected T-score (including sex assigned at birth, age, education [or mother's level of education for examinees below 18 years], and ethnicity). To obtain norm-referenced composite scores, all tests that comprise that composite must be administered.

In V2, the uncorrected standard scores compared the examinee's response to everyone in the normative sample (i.e., almost 5,000 people aged 3–85 years). The advantage of the standard score, compared to a raw score, is that it provides a common metric and distribution, allowing for comparison across any of the NIH Toolbox tests. When selecting scores, the uncorrected standard score would be most useful in a longitudinal study to track change over the lifespan or comparisons with a matched control group. However, a disadvantage of the uncorrected standard score is that it can be difficult to determine whether scores are unusual for an individual or group; it may even be difficult to understand whether a change over time was greater than or less than expected.

Rapid Reference 1.6

Standard scores have a mean of 100 and a standard deviation of 15.

The age-corrected standard scores reflect the level of a characteristic the examinee had compared to others in that age band. These scores can be useful in showing the measure of a characteristic in relation to the examinee's age. They can be used to interpret the change over time expected for that age group. For example, one might expect that a typically developing child would acquire a larger volume of vocabulary words over a five-year period than a typical adult. So, in this example, the age-corrected score for this child on the Picture Vocabulary test would likely show a greater change over time for an adult. One disadvantage of age-corrected standard scores is that they do not consider other characteristics of community-dwelling individuals that could affect cognitive ability, such as education. For instance, two 30-year-old adults who took the Picture Vocabulary test; one got a high-average score on a Picture Vocabulary test, and the other had a low-average age-corrected standard score. One may theorize a

CAUTION 1.1

Remember that when reviewing an age-corrected standard score, the examiner should understand what the age band is. For example, a child may be part of a 1-year normative age band versus an adult who could be part of a 10-year normative age band.

biological cause for this difference when it could be that the high-average scorer has a doctorate in English, while the latter had only completed some high school.

In version 2, the fully demographically corrected *T*-score was based on age, education (parent education for children), sex at birth, and ethnicity, and it is useful in detecting changes or acquired impairments due to medical conditions. It also allows for a judgment to be made as to whether the resulting score is unusual for an individual. However, this score is a poor estimate of the absolute level of a characteristic being measured. Please note that some researchers and clinicians do not believe that ethnicity should be used in score calculation, suggesting that any ethnic differences are more likely due to socioeconomic differences (Gasquoine, 2009; Rea-Sandin et al., 2021).

Rapid Reference 1.7

A T-score has a mean of 50 and a standard deviation of 10.

Theoretical Framework

The development of the cognition tests was anchored in the Cattell-Horn-Carroll (CHC) theory of cognitive abilities, one of the most empirically supported frameworks in modern psychometrics (Neisser et al., 1996; Schneider & McGrew, 2018). This model conceptualizes intelligence as a hierarchical structure, where general cognitive ability (g) is subdivided into broad abilities, most notably fluid and crystallized intelligence.

Fluid cognition refers to the capacity for problem-solving, reasoning, and novel information processing. These skills tend to decline with age and are sensitive to neurological insult (Horn & Cattell, 1967). In the NIH Toolbox, fluid cognition is assessed using five tests: Dimensional Change Card Sort, Flanker Inhibitory Control and Attention, List Sorting Working Memory, Pattern Comparison Processing Speed, and Picture Sequence Memory.

In contrast, crystallized cognition encompasses accumulated knowledge and verbal abilities—skills that are more stable over the lifespan and often used to estimate premorbid functioning (Lezak et al., 2012). There are two NIH Toolbox tests in this area, which include Oral Reading Recognition (ORR) and Picture Vocabulary (PV).

Factor Structure and Construct Validity

The internal structure of the V2 Cognition Battery has been validated through multiple factor analytic studies across both normative and clinical populations. Initial studies by Mungas et al. (2013) established the construct validity of the test battery in healthy children and adults. Among children aged 3–7 years, a three-factor structure—vocabulary, reading, and fluid abilities—was supported. For older children (aged 8–15), a five-factor model emerged, encompassing vocabulary, reading, episodic memory, working memory, and executive functioning/processing speed. These findings were consistent with the intended constructs and aligned with CHC theory. In adults (aged 20–85), these researchers found support for a similar five-factor model that included reading, vocabulary, episodic memory, working memory, and executive functioning/processing speed (Mungas et al., 2014).

Moreover, studies in clinical populations have reinforced these findings. For example, Tulsky et al. (2017) confirmed the five-factor model in individuals with acquired brain injuries, including traumatic brain injury (TBI) and stroke, showing robustness of the factor structure in the presence of neurological conditions. In addition, Ma et al. (2020) evaluated the structure of the V2 Cognition Battery in a sample ranging from cognitively intact older adults to those with mild cognitive impairment and dementia. Their findings supported a two-factor model distinguishing fluid and crystallized cognition, in line with the CHC framework, and demonstrated good construct validity across clinical severity levels.

Fluid Cognitive Abilities

The NIH Toolbox Cognition Battery evaluates fluid cognition through five core tests: Dimensional Change Card Sort (DCCS), Flanker Inhibitory Control and Attention (Flanker), List Sorting Working Memory (LSWM), Pattern Comparison Processing Speed Test (PC), and Picture Sequence Memory (PSM). Each test was designed to assess foundational constructs of fluid intelligence, including attention, executive functioning, episodic memory, working memory, and processing speed. These constructs are integral to daily cognitive function and are sensitive to developmental and neurodegenerative changes across the lifespan (Carlozzi et al., 2015; Zelazo et al., 2014).

Attention

Attention, a critical cognitive construct, supports numerous higher-order functions, including learning, memory, and executive processes. The Flanker task, adapted from Eriksen and Eriksen (1974), measures selective attention and inhibitory control. Examinees determine the direction of a central stimulus flanked by congruent or incongruent distractors. This adaptation is based on the Attention Network Test and has demonstrated robust psychometric properties in both child and adult populations (Fan et al., 2002; Rueda et al., 2004). Visual and auditory cues enhance task comprehension for younger children. Performance is evaluated based on accuracy and reaction time, and the test typically requires under three minutes to administer.

Executive Function

Executive functions enable goal-directed behavior, including planning, shifting between tasks, and inhibiting inappropriate responses. The DCCS, originally derived from Luria's rule-use paradigms (Frye et al., 1995), assesses cognitive flexibility. Participants match stimuli based on changing rules (i.e., color or shape), simulating real-life task-switching demands. For young children, the task is scaffolded with additional visual and auditory cues. Zelazo et al. (2014) demonstrated convergent validity of the Flanker and DCCS tests with the D-KEFS Color-Word Interference subtest and found strong discriminant validity relative to vocabulary assessments.

Episodic Memory

Episodic memory refers to the ability to recall information tied to a specific context or experience and is vulnerable to decline from age or neurological injury (Squire, 2004). The PSM test uniquely evaluates visual episodic memory across a broad age range using culturally neutral image sequences. Performance is based on the participant's ability to recall and reproduce the correct sequence of thematically linked pictures. Bauer et al. (2013) highlighted the relevance of medial temporal structures in episodic memory, and validation studies (e.g., Carlozzi et al., 2015) showed strong correlations with standard episodic memory tasks such as the Rey Auditory Verbal Learning Test.

Working Memory

Working memory supports short-term storage and manipulation of information and is essential for reasoning, comprehension, and learning (Baddeley, 2003). The LSWM test presents visual stimuli (e.g., animals and foods) and asks participants to reorder them based on a size hierarchy. Difficulty increases as list lengths expand and as additional stimulus categories are introduced. Convergent validity has been established with traditional measures such as WAIS-IV Letter Number Sequencing and the PASAT (Tulsky et al., 2014). Research confirms that working memory capacity increases significantly from early childhood through adulthood (Alloway et al., 2006; Dempster, 1981).

Processing Speed

Processing speed is the rate at which individuals perceive, interpret, and respond to stimuli. It is sensitive to neurological dysfunction and aging (Coyle et al., 2011). The PC task requires participants to quickly judge whether two visual stimuli are identical, providing a simple but effective index of cognitive speed. The test, validated by Carlozzi et al. (2015), showed strong convergence with the WAIS-IV Processing Speed Index and demonstrated expected age-related performance declines.

