Journal of the International Neuropsychological Society (2024), **30**, 680–688
doi:10.1017/S1355617724000225

Research Article

Mobile toolbox (MTB) remote measures of executive function and processing speed: development and validation

Miriam A. Novack[1], Stephanie Ruth Young[1], Elizabeth M. Dworak[1], Aaron J. Kaat[1], Jerry Slotkin[2], Cindy Nowinski[1], Lihua Yao[1], Hubert Adam[1], Jordan Stoeger[1], Zahra Hosseinian[1], Saki Amagai[1], Sarah Pila[1], Maria Varela Diaz[1], Anyelo Almonte Correa[1], Keith Alperin[3], Sonia Carlson[4], Michael Kellen[4], Larsson Omberg[4], Monica R. Camacho[5,6], Bernard Landavazo[5,6], Rachel L. Nosheny[5,6], Michael W. Weiner[5,6] and Richard C. Gershon[1]

[1]Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL, USA, [2]Center for Health Assessment Research and Translation, University of Delaware, Newark, DE, USA, [3]Helium Foot Software, Inc, Chicago, IL, USA, [4]Sage Bionetworks, Seattle, WA, USA, [5]University of California, San Francisco, NCIRE, San Francisco, CA, USA and [6]Northern California Institute for Research and Education, San Francisco Veteran's Administration Medical Center, San Francisco, CA, USA

Abstract

Objective: The ability to remotely monitor cognitive skills is increasing with the ubiquity of smartphones. The Mobile Toolbox (MTB) is a new measurement system that includes measures assessing Executive Functioning (EF) and Processing Speed (PS): Arrow Matching, Shape-Color Sorting, and Number-Symbol Match. The purpose of this study was to assess their psychometric properties. **Method:** MTB measures were developed for smartphone administration based on constructs measured in the NIH Toolbox® (NIHTB). Psychometric properties of the resulting measures were evaluated in three studies with participants ages 18 to 90. In Study 1 ($N = 92$), participants completed MTB measures in the lab and were administered both equivalent NIH TB measures and other external measures of similar cognitive constructs. In Study 2 ($N = 1{,}021$), participants completed the equivalent NIHTB measures in the lab and then took the MTB measures on their own, remotely. In Study 3 ($N = 168$), participants completed MTB measures twice remotely, two weeks apart. **Results:** All three measures exhibited very high internal consistency and strong test-retest reliability, as well as moderately high correlations with comparable NIHTB tests and moderate correlations with external measures of similar constructs. Phone operating system (iOS vs. Android) had a significant impact on performance for Arrow Matching and Shape-Color Sorting, but no impact on either validity or reliability. **Conclusions:** Results support the reliability and convergent validity of MTB EF and PS measures for use across the adult lifespan in remote, self-administered designs.

Keywords: Cognition; executive function; processing speed; mobile assessment; NIH Toolbox; validation

(Received 28 September 2023; final revision 29 March 2024; accepted 1 May 2024; First Published online 11 July 2024)

Introduction

Executive Functioning (EF) and Processing Speed (PS) are foundational for many complex cognitive functions, and worsening performance in these domains has been hypothesized to play a dominant role in cognitive decline (e.g., Reynolds et al., 2009; Salthouse, 1996, 2000). As such, longitudinal monitoring of both EF and PS has important implications for early detection, intervention, and management of pathological cognitive impairment. Neuropsychological batteries and standardized assessments typically include measures of EF and PS; however, most standardized measures are designed for one-on-one, in-person administration, which can be costly, burdensome, and often impractical for both researchers and participants.

Luckily, recent advancements in connected technologies have provided opportunities to perform remote health monitoring in a manner that can support research advances toward healthy aging and early disease detection. Mobile app-based assessments offer a particularly appealing mechanism for administration, as they can be completed at any time in nearly any location (Ben-Zeev & Atkins, 2017; Koo & Vizer, 2019). By leveraging personal smartphone technologies, remote cognitive administration can offer a cost-effective and efficient alternative to in-person assessment and enable study designs that include frequent and/or longitudinal cognitive monitoring.

To address the need for reliable, standardized, remote cognitive assessments, the National Institute on Aging (NIA) has awarded multiple grants to create the "Mobile Toolbox," (MTB) (www.mobiletoolbox.org, Gershon et al., 2022) a library of cognitive tests and supplemental scales embedded within the REDCap system and their companion MyCap App (Harris et al., 2022; Harris et al., 2019; Harris et al., 2009). MTB and REDCap make it possible for researchers to design and deploy smartphone-based studies that

Corresponding author: Stephanie Ruth Young; Email: stephanieruth.young@northwestern.edu

Cite this article: Novack M.A., Young S.R., Dworak E.M., Kaat A.J., Slotkin J., Nowinski C., Yao L., Adam H., Stoeger J., Hosseinian Z., Amagai S., Pila S., Varela Diaz M., Almonte Correa A., Alperin K., Carlson S., Kellen M., Omberg L., Camacho M.R., Landavazo B., Nosheny R.L., Weiner M.W., & Gershon R.C. (2024) Mobile toolbox (MTB) remote measures of executive function and processing speed: development and validation. *Journal of the International Neuropsychological Society*, **30**: 680–688, https://doi.org/10.1017/S1355617724000225

participants can complete anywhere and anytime through an iOS- or Android-based smartphone device. Through MyCap, the MTB measures are highly accessible, affordable, and as we review here, have been validated for use across the adult lifespan. Raw data from the tasks are uploaded to servers, where they are aggregated, processed for quality, and used to generate performance metrics. All MTB assessments are designed to measure well-established constructs, using existing paradigms optimized for self-administration on a personal smartphone. Furthermore, the MTB system is designed to allow these and other measures to be combined into electronic protocols customized for the needs of a large number and variety of future studies. Originally intended for research use, the MTB is not currently suitable for individual diagnosis, but the platform provides a mechanism for future research to explore this potential use. Most of the initial core MTB cognitive tests were adapted from measures included in the NIH Toolbox® for Assessment of Neurological and Behavioral Function Cognitive Battery (NIHTB-CB) or are measures of similar constructs (Carlozzi et al., 2015; Carlozzi et al., 2014; Gershon et al., 2013; Weintraub et al., 2013; Zelazo et al., 2014). The goal of the current paper is to describe the process of developing and validating three MTB measures that assess EF and PS: Arrow Matching (inhibitory control), Shape-Color Sorting (cognitive flexibility), and Number-Symbol Match (speed of information processing). These measures share similarity in that they all rely on reaction time and are all adaptations of existing NIHTB assessments.

