




Research Article

Retest reliability and reliable change of community-dwelling Black/African American older adults with and without mild cognitive impairment using NIH Toolbox-Cognition Battery and Cogstate Brief Battery for laptop

Taylor Rigby^{1,2,3,4} , Voyko Kavcic⁵, Sarah R. Shair⁴, Tanisha G. Hill-Jarrett^{6,7}, Sarah Garcia⁸, Jon Reader^{1,9}, Carol Persad^{1,2}, Arijit K. Bhaumik^{1,2,9}, Subhamoy Pal^{1,9}, Benjamin M. Hampstead^{1,2,4} and Bruno Giordani^{1,2}

¹Michigan Alzheimer's Disease Research Center, MI, USA, ²Department of Psychiatry, University of Michigan, MI, USA, ³Department of Veterans Affairs Medical Center, Geriatric Research Education and Clinical Center, Ann Arbor, MI, USA, ⁴Department of Veterans Affairs Medical Center, Ann Arbor, MI, USA, ⁵Wayne State University, MI, USA, ⁶Department of Neurology, Memory and Aging Center, University of California San Francisco, CA, USA, ⁷Global Brain Health Institute, University of California San Francisco, CA, USA, ⁸Department of Psychology, Stetson University, FL, USA and ⁹Department of Neurology, University of Michigan, MI, USA.

Abstract

Objective: With the increased use of computer-based tests in clinical and research settings, assessing retest reliability and reliable change of NIH Toolbox-Cognition Battery (NIHTB-CB) and Cogstate Brief Battery (Cogstate) is essential. Previous studies used mostly White samples, but Black/African Americans (B/AAs) must be included in this research to ensure reliability. **Method:** Participants were B/AA consensus-confirmed healthy controls (HCs) (n = 49) or mild cognitive impairment (MCI) (n = 34) adults 60–85 years that completed NIHTB-CB and Cogstate for laptop at two timepoints within 4 months. Intraclass correlations, the Bland-Altman method, *t*-tests, and the Pearson correlation coefficient were used. Cut scores indicating reliable change provided. **Results:** NIHTB-CB composite reliability ranged from .81 to .93 (95% CIs [.37–.96]). The Fluid Composite demonstrated a significant difference between timepoints and was less consistent than the Crystallized Composite. Subtests were less consistent for MCIs (ICCs = .01–.89, CIs [–1.00–.95]) than for HCs (ICCs = .69–.93, CIs [.46–.92]). A moderate correlation was found for MCIs between timepoints and performance on the Total Composite ($r = -.40, p = .03$), Fluid Composite ($r = -.38, p = .03$), and Pattern Comparison Processing Speed ($r = -.47, p = .006$).

On Cogstate, HCs had lower reliability (ICCs = .47–.76, CIs [.05–.86]) than MCIs (ICCs = .65–.89, CIs [.29–.95]). Identification reaction time significantly improved between testing timepoints across samples. **Conclusions:** The NIHTB-CB and Cogstate for laptop show promise for use in research with B/AAs and were reasonably stable up to 4 months. Still, differences were found between those with MCI and HCs. It is recommended that race and cognitive status be considered when using these measures.

Keywords: Computerized neuropsychological assessment; computerized cognitive assessment; practice effects; psychometrics; reproducibility of results; reliability of tests

(Received 21 September 2023; final revision 10 July 2024; accepted 12 August 2024)

Introduction

As the population with dementia has grown, disparities have emerged in the prevalence of all cause dementia among different races. Older Black/African Americans (B/AAs) are disproportionately more likely than older Whites to have Alzheimer's disease (AD) and other dementias (Dilworth-Anderson et al., 2008; Power et al., 2021; Steenland et al., 2016; Yaffe et al., 2013). Further, despite the increased risk posed to B/AA older adults for developing dementia, B/AA adults are largely underrepresented

in research seeking to understand these diseases. There is also evidence that a missed or delayed diagnosis of AD and other dementia types is more common among B/AA older adults than among White older adults (Clark et al., 2005; Gianattasio et al., 2019; Lin et al., 2021), which then contributes to a delay of care that may impact disease trajectory and outcomes. Thus, it is increasingly important to identify people at risk for AD and related dementias as early as possible, in part through accurately identifying individuals with mild cognitive impairment (MCI). A diagnosis of MCI refers to cognitive decline that is not normal for

Corresponding author: Bruno Giordani; Email: giordani@umich.edu

Cite this article: Rigby T., Kavcic V., Shair S.R., Hill-Jarrett T.G., Garcia S., Reader J., Persad C., Bhaumik A.K., Pal S., Hampstead B.M., & Giordani B. Retest reliability and reliable change of community-dwelling Black/African American older adults with and without mild cognitive impairment using NIH Toolbox-Cognition Battery and Cogstate Brief Battery for laptop. *Journal of the International Neuropsychological Society*, 1–11, <https://doi.org/10.1017/S1355617724000444>

© The Author(s), 2024. Published by Cambridge University Press on behalf of International Neuropsychological Society. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

a person's age but generally does not affect that person's ability to carry out most activities of daily living (Gauthier *et al.*, 2006). MCI is classified as one of two types based on a person's symptoms: amnesic (memory issues predominate) or non-amnesic (other cognitive issues predominate; Petersen *et al.*, 2018; Alzheimer's Association, 2022). It is estimated that 10–15% of individuals with MCI go on to develop a form of dementia each year (Alzheimer's Association, 2022) and about 1/3 of people with MCI develop dementia due to AD within five years (Alzheimer's Association, 2022; 42). Others with MCI may revert to their preclinical cognition or remain clinically stable (Pandya *et al.*, 2016).

Traditionally, neuropsychological measures have been used in clinical settings and in research studies to identify and track those with cognitive decline. However, more recently introduced computerized measures have a relative ease of administration when compared to traditional neuropsychological methods (Diaz-Orueta *et al.*, 2020; Weintraub *et al.*, 2013). Consequently, computerized assessments will likely be in increasing demand. While traditional neuropsychological methods have been well studied, less is known about practice effects and the retest reliability of computerized testing methods, particularly with different racial/ethnic groups (Diaz-Orueta *et al.*, 2020; Scott *et al.*, 2019). Practice effects refer to the expected and common improvement in test performance due to repeated exposures to test materials (Calamia *et al.*, 2012; Portney & Watkins, 2009). Retest reliability can be defined as the extent to which a measurement is consistent and free of random measurement error (the fluctuation in scores of repeated assessments due to unpredictable factors; Portney & Watkins, 2009). Retest reliability is essential to help clinicians and researchers understand how much of a measured change in score is attributable to measurement error and how much represents a true condition (Calamia *et al.*, 2012). Further, practice effects can mask actual cognitive decline in longitudinal studies of older adults and thereby give the illusion of stability or only minor change (Calamia *et al.*, 2012). Reliable change can be used to assess whether a change at retest on a given variable is "reliable" (meaning it is statistically improbable that the change is due to measurement error) and therefore represents a meaningful change (Chelune *et al.*, 1993; Iverson, 2001).

As more studies begin to incorporate computerized cognitive measures into clinical trials and longitudinal research, and scientists and clinicians explore the clinical applications of these tools, it becomes increasingly important to better understand reliability and retest issues for these methods. To ensure that measures and treatments are valid and reliable for B/AAs, B/AAs must be included in the research exploring these subjects. NIH Toolbox-Cognition Battery (NIHTB-CB) and the Cogstate Brief Battery (Cogstate) are computerized cognitive assessment batteries frequently used in clinical research. In a previous study the NIHTB-CB has been shown to have retest concordance correlation coefficients in healthy older adults ages 60–80 years of .73 for the Fluid Composite and .92 for the Crystallized Composite with individual subtests ranging between .46 and .88 (Scott *et al.*, 2019). The NIHTB-CB has been shown to have interclass correlations in healthy adults ages 20–85 years of .79 for the Fluid Composite and .92 for the Crystallized Composite (Heaton *et al.*, 2014) with individual subtests ranging between .72 and .94 (Weintraub *et al.*, 2013). The Cogstate individual subtest retest reliability using interclass correlations has been shown to range from .22 to .94 with healthy adults aged 18–96 years (Cole *et al.*, 2013; Faletti *et al.*, 2006; Fredrickson *et al.*, 2010; Lim *et al.*, 2013), .79–.95 for those with amnesic MCI aged 60–96 years, and .68–.93 for those with Alzheimer's disease aged 60 to 96 (Lim *et al.*, 2013).