Crystallized Cognitive Abilities

Crystallized cognition encompasses language-based knowledge and skills acquired through experience, which are assessed by two tests—Oral Reading Recognition (ORR) and Picture Vocabulary (PV). These measures assess expressive and receptive language, respectively, using IRT-calibrated item banks and CAT to enhance efficiency and precision (Gershon et al., 2014).

Expressive Language

Reading and expressive language are measured through ORR, which is a test of single-word reading ability. Word selection was based on psycholinguistic properties such as frequency and phonological complexity (Gershon et al., 2014). Validation research demonstrated high correlations with the WRAT-4 Reading subtest and low correlations with nonlanguage cognitive tasks, supporting the test's construct specificity.

Receptive Language

Receptive language involves the comprehension of structured verbal information. PV assesses receptive vocabulary by presenting a spoken word and requiring the participant to select the corresponding image from four pictured options. Items are carefully constructed based on linguistic difficulty, translatability, and age appropriateness (Dale & O'Rourke, 1976; Gershon, 1988). Distractors vary in semantic and phonemic similarity to ensure a range of difficulty. Validity studies confirmed strong correlations with the PPVT-4 and low correlations with unrelated measures, such as visuospatial memory (Gershon et al., 2014).

VERSION 3 (V3) DEVELOPMENT

The NIH Toolbox Version 3 iPad app was released in 2023 to researchers and clinicians. Compared to V2, the updated V3 app provides users with improved UI and UX capabilities, user-friendly administration features, streamlined test instructions, and updated scoring algorithms.

The NIH Toolbox initiated a normative update beginning in 2021 to enhance population representativeness of the normative sample in response to the 2020 US Census data (US Census Bureau, 2021). The V3 app includes the integration of an additional composite score and several new measures, including delayed recall tasks, a visual reasoning (VR) measure, and a processing speed test adapted for young children. Previously unnormed supplementary tests, such as the Oral Symbol Digit and Auditory Verbal Learning Test (AVLT), were normed and included in the updated Cognition domain.

Overview: Renorming V3 Cognition Tests

As part of the comprehensive update to the Cognition domain, the team prioritized renorming to address changes in cognitive performance across generations—an effect widely known as the Flynn effect. This gradual rise in average cognitive test scores over time is thought to result from improvements in factors such as education, healthcare, and environmental stimulation. Without updated norms, test scores can become misaligned with current population benchmarks, leading to overestimated or inaccurate interpretations. Moreover, transitioning from different modes of administration may introduce differences in UX that could affect performance. In keeping with the *Standards for Educational and Psychological Testing* (AERA et al., 2014), test publishers are expected to update normative data regularly to maintain the accuracy and relevance of score interpretations as populations and testing conditions evolve.

The NIH Toolbox V3 Cognition domain norming study was conducted between June and September 2021, involving a nationally representative sample of 3,956 participants.

This included 2,248 children aged 3–17 and 1,708 adults aged 18–90 and older. The primary objectives of the study were to:

- Expand the Cognition domain by introducing new tests to enhance construct coverage;
- Revise existing measures to address user-identified concerns related to floor and ceiling effects, as well as age-specific reliability;
- Update normative data to better reflect the current US population in terms of sex assigned at birth, race, ethnicity, and education;
- Implement continuous norming procedures to reduce limitations associated with broad age-group "binning;"
- Recalibrate as many tests as possible using a unified, IRT-based change-sensitive score (CSS) metric to improve sensitivity to developmental and clinical change over time.

The norming study was conducted in collaboration with a private market research firm under the supervision of researchers at Northwestern University. Data collection took place at 12 sites across all four US census regions, ensuring national representation by including at least one location per census division.

The market research firm was responsible for hiring and training examiners, recruiting participants, and conducting assessments in their office facilities. A "train-the-trainer" model was used: Northwestern University staff conducted a five-day in-person training with lead trainers, who were then certified through video-recorded practice sessions. These certified trainers went on to train and certify local examiners at each testing site.

Examiners collecting external validity data received additional training in administering both the NIH Toolbox and standardized measures available through Pearson's Q-Interactive platform. Prior experience with Q-Interactive was required, and a refresher session was provided by Northwestern University staff.

Participants were recruited from the firm's national database via phone and email, based on demographic quotas specified by Northwestern University. These quotas ensured that the sample reflected the US population in terms of age, sex assigned at birth, race, ethnicity, and education (or parent education for children). All testing was conducted in controlled office environments across the participating sites (Table 1.3).

Updates to the V3 App

Although the primary objectives of NIH Toolbox V3 were to streamline the Cognition tests and collect new normative data, the NIH Toolbox team embraced the opportunity to improve the tests and the platform overall.

To this end, the platform underwent a full restructure of the examiner interface, with more efficient workflows and a new sidebar that allows examiners to quickly and easily navigate the main sections of the V3 app. See Figure 1.1.

Table 1.3 Version 3 NIH Toolbox Cognitive Tests by Age and Subdomain

Test Name	Subdomain	Ages	Normed Composite Scores				
			Early Childhood CFPS (3–6 years)[a]	Early Childhood (4–6 years)	Crystallized (7+ years)	Fluid (7+ years)	Total (7+ years)
Dimensional Change Card Sort	Executive function	4–85		X		X	X
Flanker Inhibitory Control and Attention	Attention and executive function	4–85		X		X	X
List Sorting Working Memory	Working memory	5–85				X	X
Pattern Comparison	Processing speed	5–85				X	X
Picture Sequence Memory[a]	Episodic memory	3–85		X		X	X
Picture Vocabulary	Language	3–85	X	X	X		X
Oral Reading Recognition	Language	7–85			X		X
Speeded Matching[a]	Processing speed	3–6	X	X			
Visual Reasoning[a]	Executive function	3–85	X				
Face Name Associative Memory Exam[b]	Associative and Delayed Memory	18–85					
Rey Auditory Verbal Learning[b]	Immediate Recall and Delayed Memory	5–85	Not currently included in a composite				
Oral Symbol Digit[b]	Processing Speed	5–85					

[a] Tests or composites that were newly added or had a significant update in V3.
[b] Tests that were normed in V3, but were not previously normed.

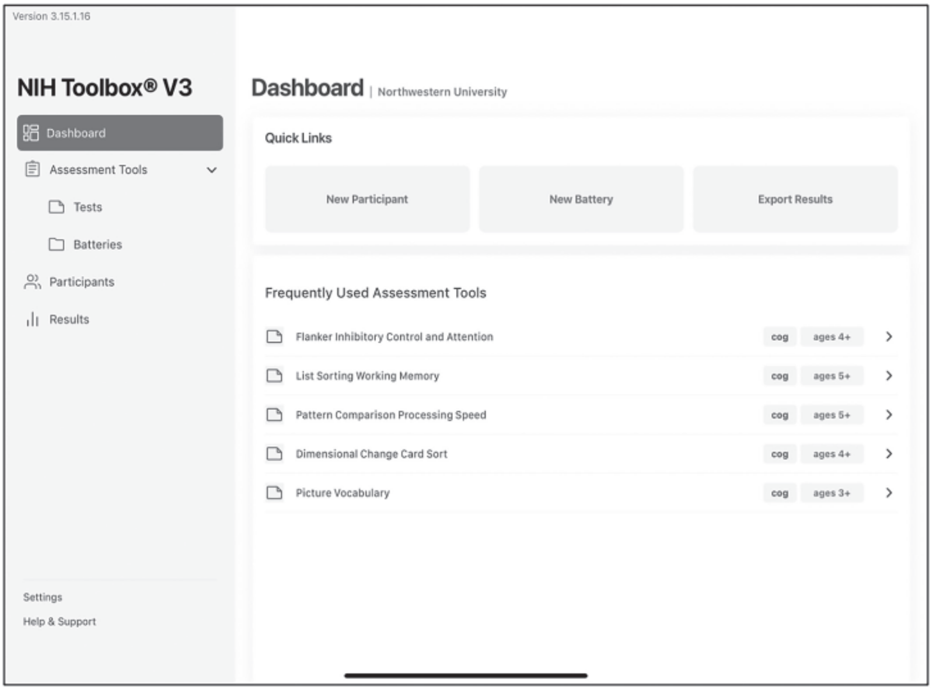

Figure 1.1 NIH Toolbox V3 Examiner Interface. Source: From NIH Toolbox V3. Copyright 2025 Toolbox Assessments, Inc. Reproduced with permission from NIH Toolbox V3/ Toolbox Assessments, Inc.