Data from three distinct samples were used to evaluate the psychometric properties of the MTB measures. In Study 1, participants completed the MTB measures in a lab on a study-provided smartphone and were administered external measures of similar and dissimilar cognitive constructs. The goal of Study 1 was to evaluate the convergent and divergent validity, internal consistency (split half), and correlations with age for MTB measures when completed in a controlled laboratory setting. In Study 2, participants completed MTB measures remotely on their own smartphones and were administered measures of similar cognitive constructs in the lab. The goal of Study 2 was to replicate the results from Study 1 and to evaluate the psychometric properties of the MTB measures when taken remotely on a personal smartphone. Study 2 also allowed us to compare results across Android and iOS devices. In Study 3, participants completed the MTB twice on their own smartphone, two weeks apart. The goal of Study 3 was to examine the test-retest reliability of the MTB measures when taken remotely.

Method

Measure development

Two EF measures from the NIHTB-CB, Flanker Inhibitory Control and Attention Test and Dimensional Change Card Sort Test (DCCS; Weintraub et al., 2013; Zelazo et al., 2014), and one PS measure from the NIHTB supplemental tests, the Oral Symbol Digit Test (See Carlozzi et al., 2015; Healy & Fernald, 1911), were selected for adaptation for self-administration on the Mobile Toolbox app. A preliminary version of each measure was created for usability and pilot testing. The results of those initial tests informed revisions, which were incorporated into the versions of the measures that were used for the validation studies.

Arrow matching

Arrow Matching assesses the inhibitory control component of executive functioning. Based on the original Eriksen flanker task (Eriksen & Eriksen, 1974) as well as the NIHTB version (Flanker Inhibitory Control and Attention Test; Zelazo et al., 2014) participants 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 in a line (See Figure 1A). Four flanking arrows appear for a fraction of a second (100 ms) prior to a central arrow. Examinees then have 2000ms to respond with the direction of the central arrow, selecting from two buttons. Participants complete 50 trials in a pseudo-random order, of which approximately one third of the central stimuli are incongruent with the flankers. A centrally located star rotates during a variable (500 ms, 1250 ms, or 2000 ms) inter-stimulus-interval (ISI). The movement of the star was chosen to help participants maintain attention and provide a sense of system status, communicating that another trial is soon to appear.

One difference between MTB Arrow Matching and its NIHTB counterpart (Flanker) was the addition of more trials (50 vs 20), with less time allotted for each item (2000 ms vs. 10,000 ms). This faster auto-advance, combined with a variable ISI, was implemented to increase task difficulty, with the goal of expanding distribution of performance.

Shape-color sorting

Shape-Color Sorting measures the cognitive flexibility component of executive function. Based on the Dimensional Change Card Sort Test (Zelazo et al., 2014), participants are cued to match a bivalent central test stimulus to one of two target stimuli based on one of two dimensions (Figure 1B). Trials vary in the relevant dimension, requiring participants to shift their matching rules.

In this test, which is taken in portrait orientation, trials switch between cueing “color” and “shape.” The measure begins with five mixed-practice items, followed by 30 test trials, 20 percent of which cue “color”. The cued word is presented in lowercase text because words in lowercase font are more easily recognizable (Tinker, 1963). There is a variable-length ISI (either 300 or 1000ms) between each trial and participants have 2500ms to respond to each trial.

Whereas the NIHTB version (DCCS) uses a bunny 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.

Number-symbol match

Number-Symbol Match is an electronic adaptation of the many extant “coding” types of tests that originated in the early 20th century (Healy & Fernald, 1911), and shares similarities with the NIH Toolbox Oral Symbol Digit Test (Carlozzi et al., 2014), which was similarly adapted from this original source. This measure assesses processing speed by instructing participants to use a reference key to pair numbers with symbols in a constrained time.

Number-Symbol Match, completed in landscape orientation, presents a “key” at the top, showing the numbers one through nine with a unique symbol connected to each number (See Figure 1C). Below the key, nine symbols are presented per screen, and the participant must tap the correct number for each symbol presented, according to the key at the top. Symbol order is pseudo-random, with the condition that no identical symbols appear contiguously. The test includes 16 successive screens of 9 items each (total items = 144) and participants are given 90 s to complete as many items as possible.