Despite the findings that older B/AAs are disproportionately more likely than older Whites to have all type dementia, previous studies examining the retest reliability of NIHTB-CB and Cogstate for laptop were conducted using mostly White samples (Cole *et al.*, 2013; Faletti *et al.*, 2006; Fredrickson *et al.*, 2010; Hammers *et al.*, 2011; Heaton *et al.*, 2014; Lim *et al.*, 2013; Scott *et al.*, 2019; Weintraub *et al.*, 2013). Thus, the current study aimed to assess the retest reliability of NIHTB-CB and the Cogstate Brief Battery for laptop up to 4 months in healthy controls and those with MCI in a B/AA sample. The differences in scores between testing timepoints were calculated to examine practice effects and provide cut scores to determine reliable change. The relationship between testing interval and performance was also examined.

It was hypothesized that the NIHTB-CB retest reliabilities for the healthy controls in an all B/AA sample would be similar to previous findings using non-impaired majority White samples (Heaton *et al.*, 2014; Weintraub *et al.*, 2013), as the scores used were a priori adjusted for age, sex, race/ethnicity, and education (available through NIHTB-CB for laptop). We hypothesized that the Crystallized Composite would be more reliable than the Fluid Composite, but that all three composites, and the subtests that comprise them, would demonstrate moderate to excellent reliability and small to medium practice effects up to 4 months in healthy controls. Less is known about the retest reliability and practice effects in those with MCI or AD when using the NIHTB-CB; however, those with MCI have demonstrated significantly attenuated learning performance on accuracy and reaction time tasks with repeated computerized testing when compared to healthy controls (Darby *et al.*, 2002). Thus, we hypothesized that those with MCI would demonstrate moderate to excellent reliability but be less susceptible to practice effects than healthy controls, particularly on Fluid tasks requiring a memory component. Less is known about the impact of shorter versus longer test intervals in the NIHTB-CB for either healthy controls or those with MCI; thus, findings should be viewed as exploratory.

On the Cogstate, no demographic-adjusted norms have been provided by the manufacturers. Still, it was hypothesized that all subtests would demonstrate moderate to excellent reliability and small to medium practice effects in healthy controls up to 4 months in an all B/AA sample based on the performance of majority White samples in prior studies (Cole *et al.*, 2013; Faletti *et al.*, 2006; Fredrickson *et al.*, 2010; Lim *et al.*, 2013). Based on a previous study with a majority White sample, it was hypothesized that those with MCI would demonstrate similar retest reliability and susceptibility to practice effects as healthy controls on Cogstate subtests (Lim *et al.*, 2013). Results have been mixed in studies exploring the length between retest intervals in Cogstate (Faletti *et al.*, 2006; Fredrickson *et al.*, 2010; Hammers *et al.*, 2011), so no a priori prediction was made.

Method

Participants

Participants were recruited through the Healthy Black Elders Center, the community engagement core for the Michigan Center for Urban African American Aging Research, a joint program through the Wayne State University Institute of Gerontology and the University of Michigan Institute of Social Research, and through the Michigan Alzheimer's Disease Research Center (MADRC). This research was completed in accordance with the Helsinki Declaration. This study was reviewed and approved by the human subjects Institutional Review Board at Wayne State

University in Detroit, MI, USA, and the human subjects Institutional Review Board at the University of Michigan Medical School in Ann Arbor, MI, USA. Participants were evaluated for decision making capacity at the time of the informed consent process. All participants signed consent as per the human subjects Wayne State University Institutional Review Board in Detroit, MI, USA, and the human subjects University of Michigan Medical School Institutional Review Board in Ann Arbor, MI, USA, prior to participation in the study. All participants completed the National Alzheimer's Coordinating Center (NACC) – Uniform Data Set (UDS) version 2 evaluation which included a multi-domain medical, neurological, social, and neuropsychological evaluation; participants were then diagnosed at the MADRC using NACC consensus conference criteria (Weintraub et al., 2009). NIHTB-CB and Cogstate results were not available to the consensus panel. The initial NIHTB-CB and Cogstate assessments were conducted up to 8 days before UDS visits and up to 117 days after UDS assessments with 71.1% of assessments taking place on the same day. The NIHTB-CB and Cogstate retest was conducted between 6 and 139 days (or within 4 months) after the initial administration with the mean being 46.9 days and the median being 33 days. Participants also completed a Computer Anxiety Survey (Wild et al., 2012) to assess their level of comfort with computers.

Participants were B/AA community-dwelling older adults between 60 and 85 years of age that reported having either male or female biological sex. Participants included in the analyses completed NIHTB-CB and Cogstate at two testing timepoints within four months of the initial administration and were classified by consensus diagnosis (Weintraub et al., 2009) as either having no clinically significant cognitive impairment (healthy control; $n = 49$) or as having MCI ($n = 34$). Those with MCI were further classified at consensus as MCI with amnesic features (aMCI; $n = 24$) or MCI with non-amnesic features (naMCI; $n = 10$). Due to the low incidence of naMCI observed in this sample and the statistical equivalence on demographic variables (see Results section, *Demographics*) to those with aMCI, the aMCI and naMCI subsamples were combined and are described hereafter as the MCI group ($n = 34$).

Assessment measures

National Institutes of Health Toolbox-Cognition Battery (NIHTB-CB): The NIHTB-CB was designed to be a brief (30-min), computerized, widely accessible, and easily administered cognitive screener for ages 3–85 that is available in both English and Spanish (Gershon et al., 2013). It was originally designed for the purpose of creating a “common currency” among different research studies (Weintraub et al., 2013). The battery consists of seven tests measuring five cognitive domains, which are separated broadly into “fluid” or dynamic thinking skills (executive functions, episodic memory, processing speed, working memory) and “crystallized” or skills that remain relatively stable in adulthood (language – vocabulary knowledge and oral reading proficiency; Heaton et al., 2014; Weintraub et al., 2013). Individual subtest performances as well as composite summary scores of crystallized cognitive abilities, fluid cognition, and total cognition are provided. The Crystallized Cognition Composite includes the Oral Reading Recognition and Picture Vocabulary subtests. Measures of fluid abilities include the Dimensional Change Card Sort task, Flanker Inhibitory Control and Attention, List Sorting

Working Memory, Pattern Comparison Processing Speed, and Picture Sequence Memory subtests. Specific test details, procedures, and extensive psychometric evaluation are available elsewhere (Weintraub et al., 2013).

Cogstate Brief Battery (Cogstate): Cogstate is a computerized cognitive assessment that provides measures of four different cognitive domains using playing card paradigms: visual learning, working memory, processing speed, and attention. Briefly, the core tests include a Detection Task (a simple reaction time task), Identification Task (a choice reaction time test of visual attention), One Card Learning Task (a continuous visual recognition learning task), and One Back Task (a test of working memory). These separate tests and their psychometric properties have been described previously (Falletti et al., 2006; Lim et al., 2013; Maruff et al., 2013).

Computer Anxiety Survey: Computer anxiety was measured using the Wild et al. (2012) Computer Anxiety Survey, a 16-item measure on which participants rate their level of anxiety when using computers (e.g., “I feel relaxed when I am working on a computer”). Responses are rated on a five-point, Likert-type scale and range from “Strongly Disagree” to “Strongly Agree.” Total scores range from 16 to 80, with higher scores indicating greater levels of computer anxiety. Computer anxiety summary scores are derived by totaling the rating for each item. Specific survey details and psychometric properties have been described previously (Wild et al., 2012).

Mini-Mental State Examination (MMSE): The MMSE is a brief objective measure of cognitive functioning that quantitatively estimates the severity of cognitive impairment (Folstein et al., 1975). However, it is important to note that the MMSE is a brief cognitive screening tool and it is not meant to be used as a means of diagnosis, but rather as a path to referral for more comprehensive testing if needed (Arevalo-Rodriguez et al., 2021; Ranson et al., 2019; Tombaugh & McIntyre, 1992). The MMSE can typically be administered in 5–10 minutes. It consists of a variety of questions with total scores ranging from 0 to 30, and lower scores representing poorer cognitive function. Cut scores <24 and <25 are the most commonly used to suggest possible cognitive impairment (Tsoi et al., 2015); however, a cut score of 27 (≤ 26) has been shown to maximize the diagnostic accuracy of the MMSE in B/AA individuals with more education (Spering et al., 2012). Specific test details, procedures, and psychometric evaluations have been compiled in the form of review articles (Tombaugh & McIntyre, 1992; Tsoi et al., 2015).

