Technical Manual

Release 1.0
May 6, 2024

Edited by
Jerry Slotkin

Table of Contents

Introduction

The NIH Toolbox® for Assessment of Neurological and Behavioral Function was introduced in a commentary published in Lancet Neurology in 2010 by Gershon and colleagues. NIH Toolbox (NIHTB) was developed with support from the NIH Blueprint for Neuroscience Research and was designed to serve as a "common currency" for in-person neurological and behavioral assessment across research and clinical applications. Since that time, the NIH Toolbox, released in 2011 for computer administration and in 2015 for administration on an iPad tablet, has been used in hundreds of studies and is currently being used at approximately 1,100 institutions worldwide. Technological advances over the past decade, however, have now made it possible to bring NIHTB-style measurement tools away from traditional approaches that rely on in-person testing, and instead leverage mobile technology to bring assessment into participants' own homes. As smartphones become increasingly pervasive across age, race, and economic status, the capacity to reach participants on their phones becomes increasingly possible. By utilizing mobile devices to assess participants, researchers can reach population-level "samples" without the need to bring participants in to clinical or research laboratory settings to complete cognitive test protocols. Although the availability of smartphone tests does not eliminate the need for in-person visits in many contexts (e.g., for studies and clinical applications where individual-level reliability is paramount, for those without access to a smartphone, or for those who are unable to complete a smartphone test due to impairment), remote assessment offers the potential to transform clinical research assessments of neuropsychological function in medical care.

To address the challenge of providing smartphone-based assessments, the National Institute on Aging (NIA) made multiple awards to create the "Mobile Toolbox" (MTB), a library of cognitive tests and, ultimately, measures assessing other domains of health and function, as well as a platform for study management, data collection, and data management (www.mobiletoolbox.org; funded under NIH awards U2CAG060426 and U02AG060408). Mobile Toolbox was designed to enable the integration of mobile cognitive (and other) testing into clinical and epidemiological research studies (Gershon et al., 2022).

The MTB currently consists of a library of cognitive tests, with administration, data collection, and data and study management integrated into REDCap (Harris et al., 2019; Harris et al., 2009) and its MyCap (Harris et al., 2022) mobile assessment application. MTB is designed to enable the integration of mobile cognitive testing (and, eventually, measures of behavioral, psychological, and social functioning) into clinical and epidemiological research studies. Evidence for the validity of these measures is rapidly expanding in the general population and in various clinical conditions. Clinical researchers may choose to use the MTB system to (1) design smartphone-based test batteries using validated measures; (2) deploy and manage mobile data collection in their research studies; (3) use smartphone technology to interact with and engage research participants; and (4) aggregate and analyze results in the context of large-scale data that is openly available through the same system. Given that all MTB components are open source, researchers and developers may also choose to integrate tests and supplemental scales into external development efforts.

MTB includes cognitive measures adapted for smartphone administration from NIHTB tests or developed *de novo* to measure comparable areas of cognition within the NIHTB Cognitive Battery. Upcoming releases will include measures adapted from the International Cognitive Ability Resource (ICAR; Condon & Revelle, 2014; Dworak et al., 2021; Revelle et al., 2020; Young & Keith, 2020), measures of psychosocial and emotional functioning, and Spanish versions of all cognitive measures.

This manual describes the development and technical data that underlies each of the published MTB measures of cognitive functioning. The measures developed and published for MTB include the following:

- Word Meaning
- Spelling
- Arrow Matching
- Shape-Color Sorting
- Number-Symbol Match
- Sequences
- Faces and Names
- Arranging Pictures

Psychometric data on validity and reliability for each MTB measure described in this manual were obtained through three separate, related studies.

- *Study 1*: The goal of this study was to evaluate each MTB measure's internal consistency, convergent and discriminant validity, and relationship to age when completed in ideal circumstances - in a lab setting on a study-provided device. Participants were recruited from a third-party market research firm. Ninety-two participants completed the MTB measures on study-provided iPhones (which exclusively use Apple's iOS operating system), unproctored in a private room in the lab. In addition to MTB tests, these participants were administered cognition tests from the NIHTB Version 3 (NIHTB V3), which was published in 2023 (Laforte et al., 2024). Participants also completed external measures of similar constructs to those purportedly assessed by the MTB measures. Both NIHTB V3 and external measures were administered to evaluate convergent validity; in addition, selected NIHTB V3 measures were chosen *a priori* and used to evaluate discriminant validity with MTB measures. Because of the small number of education level subgroups in the Study 1 sample (only 2 participants in the less than high school group), correlations with education were not examined.
- *Study 2:* The goal of this study was to evaluate each MTB measure's internal consistency, convergent and discriminant validity, relationship to age and education, and the influence of operating system (OS) when taken remotely on a personal smartphone (either iOS or Android). The recruitment and testing were conducted by the same third-party market research firm as Study 1. A sample of 1,021 participants was administered cognition tests from the NIHTB V3 in the lab, and then asked to complete MTB tests on their own iOS or Android smartphone remotely within 14 days. Specific NIHTB measures were used to evaluate convergent and discriminant validity with corresponding MTB measures.

Relationship to education was evaluated only for MTB measures of "crystallized" ability (Word Meaning and Spelling).

- *Study 3:* The goal of this study was to examine the test-retest reliability of MTB measures when taken remotely on a personal smartphone (iOS only). Participants were recruited through the Brain Health Registry, an online database that facilitates aging research (Weiner et al., 2018) . Participants were excluded from the sample if they either (a) used an inappropriate device (i.e., an iPad) or (b) switched operating systems between the first and second administration (i.e., from an Android to an iOS device, or vice versa). Participants completed each MTB test twice remotely, with a two-week delay interval. In total, 168 participants completed at least one MTB test and corresponding retest; therefore, actual sample sizes reported for each test will vary. In this sample, the test-retest intraclass correlation was examined.

In addition, MTB Word Meaning and Spelling tests first conducted item calibrations as part of the test development process (described in more detail for each test, below). Demographic information on the pilot/calibration and validation study samples is provided in Tables 1 and 2.

Specific scores provided are described below in each MTB test's section. It should be noted that normative scores are not currently available for MTB but may be derived in the future. In addition, given the self-administered nature of the MTB tests, *it is recommended that MTB scores only be evaluated and analyzed in aggregate, for the purposes of research or program evaluation. Reporting or interpreting individual scores is strongly discouraged*.

Table 1. *Demographic Characteristics of the MTB Calibration and Pilot Study Samples*

	Word Meaning Calibration (N = 7,525)	**Spelling Online Pilot Study** (N = 1,950)	**Spelling On-App Calibration Study** (N = 2,099)
Age			
Mean (SD)	53.24 (16.41)	23.87 (19.22)	23.61 (18.95)
Range	[18, 89]	[8, 88]	[8, 88]
	n (%)	**n (%)**	**n (%)**
Gender			
Female	4,285 (56.94)	Not Asked	Not Asked
Male	3,227 (42.88)	Not Asked	Not Asked
Not Identified	4 (0.05)	Not Asked	Not Asked
Racial Identity			
Asian	175 (2.33)	86 (4.4)	90 (4.3)
Black or African American	872 (11.59)	278 (14.3)	304 (14.5)
Multiracial or More Than One Race	240 (3.19)	Not Asked	Not Asked
Native American or Alaska Native	36 (0.48)	21 (1.1)	23 (1.1)
Native Hawaiian or Pacific Islander	21 (0.28)	1 (0.1)	1 (0.05)
White or Caucasian	5,695 (75.68)	1,309 (67.1)	1,415 (67.4)
Other	486 (6.46)	Not Asked	Not Asked
Multiracial or More Than One Race/Other	Not Asked	203 (10.4)	211 (10.1)
Not Identified	0 (0)	52 (2.7)	55 (2.6)
Ethnic Identity			
Not Hispanic/Latino (Any Race)	240 (3.19)	1,657 (85.0)	1,786 (85.1)
Hispanic/Latino (Any Race)	486 (6.46)	293 (15.0)	313 (14.9)
Not Identified	0 (0)	0 (0)	0 (0)
Education Level or Mother's Education Level			
Less than HS	39 (0.52)	140 (7.2)	163 (8.4)
HS Diploma or GED	808 (10.74)	739 (37.9)	799 (41)
Some College	2,855 (37.94)	363 (18.6)	375 (19.2)
4-year College Degree	2,466 (32.77)	Not Asked	Not Asked
Graduate or Professional Degree	1,357 (18.03)	Not Asked	Not Asked
Bachelor’s Degree or Higher	Not Asked	642 (32.9)	691 (35.4)
Not Identified	—	66 (3.4)	71 (3.6)

Table 2. *Demographic Characteristics of the MTB Validation Study Samples*

	Study 1 (N = 92)	**Study 2** (N = 1,021)	**Study 3** (N = 168)
Age			
Mean (SD)	49.27 (17.65)	43.97 (21.24)	63.54 (12.10)
Range	[20, 84]	[18, 90]	[28, 87]
	n (%)	**n (%)**	**n (%)**
Device Type (OS)			
iOS	92 (100)	650 (63.66)	168 (100)
Android	0 (0)	371 (36.34)	0 (0)
Gender			
Female	62 (67.39)	568 (55.63)	141 (83.93)
Male	30 (32.61)	453 (44.37)	27 (16.07)
Not Identified	0 (0)	0 (0)	0 (0)
Racial Identity			
Asian	9 (9.78)	64 (6.27)	5 (2.98)
Black or African American	30 (32.61)	142 (13.91)	7 (4.17)
Middle Eastern or North African	0 (0)	9 (0.88)	0 (0)
Multiracial or More Than One Race	4 (4.35)	22 (2.15)	5 (2.98)
Native American or Alaska Native	0 (0)	7 (0.69)	0 (0)
Native Hawaiian or Pacific Islander	1 (1.09)	5 (0.49)	0 (0)
White or Caucasian	48 (52.17)	752 (73.65)	149 (88.69)
Other	0 (0)	0 (0)	1 (0.60)
Not Identified	0 (0)	20 (1.96)	0 (0)
Ethnic Identity			
Not Hispanic/Latino (Any Race)	91 (98.91)	871 (85.31)	156 (92.86)
Hispanic/Latino (Any Race)	1 (1.09)	150 (14.69)	12 (7.14)
Not Identified	0 (0)	0 (0)	0 (0)
Education Level			
Less than HS	2 (2.17)	17 (1.67)	0 (0)
HS Diploma or GED	50 (54.35)	327 (32.03)	1 (0.60)
Some College	19 (20.65)	359 (35.16)	43 (25.60)
4-year College Degree	14 (15.22)	207 (20.27)	55 (32.74)
Graduate or Professional Degree	7 (7.61)	111 (10.87)	69 (41.07)