Additional features were added to the examiner interface to make the selection of tests easier, such as the addition of test cards (i.e., an expandable box/card describing the test). Also, outside materials or equipment needed to properly administer the tests are identified in multiple places to ensure examiners can prepare before seeing the examinee.

In addition to the examiner interface, the instructions and workflows of the Cognition tests were streamlined. These changes were tested in a pilot study and were revised as needed to support the psychometric quality of the Cognition tests. After the test workflows were updated, the NIH Toolbox team evaluated and updated the UX for a more consistent administration experience across tests.

- V3 test names were shortened by removing "NIH Toolbox," "Test," and the age (except for tests that have age-specific versions). For instance, the V2 test name "NIH Toolbox Oral Reading Recognition Test Age 3+" was renamed "Oral Reading Recognition" in V3.
- When psychometrically justifiable, multiple test versions were evaluated, streamlined, and merged into one version for all relevant ages (e.g., the three age-specific versions of Dimensional Change Card Sort in V2 are now one DCCS test that is appropriate for ages 4+).

- Instructions were updated to be clearer and easily understandable, even to the youngest examinees. Instructions, which were sometimes printed on the screen, sometimes played via prerecorded audio, and sometimes read aloud by the examiner, are now presented as text on the screen and played via prerecorded audio (with a few exceptions for unique administration procedures). Presenting the instructions via the test's audio additionally bolsters a uniform examinee experience.

When administering NIH Toolbox V3 tests, a title screen appears before that test's first instruction screen. V3 title screens introduced colored banners to indicate that test's domain (e.g., all Cognition tests have a red banner), as well as icons to indicate if/what additional materials or equipment is required to administer that test. This is a helpful reminder for examiners to arrange the equipment as needed prior to beginning that test.

The test interface itself has been made consistent where possible, including one background color and text in the same font and size across all tests. "Next," "Play again," and "Back" buttons were added where appropriate to help the examiner navigate each test. The "Touch and hold to continue" button was changed to "Slide to continue." The administrator's gesture, which is used during test administration to pause the test, was changed to address user feedback that V2 examinees, on occasion, accidentally performed the administrator's gesture when responding to live items.

Updates to the V3 Cognition Tests

For the V3 Cognition tests, there were updates and changes made to the existing tests from V2 to improve the workflows and psychometric properties. A listing of each test follows.

Dimensional Change Card Sort (DCCS)

DCCS measures executive functioning. Examinees are shown two cards and are asked to select the one that is either the same shape or the same color as the card in the middle of the screen.

In NIH Toolbox V2, DCCS had three separate age-specific tests (NIH Toolbox Dimensional Change Card Sort Test Ages 3–7, NIH Toolbox Dimensional Change Card Sort Test Ages 8–11, and NIH Toolbox Dimensional Change Card Sort Test Age 12+), which were restructured into a single Dimensional Change Card Sort test that is appropriate for ages 4+. The instructions and routing workflows were simplified so that all examinees, regardless of age, see the same instructions, navigate the same pre-switch and post-switch items, and take the same live items. According to LaForte et al. (2024), the V3 pilot study collected data down to 36 months of age, and it was found that many of the three-year-olds' response accuracy was at chance or lower, thus justifying the lower recommended age limit to change from three to four years and the removal of the V2 experimental downward extensions of DCCS.

The test images were redrawn to update the look and feel, but the content itself remained the same (i.e., balls and trucks). When needed, image colors were changed to avoid any issues for examinees with colorblindness. For example, the blue and yellow images were changed to blue and orange to address blue-yellow colorblindness.

Rapid Reference 1.8

Color blindness is the inability to differentiate certain colors. In the United States, red-green is the most common color blindness, followed by blue-yellow.

The DCCS test was also updated so that the target dimension is now presented visually in the middle of the screen and via audio (e.g., the audio says the word "Shape," while "SHAPE" is displayed as text on the screen). The star previously presented in the middle of the screen for V2 was removed. In addition, "Home Base"—a piece of paper with a small circle on it that was placed in front of the iPad for V2 DCCS—was removed and is no longer used in V3.

To improve the usability of the tests, user feedback was also incorporated into the V3 redesign. One example of this incorporation was offering a further drill-down of performance scores on timed tests. Although this information was available in V2, it was not easily accessible, as the examiner would need to export a data file from the app, import that data into another database, and then calculate these scores. To make this more user-friendly in V3 for applicable tests, such as the DCCS, new process scores were added to the score report at the test level. These scores include:

- median response time per item,
- commission errors (i.e., number of incorrect responses on live items), and
- omission errors (i.e., number of items on which the examinee does not respond/ times out).

Face Name Associative Memory Exam (FNAME)

FNAME is an assessment of delayed visual and associative memory. FNAME and FNAME Delay were marked as experimental measures in V2, were not included in the Cognition domain, and only offered raw scores. In V3, FNAME and FNAME Delay are supplemental tests in the Cognition domain and produce normative scores derived from the NIH Toolbox V3 norming sample, as well as process scores for each of the FNAME Delay sections (i.e., Faces Seen Before, First Name Letter, and Face-Name Matching).

FNAME assesses a person's ability to learn and recall the association between unfamiliar faces and names—an ecologically valid and cognitively demanding task that heavily relies on the integrity of the medial temporal lobes, especially the hippocampus (Rentz et al., 2011). This memory test has two parts. The first is the learning trial, in which the

CAUTION 1.2

Even though FNAME is one test, in the NIH Toolbox V3 app, the FNAME test will appear as if it is two separate (but associated) tests: FNAME and FNAME Delay. This was done in order to allow the examiner to add the learning trial first and then add the delayed recall portion of the test later in the assessment.

examinee is presented with 12 pictures of faces paired with names, one at a time. After a 5- to 25-minute delay, FNAME Delay asks examinees about those face-name pairs to test their memory of the prior learning trial.

From V2 to V3, other changes to the FNAME tests included an overall UI update to make it consistent with the other Cognition tests. Instructions and workflows were streamlined. The response options for FNAME were also changed from "Fits" and "Doesn't Fit" (as in, does this name "fit" the face in the picture?) to "Easy" or "Hard" (as in, will it be easy or hard to remember the face-name pair?).

Flanker Inhibitory Control and Attention (Flanker)

Flanker is a test of executive function. This test shows five fish that point either left or right, and examinees must quickly and accurately tap the button (either right or left) that matches the way the fish in the middle is pointing.

The three age-specific versions of V2 Flanker (NIH Toolbox Flanker Inhibitory Control and Attention Test Ages 3–7, NIH Toolbox Flanker Inhibitory Control and Attention Test Ages 8–11, and NIH Toolbox Flanker Inhibitory Control and Attention Test Age 12+) were combined into one V3 Flanker Inhibitory Control and Attention test. Streamlined instructions and routing workflows allow examinees aged 4 and older to take the same version of the test, with the same practice items and the same live items. LaForte et al. (2024) found that many three-year-olds in the V3 pilot study exhibited an at-chance or lower response accuracy, thus justifying the lower recommended age limit to change from 3 to 4 and the removal of the V2 experimental downward extension of Flanker.

Depending on which version of Flanker V2 examinees took, they either saw pictures of fish, black arrows, or both; V3 streamlined this so that all examinees see pictures of fish with black arrows over them. The number of live items increased from 20 to 30, with the first 20 items being identical to V2, and eight congruent (i.e., all five fish point the same way) and two incongruent (i.e., the middle fish points in the opposite direction) items were added. To make the test a bit less predictable, the delays between the fixation star and the live item were made variable (e.g., sometimes 0.3 second, sometimes 1.0 second; LaForte et al., 2024). V3 also implemented a time limit for live items; if there is no response within 10 seconds, the test automatically moves to the next item.

As with DCCS, the "Home Base"—a piece of paper with a small circle on it that was placed in front of the iPad for V2 Flanker—was eliminated and is no longer used in V3.