Statistical analyses

All statistical analyses were conducted using SPSS V.28. Scores used in the analyses for NIHTB-CB were a priori adjusted (age, sex, race/ethnicity, and education) t-scores ($M = 50$, $SD = 10$) available through NIHTB-CB for laptop. Per manufacturer recommendations, scores used in the analyses for Cogstate were log transformed and derived from raw scores available through Cogstate; specifically, accuracy scores (correct vs. incorrect responses) for One Card Learning and One Back, and reaction time (in milliseconds) for Detection and Identification. Prior to analysis, all measures were screened for univariate and multivariate outliers and seven individual scores were identified as being extreme outliers (i.e., z-score >3.29) across measures. These seven individual outlying scores were then winsorized and subsequently included in the analyses. Accuracy scores for One Back were noted to be negatively skewed; specifically, healthy controls were found to be quite accurate

Table 1. Sample characteristics

	Total sample	HC	MCI	<i>p</i> -value
	<i>N</i> = 83	<i>n</i> = 49	<i>n</i> = 34	
Education (M/SD)	14.66(2.39)	15.12(2.32)	14(2.35)	.04
Age (M/SD)	71.25(5.72)	70.27(5.22)	72.68(6.17)	.07
Female <i>n</i> (%)	73(88%)	44(89.8%)	29(85.3%)	.54
MMSE	28.3(1.36)	28.76(1.03)	27.65(1.52)	< .001
Computer anxiety	42.29(14.02)	41.18(13.81)	43.88(14.37)	.40

Note: Chi-square test was used for categorical variables and *t*-scores for continuous variables. HC = healthy controls; MCI = mild cognitive impairment; MMSE = Mini-Mental State Evaluation; *p*-value = level of significance; M/SD = mean/standard deviation; *n* = number of participants.

when performing the One Back test. Those with MCI were also skewed negatively, but to a lesser extent.

Demographics

Demographic data were examined for group differences using independent measures *t*-test on continuous variables and chi-square statistics on categorical variables (see Table 1 for sample characteristics).

Intraclass correlation coefficients

To assess the degree of correlation and the agreement between measurement timepoints (or retest reliability), two-way mixed intraclass correlation coefficients (ICCs) with 95% confidence intervals (CIs) were run using an absolute definition of agreement for the total sample, healthy controls, and those with MCI (McGraw & Wong, 1996). ICCs were interpreted using the guidelines set forth by Koo and Li (2016) who recommended that CIs be interpreted rather than the singular correlation coefficient and defined values less than .40 as low reliability, values between .40 and .74 as moderate, values between .75 and .89 as good, and values greater than .89 as excellent (Table 2, Table 3).

Bland-Altman method

The Bland-Altman method was used to test for changes in the mean between the two test occasions and inspect for systematic bias and limits of agreement. Specifically, mean differences using the Bland-Altman method and 95% estimation of CIs for limits of agreement were calculated (Bland & Altman, 1995); bias was defined as having all observations lie to one side of the line of equality (i.e., observations that did not include zero (the line of equality) within the 95% CI) (Table 4).

Paired sample *t*-tests

Paired sample *t*-tests with 95% CIs were used to evaluate for practice effects upon retest in the total sample, healthy controls, and those with MCI. Cohen's *d* was used to measure effect size and results were interpreted using benchmarks suggested by Cohen (1988) with ± 0.2 as small, ± 0.5 as medium, and ± 0.8 as large (Table 5).

Reliable change

Reliable change methodology was used to calculate reliable change CIs that can be used to assess whether a change in score after retest is reliable and meaningful (Chelune et al, 1993, Iverson, 2001). Specifically, the standard error of difference score (SE_{diff}) was calculated using the standard error of measurement at initial

testing (SEM_1) and at retest (SEM_2) as per guidelines set by Iverson (2001). That is: $SE_{diff} = \sqrt{((SEM_1)^2 + (SEM_2)^2)}$, or $SE_{diff} = \sqrt{((SD_1\sqrt{1-r_{12}})^2 + (SD_2\sqrt{1-r_{12}})^2)}$, with SD_1 and SD_2 referring to the standard deviation at test and retest, respectively, and r_{12} referring to test-retest correlation between time 1 and time 2. The SE_{diff} was then multiplied by a *z*-score to arrive at CIs (70, 80, 90%).

Reliable change CIs can be directly referenced to determine if a change in score after retest on a given variable is reliable and meaningful. For use, examiners calculate a difference score (i.e., *t*-score at retest minus *t*-score at initial testing) using their patient or participant score(s). Norm-adjusted *t*-scores provided by NIHTB-CB should be used to calculate these difference scores, and log-transformed raw scores should be used for Cogstate. Difference scores for a given variable that change by the amount presented in Table 6 or more (either positive or negative) are indicative of worsening or improvement. For example, a person who changes by the amount provided or greater would be exceeding the change in scores experienced by 85, 90, or 95% of the sample, respectively.

Pearson correlation coefficient

The relationship between testing interval and performance was examined using the Pearson correlation coefficient. Specifically, the testing performance difference scores (derived by subtracting testing timepoint two from timepoint one for a given variable) and the length in days between testing timepoints were used. Results were interpreted using benchmarks suggested by Cohen (1988) with ± 0.1 as small, ± 0.3 as medium, and ± 0.5 as large.

Results

Demographics

Sample characteristics for the total sample, healthy controls, and those with MCI (combined variable aMCI and naMCI) can be found in Table 1. All participants identified as B/AA. No significant differences were found between healthy controls and those with MCI for age, sex, or level of computer anxiety. A significant difference was found between healthy controls and those with MCI for education, with healthy controls having approximately one more year of schooling on average than those with MCI. A significant difference in the total Mini-Mental State Exam score between healthy controls and those with MCI was found; specifically, healthy controls scored approximately one point higher on average than those with MCI. While there was no significant difference between healthy controls and those with MCI in terms of sex, the majority of the sample was female. Those with a diagnosis of aMCI were compared to those with a diagnosis of naMCI on demographic variables and no significant differences were noted in age, sex, or education between the two groups. Additionally, there was no statistical difference between those with aMCI versus those with naMCI on the Mini-Mental State Examination total score or level of computer anxiety in our sample.

NIHTB-CB

Intraclass correlation coefficients

When examining retest reliabilities in the total sample, healthy controls, and those with MCI on the NIHTB-CB, the Crystallized Composite (ICCs = .87–.93, 95% CIs [.74–.96]) demonstrated the highest reliability followed by the Total Composite (ICCs = .81–.91, ICs [.42–.96]) and the Fluid Composite, respectively (ICCs = .81–.85, ICs [.37–.93]; see Table 2). The reliability of individual

Table 2. Intraclass correlation coefficients examining retest reliability up to 4 months for NIH Toolbox-Cognition Battery and Cogstate Brief Battery for laptop

Variable	Total sample			Healthy controls			Mild cognitive impairment		
	ICC	<i>p</i> -value	95% CI	ICC	<i>p</i> -value	95% CI	ICC	<i>p</i> -value	95% CI
NIHTB-C Total Composite	.91	< .001	[.73, .96]	.90	< .001	[.71, .96]	.81	< .001	[.42, .92]
Crystallized Composite	.93	< .001	[.89, .95]	.93	< .001	[.87, .96]	.87	< .001	[.74, .94]
Picture Vocabulary	.81	< .001	[.71, .88]	.86	< .001	[.75, .92]	.66	.001	[.32, .83]
Oral Reading Recognition	.91	< .001	[.86, .94]	.86	< .001	[.76, .92]	.89	< .001	[.79, .95]
Fluid Composite	.85	< .001	[.57, .93]	.83	< .001	[.53, .93]	.81	< .001	[.37, .92]
Flanker Inhibitory Control and Attention	.73	< .001	[.53, .84]	.76	< .001	[.56, .87]	.69	< .001	[.33, .85]
Dimensional Change Card Sort	.75	< .001	[.59, .84]	.69	< .001	[.46, .83]	.77	< .001	[.47, .89]
List Sorting Working Memory	.62	< .001	[.42, .76]	.70	< .001	[.46, .83]	.01	.49	[-1.00, .51]
Pattern Comparison Processing Speed	.85	< .001	[.75, .91]	.85	< .001	[.73, .92]	.80	< .001	[.60, .90]
Picture Sequence Memory	.65	< .001	[.46, .78]	.73	< .001	[.49, .85]	.21	.26	[-.62, .62]
Cogstate									
Detection RT	.83	< .001	[.75, .89]	.76	< .001	[.57, .86]	.89	< .001	[.79, .95]
Identification RT	.72	< .001	[.54, .83]	.60	< .001	[.28, .77]	.83	< .001	[.65, .92]
One Card Learning Accuracy	.73	< .001	[.58, .82]	.59	< .001	[.27, .77]	.79	< .001	[.58, .90]
One Back Accuracy	.62	< .001	[.38, .74]	.47	.02	[.05, .70]	.65	.002	[.29, .83]
One Back RT	.84	< .001	[.75, .89]	.82	< .001	[.68, .90]	.82	< .001	[.63, .91]

Note: Scores used were the norm a priori adjusted (age, sex, race/ethnicity, and education) *t*-scores ($M = 50$, $SD = 10$) available through NIH Toolbox-Cognition Battery for laptop and log-transformed scores derived from raw scores for Cogstate Brief Battery for laptop. ICC = intraclass correlation coefficient; CI = confidence interval; *p*-value = level of significance; RT = reaction time.