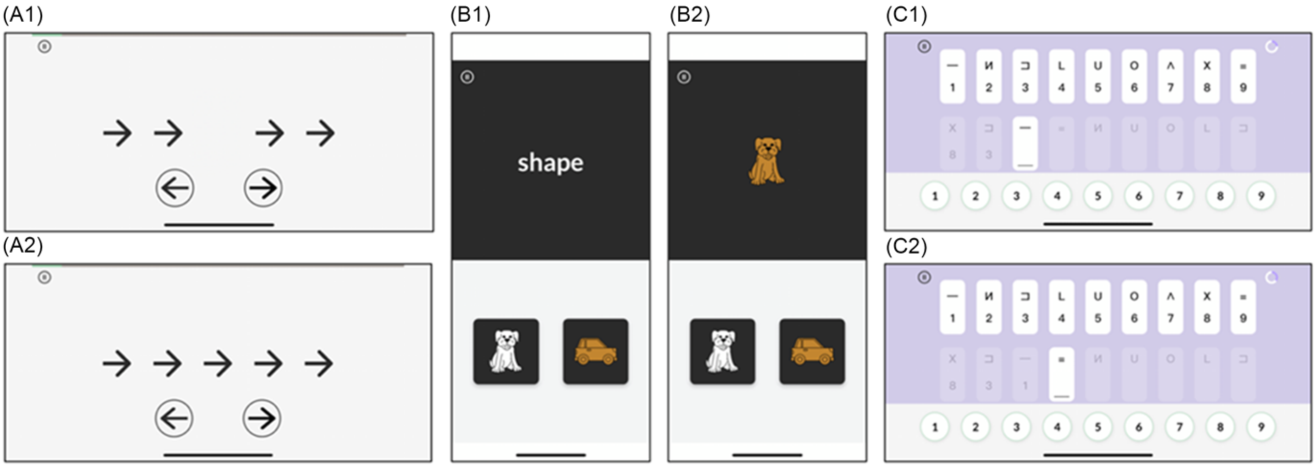

Figure 1. Screenshots of the EF and PS MTB measures. In Arrow Matching, participants see four flanking arrows (A1) before the central stimulus appears (A2), and they indicate the orientation of the central arrow using the buttons below. In Shape-Color Sorting, participants are cued to one of two dimensions (B1), after which they are presented with a bivalent target and use the buttons below to match with the corresponding category (B2). In Number-Symbol Match (far right figures), the top row contains a "key" and the bottom row contains nine symbols. The participant uses the number buttons at the bottom to select the correct number for each symbol. Each symbol is sequentially highlighted (e.g., after indicating one's response for C1, the measure automatically advances to C2).

One difference between the MTB version of Number-Symbol Match and other similar tests, including the NIHTB version of the Oral Symbol Digit Test, is that self-corrections are not permitted with the MTB design. Moreover, the Oral Symbol Digit Test uses oral responses whereas Number-Symbol Match uses a motoric response (tapping a button), as the latter was expected to reduce potential effects of background noise when collecting data in unknown environments and obviates the need for an examiner or speech recognition processing of responses.

Validation studies

Participants

92 participants from Study 1 ($M_{age} = 49.27$ years, SD = 17.65) and 1021 participants from Study 2 ($M_{age} = 43.97$ years, SD = 21.24) were enrolled in the NIHTB version 3 re-norming study and had been recruited by a third-party market research firm. They were racially and ethnically diverse and represented a range of age groups and education levels (albeit few participants had less than a high school education). The only inclusion criteria for participants were: 1) age 18 or older; 2) ownership of an iOS or Android smartphone; 3) ability to consent to participation in English. Participants were not screened for cognitive impairments prior to participation. Participants in Study 3 ($N = 168$, $M_{age} = 63.54$, SD = 12.10) were enrolled as part of a larger independent validation study through the Brain Health Registry (BHR), an online, longitudinal platform with over 100,000 members (Michael W Weiner et al., 2023). BHR consists of a public-facing website and a participant portal, where participants over the age of 18 can register, create an online profile, complete an online informed consent form, and complete study tasks. Participants are recruited to BHR through different methods (Weiner et al., 2023; Weiner et al., 2018) and advertising themes and messages include those tailored towards older adults with normal cognition, as well as those likely to have subjective cognitive decline and cognitive impairment.

Study 3 participants were required to be fluent in English, have previously opted-in to learning about additional research opportunities within BHR, and were required to have a compatible smartphone device. Participants were not screened for cognitive impairment. Due to an unexpected technical issue that corrupted data from Android devices, only users of iOS were included in this sample. See Table 1 for full demographic breakdown of participants in the three studies.

Procedure

Study 1. Participants self-administered the MTB measures on study-provided iOS smartphones (iPhones), unproctored in the lab. They were also administered the NIHTB Version 3 measures on study-provided tablets (iPads), which included measures of interest for validation: Flanker, DCCS, Oral Symbol Digit Test, and Pattern Comparison Processing Speed Test. Participants were also administered several external measures of similar constructs, including the Delis Kaplan Executive Function System (D-KEFS) Color Word Interference Test (Delis et al., 2001), Wisconsin Card Sorting Test (WCST-64; Heaton, 1981), and the Coding and Symbol Search subtests from the Wechsler Adult Intelligence Scale, 4th edition (Wechsler, 2008), as well as two measures for divergent validity – the Peabody Picture Vocabulary Test, 5th edition (PPVT-5; Dunn, 2018) and the NIH Toolbox Picture Vocabulary Test (TPVT; Gershon et al., 2014), both of which measure receptive vocabulary, a construct that is distinct from EF and PS. The PPVT has previously been used as a measure of divergent validity vis-à-vis the NIHTB EF measures (Zelazo et al., 2014). We expected the MTB EF measures to correlate at least moderately with respective measures of similar constructs (r > .3) and weakly with a measure of a divergent construct (i.e., the PPVT-5; $r < .3$).

Study 2. Participants were administered the NIHTB measures used on Study 1 in the lab, and then completed the MTB measures on their own iOS or Android smartphone remotely, no more than 14 days later.

Study 3. BHR participants were invited by email, screened for eligibility (access to a compatible smartphone), and provided online instructions for MTB app download. Participants self-administered the MTB measures on their own iOS smartphone remotely twice - once at baseline, and once 14 (± 3) days later. Participants were only included in final analyses if they completed the measures at both timepoints, and if they did not switch devices between sessions (i.e., from iOS to Android, iPhone to iPad). Of the