Similarly, based on user feedback, V3 Flanker introduced new process scores for congruent items [median response time per item and commission errors (i.e., number of congruent trials that are answered incorrectly)] and for incongruent items (i.e., median response time per item).

List Sorting Working Memory (LSWM)

LSWM is a measure of working memory. The first part (1-List) presents a series of foods or animals, then asks the examinee to say all of them back in size order, from smallest to biggest. The second part (2-List) presents lists of food and animals, which examinees must say back in size order, starting with the foods first, then the animals.

The two age-specific versions of LSWM in V2 (NIH Toolbox List Sorting Working Memory Test Ages 3–6 and NIH Toolbox List Sorting Working Memory Test 7+) were combined into one LSWM test. To facilitate a single version that is appropriate for examinees aged 5 and older, instructions were streamlined, workflows were reorganized, and age-based start points were introduced. Age-based start points paired with streamlined basal and ceiling rules help reduce administration time. As part of the streamlining of this test for inclusion in V3, the practice items were standardized, so regardless of age, everyone sees the same two practice items per List, and they have two attempts to answer each practice item correctly.

Rapid Reference 1.9

Basal and ceiling rules are standardized testing procedures used in psychological and educational assessments to determine starting and stopping points within a test, ensuring efficiency and accuracy in measuring an individual's abilities.

- Basal rule refers to the point at which it is assumed the test-taker would have answered all previous (easier) items correctly. Once a basal is established—typically by answering a set number of consecutive items correctly—the examiner can skip earlier items and credit them as correct, reducing unnecessary testing of clearly mastered content.
- Ceiling rule indicates the point at which it is assumed the test-taker would answer all subsequent (more difficult) items incorrectly. Once a ceiling is reached—often defined as a set number of consecutive incorrect responses—the examiner discontinues the subtest to prevent frustration and preserve time.

LSWM has two items at each sequence length (e.g., two items in 1-List that contain four pictures). In V2, answering the first item of each sequence length correctly moved the examinee immediately to the next sequence length; the second item of any sequence length was only administered if the first was answered incorrectly (e.g., answering the first three-picture item correctly moves to the four-picture item; an incorrect response moves to the second three-picture item). The V3 norming study showed that the difficulty often varied between the items of each sequence length, which suggested that administering both items of the sequence may improve the test's precision. This led to the change in V3 that, regardless of the examinee's response, both items of each sequence length are administered (LaForte et al., 2024).

A cross-cultural comprehensibility review was also done for the V3 update. As part of this review, two images were identified as not being easily recognized globally. Therefore, the Blueberry in V2 was replaced with Bean in V3, and the Pumpkin in V2 was replaced with Cake in V3. Notably, all images were redrawn in a uniform art style to ensure visual continuity throughout the test.

During the V2 redesign process, users reported that examinees were often unclear about when a list had ended. Some participants began responding before all the food or animal items had been presented, while others waited unnecessarily, expecting additional items. To resolve this issue in V3, a chime was added at the end of each list to clearly signal that the list is complete, and the participant may begin responding.

Oral Reading Recognition (ORR)

ORR measures reading decoding skills. Words are displayed as text on the iPad screen, one at a time, and the examinee must say the word aloud. The examiner manually scores the pronunciation as correct or incorrect by typing "1" or "0" on a wireless keyboard.

The ORR instructions underwent minor edits, and they are now presented as text on the screen and via prerecorded audio. ORR items were recalibrated, but the words themselves remain the same. The recommended age range for ORR was raised from ages 3+ to ages 7+.

As a CAT, the start points in V2 were changed from routing by age through 17 years to routing by age for examinees 7–18 years old; examinees 19 years and older are routed by education.

Oral Symbol Digit (OSD)

OSD measures processing speed without a manual motor component (i.e., only a verbal response is required). Examinees are given a two-sided, laminated "Oral Symbol Digit Examination Sheet." Side 1, used for practice, displays a grid pairing the numbers one through nine with corresponding symbols, along with a separate grid of unpaired symbols. Examinees use this side to become familiar with the task before starting the actual test. Side 2 contains the live test: a grid consisting of 13 rows, each with 11 symbols.

Rapid Reference 1.10

When examiners are approved for C-level access to Cognition tests via the NIH Toolbox website, the Oral Symbol Digit Examination Sheet will be available online for download.

OSD in V3 includes normed scores (unlike V2), and the age range was extended from ages 8 and older to ages 5 and older.

OSD administration differs from other Cognition tests; instead of seeing each item on the iPad screen, examinees look at the Oral Symbol Digit Examination Sheet for practice and live items. As a more hands-on test, instruction audio is not prerecorded for OSD. Instead, the examiner refers to the iPad screen for examiner-specific directions and instructions. Black text on the screen indicates instructions that the examiner should read aloud to the examinee (e.g., "Now you try the rest for practice. Say each number out loud so I can hear you."). Blue, italicized text represents examiner-specific notes or directions for the examiner (e.g., "Tap each word the participant says to mark it as correct.").

V2 had 144 items, with 18 symbols per row across 8 rows, which did not align with the examiner's view of 9 items across each row in the app. Unless the examiner was closely following the examinee, and the examinee made very few mistakes, it could be difficult for V2 examiners to track examinees' progress and accurately report which items were correct versus incorrect. V3 addressed the misalignment of items by having the same 11 by 13 grid on the OSD Examination Sheet and the examiner's view in the app. This equivalency makes it much easier for V3 examiners to follow along.

Pattern Comparison (PC)

PC is a test of processing speed. In this timed test, examinees are shown two pictures at a time and must quickly tap the "Yes" button if the two images are the same or tap "No" if the two pictures are not the same.

For V2, the two age-specific versions of PC (NIH Toolbox Pattern Comparison Processing Speed Test Ages 3–6 and NIH Toolbox Pattern Comparison Processing Speed Test Age 7+) were restructured into one PC test. The smile and frown face buttons children aged 3–6 used in V2 were replaced, and in V3, all PC examinees have "Yes" and "No" response buttons. Instructions were streamlined and are read by prerecorded audio in V3. Further, the V3 pilot study indicated that the youngest examinees did not perform better than chance, so the age range for PC in V3 is ages five and older. For a replacement processing speed test for younger children, please refer to the Speeded Matching test.

To ensure that the PC test remains a pure measure of processing speed (minimizing the influence of other cognitive abilities such as memory, language, or executive function), the test was carefully redesigned with simplicity and clarity in mind. Specifically, the visual items were redrawn and enhanced in both quality and resolution to make it quicker and easier for examinees to determine whether two images are the same or different (LaForte et al., 2024). This refinement reduces perceptual ambiguity and cognitive load, allowing the task to focus solely on rapid visual discrimination. Importantly, while the artwork was improved for visual clarity, the underlying item content remained consistent across versions (e.g., a tree in V2 remains a tree in V3), ensuring continuity in test structure while optimizing ease of use. This design principle reflects the broader goal of a processing speed task: to assess how quickly individuals can perform simple, automatic decisions without engaging (or at least minimizing the engagement of) higher-level reasoning or interpretation.

The audio prompt preceding each practice item was shortened in V3 to a concise and direct question: "Are these the same?" In V2, the practice phase included seven items for examinees aged 3–6 and six items for those aged 7 and older. A discontinue rule was applied, whereby examinees who responded incorrectly to two trials on two separate items did not proceed to the live test. In V3, this structure was streamlined: all examinees, regardless of age, complete the same three practice items, with up to two attempts permitted for each. Importantly, the discontinue rule was eliminated—regardless of performance on the practice items, all examinees proceed to the live items.

Three of the practice items removed in V3 were incorporated into the live item pool, increasing the total number of test items from 130 in V2 to 133 in V3. Additionally, the time limit for the test was reduced from 85 seconds in V2 to 80 seconds in V3.

Scoring procedures were also updated: the examinee's dominant hand no longer influences scoring. A new process score was introduced in V3, capturing the median response time per item to provide a more nuanced index of processing speed.

Picture Sequence Memory (PSM)

PSM is a computerized measure of episodic memory, designed to assess the ability to learn and recall the temporal order of visual information. During each trial, a sequence of images is presented one at a time in the center of the screen, accompanied by brief prerecorded phrases that describe the scenes. As each image is shown, it moves to a small box surrounding the display area. Once the full sequence has been presented, the images are scrambled and returned to the large center box. The examinee is then prompted to reconstruct the sequence by dragging the images back into their original order.