NIHTB-CB measures for the total sample varied from moderate to excellent. For healthy controls, the individual tests comprising the Crystallized Composite both demonstrated good to excellent reliability and were more consistent (both ICCs = .86, CIs [.75–.92]) when compared to the Fluid Composite which ranged from moderate to good (ICCs = .69–.85, CIs [.46–.92]). Individual subtests were less consistent within and across the composites for those with MCI, with reliabilities ranging from low to excellent (ICCs = .01–.89, CIs [-1.00–.95]). Retest reliabilities were also calculated for aMCI and naMCI separately. Though caution must be stated given the lack of power, we found that reliabilities for those with aMCI and naMCI were similar on the Total Composite and Crystallized Composite, as well as many of the subtests comprising the Composites. However, those with aMCI were somewhat less reliable than those with naMCI on Picture Vocabulary (aMCI ICC = .66; naMCI ICC = .81) and much less reliable on List Sorting Working Memory (aMCI ICC = .15; naMCI ICC = .63) and Picture Sequence Memory (aMCI ICC = .10; naMCI ICC = .64). Thus, it appears that those with aMCI were driving the poor reliabilities on these measures reported in Table 2.

Bland-Altman method

Applying the Bland-Altman method to the three NIHTB-CB composites revealed that for the total sample, healthy controls, and those with MCI only the Crystallized Composite was not significant, and the 95% CI included zero; these findings indicated that there was no proportional or systematic bias in the Crystallized Composite across samples. Of the seven individual NIHTB-CB subtests Picture Vocabulary, Oral Reading Recognition, and List Sorting Working Memory were the only subtests that were not significant and included zero within the 95%CI when applying the Bland-Altman method to the total sample and healthy controls. For those with MCI, the pattern was similar, but Picture Sequence Memory was also not significant and included zero within the 95% CI for those with MCI.

Paired sample *t*-tests

On the paired sample *t*-tests comparing NIHTB-CB testing timepoint two to testing timepoint one, the Total Composite and the Fluid

Composite were significant for the total sample, healthy controls, and those with MCI (see Table 3); this demonstrates significant practice effects (through an increase in scores) with medium effect sizes. Conversely, the NIHTB-CB Crystallized Composite was not significant across samples and had small effect sizes, thereby demonstrating less susceptibility to practice effects. Of the individual NIHTB-CB subtests, the two subtests comprising the Crystallized Composite (Picture Vocabulary and Oral Reading Recognition) were not significant across samples and had small effect sizes. In the total sample and the healthy controls, all Fluid Composite subtests, except List Sorting Working Memory (which was insignificant and had small effect sizes), did differ significantly; specifically, improved performances (or practice effects) were seen on Flanker Inhibitory Control and Attention, Dimensional Change Card Sort, Pattern Comparison Processing Speed, and Picture Sequence Memory for healthy controls. For those with MCI, the fluid measures Flanker Inhibitory Control and Attention, Dimensional Change Card Sort, and Pattern Comparison Processing Speed did significantly improve, while Picture Sequence Memory and List Sorting Working Memory did not differ significantly between testing timepoints. Effect sizes for fluid subtest measures across samples ranged from small to medium.

Reliable change

Values used to calculate reliable change between timepoint 1 and 2 are listed in Tables 3 and 4. Reliable change CIs that can be used as cut scores to interpret reliable change are provided in Table 5. See the Statistical Analysis section of this paper for more information regarding the interpretation of reliable change.

Pearson correlation coefficient

Table 6 shows correlations between difference in days between test administrations and difference in testing performance between testing timepoints for the total sample, healthy controls, and those with MCI on the NIHTB-CB. Though the Fluid Composite and Pattern Comparison Processing Speed were considered significant at the < .05 level in the total sample, a low degree of correlation was noted with correlations ranging from -.28 to .06. For healthy controls, no correlations were significant, and correlations ranged from -.27 to .21. All correlations for those with MCI were negative

Table 3. Paired-sample *t*-tests examining practice effects for NIH Toolbox-Cognition Battery and Cogstate Brief Battery for laptop

Variable	Total sample						Healthy controls						Mild cognitive impairment					
	M(SD)1	M(SD)2	<i>r</i> ₁₂	<i>t</i>	<i>p</i> -values	<i>d</i>	M(SD)1	M(SD)2	<i>r</i> ₁₂	<i>t</i>	<i>p</i> -values	<i>d</i>	M(SD)1	M(SD)2	<i>r</i> ₁₂	<i>t</i>	<i>p</i> -values	<i>d</i>
NIHTB-C Composite	52.15(9.63)	55.45(9.78)	.88	6.21	< .001	0.70	56.27(9.03)	59.42(9.17)	.87	4.65	< .001	0.67	45.77(6.66)	49.28(7.24)	.76	4.06	< .001	0.73
Crystallized Abilities Composite	54.46(8.33)	54.84(7.71)	.87	0.81	.42	0.09	57.65(7.22)	57.93(6.47)	.86	0.52	.61	0.07	49.57(7.57)	50.10(6.68)	.78	0.62	.54	0.11
Picture Vocabulary	54.10(10.19)	55.20(8.54)	.69	1.33	.19	0.15	57.65(8.30)	58.04(7.46)	.75	0.49	.63	0.07	48.85(10.57)	50.99(8.38)	.51	1.28	.21	0.22
Oral Reading Recognition	53.43(8.05)	53.23(8.25)	.83	-0.37	.72	-0.04	56.42(6.44)	56.53(6.71)	.76	0.17	.87	0.02	48.98(8.24)	48.34(7.96)	.81	-0.73	.47	-0.13
Fluid Abilities Composite	49.19(11.60)	54.49(12.35)	.82	6.54	< .001	0.73	53.22(11.49)	58.46(11.55)	.79	4.83	< .001	0.70	43.14(8.93)	48.53(11.20)	.79	4.40	< .001	0.78
Flanker Inhibitory Control and Attention	50.10(8.16)	53.59(8.85)	.62	4.28	< .001	0.47	51.38(7.78)	54.09(8.28)	.65	2.8	.007	0.40	48.26(8.46)	52.86(9.70)	.60	3.26	.003	0.56
Dimensional Change Card Sort	54.51(13.98)	59.34(15.61)	.63	3.44	< .001	0.38	58.52(14.79)	62.81(15.26)	.54	2.09	.04	0.30	48.71(10.42)	54.35(14.95)	.72	3.16	.003	0.54
List Sorting Working Memory	49.30(8.97)	50.16(9.07)	.45	0.83	.41	0.09	52.20(8.91)	53.77(7.58)	.54	1.38	.17	0.20	45.13(7.34)	44.96(8.56)	.003	-0.09	.93	-0.02
Pattern Comparison Processing Speed	48.81(12.22)	52.11(14.21)	.77	3.27	.002	0.36	51.64(12.59)	55.10(14.89)	.78	2.56	.01	0.37	44.73(10.57)	47.81(12.14)	.69	2.00	.05	0.34
Picture Sequence Memory	45.30(12.63)	49.25(12.93)	.51	2.77	.007	0.31	48.20(13.94)	53.17(11.58)	.62	3.04	.004	0.44	40.95(8.91)	43.36(12.76)	.13	0.93	.36	0.17
Cogstate																		
Detection RT	2.61(0.11)	2.62(0.10)	.72	1.39	.17	0.15	2.59(0.10)	2.61(0.09)	.62	1.49	.14	0.21	2.64(0.12)	2.64(0.11)	.81	0.31	.76	0.05
Identification RT	2.78(0.09)	2.75(0.07)	.63	-3.65	< .001	-0.40	2.77(0.10)	2.73(0.05)	.55	-2.92	.005	-0.42	2.80(0.08)	2.78(0.08)	.73	-2.19	.04	-0.38
One Card Learning Accuracy	0.94(0.10)	0.96(0.09)	.57	1.07	.29	0.12	0.97(0.10)	0.99(0.09)	.42	0.86	.39	0.12	0.90(0.09)	0.91(0.09)	.65	0.63	.53	0.11
One Back Accuracy	1.32(0.19)	1.34(0.17)	.43	0.57	.57	0.06	1.36(0.18)	1.38(0.14)	.31	0.45	.66	0.07	1.27(0.20)	1.28(0.20)	.47	0.35	.73	0.06
One Back RT	2.97(0.11)	2.95(0.10)	.72	-1.38	.17	-0.15	2.94(0.10)	2.92(0.09)	.72	-2.13	.04	-0.31	3.00(0.10)	3.00(0.11)	.68	.21	.83	0.04

Note: Scores used were the norm a priori adjusted (age, sex, race/ethnicity, and education) *t*-scores (*M* = 50, *SD* = 10) available through NIH Toolbox-Cognition Battery for laptop and log-transformed scores derived from raw scores for Cogstate Brief Battery for laptop. *t* = *t*-test statistic; *M*(*SD*)1 = *M* = mean of testing timepoint 1, *SD* = standard deviation of testing timepoint 1; *M*(*SD*)2 = *M* = mean of testing timepoint 2, *SD* = standard deviation of testing timepoint 2; *r*₁₂ = Pearson correlation between timepoint 1 and timepoint 2; *p*-value = level of significance; *d* = Cohen's measure of sample effect size; RT = reaction time.