Word Meaning

Description of the Task

The MTB Word Meaning test was designed to measure receptive vocabulary – the same construct as the NIH Toolbox Picture Vocabulary Test (Gershon et al., 2014) – and to provide a broad estimate of general intelligence. The test is a multiple-choice synonym test optimized for self-administration on a smartphone. A simple presentation was chosen that includes written words presented as the target and distractors, which maximizes accessibility and eases administration on a mobile phone in a variety of potential settings (because there is no audio component). For each item, a target word is presented, followed by 5 answer choices. Participants tap or touch one of the answer choices to indicate their response. The test is administered as a computer-adaptive test (CAT). The CAT algorithm includes maximum likelihood item response theory (IRT) scoring to compute the participant's interim ability estimate after each completed trial and uses that score to select each subsequent item based on the maximum information item selection criterion with Sympson-Hetter item exposure control. All participants must complete a minimum of 15 items. Once the item minimum is met, the test ends when a sufficiently small standard error is obtained (approximately equivalent to a person-specific reliability of 0.75) or when the participant has completed a maximum of 25 items.

Item Pool Development

The item pool for Word Meaning was based on a word corpus from the Johnson O'Connor Foundation; the same corpus that served as one of the sources for the NIH Toolbox Picture Vocabulary Test. The Johnson O'Connor Foundation has curated a list of English vocabulary words with associated distractors for decades (Shand, 1994). To develop the item pool for MTB Word Meaning, this list of target words and distractors was reviewed by a team of measure development experts and reduced to a subset aimed at the average adult vocabulary ability. Any words that were deemed culturally insensitive or that could trigger strong emotional responses were also eliminated during the review. Although the Johnson O'Connor item bank as originally developed was calibrated using the Rasch model, the MTB item bank was recalibrated given the reduction in items and change in presentation format to mobile phone. The MTB recalibration used an overlapping block design administration (75 forms) with 7,525 participants, each completing 100-103 items. The original calibrations from Johnson O'Connor were used as "precalibration" difficulties to create the overlapping forms, then final calibrations were calculated using this new data. Calibration results were then reviewed, and additional items were removed if they had fewer than 15 percent correct responses or greater than 85 percent correct responses. The final item pool contains 1,948 items across two overlapping (statistically equivalent) pools, with 1,364 items each. Difficulty parameters range from -1.89 to 2.41.

Available Scores

The MTB Word Meaning test provides the following score:

- *Theta score:* This score underlies the CAT algorithm and the final theta score value estimates the person's "true" ability for vocabulary. Higher scores indicate higher levels of ability.

- *SE of Theta:* This is the standard error associated with the final theta score of an individual's estimated vocab ability.

Reliability and Validity

- *Content validity:* The evidence for content validity was primarily established during the original creation and calibration of the Johnson-O'Connor Foundation item bank (Shand, 1994). As noted, a team of MTB measurement experts reviewed all items to identify those most appropriate for adults. Items were then evaluated for sensitivity and fairness, and any items of concern were removed.
- *General procedures for validation studies are described above. Specific analyses conducted for each validation study of Word Meaning were as follows:*
    - *Study 1*: In addition to Word Meaning, 92 participants were administered the NIH Toolbox Picture Vocabulary Test, Version 3 (NIHTB V3 PVT). Seventy-eight of these participants also completed the Peabody Picture Vocabulary Test, Fifth Edition (PPVT-5; Dunn & Dunn, 2019) in the same testing session. These tests were administered to evaluate convergent validity.

      In addition, 92 participants were administered the Rey Auditory Verbal Learning Test (RAVLT) from the NIH Toolbox Version 3, to evaluate the discriminant validity of Word Meaning; however, only 83 of these participants completed the delayed RAVLT.

      In the Study 1 sample, the reliability of Word Meaning was examined through empirical and person-specific reliability statistics. Convergent validity was examined through Spearman Rho correlations with the NIHTB V3 PVT change-sensitive scores and PPVT-5 growth scale values (both of which are person-ability scores). Likewise, discriminant validity was examined through Spearman Rho correlations with the RAVLT immediate and delayed change-sensitive scores. Spearman Rho correlations were also examined with age.
    - *Study 2:* A sample of 1,008 participants were administered the NIHTB V3 PVT and the RAVLT in the lab, and then asked to complete Word Meaning on their own iOS or Android smartphone remotely within 14 days. Reliability was examined through empirical and mean person-specific reliability statistics and convergent validity using Spearman Rho correlations with change-sensitive scores on the NIHTB V3 PVT and discriminant validity with the RAVLT immediate and delayed change-sensitive scores. Spearman Rho correlations with age and education level were also examined. Because participants used their personal devices in this group, a t-test to compare mean theta scores and reliability across and between operating systems (Android vs. iOS) was also conducted.
    - *Study 3:* A sample of 141 participants completed Word Meaning twice remotely, with a two-week delay interval. In this sample, the test-retest intraclass correlation was examined.

Demographic information on the study samples, including the calibration study, is provided in Tables 1 and 2.

All reliability and validity data for Word Meaning are shown in Table 3 below (Young et al., 2024).

- *Convergent validity*: Word Meaning correlated .70 with NIHTB V3 PVT and .75 with PPVT-5 in an in-person sample. In a remote data collection, Word Meaning correlated .67 with NIHTB V3 PVT. All correlations show strong support that Word Meaning measures the target construct successfully; moreover, there is little difference when Word Meaning is administered remotely and unproctored, which suggests it can be validly administered in such a context.
- *Discriminant validity*: Word Meaning showed no significant correlation with RAVLT, whether administered in person or remotely, and whether looking at immediate or delayed RAVLT recall. These near-zero correlations between ostensibly unrelated constructs provide further support for Word Meaning construct validity.
- *Age-related validity*: Word Meaning correlated with age .35 in an in-person data collection and .45 in a large remote data collection. These are similar but a bit higher than that reported on the original NIHTB PVT for adults (.26). A small but significant correlation with level of education was also found. Overall, results suggest a slight age effect, but as expected, much lower than a measure of fluid ability (e.g., processing speed, working memory).
- *Reliability*: Test-retest reliability ICC was .65 when Word Meaning was re-administered 2 weeks after the first administration. This is a reasonably high correlation for a completely remote assessment implementation, given that participant behavior (e.g., dishonesty, inattention) can neither be monitored nor influenced. Empirical and person-specific reliability coefficients were somewhat higher. Note that the CAT is designed to stop administering items once a person-specific reliability of 0.75 has been achieved, and thus these results are consistent with the CAT settings.

Table 3. *Reliability and Validity Statistics for MTB Word Meaning Across Three Studies*

	Study 1		Study 2		Study 3	
	Value	***N***	**Value**	***N***	**Value**	***N***
Mean Theta Score (SD)	-.94 (.98)	92	-.77 (1.01)	1,008	.40 (.77)	141
Empirical Reliability						
Total Sample	—	—	.81	1,008	—	—
iOS	.78	92	.79	643	—	—
Android	—	—	.83	365	—	—
Person-Specific Reliability						
Total Sample	—	—	.76	1,008	—	—
iOS	.74	92	.73	643		
Android	—	—	.81	365	—	—
Test-Retest Reliability	—	—	—	—	.65	141
Convergent Validity						
PPVT-5	.75	78	—	—	—	—
NIHTB V3 Picture Vocabulary	.70	92	.67	1,001	—	—
NIHTB V3 Crystallized Cognition	.70	92	.65	1,001	—	—
Discriminant Validity						
NIHTB V3 RAVLT Immediate	.14 (ns)	92	-.01 (ns)	992	—	—
NIHTB V3 RAVLT Delayed	.13 (ns)	83	-.05 (ns)	932	—	—
Demographic Correlations						
Age	.35	92	.44	1,008	—	—
Education	—	—	.26	1,008	—	—

Note: All correlation values are significant at $p < .001$ unless otherwise noted. ns = non-significant ($p \geq .05$).

Spelling

Description of the Task

The MTB Spelling test was created *de novo* and was designed to measure spelling skill, given that research suggests the construct is sensitive to condition-related cognitive change (e.g., Alzheimer's disease and related dementias). The Spelling items are presented via audio-recorded sound clips in a traditional dictation format: the word is presented orally, followed by the word used in a short sentence, followed by the word again (e.g., *Work. I had to go to work. Work.*). Immediately after the audio stimulus is presented, the participant is asked to spell the word, using an on-screen keyboard. When the participant has finished typing the word and is satisfied with their response, they tap the *done* button to submit the response and move to the next item. The keyboard used for responding is customized to remove the autocorrect and speech-to-text features available in a native smartphone keyboard; Shift, Number, and Space keys were also removed. Examinees can self-correct errors using the backspace key before submitting responses. Examinees may listen to the audio stimulus as many times as they wish before responding. Spelling items are administered via a CAT algorithm. If the individual has a previous Spelling administration on the same device, the CAT begins at a level slightly easier than the endpoint for the previous administration. If there wasn't a previous administration but the CAT program is aware of an examinee's education level (e.g., through the research team's other information), that level informs selection of the first item in the CAT administration. If there was not a previous administration or a known education level, the first item is given at the population average across education levels, and then the algorithm selects subsequent items using the Maximum Information item selection rule based on the examinee's interim ability estimate computed using the Maximum Likelihood method after each item. All examinees must complete a minimum of 10 items. Once the item minimum is met, the test ends when a sufficiently small standard error is obtained (approximately equivalent to a reliability of 0.75) or when the examinee has completed a maximum of 30 items.