Table 1. Sample descriptives for validation studies

	Study 1	Study 2	Study 3
	(*N* = 92)	(*N* = 1021)	(*N* = 168)
Missing data	% (n)	% (n)	% (n)
Arrow matching	0.00 (0)	3.82 (39)	15.48 (26)
Shape-color sorting	0.00 (0)	5.58 (57)	14.29 (24)
Number-symbol match	0.00 (0)	8.81 (90)	16.07 (27)
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			
iPhone	100.00 (92)	63.66 (650)	100.00 (168)
Android	0.00 (0)	36.34 (371)	0.00 (0)
Gender			
Female	67.39 (62)	55.63 (568)	83.93 (141)
Male	32.61 (30)	44.37 (453)	16.07 (27)
Other	0.00 (0)	0.00 (0)	0.00 (0)
Not identified	0.00 (0)	0.00 (0)	0.00 (0)
Racial Identity			
White or Caucasian	52.17 (48)	73.65 (752)	88.69 (149)
Black or African American	32.61 (30)	13.91 (142)	4.17 (7)
Asian	9.78 (9)	6.27 (64)	2.98 (5)
Native American or Alaska Native	0.00 (0)	0.69 (7)	0.60 (1)
Native Hawaiian or Other Pacific Islander	1.09 (1)	0.49 (5)	0.00 (0)
Middle Eastern or North African	0.00 (0)	0.88 (9)	0.00 (0)
Multiracial or more than one race	4.35 (4)	2.15 (22)	2.98 (5)
Other	0.00 (0)	0.00 (0)	0.60 (1)
Prefer not to say or not identified	0.00 (0)	1.96 (20)	0.00 (0)
Ethnic identity			
Hispanic / Latino (Any Race)	1.09 (1)	14.69 (150)	7.14 (12)
Not Hispanic / Latino (Any Race)	98.91 (91)	85.31 (871)	92.86 (156)
Prefer not to say or not identified	0.00 (0)	0.00 (0)	0.00 (0)
Education level			
Less than HS	2.17 (2)	1.67 (17)	0.00 (0)
HS Diploma or GED	54.35 (50)	32.03 (327)	0.60 (1)
Some college	20.65 (19)	35.16 (359)	25.60 (43)
College or bachelor's degree (4-year degree)	15.22 (14)	20.27 (207)	32.74 (55)
Graduate or professional degree (Any Level)	7.61 (7)	10.87 (111)	41.07 (69)
Prefer not to say or not identified	0.00 (0)	0.00 (0)	0.00 (0)

168 participants enrolled, 144 participants provided data for test-retest reliability for Shape-Color Sorting, 142 participants for Arrow Matching, and 141 participants for completed Number-Symbol Match.

Studies 1 and 2 were conducted in compliance with, and approved by, the Internal Review Board (IRB) at Northwestern University (IRB STU00207455) and Study 3 was conducted in compliance with, and approved by, the IRB at the University of California, San Francisco (IRB 20-30058). All data was obtained in accordance with Helsinki Declaration.

Analyses

Scores for Arrow Matching and Shape-Color Sorting use a rate-based score - the number of correct trials completed per second, which matches the scoring model used for NIHTB version 3 and is taken from prior literature (Woltz & Was, 2006). The score for Number-Symbol Match uses the number of correct responses completed in the allocated time (90 s). All MTB analyses reported here used raw scores. While these reflect the primary scores, similar to other EF tests (e.g., the DEFKS), additional metrics are also available for Arrow Matching and Shape-Color Soring including error rate, anticipation errors, median correct and incorrect, etc.

Spearman correlations were conducted to explore convergent validity against external measures of similar cognitive constructs (Study 1), and NIHTB equivalent measures (Studies 1 & 2), as well as divergent validity against the NIH Toolbox PVT (Studies 1 & 2) and PPVT (Study 2). Note that TPVT is an interval scale based on IRT models, and the remaining convergent measures are each ratio scales. Parametric tests are appropriate for the comparison of interval scales to ratio scales, as well as between ratio scales.

Tests of independence assessed whether MTB correlations with NIHTB varied as a function of testing environment (in-person vs. remote—i.e., Study 1 vs. Study 2). Performance across age was also computed with Spearman correlations (Studies 1 & 2). We anticipated that all three measures would correlate negatively with age, as EF and PS are known to decline across the adult lifespan.

Internal consistency reliability (Studies 1 & 2) was calculated using a median correlation with Spearman-Brown correction between bootstrapped random split-half coefficients. Test-retest reliability (Study 3) was calculated using intraclass correlation coefficients, and practice effects were analyzed using linear mixed-effects models with two time points and a random intercept for participant (which is statistically equivalent to a paired samples *t*-test).

Finally, because subtle differences in operating system can influence the timing of stimuli presentation and response recording we considered the effect of operating system (iOS vs. Android) on standardized scores, using linear regression models and controlling for age in each model. In addition to evaluating differences in performance scores across device type, we also compared validity and reliability metrics across device types considering overlapping 95% Confidence Intervals (for validity estimates) and overlapping 25%–75% percentiles (for reliability estimates).

Analyses for all validation studies were conducted in R (R Core Team, 2023). Due to the number of comparisons, p-values were only considered significant if they were less than .01.

Results

Study 1 (In-person sample)

Examination of score distributions did not suggest floor or ceiling effects of the measures (i.e., there were very few cases with perfect or zero scores). Validity estimates comparing scores between MTB measures and NIHTB counterparts were strong, ranging from $r = .58$ to $r = .74$ (See Table 2). Validity estimates compared to external measures were more variable. Number-Symbol Match had a strong correlation with the WAIS-IV Coding score ($r = .68$) and Symbol Search ($r = .63$). Shape-Color Sorting correlated moderately in the expected negative direction with all D-KEFS subscores ($-.54 < r < -.38$), as well as moderately in the expected positive direction with the WCST-64 ($r = .41$). Similarly, Arrow Matching correlated moderately in the expected negative direction with all D-KEFS subscores ($-.45 < r < -.39$), as well as in the expected positive direction with the WCST-64 Test ($r = .32$).

All three measures showed no significant relationship to scores on the two vocabulary measures: NIHTB PVT ($r < .15$) and PPVT ($r < .29$), demonstrating evidence of divergent validity. Measures in this study also demonstrated very strong internal consistency reliability (r's $\geq .93$) and expected negative correlations with age (r's $\leq -.39$).