Table 4. Values used to calculate reliable change between timepoints 1 and 2 for NIH Toolbox-Cognition Battery and Cogstate Brief Battery for laptop

Variable	Total sample				Healthy controls				Mild Cognitive Impairment			
	SEM1	SEM2	M(SD)diff	SEdiff	SEM1	SEM2	M(SD)diff	SEdiff	SEM1	SEM2	M(SD)diff	SEdiff
NIHTB-C Total Composite	9.63	9.78	3.30(4.72)	6.72	9.03	9.17	3.16(4.70)	4.64	6.66	7.24	3.51(4.82)	4.82
Crystallized Composite	8.33	7.71	0.38(4.17)	4.09	7.23	6.75	0.27(3.71)	3.70	7.57	6.68	0.53(4.85)	4.74
Picture Vocabulary	10.19	8.54	1.10(7.47)	7.28	8.30	7.46	0.40(5.63)	5.58	10.58	8.38	2.14(9.59)	9.44
Oral Reading Recognition	8.05	8.25	-0.19(4.78)	4.75	6.44	6.71	0.11(4.6)	4.56	8.24	7.96	-0.64(5.06)	4.99
Fluid Composite	11.60	12.35	5.30(7.25)	7.19	11.50	11.55	5.25(7.53)	7.47	8.93	11.20	5.39(6.93)	6.56
Flanker Inhibitory Control and Attention	8.16	8.85	3.48(7.42)	7.42	7.78	8.28	2.71(6.78)	6.72	8.46	9.70	4.60(8.23)	8.14
Dimensional Change Card Sort	13.98	15.61	4.84(12.83)	12.75	14.79	15.26	4.28(14.34)	14.26	10.42	14.95	5.64(10.41)	9.64
List Sorting Working Memory	8.97	9.07	0.86(9.43)	9.46	8.91	7.58	1.57(7.96)	7.94	7.34	8.57	-0.17(11.26)	11.26
Pattern Comparison Processing Speed	12.22	14.21	3.30(9.21)	8.99	12.59	14.89	3.45(9.44)	9.15	10.57	12.14	3.08(9.00)	8.96
Picture Sequence Memory	12.63	12.93	3.94(12.72)	12.65	13.94	11.58	4.97(11.32)	11.17	8.91	12.76	2.41(14.61)	14.52
Cogstate												
Detection RT	0.11	0.09	0.01(0.08)	0.08	0.10	0.09	0.02(0.08)	0.08	0.13	0.11	0.004(0.07)	0.07
Identification RT	0.09	0.07	-0.03(0.07)	0.07	0.10	0.05	-0.03(0.08)	0.07	0.08	0.08	-0.02(0.06)	0.06
One Card Learning Accuracy	0.10	0.09	0.01(0.09)	0.09	0.10	0.09	0.01(0.10)	0.10	0.09	0.09	0.008(0.07)	0.07
One Back Accuracy	0.19	0.17	0.01(0.19)	0.19	0.18	0.15	0.01(0.19)	0.19	0.20	0.20	0.01(0.20)	0.20
One Back RT	0.11	0.10	-0.01(0.08)	0.08	0.10	0.09	-0.02(0.07)	0.07	0.10	0.11	0.003(0.08)	0.08

Note: Scores used were the norm a priori adjusted (age, sex, race/ethnicity, and education) *t*-scores (*M* = 50, *SD* = 10) available through NIH Toolbox-Cognition Battery for laptop and log transformed scores derived from raw scores for Cogstate Brief Battery for laptop. SEM1 = standard error of measurement at testing timepoint 1; SEM2 = standard error of measurement at testing timepoint 2; M(SD)diff = with *M* being the mean difference and *SD* being the standard deviation of difference; SE = standard error of difference; RT = reaction time.

Table 5. Reliable change confidence intervals for NIH Toolbox-Cognition Battery and Cogstate Brief Battery for laptop

Variable	Total sample			Healthy control			Mild cognitive impairment		
	70% CI	80% CI	90% CI	70% CI	80% CI	90% CI	70% CI	80% CI	90% CI
NIHTB-C Total Composite	6.99	8.61	11.03	4.83	5.94	7.61	5.01	6.17	7.90
Crystallized Composite	4.25	5.24	6.71	3.85	4.73	6.06	4.93	6.06	7.77
Picture Vocabulary	7.57	9.32	11.94	5.80	7.14	9.15	9.82	12.09	15.49
Oral Reading Recognition	4.94	6.08	7.79	4.74	5.83	7.47	5.19	6.39	8.19
Fluid Composite	7.48	9.20	11.78	7.76	9.56	12.24	6.83	8.40	10.76
Flanker Inhibitory Control and Attention	7.72	9.50	12.17	6.99	8.60	11.02	8.47	10.42	13.35
Dimensional Change Card Sort	13.26	16.31	20.90	14.82	18.25	23.38	10.03	12.34	15.81
List Sorting Working Memory	9.83	12.10	15.51	8.25	10.16	13.02	11.71	14.42	18.47
Pattern Comparison Processing Speed	9.35	11.51	14.75	9.51	11.71	15.00	9.32	11.49	14.69
Picture Sequence Memory	13.16	16.19	20.75	11.62	14.30	18.32	15.10	18.58	23.81
Cogstate									
Detection RT	0.08	0.10	0.13	0.09	0.11	0.14	0.07	0.09	0.12
Identification RT	0.07	0.09	0.11	0.08	0.09	0.12	0.06	0.08	0.10
One Card Learning Accuracy	0.09	0.11	0.15	0.10	0.13	0.16	0.08	0.09	0.12
One Back Accuracy	0.20	0.25	0.32	0.20	0.24	0.31	0.21	0.26	0.33
One Back RT	0.08	0.10	0.13	0.08	0.09	0.12	0.09	0.11	0.14

Note: Scores used were the a priori norm adjusted (age, sex, race/ethnicity, and education) t-scores ($M = 50$, $SD = 10$) available through NIH Toolbox Cognition Battery for laptop and log transformed scores derived from raw scores for Cogstate Brief Battery for laptop. Confidence intervals (CI) were calculated by multiplying the Standard Error of Difference of performance on testing timepoint 1 and testing timepoint 2 by a z-score to arrive at the confidence intervals (70%, 80%, 90%). If a retest score for a given variable changes by the provided amount or more (either positive or negative), that score is indicative of worsening or improvement. For example, a person whose score has gotten worse with retesting by the amount shown above (or greater) for a given variable, would be exceeding the worsening in scores experienced by 85, 90, or 95% of the sample, respectively. CI = Confidence Interval; RT = reaction time. CI = confidence interval.

Table 6. Correlation between difference scores and the difference in days between testing timepoints on the NIH Toolbox-Cognition Battery and Cogstate Brief Battery for laptop

Variable	Total sample		Healthy controls		Mild cognitive impairment	
	<i>r</i>	<i>p</i> -value	<i>r</i>	<i>p</i> -value	<i>r</i>	<i>p</i> -value
NIHTB-C Composite	-.21	.06	-.12	.43	-.40	.03
Crystallized Abilities Composite	.06	.57	.21	.15	-.13	.47
Picture Vocabulary	.04	.74	.15	.32	-.09	.60
Oral Reading Recognition	.02	.88	.13	.38	-.14	.45
Fluid Abilities Composite	-.28	.01	-.22	.13	-.38	.03
Flanker Inhibitory Control and Attention	-.16	.16	-.17	.25	-.18	.31
Dimensional Change Card Sort	-.21	.06	-.27	.06	-.09	.61
List Sorting Working Memory	-.10	.39	-.05	.76	-.14	.43
Pattern Comparison Processing Speed	-.25	.02	-.12	.42	-.47	.006
Picture Sequence Memory	-.01	.92	.12	.44	-.16	.38
Cogstate						
Detection RT	.08	.47	-.01	.93	.28	.12
Identification RT	-.05	.65	-.05	.74	-.08	.65
One Card Learning Accuracy	.02	.84	-.04	.78	.17	.33
One Back Accuracy	.004	.97	.17	.24	-.23	.19
One Back RT	.03	.77	-.12	.41	.18	.30

Note: Scores used were the norm a priori adjusted (age, sex, race/ethnicity, and education) t-scores ($M = 50$, $SD = 10$) available through NIH Toolbox-Cognition Battery for laptop and log transformed scores derived from raw scores for Cogstate Brief Battery for laptop. Difference scores were calculated by subtracting the score from testing timepoint two from the score from testing timepoint one for a given variable per individual participant. Pearson correlation coefficient was used to evaluate the correlation between testing timepoint difference scores and the difference in days between testing timepoints. *r* = Pearson correlation coefficient; *p*-value = level of significance; RT = reaction time.

and many were low and not significant (ranging from $-.18$ to $-.09$), but a moderate degree of correlation was found between the difference in days between administrations and difference in testing performance between testing timepoints for the NIHTB-CB Total Composite, Fluid Composite, and Pattern Comparison Processing Speed.