Item Pool Development

For the Spelling test, 500 new items were developed, with almost half targeted at grades 7-12, and the remainder split roughly evenly at grade 6 and below or college level, to ensure a broad range of difficulties were assessed. Words were included in the item pool representing all possible American English phoneme/grapheme combinations, both common and uncommon; in addition, slightly less than 10 percent of the newly developed word list were homophones.

Two preliminary studies on the newly developed items were conducted. The first was a pilot study designed to preliminarily estimate the Spelling item difficulties and to assess their distribution. The initial pilot study was conducted using an online market research platform, in which 30 overlapping forms of 50 items each were administered. Each item appeared on between 1 and 5 forms.

Participants in the first pilot study were 1,950 individuals ages 8 through 88 who were recruited by a third-party market research firm. The sample included 1,229 children ages 8 through 17 (M_{age} = 12.4 years; *SD* = 2.8 years) and 721 adults ages 18 and older (M_{age} = 43.5 years; *SD* = 19.4

years). Children were included in the pilot study to ensure information about the difficulty of the easiest items in the pool was attained.

Items on each form were presented in random order, one item per screen. The test ended when the examinee completed all 50 items or when the administration time reached 30 minutes, whichever occurred first. The number of items completed by each examinee ranged from 26 to 50 (Mean = 44 items, *SD* = 4.5, Median = 45 items). Approximately 3% of the examinees completed all 50 items. Items were preliminarily calibrated to create more accurately targeted forms for the second pilot (calibration) study.

Following the first pilot, a second calibration study was conducted to estimate the difficulty of the items when administered through the MTB app. Again, 30 overlapping fixed forms of 50 items each were used, but with some items moved across forms based on the first pilot study. Each item appeared on two to four forms.

For calibration, the Spelling test was administered via smartphone on the MTB app to 1,335 children ages 8 to 17 (M_{age} = 12.4, *SD* = 2.8) and 764 adults ages 18 and older (M_{age} = 43.3, *SD* = 19.2). Items on each form were presented in random order, one item per screen. The test ended when the examinee completed all 50 items or when the administration time reached 30 minutes, whichever occurred first. The number of items completed by each examinee ranged from 26 to 50 (*M* = 45 items, *SD* = 6.5) items (Median = 46 items). Approximately 5 percent of the examinees completed all 50 items. The final scoring model selected from the calibration study was a 2-Parameter Logistic IRT model. Results were used to create the final item pool used in the validation and reliability studies, described below.

Available Scores

The MTB Spelling test provides the following score:

- *Theta score:* This score underlies the CAT algorithm and the final theta score value estimates the person's "true" ability for spelling. Higher scores indicate higher levels of ability.
- *SE of Theta:* This is the standard error associated with the final theta score of an individual's estimated spelling ability.

Reliability and Validity

- *Content validity:* Content validity evidence primarily consisted of following a detailed blueprint for item (word) development and dictated sentences, with subsequent review by a test development expert for accuracy, fairness, and cultural sensitivity.
- *General procedures for validation studies are described above. Specific analyses conducted for each validation study of Spelling were as follows:*
    - *Study 1*: In addition to Spelling, 92 participants were administered the NIH Toolbox Oral Reading Recognition Test, Version 3 (NIHTB V3 ORRT). Sixty-nine of these participants also completed the Spelling subtest from the *Wechsler Individual Achievement Test, 4th Edition* (WIAT-IV; Pearson, 2020). The WIAT-IV Spelling subtest is a measure of spelling dictation accuracy in written expression; it is

administered to an individual examinee by a trained examiner. NIHTB ORRT and WIAT-IV Spelling were administered to evaluate convergent validity.

In addition, participants were administered the RAVLT and the Visual Reasoning Test (VR) from the NIH Toolbox Version 3, to evaluate the discriminant validity of Spelling.

In the Study 1 sample, the reliability of Spelling was examined through empirical and person-specific reliability statistics. Convergent validity was examined through Spearman Rho correlations with the ORRT change-sensitive scores and WIAT-IV Growth-Score Value scores. Likewise, discriminant validity was examined through Spearman Rho correlations with the RAVLT immediate and delayed change-sensitive scores, as well those with VR. Spearman Rho correlations were also examined with age.

- *Study 2:* A sample of 977 participants were administered the ORRT, the RAVLT, and VR in the lab, and then asked to complete Spelling on their own iOS or Android smartphone remotely within 14 days. In this study, reliability was examined through empirical and mean person-specific reliability statistics and convergent validity using Spearman Rho correlations with change-sensitive scores on the ORRT and discriminant validity with VR and with the RAVLT immediate and delayed change-sensitive scores. Spearman Rho correlations were also examined with age and education level. Because participants used their personal devices in this group, a t-test to compare mean theta scores and reliability across and between operating systems (Android vs. iOS) was also conducted.
- *Study 3:* A sample of 132 participants completed Spelling twice remotely, with a two-week delay interval. In this sample, the test-retest intraclass correlation was examined.

Demographic information on the study samples, including the pilot and calibration studies, is provided in Tables 1 and 2.

All reliability and validity data for Spelling are shown in Table 4 below.

- *Convergent validity*: Spelling correlated .67 with NIHTB V3 ORRT and .81 with WIAT-IV in an in-person sample. In a remote data collection, Spelling correlated .56 with NIHTB V3 ORRT. All correlations show strong support that Spelling measures the target construct successfully; however, there is some difference when Spelling is administered remotely and unproctored vs. in-person. While it can be validly administered in a remote context, a focus on group results will be important.
- *Discriminant validity*: Spelling showed a small but significant correlation with RAVLT immediate recall and VR, whether administered in person or remotely. However, these correlations were far lower than convergent validity correlations, as one would expect in measures of discriminant validity. The correlation for RAVLT delayed recall was slightly higher in the in-person sample but was significant in the remote sample due to the much larger sample size. These lower correlations with RAVLT and VR provide supporting evidence of discriminant validity.

- *Age-related validity*: Spelling had a near zero correlation with age for both in-person and remote data collection. A small but significant correlation with level of education was found, which would be expected. The lack of an age effect suggests relatively consistent performance throughout adulthood, which may make Spelling a useful premorbid functioning measure.
- *Reliability*: Test-retest reliability ICC was .63 when Spelling was re-administered 2 weeks after the first administration. This is a reasonably high correlation for a completely remote assessment implementation, given that participant behavior (e.g., dishonesty, inattention) can neither be monitored nor influenced. Empirical and Person-Specific reliability coefficients were substantially higher.

Table 4. *Results from the MTB Spelling Validation Studies*

	Study 1		**Study 2**		**Study 3**	
	Value	***N***	**Value**	***N***	**Value**	***N***
Mean Theta Score (SD)	.05 (.81)	92	.24 (.72)	977	.65 (.76)	132
Empirical Reliability						
Total Sample	—	—	.79	977	—	—
iOS	.83	92	.78	627	—	—
Android	—	—	.80	350	—	—
Person-Specific Reliability						
Total Sample	—	—	.86	977	—	—
iOS	.86	92	.86	627		—
Android	—	—	.86	350	—	—
Test-Retest Reliability	—	—	—	—	.63	132
Convergent Validity						
WIAT-IV Spelling Subtest	.81	69	—	—	—	—
NIHTB V3 Oral Reading Recognition	.67	92	.56	968	—	—
Discriminant Validity						
NIHTB V3 Visual Reasoning	.36	92	.25	967	—	—
NIHTB V3 RAVLT Immediate	.33	92	.23	961	—	—
NIHTB V3 RAVLT Delayed	.22	83	.14	905	—	—
Demographic Correlations						
Age	-.06	92	-.04	977	—	—
Education	—	—	.27	977	—	—

Note*:* All correlation values are significant at $p < .001$ unless otherwise noted.

Arrow Matching

Description of the Task

The MTB Arrow Matching test was designed to assess inhibitory control, a component of executive functioning. Based on the original Eriksen flanker task as well as the NIHTB Flanker Inhibitory Control and Attention Test (Zelazo et al., 2013), examinees indicate whether a central stimulus is oriented to the left or right, while inhibiting focus on potentially incongruent flanking stimuli on either side.

Designed to be taken in landscape orientation on a smartphone screen, Arrow Matching presents five arrows horizontally. Four flanking arrows, all of which are oriented in the same direction, appear 100 milliseconds (ms) prior to a central arrow. Examinees then have 2,000 ms to respond by indicating the direction of the central arrow, tapping one of two buttons. A centrally located star rotates during a variable inter-stimulus-interval (ISI) of 500 ms, 1,250 ms, or 2,000 ms.

One difference between MTB Arrow Matching and NIHTB Flanker, Version 3 (NIHTB V3 Flanker) was the addition of more trials (50 vs 30), with less time allotted for each item (2,000 ms vs. 3,000 ms). This faster auto-advance, combined with a variable ISI, was implemented to increase task difficulty, with the goal of expanding distribution of performance.

Item Pool Development

For the Arrow Matching test, MTB built off the design of the NIHTB Flanker Test. Overall, 100 arrow items were developed for the pool, of which 50 are administered on a given form (100 items allow for multiple forms administration). Of the 50 items administered in pseudo-random order, 60 percent are congruent, and 40 percent are incongruent trials. ISI length is roughly equally distributed among the items, but also in pseudo-random order.

Available Scores

The MTB Arrow Matching test provides the following score:

- *Rate Correct per Second (RCS):* Arrow Matching score uses the number of correct trials per second as a derived score. Doing so creates a varied distribution of possible scores, with up to 50 correct answers and up to 2,000 ms allowed per response. In this way, both accuracy and reaction time are valued. RCS varies from 0 to 10 (i.e., 1 correct per 100ms - about the fastest a human can work) with most scores falling somewhere between 0.7 and 2.5. Note that RCS is the primary score, and the one for which validity evidence (below) is reported. In REDCap, RCS is reported as “rateScore.”
- *Raw Score:* Sum of correct trials. This score reflects the accuracy of a participant, regardless of speed. It provides the numerator of the primary RCS score, with a range of 0-50.
- *Number of Anticipation Errors:* This is the sum of errors a person commits by pressing a response button on an interim screen when no test stimuli are presented. This is tracked on live items only, with a range from 0 to 100.