Study 2 (Fully remote sample)

Examination of score distributions did not suggest floor or ceiling effects of the measures. Reliability and validity estimates, as well as correlations with age, were largely similar to those seen in Study 1 (See Table 3 for all results on the Full Sample). Tests of independent correlations comparing NIHTB convergent validity correlations between the in-person (Study 1) and remote (Study 2) samples indicated no significant differences between the estimates for Arrow Matching vs. NIHTB Flanker ($z = 1.56$, $p = .12$), Shape-Color Sorting vs. NIHTB DCCS ($z = 0.18$, $p = .86$), Number-Symbol Match vs. NIHTB Pattern Comparison ($z = 0.21$, $p = .83$), or Number-Symbol Match vs. NIHTB Oral Symbol Digit Test ($z = 0.14$, $p = .89$). Finally, all three measures again showed no significant correlation with the divergent measures of NIHTB PVT (r's $\leq .15$) and PPVT (r's $\leq .27$). Like in Study 1, all three measures demonstrated very strong internal consistency reliability (r's $\geq .94$)

Unlike Study 1, in which all individuals completed the MTB measures on a study-provided iOS smartphone (iPhone), in Study 2, participants completed the measures on their own phones, 31% of which were Android devices. Therefore, Study 2 allowed us to compare MTB reliability and validity between operating systems.

First, we considered the effect of operating system on scores. Linear regressions demonstrated that operating system did indeed have a significant effect on standardized scores, controlling for age, for both Arrow Matching ($\beta = 0.23$, $p < .001$) and Shape-Color Sorting ($\beta = 0.21$, $p < .001$), but not for Number-Symbol Match ($\beta = 0.005$, $p = .93$). For both Arrow Matching and Shape-Color Sorting, scores were higher on iOS than on Android. See Table 4 for full regression results.

Importantly, despite the small effect of operating system on scores, there were no significant differences in convergent validity, divergent validity, or internal reliability between operating systems, as evidenced by overlapping confidence intervals (See Table 3 for estimate comparisons by operating system).

Study 3 (Fully remote test-retest sample)

Again, examination of score distributions did not suggest floor or ceiling effects of the measures. Shape-Color Sorting showed good test-retest reliability ($N = 144$, ICC = .78, 95% CI: [.71–.84]) with no significant practice effects from baseline to retest (average increase of 0.03 items correct per second, $t\,(143) = 1.8$, $p = .08$). Number-Symbol Match exhibited excellent test-retest reliability ($N = 141$, ICC = .83, 95% CI: [.74–.89]), with a small but significant increase of 2.73 additional items correct the second time it was taken ($t(140) = 4.8$, $p <$ .0001). Finally, Arrow Matching showed good test-retest reliability ($N = 142$, ICC = 0.69, 95% CI: [.59–.76]), and unexpectedly showed a small but significant decrease in performance between baseline and retest. On average, scores changed by -0.05 items correct per second ($t(141) = -2.3$, $p = .02$) on the second administration.

Discussion

This paper describes the results of a multi-part validation effort demonstrating the psychometric properties of three MTB measures that assess EF and PS. Results from the in-person sample (Study 1) provide convergent validity evidence compared to both NIHTB measure equivalents as well as external measures, under "optimal" circumstances. Results from a larger and fully remote sample (Study 2) replicate the validity and reliability results of Study 1 while demonstrating an effect of phone operating system on results for two of the tests. Study 3 provides evidence of test-retest reliability and practice effects.

All three measures demonstrated good evidence of convergent and divergent validity on both iOS and Android devices, supporting their effectiveness in assessing the specified constructs. Scores on Arrow Matching correlated strongly with those on the NIHTB Flanker and moderately with D-KEFS Color Word Inhibition including the raw score, Color Naming, Word Reading, and Inhibition/Switching. Shape-Color Sorting scores correlated strongly with the NIHTB DCCS measure, as well as moderately with the D-KEFS Inhibition/Switching. Number-Symbol Match correlated strongly with NIHTB Pattern Comparison and Oral Symbol Digit tests, and with the WAIS-IV Coding and Symbol Search scores. Additionally, all three measures had small, non-significant correlations with both the NIHTB PVT and PPVT, measures of vocabulary knowledge that tend to be a proxy for general abilities. Together, these results suggest that Arrow Matching, Shape-Color Sorting, and Number-Symbol Match assess the targeted constructs of EF and PS.

Correlations between MTB scores and age are similar to those reported for the original NIHTB (−0.50 to −0.55; Carlozzi et al., 2014; Zelazo et al., 2014). Additionally, correlations between MTB and NIHTB equivalents reported here were quite strong (ranging from .58 to .70), and correlations between MTB measures and external measures of similar cognitive constructs reflected a similar range to those seen for the original NIHTB validation study (.52 between NIHTB Flanker and D-KEFS Color-Word Interference Inhibition, .55 between NIHTB DCCS and D-KEFS Color-Word Interference Inhibition). This level of correlation is impressive given that MTB measures are self-administered (and, for two of our three samples, were completed in an unproctored setting).

Table 2. Validity and reliability estimates for Mobile Toolbox measures: Study 1 (in-person sample)

	Arrow matching		Shape-color sorting		Number-symbol match	
	Value	*N*	Value	*N*	Value	*N*
Age correlation	−.49^	92	−.39^	92	−.67^	92
Convergent validity						
NIHTB flanker	.74	84				
NIHTB DCCS	–	–	.69	92	–	–
NIHTB pattern comparison	–	–	–	–	.58	92
NIHTB oral symbol digit	–	–	–	–	.71	92
WCST-64 correct	.32	86	.41	86	–	–
D-KEFS color word inhibition raw score	−.41^	76	−.54^	76	–	–
D-KEFS raw score of color naming	−.45^	76	−.43^	76	–	–
D-KEFS raw score of word reading	−.39^	76	−.38^	76	–	–
D-KEFS raw score of inhibition/switching	−.45^	76	−.49^	76	–	–
WAIS-IV coding	–	–	–	–	.68	76
WAIS-IV symbol search	–	–	–	–	.63	76
Divergent validity						
NIHTB PVT	.15 (ns)	92	.13 (ns)	92	.06 (ns)	92
PPVT–5	.22 (ns)	78	.28 (ns)	78	.20 (ns)	78
Internal reliability: 50%, [25–75 percentile]	.97 [.96, .97]	92	.93 [.92, .94]	92	.99 [.97 .98]	92

Note. All correlations significant at $p < .01$ unless otherwise noted.
^Negative correlation expected due to the inverse relationship between scores (e.g., performance vs. age; speed vs. accuracy).