Cogstate

Intraclass correlation coefficients

When examining retest reliabilities for the total sample, healthy controls, and those with MCI, the individual Cogstate subtests had CIs that ranged from low reliability to excellent (see Table 2). Healthy controls demonstrated consistently lower retest reliability (ranging from low to good reliability (ICCs = .47-.76, CIs

[.05-.86])) relative to those with MCI (ranging from low to excellent (ICCs = .65-.89, CIs [.29-.95])) on all Cogstate subtests. One Back Accuracy was the least reliable of the Cogstate measures for both those with MCI and healthy controls. An ad hoc analysis of One Back reaction time was conducted due to the skewed distribution of the One Back Accuracy subtest. It was found that One Back reaction time had a more normal distribution and significant reliabilities (all $p < .001$) ranging from moderate to excellent across samples (ICCs = .82-.84, CIs [.63-.91]; see Table 2). Reliabilities were also calculated for nMCI and aMCI separately. Though caution must be stated given the lack of power, we found that those with aMCI demonstrated similar reliabilities on Cogstate subtests with the exception of One Back Accuracy (aMCI ICC = .58; nMCI ICC = .81) and One Back reaction time (aMCI ICC = .73; nMCI ICC = .91).

Bland-Altman method

Of the four subtests that comprise Cogstate, only Identification reaction time was significant and did not contain zero within the 95% CI when applying the Bland-Altman method across samples (total sample, healthy controls, MCI). This indicates that there was proportional or systematic bias in this subtest but not the other three subtests across samples.

Paired sample t-tests

When comparing Cogstate testing timepoint two to testing timepoint one using paired sample *t*-tests it was found that Identification reaction time differed significantly with medium effect sizes for the total sample, healthy controls, and those with MCI (see Table 3). Specifically, participants demonstrated practice effects when they completed the task faster on average on the second administration across samples. Healthy controls were found to significantly improve their performance on One Back reaction time, whereas those with MCI did not. Detection reaction time, One Card Learning Accuracy, and One Back Accuracy did not differ significantly between testing timepoints or across samples.

Reliable change

Values used to calculate reliable change between timepoints 1 and 2 are listed in Tables 3 and 4. Reliable change CIs that can be used as cut scores to interpret reliable change are provided in Table 5. See the Statistical Analysis section of this paper for more information regarding the interpretation of reliable change.

Pearson correlation coefficient

A low degree of correlation was noted in the total sample, for healthy controls, and those with MCI when correlating difference in days between test administrations and difference in testing performance between testing timepoints (see Table 4). Correlations ranged from $r = -.23$ to $.28$ and none were significant.

Discussion

NIHTB-CB

Across samples (total sample, healthy controls, those with MCI), the NIHTB-CB composite scores demonstrated good to excellent reliability up to 4 months in a B/AA sample. These findings are similar to prior research using healthy adults ages 20 to 85 with a majority White sample that retested participants between 7 and 21 days (Heaton et al., 2014). As hypothesized, and consistent with previous research, the NIHTB-CB Crystallized Composite was found to be more stable between testing timepoints and across samples when compared to the Fluid Composite (Heaton et al., 2014; Scott et al., 2019). The Crystallized Composite was also shown to be the only composite free of systematic bias across samples when applying the Bland-Altman method (Bland & Altman, 1995). These findings were not unexpected as measures of vocabulary and reading are often found to be less susceptible to cognitive changes and age in adulthood (Heaton et al., 2004).

Consistent with a previous finding using healthy adults ages 20 to 85 with a majority White sample that retested participants between 7 and 21 days, the NIHTB-CB individual subtests demonstrated moderate to excellent reliability for the total sample and healthy controls (Weintraub et al., 2013) but were less consistent for those with MCI. Healthy controls were the most reliable on crystallized measures and the least reliable on fluid skills – adding further support to previous findings describing “fluid” skills on the NIHTB-CB as more susceptible to practice effects in

healthy adults than “crystallized” skills (Heaton et al., 2014; Scott et al., 2019). Those with MCI demonstrated a similar pattern on composites. However, those with MCI showed greater variability than healthy controls when looking at the reliabilities of individual tests that make up the composites. For example, those with MCI demonstrated poor reliabilities and no significant benefit with retesting on a task of working memory and a test of episodic memory. While this finding is not consistent with our hypothesis that those with MCI would demonstrate moderate to excellent reliabilities, it is consistent with our hypothesis that those with MCI would be less susceptible to practice effects than healthy controls. The poorer reliabilities seen in those with MCI compared to healthy controls may be due to the heterogeneity of the sample. For example, our preliminary findings showed that those with aMCI were less reliable than those with naMCI on the working memory and episodic memory tasks in particular – suggesting that those with aMCI were driving the poor reliabilities found on these measures. Working memory deficits have been seen in both aMCI and naMCI when compared to healthy controls (Saunders & Summers, 2010; Klekociuk & Summers, 2014). However, differences between those with aMCI and naMCI on working memory tasks may depend on the type of working memory task and the level of impairment of the individual (Klekociuk & Summers, 2014). In a recent study exploring how well the NIHTB-CB and Cogstate differentiate those with aMCI from those with naMCI, working memory was not a significant predictor of disease type (Garcia et al., 2023). Thus, our finding that a task of working memory was notably less reliable for those with aMCI than for those with naMCI should be viewed with caution as it may be due to individual differences in our sample or the relatively low sample size of those with naMCI. The finding that those with aMCI were less reliable than both healthy controls and those with naMCI on an episodic memory task is unsurprising, as memory is one of the earliest cognitive domains negatively impacted by cognitive decline (Bastin & Salmon, 2014) and predominate memory dysfunction is the criteria that differentiates those with aMCI from those with naMCI (Peterson et al., 2018).

We found that the length of testing interval was not significantly associated with changes between testing timepoints for healthy controls on the NIHTB-CB up to 4 months. Though not significant, we did find that as length of test interval increased, the association with test performance decreased on all fluid measures across samples. For those with MCI, the association decreased on crystallized measures as well, though not significantly. The only significant finding was that, for those with MCI, the association between performance on a visual processing speed test significantly decreased as time between testing intervals increased. These findings are generally consistent with previous findings showing as length of time increases the difference between scores decreases (Calamia et al., 2012; Hausknecht et al., 2007; Salthouse et al., 2004; Scharfen et al., 2018).

Cogstate

We found that all subtests demonstrated moderate to good reliability in the total sample up to 4 months in a B/AA sample, which is generally consistent with prior research using predominantly White samples (Cole et al., 2013; Faletti et al., 2006; Fredrickson et al., 2010; Lim et al., 2013). Similar to prior research with a majority White sample aged 60–96 years that compared healthy older adults to those with MCI and Alzheimer’s dementia with retesting at 1, 2, and 3 months (Lim et al., 2013), we found that

healthy controls demonstrated lower retest reliability than those with MCI on all Cogstate subtests. Further, the finding that One Back Accuracy was the least reliable of the Cogstate measures for both healthy controls (Falleti et al., 2006; Lim et al., 2013) and those with MCI was replicated (Lim et al., 2013). As hypothesized, healthy controls and those with MCI demonstrated similar practice effect profiles upon retest. Though participants did not significantly improve on a simple reaction time task, a measure of choice reaction time did show significant improvement across samples. This finding is not surprising considering that reaction time is generally known to improve with practice and larger improvements are observed in more complex mental speed tasks compared to simple ones (Scharfen et al., 2018). Though caution should be used in interpretation due to a lack of power, preliminary findings demonstrated similar reliabilities on Cogstate subtests for both subtypes of MCI, apart from accuracy on a memory task. Retest reliability was poorer for those with aMCI on this memory task than for healthy controls and those with naMCI. This finding is consistent with the greater decline in memory expected in those with aMCI relative to both unimpaired individuals and those with naMCI (Bastin & Salmon, 2014; Peterson et al., 2018).