- *Median Correct Reaction Time:* For the items on which the examinee responds correctly, this is the median speed in milliseconds. To calculate this score, at least one item must be responded to correctly (i.e., raw score ≥ 1).
- *Median Incorrect Reaction Time:* For the items on which the examinee responses incorrectly, this is the median speed in milliseconds. To calculate this score, at least one item must be responded to incorrectly (i.e., raw score < 50).

Reliability and Validity

- *Content validity:* The evidence for content validity was primarily established through both the original Eriksen task and in the development of the NIHTB Flanker Inhibitory Control and Attention Test. The stimuli and mode of responding are well-established for this paradigm, so no additional content validity evidence was sought. However, a small user experience evaluation was conducted, which confirmed that the items could be easily seen, and stimuli differentiated on a mobile phone.
- *General procedures for validation studies are described above. Specific analyses conducted for each validation study of Arrow Matching were as follows:*
    - *Study 1*: Ninety-two participants completed Arrow Matching on study-provided iPhones, unproctored in a private room in the lab. In addition to Arrow Matching, participants were administered the following measures to evaluate convergent validity: NIHTB V3 Flanker (n = 84); the Delis-Kaplan Executive Function System (D-KEFS) Color Word Interference Test (n = 76); and the Wisconsin Card Sorting Test (WCST-64; n = 86).

      In addition, 76 of these participants also completed the PPVT-5 in the same testing session, and 92 participants completed the NIHTB V3 PVT, to evaluate the discriminant validity of Arrow Matching.

      In the Study 1 sample, the reliability of Arrow Matching was examined through split half statistics (median correlation with Spearman-Brown correction between bootstrapped random split-half coefficients). Convergent validity was examined through Spearman Rho correlations using raw scores for all tests. Likewise, discriminant validity was examined through Spearman Rho correlations using raw scores. Spearman Rho correlations were also examined with age. Because of the small number of education level subgroups, correlations with education were not examined in this sample.
    - *Study 2:* A sample of 837 participants completed NIHTB V3 Flanker and 975 completed NIHTB V3 PVT in a lab setting. Separately, 982 participants completed the remote Arrow Matching assessment. In this study, reliability was examined through split half statistics (median correlation with Spearman-Brown correction between bootstrapped random split-half coefficients), while convergent and discriminant validity used Spearman Rho correlations with raw scores for all measures. Spearman Rho correlations were also examined with age. Because participants used their personal devices in this group, reliability across and between operating systems (Android vs. iOS) was also evaluated.

- *Study 3:* A sample of 141 participants completed Arrow Matching twice remotely, with a two-week delay interval. In this sample, the test-retest intraclass correlation was examined.

Demographic information on the study samples is provided in Table 2.

All reliability and validity data for Arrow Matching are shown below in Table 5.

- *Convergent validity*: Arrow Matching correlated .74 with NIHTB V3 Flanker, -.39 to -.48 with different components of D-KEFS Color-Word Interference, and .32 with WCST-64 in an in-person sample. In a remote data collection, Arrow Matching correlated .65 with NIHTB V3 Flanker. Correlations with NIHTB Flanker in both studies were strong, which should be expected given the similarity of the measures; correlations with external measures were moderate and in the expected direction. There is overall strong support that Arrow Matching measures the target construct successfully; moreover, the relatively small difference when Arrow Matching is administered remotely and unproctored vs. in-person is encouraging for its administration in a remote context.
- *Discriminant validity*: Arrow Matching showed small and non-significant correlations with NIHTB V3 PVT in both the in-person and remote samples, and a small and non-significant correlation with PPVT-5 in the in-person sample. These findings match expectations and provide supporting evidence of discriminant validity.
- *Age-related validity*: Arrow Matching had a sizable negative correlation with age for both in-person and remote data collection. This significant reduction in executive functioning with age is typically observed and provides further support for the validity of this test.
- *Reliability*: Test-retest reliability ICC was .69 when Arrow Matching was re-administered 2 weeks after the first administration. This is a reasonably high correlation for a completely remote assessment implementation, given that participant behavior (e.g., inattention) can neither be monitored nor influenced. Internal consistency reliability coefficients were substantially higher (.97) and overall excellent.
- *Operating system differences*: In Study 2, participants completed the measures on their own phones, 37 percent of which were Android devices, allowing for a comparison between Android and iOS. First, the effect of OS on scores was evaluated using linear regressions, controlling for age. Regressions showed that OS did have a significant effect on scores for Arrow Matching ($\beta = 0.23$, $p < .001$). However, despite the small effect of OS on scores, there were no significant differences in convergent validity, divergent validity, or internal consistency reliability between operating systems.

Table 5. *Results from the MTB Arrow Matching Validation Studies*

	Study 1		Study 2		Study 3	
	Value	***N***	**Value**	***N***	**Value**	***N***
Internal Reliability: 50% [25% - 75%]						
Total Sample	—	—	.97 [.97, .98]	982	—	—
iOS	.97 [.96, .97]	92	.97 [.97, .97]	626	—	—
Android	—	—	.98 [.98, .98]	356	—	—
Test-Retest Reliability	—	—	—	—	.69	142
Convergent Validity						
WCST-64 Correct	.32	86	—	—	—	—
D-KEFS Color Word Inhibition Raw Score	-.41	76	—	—	—	—
D-KEFS Raw Score of Color Naming	-.48	76	—	—	—	—
D-KEFS Raw Score of Word Reading	-.39	76	—	—	—	—
D-KEFS Raw Score of Inhibition/Switching	-.45	76	—	—	—	—
NIHTB V3 Flanker	.74	84	.65	837	—	—
Discriminant Validity						
PPVT-5	.15 (ns)	92	—	—	—	—
NIHTB V3 Picture Vocabulary	.24 (ns)	76	.003 (ns)	975	—	—
Demographic Correlations						
Age	-.49	92	-.57	982	—	—

Note: All correlation values are significant at $p < .01$ unless otherwise noted. ns = non-significant ($p \geq .05$).

Shape-Color Sorting

Description of the Task

Shape-Color Sorting measures cognitive flexibility as part of executive function. Based on the NIHTB Dimensional Change Card Sort Test (NIHTB DCCS; Zelazo et al., 2013), examinees are cued to match a bivalent central test stimulus to one of two target stimuli based on one of two dimensions – shape or color. Trials vary on the relevant dimension, which requires examinees to shift their rules amidst trials ("set shifting").

In this test, which is taken in portrait orientation on a smartphone, trials switch between "color" and "shape" in a pseudo-random order. The measure begins with five mixed-practice items, followed by 30 test trials. There is a variable ISI of either 300 or 1,000 ms between trials, and participants have 2,000 ms to respond to each trial (vs. 3,000 ms on the NIHTB V3 DCCS).

Whereas the NIHTB DCCS uses a rabbit and sailboat for practice items and a ball and truck for live items, the MTB version uses a dog and car for practice items, and balloon and house for live items. These new stimuli use the same colors, general shape, and a similar style of drawing as those in the NIHTB version.

Item Pool Development

For the Shape-Color Sorting test, MTB built off the design of the NIHTB DCCS Test. Overall, 60 items were developed for the pool, of which 30 are administered on a given form (60 items allow for multiple forms administration). Of the 30 items administered in pseudo-random order, 23 percent cue color and 77 percent cue shape. ISI length is roughly equally distributed among the items, but also in pseudo-random order. Specific images were changed from DCCS to MTB Shape-Color Sorting to delineate a clear difference among measures, but similar shapes were sought, using simple figures, and identical color schemes were used, as noted.

Available Scores

The MTB Shape-Color Sorting test provides the following score:

- *Rate Correct per Second (RCS):* Shape-Color Sorting uses the number of correct trials per second as a derived score. Doing so creates a varied distribution of possible scores, with up to 30 correct answers and up to 2,000 ms allowed per response. In this way, both accuracy and reaction time are valued. RCS varies from 0 to 10 (i.e., 1 correct per 100ms - about the fastest a human can work) with most scores falling somewhere between 0.7 and 2.5. Note that this is the primary score, and the one for which validity evidence (below) is reported. In REDCap, RCS is reported as "rateScore."
- *Raw Score:* Sum of correct trials. This score reflects the accuracy of a participant, regardless of speed. It provides the numerator of the primary RCS score, with a range of 0-30.
- *Number of Anticipation Errors:* This is the sum of errors a person commits by pressing a response button on an interim screen when no test stimuli are presented. This is tracked on live items only, with a range from 0 to 60.

- *Median Correct Reaction Time:* For the items on which the examinee responds correctly, this is the median speed in milliseconds. To calculate this score, at least one item must be responded to correctly (i.e., raw score ≥ 1).
- *Median Incorrect Reaction Time:* For the items on which the examinee responds incorrectly, this is the median speed in milliseconds. To calculate this score, at least one item must be responded to incorrectly (i.e., raw score < 30).

Reliability and Validity

- *Content validity:* The evidence for content validity was primarily established through the development of the NIHTB DCCS Test. The color scheme chosen largely eliminates any confounds of color blindness/differentiation. The new stimuli shapes developed for the MTB version of the test were reviewed by content experts to assure no new concerns were introduced. In addition, a small user experience evaluation confirmed that the items could be easily seen, and stimuli differentiated on a mobile phone.
- *General procedures for validation studies are described above. Specific analyses conducted for each validation study of Shape-Color Sorting were as follows:*
    - *Study 1*: Ninety-two participants completed Shape-Color Sorting on iPhones, unproctored in a private room in the lab. In addition to Shape-Color Sorting, participants were administered the following measures to evaluate convergent validity: NIHTB V3 DCCS (n = 92); the D-KEFS Color Word Interference Test (n = 76); and the WCST-64 (n = 86).
      In addition, 76 of these participants also completed the PPVT-5 in the same testing session, and 92 participants completed the NIHTB V3 PVT, to evaluate the discriminant validity of Shape-Color Sorting.
      In the Study 1 sample, the reliability of Shape-Color Sorting was examined through split half statistics (median correlation with Spearman-Brown correction between bootstrapped random split-half coefficients). Convergent validity was examined through Spearman Rho correlations using raw scores for D-KEFS and WCST-64 and change sensitive scores on the NIHTB V3 DCCS Test. Likewise, discriminant validity was examined through Spearman Rho correlations using the NIHTB V3 PVT change-sensitive scores and the growth scale value from the PPVT-5. Spearman Rho correlations were also examined with age.
    - *Study 2:* A sample of 884 completed NIHTB V3 DCCS and 958 completed NIHTB V3 PVT. Separately, 964 participants completed Shape-Color Sorting on their own iOS or Android smartphone remotely within 14 days. In this study, reliability was examined through split half statistics (median correlation with Spearman-Brown correction between bootstrapped random split-half coefficients), while convergent and discriminant validity used Spearman Rho correlations with raw scores for all measures. Spearman Rho correlations were also examined with age. Because participants used their personal devices in this group, reliability across and between operating systems (Android vs. iOS) was also evaluated.
    - *Study 3:* A sample of 144 participants completed Shape-Color Sorting twice remotely, with a two-week delay interval. In this sample, the test-retest intraclass correlation was examined.