Table 3. Validity and reliability estimates for Mobile Toolbox measures: study 2 (fully remote sample)

		Arrow matching		Shape-color sorting		Number-symbol match	
		Value	*N*	Value	*N*	Value	*N*
Age correlation	Full sample	−.57^	982	−.50^	964	−.61^	931
	iOS	−.55 [−.61, −.49]^	626	−.46 [−.52, −.39]^	625	−.58 [−.63, .52]^	622
	Android	−.44 [−.52, −.35]^	356	−.41 [−.49, −.31]^	339	−.62 [−.69, .54]^	309
Convergent validity							
NIHTB flanker	Full sample	0.65	837	–	–	–	–
	iOS	.66 [.61, .71]	551	–	–	–	–
	Android	.57 [.48, .65]	286	–	–	–	–
DCCS	Full sample	–	–	.70	884	–	–
	iOS			.68 [.63, .72]	584		
	Android			.70 [.62, .76]	300		
Pattern comparison	Full sample	–	–	–	–	.57	923
	iOS					.58 [.52, .63]	616
	Android					.47 [.37, .55]	307
Oral symbol digit	Full sample	–	–	–	–	.70	917
	iOS					.69 [.64, .73]	613
	Android					.68 [.61, .74]	304
Divergent validity							
NIHTB PVT	Full sample	.003 (ns)	975	−.002 (ns)	958	−0.07 (ns)	924
	iOS	−.02 [−.1, .06]	621	−.01 [−.09, .06]	620	−.08 [−.16, −.004]	617
	Android	.14 [.04, .24]	354	.12 [.01, .22]	338	.009 [−.10, .12]	307
Reliability							
Internal reliability 50%, [25, 75 percentile]	Full sample	.97 [.97, .98]	982	.94 [.93, .94]	964	.98 [.97, .98]	931
	iOS	.97 [.97, .97]	626	.93 [.93, .94]	629	.98 [.97, .98]	626
	Android	.98 [.98, 0.98]	356	.94 [.93, .94]	335	.98 [0.97, .98]	305

Note. All correlations significant at $p < .01$ unless otherwise noted.
^Negative correlation expected due to the inverse relationship between scores (e.g., performance vs. age; speed vs. accuracy) All values, with the exception of Internal reliability report 95% CIs.

It is also notable that there were no differences in validity or internal consistency reliability coefficients between the sample that completed the MTB measures in person (Study 1) and those that completed the measures remotely (Study 2). Despite potential challenges facing remote assessment, the consistency in reliability and validity estimates offers further confidence in the utility of this tool as intended.

Finally in Study 3, we considered test-retest reliability on iOS devices only, as participants in this sample completed each measure twice, 14 days apart. All measures exhibited strong test-retest reliability with generally stable performance after two weeks. Nevertheless, we did see slight practice effects in the positive direction for Number-Symbol Match, which is comparable to practice effects on similar processing speed tests (Carlozzi et al., 2014) and in the negative direction for Arrow Matching, which differs from the positive practice effects previously found for NIHTB Flanker (Zelazo et al., 2014). Although the practice effects were minimal, they suggest that those interested in using the MTB EF and PS measures for high-frequency testing or Ecological Momentary Assessment designs use caution in interpreting changes in scores over relatively short periods of time. One challenge in developing cognitive assessments for remote administration on individuals' smartphones is that devices vary in their timing precision, which is problematic for tests that depend

Table 4. Regression results of device type predicting scores with and without age covariates from Study 2

	Arrow Matching		Shape-Color Sorting		Numbers-Symbol Match	
Model 1	***β***	***t***	***β***	***t***	***β***	***t***
Device type	0.56**	8.83	0.51**	7.79	0.38**	5.51
F(df)	77.94 (1, 980)		60.63 (1, 962)		30.40 (1, 929)	
R^2_{adj}		0.07**		.06**		0.03**
Model 2	***β***	***t***	***β***	***t***	***β***	***T***
Device type	0.23**	3.98**	0.21**	3.44	0.01	0.09
Age	−0.02**	−18.59**	−0.02**	−15.69	−0.03**	−21.35
F(df)	225.40 (2, 979)		161.10 (2, 961)		250.50 (2, 928)	
R^2_{adj}		0.31**		0.25**		0.35**
Model 3	***β***	***t***	***β***	***t***	***β***	***t***
Device type	0.22**	3.73	0.21**	3.39	0.00	0.05
Age	−16.26**	−18.64	−14.30**	−15.67	−18.04**	−21.35
Age^2	1.17	1.40	0.07	0.08	0.23	0.78
F(df)	151.10 (3, 978)		107.3 (3, 960)		166.80 (3, 927)	
R^2_{adj}		0.31**		0.25**		0.35**

Note. ***p* < .001.

on precise stimuli presentation and response timing (Germine et al., 2019; Passell et al., 2021). In comparing performance across users who completed the measures on an iOS vs. Android device in Study 2, we indeed found that operating system made a difference on scores for the two measures that are highly time dependent (Arrow Matching and Shape-Color Sorting), whereas it did not impact Number-Symbol Match, for which precise timing matters less. Given that the effect of operating system emerged only for the two measures that rely on precise stimuli presentation timing and response recording, these results suggest that the effect of operating system is likely due to software or hardware differences rather than a third, person-specific variable. Note also that there are differences beyond operating system that exist across different devices – Android devices in particular are manufactured by a wide range of companies using different hardware components that could affect the measurement of fine-grained timing events during tests. However, the diversity of devices used precluded any formal study of the effect of device hardware on results. Future work will determine whether there is a subset of hardware devices, such as older or less expensive models, that yield differing results. In the absence of this additional research, for studies proposing multiple assessments of the same individual on time-dependent tests, we strongly recommend ensuring that they complete the measures on the same device over time so that any device effect is consistent within an examinee, and that operating system is included as a covariate in analyses.