In V2, PSM included three forms (A, B, and C), each with four age-specific versions tailored for ages 3–4, 5–6, 7, and 8+, resulting in a total of 12 separate PSM tests. In V3, the structure was streamlined by collapsing the age-specific versions within each form. Consequently, each form now exists in a single version—Form A and Form B—while Form C was retired.

A primary enhancement in V3 was the automation of test routing to more closely align sequence difficulty with the examinee's developmental level. Routing is initially determined by the examinee's age, which assigns age-appropriate practice item(s) and an initial live test item. Subsequent routing is based on the examinee's adjacent-pairs score (i.e., the number of correctly ordered pairs in the sequence). For example, a four-year-old taking PSM Form A will view two practice items, followed by the Fair 6 item (a six-picture sequence with a fairground theme). Based on performance:

- An adjacent-pairs score of 0 leads to the Park 3 item (three-picture sequence) presented twice,
- An adjacent-pairs score of 1–3 results in a second administration of Fair 6, and
- An adjacent-pairs score of 4–5 advances the examinee to the Park 9 item (nine-picture sequence) presented twice.

Thus, the initial test item is determined by age, while the subsequent item(s) and sequence length are tailored to performance.

In V2, each form featured a unique thematic narrative—Form A (Park), Form B (Fair), and Form C (Farm). In V3, themes are standardized for routing purposes: Form A begins with the Fair theme, and Form B begins with the Park theme. The examinee's performance on the first item determines whether they continue with the same theme or transition to another. Forms are counterbalanced to ensure comparable difficulty levels across administrations.

Unlike other tests in the Cognition domain that rely on fixed response options, PSM is highly interactive, requiring examinees to tap and drag images into their chosen sequence order. In V2, training was examiner-led, involving scripted instructions and physical demonstrations. In V3, this training was standardized through animated instructional sequences with synchronized audio, which can be replayed as needed until the examinee demonstrates comfort with the task demands.

Additional refinements in V3 include streamlined audio instructions and a simplified post-shuffle prompt. Examinees aged 3–5 receive two practice items, while those aged 6 and older receive one practice item. Regardless of age, each practice item allows for two trials. If the examinee responds incorrectly on all trials of all practice items, the test is discontinued.

Picture Vocabulary (PV)

PV is a measure of receptive vocabulary, administered as a CAT. During each item, the examinee hears a target word through the iPad's speaker and is asked to select the image—among four options—that best represents the meaning of the word.

The word pool in V3 remains unchanged from V2, and most of the corresponding images were retained. However, several outdated visuals were updated with more contemporary images to enhance relevance and clarity. One item was removed due to ambiguity, as it contained two plausible correct responses. The child-directed audio prompt used in V2 (e.g., "Touch the picture of . . .") was eliminated in V3; live items now present only the target word, streamlining the task and reducing linguistic scaffolding.

All items were recalibrated in V3, and difficulty parameters were updated using refreshed response data. For examinees aged 18 years and younger, the test uses age to determine the starting item. For those aged 19 and older, routing is based on the examinee's education level—a slight revision from V2, which used education level for examinees 18 and older.

Rey Auditory Verbal Learning Test (RAVLT)

RAVLT assesses immediate verbal learning and recall. Examinees listen to a prerecorded list of 15 words and are then prompted to verbally recall as many words as they can remember. This process is repeated over three trials. The RAVLT Delay task, introduced in V3, measures delayed recall by asking the examinee to recall the same 15 words following a delay of 5–25 minutes.

Rapid Reference 1.11

Similar to the FNAME test, RAVLT appears as two separate (but related) tests in the V3 app—RAVLT as the learning trial and RAVLT Delay as the delayed recall trial. This was done to assist in administration flexibility when using the app, and this should be treated as one test.

Both RAVLT and RAVLT Delay now generate norm-referenced scores. Additionally, the age range for RAVLT was extended in V3 to include examinees as young as five years old (compared to ages 8 and older in V2), enhancing developmental coverage. RAVLT Delay in V3 produces normed scores for examinees aged 12 and older.

To improve task clarity, a chime was added at the end of each word list to indicate when examinees may begin responding. This change addresses prior feedback from V2, where participants occasionally responded prematurely or delayed their responses due to uncertainty about the list's completion.

In V2, responses were stored in a bitwise format, complicating data analysis. In V3, each word is scored individually as correct or incorrect for each trial, facilitating cleaner item-level data analysis.

Speeded Matching (SM)

SM is a test of processing speed for children aged 3–6 years, in which examinees view a target image centered on the screen and must quickly identify which of the four comparison images matches the target.

In V2, SM was considered experimental and was not formally included in the Cognition domain. In V3, SM was fully integrated into the Cognition domain and contributes to two composite scores: the Early Childhood Cognition Composite and the Early Childhood Cognition—Crystallized-Fluid-Processing Speed—Composite. The test now generates normed scores as well as a median response time metric.

Several updates were made to enhance usability and scoring. Entry of the examinee's dominant hand is no longer required for administration. Instructions are now delivered via audio, with only minor modifications to wording. The practice item fail rule was revised: the test now discontinues after eight incorrect responses to practice items, up from four in V2. Additionally, the response window for live items was reduced from 120 seconds to 90 seconds, reflecting a more efficient design aligned with the test's focus on speed.

Visual Reasoning (VR)

VR is a newly introduced measure in V3 that assesses executive functioning and nonverbal reasoning. Examinees aged 3 and older are shown a visual stimulus in the center of the screen and asked to select from multiple choices the image that best completes or corresponds to it. Item formats vary and may include selecting the missing piece in a matrix, completing a visual sequence, or choosing the most similar image.

VR was developed specifically for V3 to expand the Cognition domain's coverage of higher-order cognitive processes, particularly in reasoning and problem-solving. For the youngest examinees aged 3–6, VR is included in the Early Childhood Cognition Battery—Crystallized-Fluid-Processing Speed (CFPS).

Updates to V3 Cognition Test-Level and Composite-Level Scores

Whether used in education, clinical, or research settings, score selection should be guided by the assessment's intended purpose. Many of the scores and composites used in V2 were pulled forward into V3, though there were also some changes that are reviewed here.

Like in V2, the V3 raw scores can differ in scale and distribution and may reflect computed values or be derived from IRT models rather than simple tallies. One key advantage of raw scores is that they closely reflect the examinee's observed performance, making them potentially useful for tracking change over time. However, raw scores are limited in that they do not allow for straightforward comparisons across different NIH Toolbox measures and can be challenging to interpret in isolation. Without a normative reference or benchmark, it is difficult to determine whether a raw score represents strong or weak performance.

The V3 age-adjusted standard scores indicate how an individual's performance compares to a nationally representative sample of peers within the same age group, based on 2020 US Census benchmarks (US Census Bureau, 2021). These scores provide valuable context by accounting for typical age-related expectations, making them especially useful for assessing developmental trends or tracking meaningful changes in performance over time.

In V3 Cognition, the fully corrected scores used in V2—which adjusted for age, sex, education, and race/ethnicity—were not retained. Instead, V3 Cognition provides age- and education-adjusted T-scores. The decision was made to focus on age and education, which remain two of the most robust predictors of cognitive performance and offer a more parsimonious yet meaningful basis for comparison across individuals. One drawback of the age- and education-adjusted T-score is that certain group-specific variations may not be fully captured, such as the impact of socioeconomic status, which is an important contextual factor known to influence cognitive outcomes.

The inclusion of a CSS is new to V3. The CSS is derived from IRT and is specifically designed to track meaningful changes in an individual's performance over time. Unlike traditional norm-referenced scores that compare an individual's performance to that of a demographic peer group, the CSS is calculated directly from the raw test performance and reflects change in the individual's underlying ability or trait. Because it is not adjusted relative to age or other demographic norms, the CSS provides a clearer picture of actual improvement or decline.

Norm-referenced scores may only reflect a person's rank within a group, whereas the CSS captures true within-person change, independent of group performance trends.