Demographic information on the study samples is provided in Table 2.

All reliability and validity data for Shape-Color Sorting are shown below in Table 6.

- *Convergent validity*: Shape-Color Sorting correlated .69 with NIHTB V3 DCCS, -.38 to -.54 with different components of D-KEFS Color-Word Interference, and .41 with WCST-64 in an in-person sample. In remote data collection, Shape-Color Sorting correlated .70 with NIHTB V3 DCCS. Correlations with NIHTB V3 DCCS in both studies are excellent, which should be expected given the similarity of the measures. Correlations with external measures were moderate and in the expected direction. There is overall strong support that Shape-Color Sorting measures the target construct successfully; moreover, the comparable correlations with DCCS when Shape-Color Sorting is administered remotely and unproctored vs. in-person is encouraging for its administration in a remote context.
- *Discriminant validity*: Shape-Color Sorting showed small and non-significant correlations with NIHTB V3 PVT in both the in-person and remote samples, and a small and non-significant correlation with PPVT-5 in the in-person sample. These findings match expectations and provide supporting evidence of discriminant validity.
- *Age-related validity*: Shape-Color Sorting had a sizable negative correlation with age for both in-person and remote data collection. This significant reduction in executive functioning with age is typically observed and provides further support for the validity of this test. Of interest, the negative correlation with age was stronger in the remote administration sample.
- *Reliability*: Test-retest reliability ICC was .78 when Shape-Color Sorting was re-administered 2 weeks after the first administration. This is a reasonably high correlation for a completely remote assessment implementation, given that participant behavior (e.g., inattention) can neither be monitored nor influenced. Internal consistency reliability coefficients were substantially higher (.93-.94) and overall excellent.
- *Operating system differences*: In Study 2, participants completed the measures on their own phones, 37 percent of which were Android devices, allowing for a comparison between Android and iOS. First, the effect of OS on scores was evaluated using linear regressions, controlling for age. Regressions showed that OS did have a significant effect on scores for Shape-Color Sorting ($\beta = 0.21$, $p < .001$). However, despite the small effect of OS on scores, there were no significant differences in convergent validity, divergent validity, or internal consistency reliability between operating systems.

Table 6. *Results from the MTB Shape-Color Sorting Validation Studies*

	Study 1		Study 2		Study 3	
	Value	***N***	**Value**	***N***	**Value**	***N***
Internal Reliability: 50% [25% - 75%]						
Total Sample	—	—	.94 [.93, .94]	964	—	—
iOS	.93 [.92, .94]	92	.93 [.93, .94]	629	—	—
Android	—	—	.94 [.93, .94]	335	—	—
Test-Retest Reliability	—	—	—	—	.78	144
Convergent Validity						
WCST-64 Correct	.41	86	—	—	—	—
D-KEFS Color Word Inhibition Raw Score	-.54	76	—	—	—	—
D-KEFS Raw Score of Color Naming	-.43	76	—	—	—	—
D-KEFS Raw Score of Word Reading	-.38	76	—	—	—	—
D-KEFS Raw Score of Inhibition/Switching	-.49	76	—	—	—	—
NIHTB V3 DCCS	.69	92	.70	884	—	—
Discriminant Validity						
PPVT-5	.28 (ns)	78	—	—	—	—
NIHTB V3 Picture Vocabulary	.13 (ns)	92	-.002 (ns)	958	—	—
Demographic Correlations						
Age	-.39	92	-.50	964	—	—

Note: All correlation values are significant at $p < .01$ unless otherwise noted. ns = non-significant ($p \geq .05$).

Number-Symbol Match

Description of the Task

Number-Symbol Match is an electronic adaptation of the many "coding" types of tests that originated in the early 20th century. It is similar in design to the NIH Toolbox Oral Symbol Digit Test, which was similarly adapted from early 20th century stimuli (which are now in the public domain). This measure assesses processing speed by instructing examinees to use a reference key to pair numbers with symbols in a specific amount of time.

Number-Symbol Match is completed in landscape orientation on a smartphone and shows a "key" at the top of the screen, with the numbers one through nine and a unique symbol paired with each number. Below this key, nine symbols are presented per screen horizontally, and the participant must tap the correct number for each symbol presented, according to the key. Symbols are presented in a pseudo-random order, but no identical symbols may appear consecutively. The test includes 144 total items, with up to 16 successive screens of 9 items each presented, and examinees have 90 seconds to complete as many items as possible. To facilitate speed of responding and to minimize confusion or additional error on a smartphone, self-corrections are not permitted on Number-Symbol Match. This design is different than other similar tests, including the NIHTB Oral Symbol Digit Test.

Item Pool Development

For the Number-Symbol Match test, the content was replicated to the greatest extent possible from the NIHTB Oral Symbol Digit Test. One symbol did not render well when converted for use on smartphones, and so was replaced with a similar symbol that did render properly. Otherwise, no new content was developed, but a pseudo-random order was established for the 144 items, with all symbols presented roughly equally and none repeated back-to-back.

Available Scores

The MTB Number-Symbol Match test provides the following score:

- *Raw score:* This score is the total number of items correctly answered in 90 seconds. The possible score range is 0-144. REDCap also provides something called "ItemCount score," which is the count of the number of test items completed.

Reliability and Validity

- *Content validity:* The evidence for content validity was primarily established during the original creation of this content in the early 20th century, and subsequently many times on similar types of "coding" tasks, including the NIHTB Oral Symbol Digit Test. No additional content validity evidence was sought. However, a small user experience evaluation was conducted to assure that the items could be easily seen, stimuli differentiated on a mobile phone, and that the placement of images and the key was logical. Information obtained from this study confirmed all of the above.
- *General procedures for validation studies are described above. Specific analyses conducted for each validation study of Number-Symbol Match were as follows:*

- *Study 1*: Ninety-two participants completed Number-Symbol Match on study-provided iPhones, unproctored in a private room in the lab. The same participants were administered the NIH Toolbox Pattern Comparison Processing Speed Test, Version 3 (NIHTB V3 PCPS) and the NIH Toolbox Oral Symbol Digit Test, Version 3 (NIHTB V3 OSD). Seventy-four of these participants also completed the Coding and Symbol Search subtests of the Wechsler Adult Intelligence Test, Fourth Edition (WAIS-IV; Wechsler, 2008), in the same testing session. The NIHTB and WAIS-IV tests were administered to evaluate convergent validity.
  In addition, 76 of these participants also completed the PPVT-5 in the same testing session, and 92 participants completed the NIHTB V3 PVT, to evaluate the discriminant validity of Number-Symbol Match.
  In the Study 1 sample, the reliability of Number-Symbol Match was examined through split half statistics (median correlation with Spearman-Brown correction between bootstrapped random split-half coefficients). Convergent validity was examined through Spearman Rho correlations using change sensitive scores on the NIHTB V3 PCPS and NIHTB V3 OSD, and raw scores for the WAIS-IV subtests. Likewise, discriminant validity was examined through Spearman Rho correlations using change sensitive scores on the NIHTB V3 PVT and growth scale values on the PPVT-5. Spearman Rho correlations were also examined with age.
- *Study 2:* A sample of 931 participants completed Number-Symbol Match on their own iOS or Android smartphone remotely. In addition, 923 of these participants were administered the NIHTB V3 PCPS; 917 took the NIHTB V3 OSD; and 924 completed the NIHTB V3 PVT. Reliability was examined through split half statistics (median correlation with Spearman-Brown correction between bootstrapped random split-half coefficients), while convergent and discriminant validity used Spearman Rho correlations with raw scores for all measures. Spearman Rho correlations were also examined with age. Because participants used their personal devices in this group, reliability across and between operating systems (Android vs. iOS) was also evaluated.
- *Study 3:* A sample of 141 participants completed Number-Symbol Match twice remotely, with a two-week delay interval. In this sample, the test-retest intraclass correlation was examined.

Demographic information on the study samples is provided in Table 2.

All reliability and validity data for Number-Symbol Match are shown in Table 7.

- *Convergent validity*: Number-Symbol Match correlated .58 with NIHTB V3 PCPS, .71 with the NIHTB V3 OSD, .68 with WAIS-IV Coding, and .63 with WAIS-IV Symbol Search for data collected in-person. In a remote data collection, Number-Symbol Match correlated .57 with NIHTB V3 PCPS and .70 with NIHTB V3 OSD. Correlations with NIHTB tests in both studies are excellent and nearly identical in both remote and in-person administrations; the strong correlation with OSD in particular is as expected, given the similarity of the

measures. Correlations with external measures were also excellent, especially WAIS-IV coding, which measures a similar construct. There is overall strong support that Number-Symbol Match measures the target construct successfully; moreover, there is strong support for its administration in a remote context.