During the preliminary data analysis, One Back Accuracy was noted to have a negatively skewed distribution across samples with most performances in the highly accurate range – thereby demonstrating a ceiling effect and less opportunity for change. An ad hoc analysis was conducted evaluating One Back reaction time instead of Accuracy, and reaction time was found to have a more normal distribution and produced better reliabilities. Similarly, a previous study with cognitively healthy adults reported consistently better reliabilities for One Back reaction time versus One Back Accuracy (Falleti et al., 2006), and another study of cognitively healthy older adults found excellent reliability when using One Back reaction time instead of One Back Accuracy (Fredrickson et al., 2010). This suggests that One Back reaction time may be a more reliable measure when using Cogstate over multiple testing timepoints despite the manufacturer's recommendation to use One Back Accuracy.

The amount of time between testing timepoints did not appear to be significantly related to performance on Cogstate across samples up to 4 months. This finding is consistent with a study that did not find a significant change in performance for healthy controls, those with MCI, or those with AD across 1-, 2-, and 3-month retest intervals (Lim et al., 2013), and a study of healthy controls that did not find a significant difference in reaction time on the Detection subtest at a 10-minute retest interval or a 7-day retest interval (Falleti et al., 2006). Conversely, this finding was not consistent with a study of healthy controls that found a small improvement in group performance for accuracy of performance in the One Card Learning task from baseline to the 3-month retest that persisted on the 6-, 9-, and 12-month retests (Fredrickson et al., 2010), and a study of healthy controls that found significant improvement in One Back Accuracy at a 10-min retest interval and at a 7-day retest interval (Falleti et al., 2006).

Considerations for both measures

Though NIHTB-CB composite scores demonstrated strong retest reliability, Cogstate subtests appeared less susceptible on the whole to retest effects than NIHTB-CB subtests (particularly fluid measures). Interestingly, those with MCI demonstrated better retest reliability than healthy controls on Cogstate, while healthy controls demonstrated higher retest reliability on the NIHTB-CB composites

than those with MCI. This difference rests primarily on the lower retest reliability for those with MCI on a measure of working memory and a measure of episodic memory within the NIHTB-CB. With the exception of significant improvement on an episodic memory task for healthy controls and not for those with MCI with retest, practice effect profiles were similar on NIHTB-CB subtests. Cogstate reliabilities for healthy controls were lower across each subtest when compared to those with MCI, but the practice effect profiles were similar between healthy controls and those with MCI.

Reliable change methodology was used to assess whether a change in score after retest on a given variable is reliable and meaningful (Chelune et al., 1993; Iverson, 2001). Reliable change CIs (70, 80, 90%) were provided for both NIHTB-CB and Cogstate. The resulting values serve as cutoffs indicative of reliable change that are easily translatable into research and clinical practice (indicating improvement, decline, or stability; Chelune et al., 1993; Iverson, 2001). That is, a person whose score changes by the amount (or greater) provided in Table 5 for a given variable, would be exceeding the change in scores experienced by 85, 90, or 95% of the sample, respectively. For example, if a clinician or researcher used the 70% CI, a change score exceeding the cutoff would indicate a greater change than in 85% of the present sample. The greater the CI value the more conservative (or higher specificity). Practice effects can also be taken into account if desired (see Chelune et al., 1993).

Limitations and future directions

One limitation of this study is that we were not able to include older adults over the age of 85 due to a lack of norms in the field for this age group. Other studies such as the Advancing Reliable Measurement in Alzheimer's Disease and Cognitive Aging Study are addressing this problem by extending the NIHTB norms to those over the age of 85 (Weintraub et al., 2022). Due to the low incidence of naMCI observed in this sample, those with aMCI and naMCI were combined into one group (MCI). Although we did not have enough participants with a diagnosis of naMCI to reliably make recommendations regarding the reliability, we did find differences between those with aMCI and naMCI on NIHTB-CB and Cogstate. We also found that the NIHTB-CB was better able to differentiate the two subtypes of MCI than Cogstate, which is consistent with a recent publication (Garcia et al., 2023). Future studies should attempt to recruit larger numbers of both classifications of MCI to reliably examine potential differences. Healthy controls had approximately one year more of education on average in our study than those with MCI. This finding is consistent with prior literature that has found higher education to be associated with less cognitive impairment (Heaton et al., 2014; Meng & D'Arcy, 2012; Mungas et al., 2021). Additionally, our sample, as a whole, was highly educated (average greater than 14 years), and, thus, may not be generalizable to less educated individuals; however, it should be noted that scores used in the NIHTB-CB analyses were a priori norm-adjusted scores (age, sex, race/ethnicity, and education) provided by NIHTB-CB. When such norms are used appropriately, they offer greater diagnostic accuracy (Manly, 2005) and have been recently shown to reduce the association between education and cognitive performance in a racially diverse sample (Mungas et al., 2021). Though not unusual for aging research, there were significantly more females than males that participated in this study; future studies might attempt to recruit equal numbers of males and females to study possible sex effects more directly in the context of the reliability of these measures.

Conclusions

Despite findings that older B/AAs are disproportionately more likely than older Whites to have all type dementia (Dilworth-Anderson et al., 2008; Power et al., 2021; Steenland et al., 2016; Yaffe et al., 2013), previous studies examining the retest reliability of NIHTB-CB and Cogstate for laptop were conducted using mostly White samples (Cole et al., 2013; Faletti et al., 2006; Fredrickson et al., 2010; Hammers et al., 2011; Heaton et al., 2014; Lim et al., 2013; Scott et al., 2019; Weintraub et al., 2013). Thus, this study provided retest reliabilities and reliable change cutoffs for an all B/AA sample of healthy controls and those with MCI. Though retest reliabilities found in this study are similar to previous findings using mostly White samples, differences were also noted. Differences were also found between healthy controls and those with MCI. It is, therefore, recommended that race and cognitive status be considered when using these measures. Overall, it was found that the NIHTB-CB and Cogstate for laptop both show promise for use in research with B/AA and were reasonably stable for up to 4 months.

Although differences in reliability were noted between the two measures, the choice between test measures should reflect multiple considerations and the needs of a particular study. For example, while Cogstate takes considerably less time to administer, it is heavily reliant on reaction time, and it does not offer the breadth of cognitive measurement as NIHTB-CB. NIHTB-CB also provides norm-adjusted scores for age, sex, race/ethnicity, and education, whereas Cogstate does not. Preliminary findings also demonstrated greater differentiation between MCI subtypes (aMCI and naMCI) when using the NIHTB-CB versus Cogstate.

Acknowledgments. Specific thanks are given to the reviewers for this article. Their input and recommendations helped to strengthen this paper.

Funding statement. This work was supported by the US National Institute on Aging (P30 AG053760, P30 AG072931) and the US National Institute of Health (V.K., R01 AG054484, R21 AG046637).

Competing interests. None.