Despite a small impact on scores for the timed measures (Arrow Matching and Shape-Color Sorting), we found no differences in the convergent validity, divergent validity, or internal consistency across operating systems for any of the measures. This suggests that MTB measures can be used reliably with both types of operating systems. However, researchers should use caution when comparing Arrow Matching or Shape-Color Soring scores from different devices and may want to avoid combining different operating systems in their samples when using these measures. In contrast, this should not be a concern for Number-Symbol Match, which saw no differences in reliability, validity, or mean scores between operating systems.

Limitations

Despite their strengths, our studies have some limitations. First, test-retest reliability and practice effects were only examined on iOS devices due to a technical error in the Android sample. Further research with Android phones should be conducted to understand the influence of repeated administrations on the measures' reliability before they are used with Android samples. Second, although the demographics of our three samples were reasonably diverse, they lacked representation from certain groups, for example, those with less than a high school education. Future validation studies with underrepresented groups are important for the measures to be used in research with these populations. Third, the current samples were not recruited specifically to include individuals with cognitive impairments, and no cognitive assessments were conducted prior to enrollment. As the MTB was designed to track cognitive change across the lifespan and support research in cognitive decline, it is imperative that future work determine the feasibility, reliability and validity of the MTB in samples with impairments, including cognitive impairments, and other clinical groups.

The MTB is designed to be a remote assessment tool, which comes with both strengths and limitations. Remote measures that can be self-administered on a personal smartphone can reduce the cost and participation burden of research (Naito et al., 2021). However, it is difficult to monitor for cheating or poor effort, as well as other environmental factors that may influence test performance when measures are taken in remote settings. The measures time out after 10 minutes of inactivity to protect against low engagement; however, we were not able to monitor for other types of performance validity within these measures. Although it is difficult to cheat on these EF measures as participants cannot look up answers, it is possible that some participants asked another person to complete the test for them or did not try their best on the measures. Future versions of the MTB will implement measures to monitor and control for performance validity in remote settings, as well as collect data on contextual factors such as background noise or movement, to empirically test if and how these factors impact test performance in the real world. Moreover, users can easily include their own instructions to participants through the REDCap system to address engagement concerns.

Finally, while the three samples, particularly those from Studies 1 and 2, were reasonably diverse, they do not reflect the comprehensive demographic breakdown of populations in the US. This does limit the generalizability of validity and reliability across all populations and limits the use of these data for creating normed scores. As of now, MTB can be considered useful for research purposes in the tested populations but is not appropriate for clinical use or high-stakes testing, and may not be as useful when testing populations underrepresented in this study.

Conclusion

MTB Arrow Matching, Shape-Color Sorting, and Number-Symbol Match are shown here to be reliable and valid tools for remotely assessing EF and PS in healthy adults. Future work should consider their efficacy in additional contexts including with clinical populations, however the work reported here provides a critical foundation for the expansion of the MTB in future studies. Our hope is that the MTB will enhance research on cognitive change across the lifespan and advance our knowledge of both typical and atypical cognitive decline.

Acknowledgements. This work was supported by the National Institutes of Health grant U2CAG060426. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors would like to recognize the contributions of the Sage Bionetworks UI / UX Design team in conducting the user research studies and contributing to the design of the tests, and the Sage Engineering teams that developed and managed the system for remote administration of the study.

Competing interests. Dr Weiner serves on Editorial Boards for Alzheimer's & Dementia, and the Journal for Prevention of Alzheimer's Disease. He has served on Advisory Boards for Acumen Pharmaceutical, Alzheon, Inc., Amsterdam UMC; MIRIADE, Cerecin, Merck Sharp & Dohme Corp., and NC Registry for Brain Health. He also serves on the University of Southern California (USC) ACTC grant, which receives funding from Eisai. He has provided consulting to Boxer Capital, LLC, Cerecin, Inc., Clario, Dementia Society of Japan, Dolby Family Ventures, Eisai, Guidepoint, Health and Wellness Partners, Indiana University, LCN Consulting, MEDA Corp., Merck Sharp & Dohme Corp., NC Registry for Brain Health, Prova Education, T3D Therapeutics, USC, and WebMD. He has acted as a speaker/lecturer for China Association for Alzheimer's Disease and Taipei Medical University, as well as a speaker/lecturer with academic travel funding provided by: AD/PD Congress, Amsterdam UMC, Cleveland Clinic, CTAD Congress, Foundation of Learning; Health Society (Japan), Kenes, U. Penn, U. Toulouse, Japan Society for Dementia Research, Korean Dementia Society, Merck Sharp & Dohme Corp., National Center for Geriatrics and Gerontology (NCGG; Japan), USC. He holds stock options with Alzeca, Alzheon, Inc., ALZPath, Inc., and Anven. Dr Weiner received support for his research from the following funding sources: National Institutes of Health (NIH)/NINDS/NIA, Department of Defense, California Department of Public Health, University of Michigan, Siemens, Biogen, Hillblom Foundation, Alzheimer's Association, Johnson & Johnson, Kevin and Connie Shanahan, GE, VUmc, Australian Catholic University (HBI-BHR), The Stroke Foundation, and the Veterans Administration.

All other authors have no conflicts to report.

References

Ben-Zeev, D., & Atkins, D. C. (2017). Bringing digital mental health to where it is needed most. *Nature Human Behaviour*, *1*(12), 849–851.

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.