- *Discriminant validity*: Number-Symbol Match showed small and non-significant correlations with NIHTB V3 PVT in both the in-person and remote samples, and a small and non-significant correlation with PPVT-5 in the in-person sample. These findings match expectations and provide supporting evidence of discriminant validity.
- *Age-related validity*: Number-Symbol Match had a high negative correlation with age for both in-person and remote data collection. This significant reduction in processing speed with age is typically observed and provides further support for the validity of this test, especially as a potentially sensitive marker of age-related cognitive decline.
- *Reliability*: Test-retest reliability ICC was .83 when Number-Symbol Match was re-administered 2 weeks after the first administration. This is a strong correlation for a completely remote assessment implementation, given that participant behavior (e.g., inattention) can neither be monitored nor influenced. Internal consistency reliability coefficients were exceptionally high (.98).
- *Operating system differences*: In Study 2, participants completed the measures on their own phones, 37 percent of which were Android devices, allowing for a comparison between Android and iOS. First, the effect of OS on scores was evaluated using linear regressions, controlling for age. Regressions showed that OS had no effect on Number-Symbol Match ($\beta = 0.005$, $p = .93$), most likely because the test does not rely on precise timing for scoring (only a gross 90-second time limit). Moreover, there were no significant differences in convergent validity, divergent validity, or internal consistency reliability between operating systems.

Table 7. *Results from the MTB Number-Symbol Match Validation Studies*

	Study 1		Study 2		Study 3	
	Value	***N***	**Value**	***N***	**Value**	***N***
Internal Reliability: 50% [25% - 75%]						
Total Sample	—	—	.98 [.97, .98]	931	—	—
iOS	.99 [.97, .98]	92	.98 [.97, .98]	626	—	—
Android	—	—	.98 [.97, .98]	305	—	—
Test-Retest Reliability	—	—	—	—	.83	141
Convergent Validity						
WAIS-IV Coding	.68	76	—	—	—	—
WAIS-IV Symbol Search	.63	76	—	—	—	—
NIHTB V3 Pattern Comparison	.58	92	.57	923	—	—
NIHTB V3 Oral Symbol Digit	.71	92	.70	917	—	—
Discriminant Validity						
PPVT-5	.20 (ns)	78	—	—	—	—
NIHTB V3 Picture Vocabulary	.06 (ns)	92	-.07 (ns)	924	—	—
Demographic Correlations						
Age	-.67	92	-.61	931	—	—

Note: All correlation values are significant at $p < .01$ unless otherwise noted. ns = non-significant ($p \geq .05$).

Sequences

Description of the Task

The MTB Sequences test was designed to measure working memory – the same construct as the NIH Toolbox List Sorting Working Memory Test (Tulsky et al., 2014) – in an efficient and effective way on a personal smartphone. The Sequences test is completed in landscape orientation on a smartphone. A sequence of numbers and letters is shown on the screen, one per second, and the participant must enter what they saw by first tapping the letters seen in alphabetical order, and then the numbers from smallest to largest. Nine digits from 1-9 and nine letters (a-b-c-q-r-s-x-y-z) were selected as stimulus and response options, to account for the limited screen area available for the task on a smartphone. Participants view a set of instructions and then practice until they understand how to correctly respond. All participants start the "live" items with three consecutive items at a sequence length of 2 (to avoid reversals and resultant confusion that could occur in a self-administered context) and, if they get at least one of three correct, proceed to a sequence length of 3. Sequence lengths of 3 through 10 are then administered serially until a participant misses all three items in a given set (note that there are two three-item sets at sequence length 3, to enhance assessment at the lower end of the performance continuum). The test is discontinued after all 3 items in a set at a given sequence length are answered incorrectly. Anywhere from 3 to 30 items may be administered, and the test takes about 5 minutes on average (but may take up to 7 minutes to complete for strong performers).

Item Pool Development

The MTB Sequences test was created *de novo*, based on a number of common published working memory paradigms, including the NIHTB List Sorting Working Memory (LSWM) test. In this paradigm, the participant is presented with a stimulus as a series with multiple categorizations, must keep it in short-term memory, mentally manipulate it and respond with the results of the manipulated sequence according to the task requirements. Prior to large-scale studies of validity and reliability, two pilots of Sequences were conducted. In the first pilot, alternate letter choices were tried as stimuli (given the limited screen sizes of smartphones, only nine letters and nine numbers could be accommodated as response choices on the screen) to assess differences in performance, and evidence of floor and ceiling effects were studied. As a result of the first pilot study, a set of three sequences of length 2 and an extra set of three sequences of length 3 were added to offset noted floor effects among participants. In addition, the letter groups a-b-c, q-r-s, and x-y-z appeared to work slightly better than using nine consecutive letters within the alphabet. Separately, a small user experience study was conducted, which resulted in a change from uppercase to lowercase letters for stimuli.

In a second pilot study, these changes were tried out in a new sample using the MTB app on iPhones. Floor effects were removed as a result of the additional items, and no ceiling effects appeared. Median testing time was 5 minutes. Given the overall success of the second pilot study, this version of Sequences was chosen for subsequent validation studies (and is the version described above, under "Description of the Task").

Available Scores

The MTB Sequences test provides the following score:

- *Raw score:* This score is the total number of items (individual sequence) correctly answered. The possible score range is 0-30. REDCap also provides something called "ItemCount score," which is the count of the number of test items completed.

Reliability and Validity

- *Content validity:* Although Sequences is a newly developed measure, the evidence for its content validity is well-established, in that Sequences uses existing, well-studied paradigms for assessing working memory. Other types of construct validity were evaluated thoroughly, given that the "novelty" of Sequences is its self-administration on a mobile phone, and are described in more detail below.
- *General procedures for validation studies are described above. Specific analyses conducted for each validation study of Sequences were as follows:*
    - *Study 1:* Ninety-two participants completed Sequences on study-provided iPhones, unproctored in a private room in the lab. In addition to Sequences, these participants were administered the NIH Toolbox List Sorting Working Memory Test, Version 3 (NIHTB V3 LSWM). Seventy-eight participants were also administered the Letter-Number Sequencing subtest from the WAIS-IV, and 77 were administered the WAIS-IV Digit Span subtest. These three tests were administered to evaluate convergent validity.
      In addition, 92 participants were administered the NIHTB V3 PVT, to evaluate the discriminant validity of Sequences.
      In the Study 1 sample, the reliability of Sequences was examined through split half statistics (median correlation with Spearman-Brown correction between bootstrapped random split-half coefficients). Convergent validity was examined through Spearman Rho correlations using change sensitive scores on the LSWM and raw scores for the WAIS-IV subtests. Likewise, discriminant validity was examined through Spearman Rho correlations using change sensitive scores on PVT. Spearman Rho correlations were also examined with age.
    - *Study 2:* A sample of 1,007 completed Sequences on their own iOS or Android smartphone remotely. In addition, 996 participants were administered the NIHTB V3 LSWM and 1,000 completed the NIHTB V3 PVT. In this study, reliability was examined through split half statistics (median correlation with Spearman-Brown correction between bootstrapped random split-half coefficients), while convergent and discriminant validity used Spearman Rho correlations with raw scores for all measures. Spearman Rho correlations were also examined with age. Because participants used their personal devices in this group, reliability across and between operating systems (Android vs. iOS) was also evaluated.
    - *Study 3:* A sample of 147 participants completed Sequences twice remotely, with a two-week delay interval. In this sample, the test-retest intraclass correlation was examined.

Demographic information on the study samples is provided in Table 2.

All reliability and validity data for Sequences are shown in Table 8.

- *Convergent validity*: Sequences correlated .64 with NIHTB V3 LSWM, .52 with WAIS-IV Letter-Number Sequencing, and .58 with WAIS-IV Digit Span in an in-person data collection. In a remote data collection, Sequences correlated .46 with NIHTB V3 LSWM. All correlations show solid support that Sequences measures the target construct successfully; however, there is some difference in convergent correlations when Sequences is administered remotely and unproctored vs. in-person. While it can be validly administered in a remote context, a focus on group results will be important.
- *Discriminant validity*: Sequences showed meaningfully smaller correlations with NIHTB V3 PVT in both the in-person (.34) and remote (.18) samples compared with convergent measures. These findings match expectations and provide supporting evidence of discriminant validity.
- *Age-related validity*: Sequences had a small, non-significant negative correlation (-.07) with age for the in-person sample and a small but significant negative correlation (-.19) in the remote data collection. This reduction in working memory with age, while small, is commonly seen and provides incremental support for the validity of this test.
- *Reliability*: Test-retest reliability ICC was .55 when Sequences was re-administered 2 weeks after the first administration. This is a moderate correlation for a completely remote assessment implementation, given that participant behavior (e.g., inattention, dishonesty) can neither be monitored nor influenced. Internal consistency reliability coefficients were very high (.90 in the total sample).
- *Operating system differences*: In Study 2, participants completed the measures on their own phones, 37 percent of which were Android devices, allowing for a comparison between Android and iOS. First, the effect of OS on scores was evaluated using linear regressions, controlling for age. Regressions showed that OS had no significant effect on Sequences, most likely because the test is not timed. Moreover, there were no significant differences in convergent validity, divergent validity, or internal consistency reliability between operating systems.

Table 8. *Results from the MTB Sequences Validation Studies*

	Study 1		Study 2		Study 3	
	Value	***N***	**Value**	***N***	**Value**	***N***
Split-Half Reliability						
Total Sample	—	—	.90	1,007	—	—
iOS	.90	92	.89	647	—	—
Android	—	—	.92	360	—	—
Test-Retest Reliability	—	—	—	—	.55	147
Convergent Validity						
WAIS-IV Letter-Number Sequencing	.52	78	—	—	—	—
WAIS-IV Digit Span	.58	77	—	—	—	—
NIHTB V3 List Sorting Working Memory	.64	92	.46	996	—	—
Discriminant Validity						
NIHTB V3 Picture Vocabulary	.34	92	.18	1,000	—	—
Demographic Correlations						
Age	-.07 (ns)	92	-.19	1,007	—	—

Note: All correlation values are significant at $p < .01$ unless otherwise noted. ns = non-significant ($p \geq .05$).

Faces and Names

Description of the Task

Faces and Names is a task derived from the Face-Name Associative Memory Exam (FNAME; Rentz et al., 2011), which is an associative memory test designed to be sensitive to amnestic processes. Previous studies have shown that the in-clinic version of FNAME is correlated with other measures of episodic memory and can differentiate clinically normal individuals from those with subjective cognitive decline and Mild Cognitive Impairment (Kormas et al., 2020). Faces and Names consists of one encoding phase and three memory phases: Face Recognition, Name Recall, and Name Recognition.