For example, an individual might demonstrate a four-point increase in their CSS over one year, from Time 1 to Time 2. However, if their reference peer group showed an average increase of five points during that same period, the individual's adjusted score at Time 2 would actually be lower relative to their peers—effectively obscuring their true individual improvement. This highlights the unique value of the CSS for monitoring personal growth, especially in developmental or intervention-based contexts. To maintain consistency across Cognition measures in the NIH Toolbox, the

CSS has a standardized reference point: a value of 500 corresponds to the median performance level of 10-year-olds from the norming sample. While not norm-referenced, this shared anchor supports interpretability across tests.

V3 COGNITION COMPOSITE SCORES

The V3 offers two composite scores geared toward younger children and three for older children through adulthood. These composite scores offer standard scores similar to the individual tests.

Early Childhood Cognition Battery—Crystallized-Fluid-Processing Speed (CFPS)

New to V3, the Early Childhood Cognition Battery—CFPS is tailored for children aged 3–6, with extended normative data available for up to age 8.5 years. This composite includes tests of executive function, processing speed, and language development:

- Picture Vocabulary
- Visual Reasoning
- Speeded Matching

Administering the complete set of live items across these assessments yields an Early Childhood Cognition CFPS Composite score, offering a concise measure of cognitive performance across these targeted domains.

Early Childhood Cognition Composite

The Early Childhood Cognition Composite is designed to be developmentally appropriate for children between the ages of 4 and 6. However, extended normative scores are available for children up to age 8.5. This composite evaluates core areas of early cognitive development, including attention, episodic memory, executive function, processing speed, and language. The following tests are included:

- Picture Vocabulary
- Flanker Inhibitory Control and Attention
- Dimensional Change Card Sort
- Picture Sequence Memory
- Speeded Matching

When all live items from each of these assessments are administered, an Early Childhood Composite score is generated, reflecting general cognitive functioning in this age group. Of note, the V2 Early Childhood Cognition Composite went down to age 3, but due to the streamlining and renorming of these tests, some of the start ages changed and were no longer determined to be appropriate for age 3.

Total Cognition Composite

The Total Cognition Composite is recommended for individuals aged 7 years and older, though it can be used with children as young as 6, generating extended normative data for that younger subset. This battery offers a broad assessment of cognitive functioning, covering domains such as attention, episodic memory, executive function, language, processing speed, and working memory.

The Total Cognition Composite comprises the Fluid and Crystallized Composites:

- Fluid Composite tests
  - Dimensional Change Card Sort
  - Flanker Inhibitory Control and Attention
  - List Sorting Working Memory
  - Pattern Comparison Processing Speed
  - Picture Sequence Memory
- Crystallized Composite tests
  - Oral Reading Recognition
  - Picture Vocabulary

Completion of all live items across the tests results in three composite scores—Fluid Cognition Composite, Crystallized Cognition Composite, and Total Cognition Composite—which provide a comprehensive profile of cognitive abilities.

V3 ACCESSIBILITY AND CROSS-CULTURE APPLICABILITY UPDATES

Keeping in line with the NIH Toolbox's core tenet of maximizing accessibility, the V3 updates were made with careful attention to reducing barriers across diverse users and settings. These updates reflect a continued commitment to inclusivity through clearer instructions, improved UIs, and refinements that support administration across varied age groups, educational backgrounds, and clinical populations.

Building on the V2 foundation, V3 placed a focus on accessibility and usability for both examiners and participants. Design and usability experts were consulted to ensure that both the examiner interface and participant-facing screens align as closely as possible with current accessibility standards. Particular attention was paid to the requirements outlined in Section 508 accessibility guidelines (Rehabilitation Act of 1973, 1998). Accessibility compliance levels—categorized as A, AA, or AAA—were considered throughout the design process. These standards were addressed through thoughtful design features such as consistent layout, appropriate contrast ratios, readable sizing, and proportional visual elements.

In the NIH Toolbox tests—particularly the DCCS, where color is one of the key matching dimensions—drawings were carefully evaluated for accessibility using color analysis tools. Colors were mapped onto a color wheel and reviewed to ensure that individuals with color vision deficiencies could still accurately distinguish between categories.

Special attention was given to avoiding color pairings that may be indistinguishable to those with red-green or blue-yellow color blindness. Adjustments to hue, saturation, and brightness were made as needed, and simulations were used to confirm that all examinees, regardless of color perception, could engage with the task as intended.

As part of the NIH Toolbox V3 development, some measures underwent expert cultural review to ensure greater sensitivity and relevance across diverse populations. Feedback from users and administrators was actively incorporated to enhance clarity, reduce cultural bias, and promote global usability. In addition, insights gained from current and prior translation efforts played a key role in shaping revisions to language, imagery, and instructional content, thus fostering broader cultural acceptance and supporting more equitable administration across varied linguistic and cultural contexts.

USE OF THE NIH TOOLBOX AND FUTURE DIRECTIONS

Since its introduction in 2012, the NIH Toolbox has become an increasingly valuable resource across a wide range of settings, including research, clinical practice, and education. Initially developed to support large-scale research studies, the NIH Toolbox has evolved into a versatile platform used in both applied and translational contexts. Its broad lifespan coverage—from early childhood through older adulthood—combined with brief administration times and digital delivery, makes it especially well-suited for diverse applications. For example, in educational environments, the Cognition tests have been used to monitor developmental trajectories and identify learning needs. In clinical settings, they support neuropsychological screening, intervention planning, and outcome tracking. In research, the NIH Toolbox has become a core component of numerous national and international studies such as the Human Connectome Project (Human Connectome Project, n.d.), the Adolescent Brain Cognitive Development (ABCD) Study® (Thompson et al., 2022; National Institutes of Health, 2015), and the Environmental Influences on Child Health Outcomes (ECHO) program (Blaisdell et al., 2022; Gillman & Blaisdell, 2018).

A recent scoping review by Fox et al. (2022) documented over 280 published studies using the NIH Toolbox in clinical populations. Among these, the Cognition Battery was the most frequently employed, appearing in about 80% of studies. Of these 225 studies that used the NIH Toolbox Cognition tests, half were focused on neurological disorders; however, clinical applications span psychiatric conditions, cancer, and other chronic diseases, with growing interest in populations with intellectual and developmental disabilities (Fox et al., 2022). This expansion is supported by targeted resources such as the *NIH Toolbox Version 3 Supplemental Administration Manual for Intellectual and Developmental Disabilities* (McKenzie et al., 2023). In addition to Cognition-focused studies, Fox et al. (2022) also identified 49 studies that referenced NIH Toolbox Emotion tests, 29 studies that used NIH Toolbox Motor tests, and 16 studies that utilized NIH Toolbox Sensation tests from 2008 to 2020. This shows that although Cognition testing is undeniably the primary focus of clinical testing when using the NIH Toolbox, tests in the Emotion, Motor, and Sensation domains have also been utilized in nearly 100 other clinical studies alone in the meta-analysis of the Fox et al. (2022) review, further emphasizing the NIH Toolbox's relevance as a versatile platform for a wide array of uses.

As demand grows for scalable, reliable, and lifespan-sensitive measures, the NIH Toolbox is uniquely positioned to meet these needs. Its flexibility and scientific rigor make it a strong candidate for continued integration into clinical trials, educational practice, and health systems aiming to monitor cognitive and behavioral function over time. With each iteration, the NIH Toolbox continues to align with emerging priorities in neuropsychology, education, neuroscience, and personalized care—bridging gaps between research and practice in a rapidly changing landscape.