Carlozzi, N. E., Tulsky, D. S., Chiaravalloti, N. D., Beaumont, J. L., Weintraub, S., Conway, K., & Gershon, R. C. (2014). NIH toolbox cognitive battery (NIHTB-CB): The NIHTB pattern comparison processing speed test. *Journal of the International Neuropsychological Society*, *20*(6), 630–641.

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

Dunn, D. Peabody picture vocabulary test fifth edition (PPVT-5) (2018).

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143–149.

Germine, L., Reinecke, K., & Chaytor, N. S. (2019). Digital neuropsychology: Challenges and opportunities at the intersection of science and software. *Clinical Neuropsychologist*, *33*(2), 271–286. https://doi.org/10.1080/13854046.2018.1535662

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

Gershon, R. C., Sliwinski, M. J., Mangravite, L., King, J. W., Kaat, A. J., Weiner, M. W., & Rentz, D. M. (2022). The mobile toolbox for monitoring cognitive function. *The Lancet Neurology*, *21*(7), 589–590.

Gershon, R. C., Wagster, M. V., Hendrie, H. C., Fox, N. A., Cook, K. F., & Nowinski, C. J. (2013). NIH toolbox for assessment of neurological and behavioral function. *Neurology*, *80*(11_supplement_3), S2–S6.

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). MyCap: A flexible and configurable platform for mobilizing the participant voice. *JAMIA Open*, *5*(2), 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. (2019). The REDCap consortium: Building an international community of software platform partners. *Journal of Biomedical Informatics*, *95*, 103208.

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*, *42*(2), 377–381.

Healy, W., & Fernald, G. M. (1911). Tests for practical mental classification. *The Psychological Monographs*, *13*(2), i–54.

Heaton, R. K. Wisconsin card sorting test manual (1981), Psychological assessment resources.

Koo, B. M., & Vizer, L. M. (2019). Mobile technology for cognitive assessment of older adults: A scoping review. *Innovation in Aging*, *3*(1), igy038.

Naito, A., Wills, A.-M., Tropea, T. F., Ramirez-Zamora, A., Hauser, R. A., Martino, D., Turner, T. H., Rafferty, M. R., Afshari, M., Williams, K. L., Vaou, O., McKeown, M. J., Ginsburg, L., Ezra, A., Iansek, R., Wallock, K., Evers, C., Schroeder, K., DeLeon, R., Yarab, N., Alcalay, R. N., & Beck, J. C. (2021). Expediting telehealth use in clinical research studies: Recommendations for overcoming barriers in North America. *npj Parkinson's Disease*, *7*(1), 34.

Passell, E., Strong, R. W., Rutter, L. A., Kim, H., Scheuer, L., Martini, P., Grinspoon, L., & Germine, L. (2021). Cognitive test scores vary with choice of personal digital device. *Behavior Research Methods*, *53*(6), 2544–2557.

R Core Team (2023). R: A language and environment for statistical computing. In *R Foundation for Statistical Computing*. https://www.R-project.org/

Reynolds, J. R., West, R., & Braver, T. (2009). Distinct neural circuits support transient and sustained processes in prospective memory and working memory. *Cerebral Cortex*, *19*(5), 1208–1221.

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review*, *103*(3), 403–428.

Salthouse, T. A. (2000). Aging and measures of processing speed. *Biological Psychology*, *54*(1-3), 35–54.

Tinker, M. A. (1963). Influence of simultaneous variation in size of type, width of line, and leading for newspaper type. *Journal of Applied Psychology*, *47*(6), 380–382.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale—Fourth Edition Administration and Scoring Manual.* Pearson.

Weiner, M. W., Aaronson, A., Eichenbaum, J., Kwang, W., Ashford, M. T., Gummadi, S., Santhakumar, J., Camacho, M. R., Flenniken, D., Fockler, J., Truran-Sacrey, D., Ulbricht, A., Scott Mackin,R., & Nosheny,R. L. (2023). Brain health registry updates: An online longitudinal neuroscience platform. *Alzheimer's & Dementia*, *19*(11), 4935–4951.

Weiner, M. W., Aaronson, A., Eichenbaum, J., Kwang, W., Ashford, M. T., Gummadi, S., Santhakumar, J., Camacho, M. R., Flenniken, D., Fockler, J., Truran-Sacrey, D., Ulbricht, A., Mackin, R. S., & Nosheny, R. L. (2023).Brain health registry updates: An online longitudinal neuroscience platform. *Alzheimer's Dementia*, https://doi.org/10.1002/alz.13077

Weiner, M. W., Nosheny, R., Camacho, M., Truran-Sacrey, D., Mackin, R. S., Flenniken, D., Ulbricht, A., Insel, P., Finley, S., Fockler, J., Veitch, D. (2018). The brain health registry: An internet-based platform for recruitment, assessment, and longitudinal monitoring of participants for neuroscience studies. *Alzheimer's & Dementia*, *14*(8), 1063–1076. https://doi.org/10.1016/j.jalz.2018.02.021

Weintraub, S., Dikmen, S. S., Heaton, R. K., Tulsky, D. S., Zelazo, P. D., Bauer, P. J., Carlozzi, N. E., Slotkin, J., Blitz, D., Wallner-Allen, K., Fox, N. A., Beaumont, J. L., Mungas, D., Nowinski, C. J., Richler, J., Deocampo, J. A., Anderson, J. E., Manly, J. J., Borosh, B., Havlik, R., Conway, K., Edwards, E., Freund, L., King, J. W., Moy, C., Witt, E., & Gershon, R. C. (2013). Cognition assessment using the NIH toolbox. *Neurology*, *80*(11_supplement_3), S54–S64.

Woltz, D. J., & Was, C. A. (2006). Availability of related long-term memory during and after attention focus in working memory. *Memory & Cognition*, *34*(3), 668–684.

Zelazo, P. D., Anderson, J. E., Richler, J., Wallner-Allen, K., Beaumont, J. L., Conway, K. P., 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.