During the encoding phase, 12 face-name pairs are serially presented. To ensure that participants pay attention to learning the face-name pairs, the participants choose a button on the screen stating whether it is "easy" or "hard" to remember the name that goes with the face. This pairing is shown on the screen for five seconds, regardless of examinee response. This initial encoding phase is not scored. Neither the faces nor the names in the face-name pairs used for this MTB version have been previously published in other versions of the original FNAME. Following a delay (5-20 minutes in the in-person version, but variable when conducted remotely due to lack of examiner control), participants complete three distinct memory phases. First, they are asked to choose (by tapping) the face they learned from among three faces presented – the correct choice, plus two distractors of matching age, sex, and race (Face Recognition); then, they are asked to select the first letter of the name previously paired with the face (Name Recall), using a standard keyboard for mobile phones; and finally, they are asked to select the correct name by tapping one of 3 face/name combinations (Name Recognition). Participants are scored on accuracy, with a maximum score of 12 points for each component. A higher score indicates higher accuracy. A total score is calculated by adding together each component score, for a total of 36 possible points. Scores may range from 0-36. This total score is then converted to a sum-score based on *expected a posteriori* score using a bifactor item response theory model.

Item Pool Development

Face-name pairs were developed by Rentz et al. (2011) in the lab, using headshots with consistent backgrounds of people representing males and females and diversity in race and ethnicity. This content development approach was consistent with that of previous FNAME versions (e.g., ARMADA study; Rentz et al., 2023). As a result of a content and sensitivity review by test development experts, the encoding task was modified slightly from previous FNAME versions (e.g., ARMADA), to focus on whether the face-name pair was easy or hard to remember. This change obviated potential sensitivity concerns. The paradigm for Faces and Names has otherwise already been well-established through previous research on FNAME.

Available Scores

The MTB Faces and Names test provides the following score:

- Faces and Names test is scored using the theta metric, taking into account the scores on the Face Recognition, Name Recall, and Name Recognition components of the test. This score is based on item response theory analyses to estimate a person's "true" ability for

associative memory. Higher scores indicate higher levels of ability. Note that this is the primary score, and the one for which validity evidence (below) is reported.

- *Face Seen Before (FSB):* The raw score for the first portion of the test is the sum of the faces correctly identified as having been seen before. FSB scores range from 0 to 12.
- *Face Name Letter (FNL):* The raw score for the second portion of the test is the sum of the faces for which the examinee correctly enters the first letter of their name. FNL scores range from 0 to 12.
- *Face Name Matching (FNM):* The raw score for the third and final portion of the test is the sum of the names correctly matched to the target face. FNM scores range from 0 to 12.
- *Raw Score*: The total raw score is the sum of FSB, FNL, and FNM scores. It ranges from 0 to 12 and is the basis for converting to the primary overall score—the theta score, described above.

Reliability and Validity

- *Content validity:* As noted above, content validity has been well established in previous published versions of FNAME, of which Faces and Names is directly derived. Although actual faces and names are novel, the content creation and stimulus presentation processes have not changed. Thus, there would appear to be ample existing evidence of content validity and further evidence was not sought. However, as noted above, a minor change was made to the encoding phase task, to mitigate any potential sensitivity concerns.
- *General procedures for validation studies are described above. Specific analyses conducted for each validation study of Faces and Names were as follows:*
    - *Study 1:* Ninety-two participants completed Faces and Names on study-provided iPhones, unproctored in a private room in the lab. In addition, these participants were administered the NIHTB V3 FNAME. The Weschler Memory Scales, 4th Edition, Verbal Paired Associates test (WMS-IV VPA; Wechsler, 2009), Immediate and Delayed, were also administered to 74 of these participants.
      In addition, 92 participants were administered the NIHTB V3 PVT, to evaluate the discriminant validity of Faces and Names.
      In the Study 1 sample, the reliability of Faces and Names was examined through empirical and mean person-specific reliability statistics. Convergent validity was examined through Spearman Rho correlations using change sensitive scores on the NIHTB V3 FNAME and raw scores on the WMS-IV VPA. Likewise, discriminant validity was examined through Spearman Rho correlations using change sensitive scores on NIHTB V3 PVT. Spearman Rho correlations were also examined with age.
    - *Study 2:* A sample of 956 participants completed Faces and Names on their own iOS or Android smartphone remotely. In addition, 946 participants were administered the NIHTB V3 FNAME and 949 completed the NIHTB V3 PVT in the lab. In this study, reliability was examined through empirical and mean person-specific reliability statistics, while convergent and discriminant validity used Spearman Rho correlations with raw scores for all measures. Spearman Rho

correlations were also examined with age. Because participants used their personal devices in this group, reliability across and between operating systems (Android vs. iOS) was also evaluated.

- *Study 3:* A sample of 123 participants completed Faces and Names twice remotely, with a two-week delay interval. In this sample, the test-retest intraclass correlation was examined.

Demographic information on the study samples is provided in Table 2.

All reliability and validity data for Faces and Names are shown in Table 9.

- *Convergent validity*: Faces and Names correlated .69 with NIHTB V3 FNAME, .54 with WMS-IV Verbal Paired Associates Immediate, and .58 with WMS-IV Verbal Paired Associates Delayed in an in-person data collection. In a remote data collection, Faces and Names correlated .56 with NIHTB V3 FNAME. All correlations show strong support that Faces and Names measures the target construct successfully. There is a small decrement in convergent correlation when Faces and Names is administered remotely and unproctored vs. in-person. While the evidence indicates it can be validly administered in a remote context, a focus on group results should be emphasized.
- *Discriminant validity*: Faces and Names showed meaningfully smaller correlations with NIHTB V3 PVT in both the in-person (.23) and remote (.10) samples. These findings match expectations and provide supporting evidence of discriminant validity.
- *Age-related validity*: Faces and Names had a moderate negative correlation (-.41) with age for the in-person sample as well as in the remote data collection sample (-.33). This reduction in memory with age is commonly seen and provides further support for the validity of this test.
- *Reliability*: Test-retest reliability ICC was .73 when Faces and Names was re-administered 2 weeks after the first administration. This is a moderate to strong correlation for a completely remote assessment implementation, given that participant behavior (e.g., inattention, dishonesty) can neither be monitored nor influenced. Empirical reliability coefficients were similar in size (.79).
- *Operating system differences*: In Study 2, participants completed the measures on their own phones, 37 percent of which were Android devices, allowing for a comparison between Android and iOS. First, the effect of OS on scores was evaluated using linear regressions, controlling for age. Regressions showed that there were some OS differences in scores on Faces and Names that could not be explained by age differences in the OS samples; however, all validity and reliability metrics were compared by device in each sample and showed no significant differences, thus allowing the OS-affected cases to be combined for analysis purposes.

Table 9. *Results from the MTB Faces and Names Validation Studies*

	Study 1		Study 2		Study 3	
	Value	***N***	**Value**	***N***	**Value**	***N***
Mean Theta Score (SD)						
Empirical Reliability						
Total Sample	—	—	.79	956	—	—
iOS	.76	92	.79	624		—
Android	—	—	.77	332	—	—
Person-Specific Reliability						
Total Sample	—	—	.78	956	—	—
iOS	.78	92	.78	624		—
Android	—	—	.78	332	—	—
Test-Retest Reliability	—	—	—	—	.73	123
Convergent Validity						
WMS-IV Verbal Paired Associates Immediate	.54	74	—	—	—	—
WMS-IV Verbal Paired Associates Delayed	.58	74	—	—	—	—
NIHTB V3 FNAME	.69	92	.56	946	—	—
Discriminant Validity						
NIHTB V3 Picture Vocabulary	.23	92	.10	949	—	—
Demographic Correlations						
Age	-.41	92	-.33	956		—

Note. All correlations significant at $p < .01$ unless otherwise noted.

Arranging Pictures

Description of the Task

The MTB Arranging Pictures task is designed to measure episodic memory and is modeled after the NIH Toolbox Picture Sequence Memory Test (NIHTB PSM; Dikmen et al., 2014). Arranging Pictures was designed to be completely self-administered on a participant's own smartphone in landscape mode. Participants are first presented with written instructions and an animated demonstration of the task, followed by an abbreviated practice sequence that provides feedback. Once practice is successfully completed, the live items are presented, consisting of two successive trials of 14 illustrations based on the theme of "A Day with a Friend." The illustrations, while loosely linked thematically, are not sequenced temporally or otherwise, making them essentially independent images. Participants are shown each picture in the sequence one at a time, with a corresponding audio description; then, the pictures are scrambled, and participants are asked to arrange them in the order in which they were originally presented. The second trial of arranging pictures is identical to the first and is presented immediately upon completion of trial 1. The primary score on MTB Arranging Pictures is the number of correctly placed adjacent pairs (i.e., pictures correctly placed next to each other) on each trial, summed together.

Item Pool Development

As noted above, Arranging Pictures was modeled directly on NIHTB PSM. However, based on typical smartphone screen sizes, it was determined that a maximum of 14 images could be displayed effectively to comprise a sequence (as opposed to 15 in NIHTB PSM). In addition, novel content was developed for Arranging Pictures to assure independence from NIHTB PSM and avoid item exposure. Arranging Pictures used the same development approach to NIHTB PSM. A theme was created, 14 images were identified, and audio recordings were made and associated with each image. Illustrations were slightly simplified for Arranging Pictures to minimize the likelihood of examinee vision challenges affecting performance.

Available Scores

The MTB Arranging Pictures test provides the following score:

- Sum of adjacent pairs correctly placed across both trials. Scores range from 0-26 (14 images results in a maximum of 13 adjacent pairs correct per trial).

Reliability and Validity

- *Content validity:* The paradigm, approach, and content development model used in Arranging Pictures matches the NIHTB PSM development model quite closely. There is ample published evidence of content validity for PSM and as such, content validity evidence can be considered successfully established for Arranging Pictures as well.
- *General procedures for validation studies are described above. Specific analyses conducted for each validation study of Arranging Pictures were as follows:*
    - *Study 1:* Ninety-two participants completed Arranging Pictures on study-provided iPhones, unproctored in a private room in the lab. In addition, these participants were administered the NIH Toolbox Picture Sequence Memory Test, Version 3

(NIHTB V3 PSM). The WMS-IV VPA, Immediate and Delayed, was administered to 74 of these participants.