TEST YOURSELF

1. **Updates to the NIH Toolbox V3 Cognition domain prioritized renorming to address changes in cognitive performance across generations.**
   True or False
2. **Which of the following tests are new to NIH Toolbox V3?**
   (a) Flanker Inhibitory Control and Attention
   (b) Visual Reasoning
   (c) Picture Sequence Memory
   (d) Oral Reading Recognition
3. **The Crystallized Composite includes the Oral Reading Recognition and Visual Reasoning tests.**
   True or False
4. **Which of the following formats was NOT an objective for the new NIH Toolbox V3 app?**
   (a) Creating a test suite available on iPads, Android tablets, and computers
   (b) Improving the user interface and user experience
   (c) Updating test scoring and implementing new normative data
   (d) Streamlining the Cognition tests and updating their workflows
5. **When was NIH Toolbox V3 released?**
   (a) 2012
   (b) 2015
   (c) 2021
   (d) 2023
6. **What are the four domains in the NIH Toolbox?**
   (a) Cognition, Emotion, Intelligence, Motor
   (b) Cognition, Intelligence, Personality, Sensation
   (c) Cognition, Motor, Sensation, Speech
   (d) Cognition, Emotion, Motor, Sensation

7. **NIH Toolbox Cognition tests utilize item response theory (IRT) and computerized adaptive testing (CAT).**
   True or False
8. **Which of the following statements is true of the NIH Toolbox?**
   (a) Version 1 was available on computers via a downloadable, local software system.
   (b) Version 1 was available as a paper-and-pencil test kit.
   (c) Version 1 was available on computers via the web.
   (d) Version 1 was available on computers and iPads.
9. **The new Early Childhood Cognition CFPS Composite includes which of the following tests?**
   (a) Picture Vocabulary, Speeded Matching, Visual Reasoning
   (b) Speeded Matching, Pattern Comparison Processing Speed, Picture Vocabulary
   (c) Flanker Inhibitory Control and Attention, Picture Vocabulary, Picture Sequence Memory
   (d) Picture Sequence Memory, Speeded Matching, Visual Reasoning
10. **The Fluid Cognition Composite is new to NIH Toolbox V3.**
   True or False

Answers: 1. True; 2. (b); 3. False; 4. (a); 5. (d); 6. (d); 7. True; 8. (c); 9. (a); 10. False

REFERENCES

Akshoomoff, N., Beaumont, J. L., Bauer, P. J., Dikmen, S. S., Gershon, R. C., Mungas, D., & Heaton, R. K. (2013). NIH toolbox cognitive battery (CB): Composite scores of crystallized, fluid, and overall cognition. *Monographs of the Society for Research in Child Development*, *78*(4), 119–132.

Alloway, P. A., Gathercole, S. E., & Pickering, S. J. (2006). Verbal and visuospatial short-term and working memory in children: Are they separable? *Child Development*, *77*(6), 1698–1716.

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

American Psychological Association (APA). (2002). *Ethical principles of psychologists and code of conduct*. https://www.apa.org/ethics/code

Baddeley, A. (2003). Working memory and language: An overview. *Journal of Communication Disorders*, *36*(3), 189–208.

Bauer, P. J., Dikmen, S. S., Heaton, R. K., Mungas, D., Slotkin, J., & Beaumont, J. L. (2013). NIH Toolbox Cognition Battery (Cb): Measuring episodic memory. *Monographs of the Society for Research in Child Development*, *78*(4), 34–48.

Beaumont, J. L., Havlik, R., Cook, K. F., Hays, R. D., Wallner-Allen, K., Korper, S. P., … Gershon, R. (2013). Norming plans for the NIH toolbox. *Neurology*, *80*(11_supplement_3), S87–S92.

Ben-Porath, Y. S. (2012). Interpreting assessment data: Ethical and practical considerations. In J. N. Butcher (Ed.), *Oxford handbook of personality assessment* (pp. 144–161). Oxford University Press.

Blaisdell, C. J., Park, C., Hanspal, M., Roary, M., Arteaga, S. S., Laessig, S., … Program Collaborators for Environmental influences on Child Health Outcomes. (2022). The NIH ECHO Program: Investigating how early environmental influences affect child health. *Pediatric Research*, *92*(5), 1215–1216.

Bush, S. S., Ruff, R. M., Troster, A. I., Barth, J. T., Koffler, S. P., Pliskin, N. H., … Silver, C. H. (2005). Symptom validity assessment: Practice issues and medical necessity. *Archives of Clinical Neuropsychology*, *20*(4), 419–426.

Carlozzi, N. E., Beaumont, J. L., Tulsky, D. S., & Gershon, R. C. (2015). The NIH toolbox pattern comparison processing speed test: Normative data. *Archives of Clinical Neuropsychology*, *30*(5), 359–368.

Casaletto, K. B., Umlauf, A., Beaumont, J., Gershon, R., Slotkin, J., Akshoomoff, N., & Heaton, R. K. (2015). Demographically corrected normative standards for the English version of the NIH toolbox cognition battery. *Journal of the International Neuropsychological Society*, *21*(S1), 378–391.

Cella, D., Gershon, R., Lai, J.-S., & Choi, S. (2007). The future of outcomes measurement: Item banking, tailored short-forms, and computerized adaptive assessment. *Quality of Life Research*, *16*(S1), 133–141.

Coldwell, S. E., Mennella, J. A., Duffy, V. B., Pelchat, M. L., Griffith, J. W., Smutzer, G., & Hoffman, H. J. (2013). Gustation assessment using the NIH Toolbox. *Neurology*, *80*(11 Suppl. 3), S20–S24.

Cook, K. F., Dunn, W., Griffith, J. W., Morrison, M. T., Tanquary, J., Sabata, D., … Gershon, R. C. (2013). Pain assessment using the NIH Toolbox. *Neurology*, *80*(11 Supplement 3), S49–S53. https://doi.org/10.1212/WNL.0b013e3182872e80

Coyle, T. R., Pillow, D. R., Snyder, A. C., & Kochunov, P. (2011). Processing speed mediates the development of general intelligence (g) in adolescence. *Psychological Science*, *22*(10), 1265–1269.

Dale, E., & O'Rourke, J. (1976). *The living word vocabulary: The words we know: A national vocabulary inventory*. Field Enterprises Educational Corp.

Dalton, P., Doty, R. L., Murphy, C., Frank, R., Hoffman, H. J., Maute, C., & Slotkin, J. (2013). Olfactory assessment using the NIH Toolbox. *Neurology*, *80*(11 Suppl. 3), S32–S36.

Delis, D. C., Kaplan, E., & Kramer, J. H. (2001). *Delis-Kaplan executive function system (D-KEFS)*. Pearson.

Dempster, F. N. (1981). Memory span: Sources of individual and developmental differences. *Psychological Bulletin*, *89*(1), 63–100.

Dunn, L. M., & Dunn, D. M. (2007). *Peabody picture vocabulary test* (4th ed.). Pearson.

Dunn, W., Griffith, J. W., Morrison, M. T., Tanquary, J., Sabata, D., Victorson, D., & Gershon, R. C. (2013). Somatosensation assessment using the NIH Toolbox. *Neurology*, *80*(11 Suppl. 3), S41–S44.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception and Psychophysics*, *16*, 143–149.

Fan, J., McCandliss, B. D., Sommer, T., Raz, A., & Posner, M. I. (2002). Testing the efficiency and independence of attentional networks. *Journal of Cognitive Neuroscience*, *14*(3), 340–347.

Fox, R. S., Zhang, M., Amagai, S., Bassard, A., Dworak, E. M., Han, Y. C., … Gershon, R. C. (2022). Uses of the NIH Toolbox® in clinical samples: A scoping review. *Neurology. Clinical Practice*, *12*(4), 307–319.

Frye, D., Zelazo, P. D., & Palfai, T. (1995). Theory of mind and rule-based reasoning. *Cognitive Development*, *10*, 483–527.

Gasquoine, P. G. (2009). Race-norming of neuropsychological tests. *Neuropsychological Review*, *19*, 250–263.

Gershon, R. (1988). *Index of words in the Johnson O'Connor Research Foundation, Inc. Vocabulary item bank*. Johnson O'Connor Research Foundation Human Engineering Laboratory.

Gershon, R. (2005). Computer adaptive testing. *Journal of Applied Measurement*, *6*(1), 109–127.

Gershon, R., Cella, D., Dineen, K., Rosenblom, S., Peterman, A., & Lai, J. S. (2003). Item response theory and health-related quality of life in cancer. *Expert Review of Pharmacoeconomics and Outcomes Research*, *3*(6), 783–791.

Gershon, R., Cella, D., Fox, N. A., Havlik, R. J., Hendrie, H. C., & Wagster, M. V. (2010). Assessment of neurological and behavioral function: The NIH Toolbox. *Lancet Neurology*, *9*(2), 138–139.

Gershon, R., Cook, K. F., Mungas, D., Manly, J. J., Slotkin, J., Beaumont, J. L., & Weintraub, S. (2014). The language measures of the NIH toolbox cognition battery. *Journal of the International Neuropsychological Society*, *20*(6), 642–651.