In addition, 74 participants were administered the PPVT-5 in the same testing session, and 92 participants completed the NIHTB V3 PVT, to evaluate the discriminant validity of Arranging Pictures.

In the Study 1 sample, the reliability of Arranging Pictures was examined through split half statistics (median correlation with Spearman-Brown correction between bootstrapped random split-half coefficients). Convergent validity was examined through Spearman Rho correlations using change sensitive scores for NIHTB V3 PSM and raw scores for the WMS-IV VPA. Likewise, discriminant validity was examined through Spearman Rho correlations using change sensitive scores for NIHTB V3 PVT and growth scale values for PPVT-5. Spearman Rho correlations were also examined with age.

- *Study 2:* A sample of 979 completed Arranging Pictures on their own iOS or Android smartphone remotely. In addition, 963 participants were administered the NIHTB V3 PSM and 972 completed the NIHTB V3 PVT in the lab. In this study, reliability was examined through split half statistics (median correlation with Spearman-Brown correction between bootstrapped random split-half coefficients), while convergent and discriminant validity used Spearman Rho correlations with raw scores for all measures. Spearman Rho correlations were also examined with age. Because participants used their personal devices in this group, reliability across and between operating systems (Android vs. iOS) was also evaluated.
- *Study 3:* A sample of 141 participants completed Arranging Pictures twice remotely, with a two-week delay interval. In this sample, the test-retest intraclass correlation was examined

Demographic information on the study samples is provided in Table 2.

All reliability and validity data for Arranging Pictures are shown in Table 10.

- *Convergent validity*: Arranging Pictures correlated .64 with NIHTB PSM V3, .65 with WMS-IV VPA Immediate Recall, and .66 with WMS-IV Delayed Recall, in an in-person data collection. In a remote data collection, Arranging Pictures correlated .45 with NIHTB PSM V3. Correlations from in-person data collection show strong support that Arranging Pictures measures the target construct successfully; however, correlations are more moderate when Arranging Pictures is administered remotely and unproctored, and there is some difference in this convergent correlation as compared with in-person administration. While Arranging Pictures can be validly administered in a remote context, a focus on group results will be important.
- *Discriminant validity*: Arranging Pictures showed meaningfully smaller correlations with NIHTB PVT V3 in both the in-person (.27) and remote (.02) samples. Similarly, a smaller correlation was found when comparing Arranging Pictures with PPVT-5 in an in-person

sample (.23). These findings match expectations and provide supporting evidence of discriminant validity.

- *Age-related validity*: Arranging Pictures had small-to-moderate, significant negative correlations with age for both the in-person sample (-.32) and the remote data collection sample (-.22). This reduction in episodic memory with age, while relatively small, is commonly seen in measures of this construct and provides further support for this test's validity.
- *Reliability*: Test-retest reliability ICC was .49 when Arranging Pictures was re-administered 2 weeks after the first administration. This is a moderate correlation for a completely remote assessment implementation, given that participant behavior (e.g., inattention, dishonesty) can neither be monitored nor influenced. Internal consistency reliability coefficients were considerably higher (.80-81).
- *Operating system differences*: In Study 2, participants completed the measures on their own phones, 37 percent of which were Android devices, allowing for a comparison between Android and iOS. First, the effect of OS on scores was evaluated using linear regressions, controlling for age. Regressions showed that there were some OS differences in scores on Arranging Pictures that could not be explained by age differences in the OS samples; however, all validity and reliability metrics were compared by device in each sample and showed no significant differences, thus allowing the OS-affected cases to be combined for analysis purposes.

Table 10. *Results from the MTB Arranging Pictures Validation Studies*

	Study 1		**Study 2**		**Study 3**	
	Value	***N***	**Value**	***N***	**Value**	***N***
Split-Half Reliability						
Total Sample	—	—	.81	979	—	—
iOS	.85	92	.81	625	—	—
Android	—	—	.80	354	—	—
Test-Retest Reliability	—	—	—	—	.49	141
Convergent Validity						
WMS-IV Verbal Paired Associates Immediate	.65	74	—	—	—	—
WMS-IV Verbal Paired Associates Delayed	.66	74	—	—	—	—
NIHTB V3 Picture Sequence Memory Test	.64	92	.45	963	—	—
Discriminant Validity						
PPVT-5	.23	74	—	—	—	—
NIHTB V3 Picture Vocabulary	.27	92	.02 (ns)	972	—	—
Demographic Correlations						
Age	-.32	92	-.22	979	—	—

Note: All correlation values are significant at $p < .01$ unless otherwise noted. ns = non-significant ($p \geq .05$).

References Cited

Condon, D. M., & Revelle, W. (2014). The international cognitive ability resource: Development and initial validation of a public-domain measure. *Intelligence, 43*, 52-64. https://doi.org/10.1016/j.intell.2014.01.004

Dikmen, S. S., Bauer, P. J., Weintraub, S., Mungas, D., Slotkin, J., Beaumont, J. L., Gershon, R., Temkin, N. R., & Heaton, R. K. (2014, Jul). Measuring episodic memory across the lifespan: NIH Toolbox Picture Sequence Memory Test. *Journal of the International Neuropsychological Society, 20*(6), 611-619. https://doi.org/10.1017/S1355617714000460

Dworak, E. M., Revelle, W., Doebler, P., & Condon, D. M. (2021). Using the International Cognitive Ability Resource as an open source tool to explore individual differences in cognitive ability. *Personality and Individual Differences, 169*, 109906. https://doi.org/10.1016/j.paid.2020.109906

Gershon, R. C., Sliwinski, M. J., Mangravite, L., King, J. W., Kaat, A. J., Weiner, M. W., & Rentz, D. M. (2022, Jul). The Mobile Toolbox for monitoring cognitive function. *Lancet Neurology, 21*(7), 589-590. https://doi.org/10.1016/S1474-4422(22)00225-3

Harris, P. A., Swafford, J., Serdoz, E. S., Eidenmuller, J., Delacqua, G., Jagtap, V., Taylor, R. J., Gelbard, A., Cheng, A. C., & Duda, S. N. (2022, Jul). MyCap: A flexible and configurable platform for mobilizing the participant voice. *JAMIA Open, 5*(2), ooac047. https://doi.org/10.1093/jamiaopen/ooac047

Harris, P. A., Taylor, R., Minor, B. L., Elliott, V., Fernandez, M., O'Neal, L., McLeod, L., Delacqua, G., Delacqua, F., Kirby, J., Duda, S. N., & Consortium, R. E. (2019, Jul). The REDCap consortium: Building an international community of software platform partners. *J Biomed Inform, 95*, 103208. https://doi.org/10.1016/j.jbi.2019.103208

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009, Apr). Research electronic data capture (REDCap) -- A metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform, 42*(2), 377-381. https://doi.org/10.1016/j.jbi.2008.08.010

Kormas, C., Zalonis, I., Evdokimidis, I., Kapaki, E., & Potagas, C. (2020). Face–name associative memory performance among cognitively healthy individuals, individuals with subjective memory complaints, and patients with a diagnosis of aMCI. *Frontiers in Psychology, 11*, 2173. https://doi.org/10.3389/fpsyg.2020.02173

Laforte, E. M., Hook, J. N., & Giella, A. K. (2024). *National Institutes of Health (NIH) Toolbox® Version 3 Technical Manual*. https://nihtoolbox.org/app/uploads/2024/04/NIHTB-V3-Technical-Manual_040524.pdf

Rentz, D. M., Amariglio, R. E., Becker, J. A., Frey, M., Olson, L. E., Frishe, K., Carmasin, J., Maye, J. E., Johnson, K. A., & Sperling, R. A. (2011, Jul). Face-name associative memory performance is related to amyloid burden in normal elderly. *Neuropsychologia, 49*(9), 2776-2783. https://doi.org/10.1016/j.neuropsychologia.2011.06.006

Rentz, D. M., Klinger, H. M., Samaroo, A., Fitzpatrick, C., Schneider, O. R., Amagai, S., & Peipert, J. D. (2023, Jul-Sep). Face Name Associative Memory Exam and biomarker status in the ARMADA study: Advancing reliable measurement in Alzheimer's disease and cognitive aging. *Alzheimers Dement (Amst), 15*(3), e12473. https://doi.org/10.1002/dad2.12473

Revelle, W., Dworak, E. M., & Condon, D. (2020). Cognitive ability in everyday life: The utility of open-source measures. *Current Directions in Psychological Science, 29*(4), 358-363. https://doi.org/10.1177/096372142092217

Tulsky, D. S., Carlozzi, N., Chiaravalloti, N. D., Beaumont, J. L., Kisala, P. A., Mungas, D., Conway, K., & Gershon, R. (2014, Jul). NIH Toolbox Cognition Battery (NIHTB-CB): List sorting test to measure working memory. *Journal of the International Neuropsychological Society, 20*(6), 599-610. https://doi.org/10.1017/S135561771400040X

Young, S. R., Dworak, E. M., Kaat, A. J., Adam, H., Novack, M. A., Slotkin, J., Stoeger, J., Nowinski, C. J., Hosseinian, Z., Amagai, S., Pila, S., Diaz, M. V., Correa, A. A., Alperin, K., Omberg, L., Kellen, M., Camacho, M. R., Landavazo, B., Nosheny, R. L., Weiner, M. W., & Gershon, R. M. (2024, Feb 27). Development and validation of a vocabulary measure in the Mobile Toolbox. *Archives of Clinical Neuropsychology*, acae010. https://doi.org/10.1093/arclin/acae010

Young, S. R., & Keith, T. Z. (2020). An examination of the convergent validity of the ICAR16 and WAIS-IV. *Journal of Psychoeducational Assessment, 38*(8), 1052-1059. https://doi.org/10.1177/0734282920943455

Zelazo, P. D., Anderson, J. E., Richler, J., Wallner-Allen, K., Beaumont, J. L., & Weintraub, S. (2013, Aug). II. NIH Toolbox Cognition Battery (CB): Measuring executive function and attention. *Monographs of the Society for Research in Child Development, 78*(4), 16-33. https://doi.org/10.1111/mono.12032