Gershon, R., Wagster, M. V., Hendrie, H. C., Fox, N. A., Cook, K. F., & Nowinski, C. J. (2013). Responding to the needs of the neurology research community: introduction to the NIH toolbox for the assessment of neurological and behavioral function. *Neurology*, *80*(11 Suppl. 3), S2–S6.

Gillman, M. W., & Blaisdell, C. J. (2018). Environmental influences on child health outcomes, a research program of the NIH. *Current Opinion in Pediatrics*, *30*(2), 260–262.

Horn, J. L., & Cattell, R. B. (1967). Age differences in fluid and crystallized intelligence. *Acta Psychologica*, *26*, 107–129.

Rehabilitation Act of 1973. (1998). *Title number 29 U.S.C. §794d*. https://www.section508.gov/

Human Connectome Project. (n.d.). *About the HCP*. Retrieved July 30, 2025, from https://www.humanconnectome.org/

LaForte, E. M., Hook, J. N., & Giella, A. K. (2024). *National Institutes of Health (NIH) Toolbox® V3 technical manual*. Toolbox Assessments. Inc.

Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). *Neuropsychological assessment* (5th ed.). Oxford University Press.

Ma, Y., Carlsson, C. M., Wahoskie, M. L., Blazel, H. M., Chappel, R. J., Hohnson, S. C., … Gleason, C. E. (2020). Latent factor structure and measurement invariance of the NIH toolbox cognition battery in an Alzheimer's disease research sample. *Journal of International Neuropsychologic Society*, *27*(5), 412–425.

McKenzie, F. J., Drayton, A., Shields, R. H., Dakopolos, A., Glassman, D., Kaat, A., Coleman, J., Thompson, T., Sansone, S., Riley, K., Berry-Kravis, E., Gershon, R. C., & Hessl, D. (2023). *National Institutes of Health Toolbox® V3 cognition battery supplemental administration manual for intellectual and developmental disabilities: A guide on administration and scoring standards [Supplemental material]. NIH Toolbox*. https://nihbabytoolbox.org/wp-content/uploads/2023/07/NIHTB-V3-CB-IDD-Supplemental-Manual_07052023.pdf

Mungas, D., Heaton, R., Tlsky, D., Zelazo, P. D., Slotkin, J., Blitz, D., … Gershon, R. (2014). Factor structure, convergent validity, and discriminant validity of the NIH toolbox cognitive health battery (NIHTB-CHB) in adults. *Journal of the International Neuropsychological Society*, *20*, 579–587.

Mungas, D., Widaman, K., Zelazo, P. D., Tulsky, D., Heaton, R. K., Slotkin, J., & Gershon, R. (2013). NIH Toolbox cognition battery (CB): Factor structure for 3–15 year olds. *Monographs of the Society for Research in Child Development*, *78*(4), 103–118.

National Institute of Neurological Disorders and Stroke. (n.d.). *NIH blueprint overview flyer* [PDF]. US Department of Health and Human Services. Retrieved July 24, 2025, from https://www.ninds.nih.gov/sites/default/files/documents/NIH_Blueprint_Overview_Flyer_508C.pdf

National Institutes of Health. (2015). *Adolescent brain cognitive development study*.

Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., & Ceci, S. J. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, *51*(2), 77–101.

Nowinski, C. J., Victorson, D., Debb, S. M., & Gershon, R. C. (2013). Input on NIH Toolbox inclusion criteria: Surveying the end-user community. *Neurology*, *80*(11 Suppl. 3), S7–S12.

Rea-Sandin, G., Korous, K. M., & Causadias, J. M. (2021). A systematic review and meta-analysis of racial/ethnic differences and similarities in executive function performance in the United States. *Neuropsychology*, *35*(2), 141–156.

Rentz, D. M., Amariglio, R. E., Becker, J. A., Frey, M., Olson, L. E., Frishe, K., … Sperling, R. A. (2011). Face-name associative memory performance is related to amyloid burden in normal elderly. *Neuropsychologia*, *49*(9), 2776–2783.

Reuben, D. B., Magasi, S., McCreath, H. E., Bohannon, R. W., Wang, Y. C., Bubela, D. J., & Gershon, R. C. (2013). Motor assessment using the NIH Toolbox. *Neurology*, *80*(11 Suppl. 3), S65–S75.

Rine, R. M., Schubert, M. C., Whitney, S. L., Roberts, D., Redfern, M. S., Musolino, M. C., & Slotkin, J. (2013). Vestibular function assessment using the NIH Toolbox. *Neurology*, *80*(11 Suppl. 3), S25–S31.

Rohling, M. L., Miller, R. M., Axelrod, B. N., Wall, J. R., Lee, A. J. H., & Kinikini, D. T. (2015). Is co-norming required? *Archives of Clinical Neuropsychology*, *30*, 611–633.

Rueda, M. R., Fan, J., McCandliss, B. D., Halparin, J. D., Gruber, D. B., Lercari, L. P., & Posner, M. I. (2004). Development of attentional networks in childhood. *Neuropsychologia*, *42*, 1029–1040.

Russell, E. W. (2004). The operating characteristics of the major HRNES-R measures. *Archives of Clinical Neuropsychology*, *19*, 1043–1061.

Salsman, J. M., Butt, Z., Pilkonis, P. A., Cyranowski, J. M., Zill, N., Hendrie, H. C., & Cella, D. (2013). Emotion assessment using the NIH toolbox. *Neurology*, *80*(11 Suppl. 3), S76–S86.

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). Guilford.

Squire, L. R. (2004). Memory systems of the brain: a brief history and current perspective. *Neurobiology of Learning and Memory*, *82*, 171–177.

Thompson, R. C., Montena, A. L., Liu, K., Watson, J., & Warren, S. L. (2022). Associations of family distress, family income, and acculturation on pediatric cognitive performance using the NIH Toolbox: Implications for clinical and research settings. *Archives of Clinical Neuropsychology*, *37*(4), 798–813.

Tulsky, D. S., Carlozzi, N., Chiaravalloti, N., Beaumont, J. L., Conway, K., & Mungas, D. (2014). NIH Toolbox cognition battery (NIHTB-CB): Measuring working memory. *Journal of the International Neuropsychological Society*, *20*(6), 599–610.

Tulsky, D. S., Holdnack, J. A., Cohen, M. L., Heaton, R. K., Carlozzi, N. E., Wong, A. W. K., & Heinemann, A. W. (2017). Factor structure of the NIH Toolbox cognition battery in individuals with acquired brain injury. *Rehabilitation Psychology*, *62*(4), 435–442.

U.S. Census Bureau. (2021). *2020 Census data*. U.S. Department of Commerce. https://www.census.gov/programs-surveys/decennial-census/decade/2020.html

Varma, R., McKean-Cowdin, R., Vitale, S., Slotkin, J., & Hays, R. D. (2013). Vision assessment using the NIH Toolbox. *Neurology*, *80*(11 Suppl. 3), S37–S40.

Wechsler, D. (2008). *Wechsler adult intelligence scale IV*. Harcourt Assessment.

Weintraub, S., Dikmen, S. S., Heaton, R. K., Tulsky, D. S., Zelazo, P. D., Bauer, P. J., & Gershon, R. C. (2013). Cognition assessment using the NIH Toolbox. *Neurology*, *80*(11 Suppl. 3), S54–S64.

Weintraub, S., Dikmen, S. S., Heaton, R. K., Tulsky, D. S., Zelazo, P. D., Slotkin, J., & Gershon, R. (2014). The cognition battery of the NIH toolbox for assessment of neurological and behavioral function: Validation in an adult sample. *Journal of the International Neuropsychological Society*, *20*(6), 567–578.

Zecker, S. G., Hoffman, H. J., Frisina, R., Dubno, J. R., Dhar, S., Wallhagen, M., & Wilson, R. H. (2013). Audition assessment using the NIH Toolbox. *Neurology*, *80*(11 Suppl. 3), S45–S48.

Zelazo, P. D., Anderson, J. E., Richler, J., Wallner-Allen, K., Beaumont, J. L., Gershon, R., … Weintraub, S. (2014). NIH toolbox cognition battery (CB): Validation of executive function measures in adults. *Journal of the International Neuropsychological Society*, *20*(6), 620–629.