References

- Alzheimer's Association. (2022). More than normal aging: Understanding mild cognitive impairment. Alzheimer's disease facts and figures. (2022). Mild Cognitive Impairment [Special Report on MCI] Retrieved December 12, 2024, from <https://www.alz.org/media/Documents/alzheimers-facts-and-figures-special-report-2022.pdf>. Alzheimer's Association. https://www.alz.org/alzheimers-dementia/what-is-dementia/related_conditions/mild-cognitive-impairment
- Arevalo-Rodriguez, I., Smailagic, N., Roqué-Figuels, M., Ciapponi, A., Sanchez-Perez, E., Giannakou, A., & Cullum, S. (2021). Mini-mental state examination (MMSE) for the early detection of dementia in people with mild cognitive impairment (MCI). *Cochrane Database of Systematic Reviews*, 7, CD010783. <https://doi.org/10.1002/14651858.CD010783.pub3>
- Bastin, C., & Salmon, E. (2014). Early neuropsychological detection of Alzheimer's disease. *European journal of clinical nutrition*, 68, 1192–1199.
- Bland, J. M., & Altman, D. G. (1995). Comparing methods of measurement: Why plotting difference against standard method is misleading. *The lancet*, 346, 1085–1087.
- Calamia, M., Markon, K., & Tranel, D. (2012). Scoring higher the second time around: Meta-analyses of practice effects in neuropsychological assessment. *The Clinical Neuropsychologist*, 26, 543–570.
- Chelune, G. J., Naugle, R. I., Luders, H., Sedlak, J., & Awad, I. A. (1993). Individual change after epilepsy surgery: Practice effects and base-rate information. *Neuropsychology*, 7, 41–52.
- Clark, P. C., Kutner, N. G., Goldstein, F. C., Peterson-Hazen, S., Garner, V., Zhang, R., & Bowles, T. (2005). Impediments to timely diagnosis of Alzheimer's disease in African Americans. *Journal of the American Geriatrics Society*, 53, 2012–2017.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Routledge Academic.
- Cole, W. R., Arrioux, J. P., Schwab, K., Ivins, B. J., Qashu, F. M., & Lewis, S. C. (2013). Test-retest reliability of four computerized neurocognitive assessment tools in an active duty military population. *Archives of clinical neuropsychology*, 28, 732–742.
- Darby, D., Maruff, P., Collie, A., & McStephen, M. (2002). Mild cognitive impairment can be detected by multiple assessments in a single day. *Neurology*, 59, 1042–1046.
- Diaz-Orueta, U., Blanco-Campal, A., Lamar, M., Libon, D. J., & Burke, T. (2020). Marrying past and present neuropsychology: Is the future of the process-based approach technology-based? *Frontiers in psychology*, 11, 361.
- Dilworth-Anderson, P., Hendrie, H. C., Manly, J. J., Khachaturian, A. S., & Fazio, S. (2008). Diagnosis and assessment of Alzheimer's disease in diverse populations. *Alzheimer's & Dementia*, 4, 305–309.
- Falletti, M. G., Maruff, P., Collie, A., & Darby, D. G. (2006). Practice effects associated with the repeated assessment of cognitive function using the CogState battery at 10-minute, one week and one month test-retest intervals. *Journal of clinical and experimental neuropsychology*, 28, 1095–1112.
- Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research*, 12, 189–198.
- Fredrickson, J., Maruff, P., Woodward, M., Moore, L., Fredrickson, A., Sach, J., & Darby, D. (2010). Evaluation of the usability of a brief computerized cognitive screening test in older people for epidemiological studies. *Neuroepidemiology*, 34, 65–75.
- Garcia, S., Askew, R. L., Kavcic, V., Shair, S., Bhaumik, A. K., Rose, E., & Giordani, B. (2023). Mild cognitive impairment subtype performance in comparison to healthy older controls on the NIH toolbox and cogstate. *Alzheimer Disease & Associated Disorders*, 37, 328–334.
- Gauthier, S., Reisberg, B., Zaudig, M., Petersen, R. C., Ritchie, K., Broich, K., & Winblad, B. (2006). Mild cognitive impairment. *The lancet*, 367, 1262–1270.
- 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. <https://doi.org/10.1212/WNL.0b013e3182872e5f>
- Gianattasio, K. Z., Prather, C., Glymour, M. M., Ciarleglio, A., & Power, M. C. (2019). Racial disparities and temporal trends in dementia misdiagnosis risk in the United States. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 5, 891–898.
- Hammers, D., Spurgeon, E., Ryan, K., Persad, C., Heidebrink, J., Barbas, N., & Giordani, B. (2011). Reliability of repeated cognitive assessment of dementia using a brief computerized battery. *American Journal of Alzheimer's Disease & Other Dementias*, 26, 326–333.
- Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: a meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology*, 92(2), 373. <https://doi.org/10.1037/0021-9010.92.2.373>.
- Heaton, R. K. (2004). *Revised comprehensive norms for an expanded halstead-reitan battery: Demographically adjusted neuropsychological norms for African American and Caucasian adults, professional manual*. Psychological Assessment Resources.
- Heaton, R. K., Akshoomoff, N., Tulsky, D., Mungas, D., Weintraub, S., Dikmen, S., & Gershon, R. (2014). Reliability and validity of composite scores from the NIH toolbox cognition battery in adults. *Journal of the International Neuropsychological Society*, 20, 588–598.
- Iverson, G. L. (2001). Interpreting change on the WAIS-III/WMS-III in clinical samples. *Archives of Clinical Neuropsychology*, 16, 183–191.
- Klekociuk, S. Z., & Summers, M. J. (2014). Lowered performance in working memory and attentional sub-processes are most prominent in multi-domain amnesic mild cognitive impairment subtypes. *Psychogeriatrics*, 14, 63–71.
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15, 155–163.

- Lim, Y. Y., Jaeger, J., Harrington, K., Ashwood, T., Ellis, K. A., Stöfler, A., & Maruff, P. (2013). Three-month stability of the CogState brief battery in healthy older adults, mild cognitive impairment, and Alzheimer's disease: Results from the Australian imaging, biomarkers, and lifestyle-rate of change substudy (AIBL-ROCS). *Archives of clinical neuropsychology*, *28*, 320–330.
- Lin, P. J., Daly, A. T., Olchanski, N., Cohen, J. T., Neumann, P. J., Faul, J. D., & Freund, K. M. (2021). Dementia diagnosis disparities by race and ethnicity. *Medical care*, *59*, 679–686.
- Manly, J. J. (2005). Advantages and disadvantages of separate norms for African Americans. *The Clinical Neuropsychologist*, *19*(2), 270–275. <https://doi.org/10.1080/13854040590945346>
- Maruff, P., Lim, Y. Y., Darby, D., Ellis, K. A., Pietrzak, R. H., Snyder, P. J., & Masters, C. L. (2013). Clinical utility of the cogstate brief battery in identifying cognitive impairment in mild cognitive impairment and Alzheimer 19's disease. *BMC psychology*, *1*, 1–11.
- McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological methods*, *1*, 30–46.
- Meng, X., & D'arcy, C. (2012). Education and dementia in the context of the cognitive reserve hypothesis: A systematic review with meta-analyses and qualitative analyses. *PLoS one*, *7*(6), e38268. <https://doi.org/10.1371/journal.pone.0038268>
- Mungas, D., Shaw, C., Hayes-Larson, E., DeCarli, C., Farias, S. T., Olichney, J., & Mayeda, E. R. (2021). Cognitive impairment in racially/ethnically diverse older adults: Accounting for sources of diagnostic bias. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *13*, e12265.
- Pandya, S. Y., Clem, M. A., Silva, L. M., & Woon, F. L. (2016). Does mild cognitive impairment always lead to dementia? A review. *Journal of the neurological sciences*, *369*, 57–62.
- Petersen, R. C., Lopez, O., Armstrong, M. J., Getchius, T. S., Ganguli, M., Gloss, D., & Rae-Grant, A. (2018). Practice guideline update summary: Mild cognitive impairment: Report of the guideline development, dissemination, and implementation subcommittee of the American Academy of Neurology. *Neurology*, *90*, 126–135.
- Portney, L. G., & Watkins, M. P. (2009). *Foundations of clinical research: applications to practice*, vol. 892, 11–15: Pearson/Prentice Hall.
- Power, M. C., Bennett, E. E., Turner, R. W., Dowling, N. M., Ciarleglio, A., Glymour, M. M., & Gianattasio, K. Z. (2021). Trends in relative incidence and prevalence of dementia across non-hispanic black and white individuals in the United States, 2000–2016. *JAMA neurology*, *78*, 275–284.
- Ranson, J. M., Kuźma, E., Hamilton, W., Muniz-Terrera, G., Langa, K. M., & Llewellyn, D. J. (2019). Predictors of dementia misclassification when using brief cognitive assessments. *Neurology: Clinical Practice*, *9*, 109–117.
- Salthouse, T. A., Schroeder, D. H., & Ferrer, E. (2004). Estimating retest effects in longitudinal assessments of cognitive functioning in adults between 18 and 60 years of age. *Developmental Psychology*, *40*, 813–822. <https://doi.org/10.1037/0012-1649.40.5.813>
- Saunders, N. L., & Summers, M. J. (2010). Attention and working memory deficits in mild cognitive impairment. *Journal of Clinical and Experimental Neuropsychology*, *32*, 350–357.
- Scharfen, J., Blum, D., & Holling, H. (2018). Response time reduction due to retesting in mental speed tests: A meta-analysis. *Journal of Intelligence*, *6*, 6.
- Scott, E. P., Sorrell, A., & Benitez, A. (2019). Psychometric properties of the NIH toolbox cognition battery in healthy older adults: Reliability, validity, and agreement with standard neuropsychological tests. *Journal of the International Neuropsychological Society*, *25*, 857–867.
- Spering, C. C., Hobson, V., Lucas, J. A., Menon, C. V., Hall, J. R., & O'Bryant, S. E. (2012). Diagnostic accuracy of the MMSE in detecting probable and possible alzheimer's disease in ethnically diverse highly educated individuals: An analysis of the NACC database. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences*, *67*, 890–896.
- Steenland, K., Goldstein, F. C., Levey, A., & Wharton, W. (2016). A meta-analysis of alzheimer's disease incidence and prevalence comparing African-americans and caucasians. *Journal of Alzheimer's Disease*, *50*, 71–76.
- Tombaugh, T. N., & McIntyre, N. J. (1992). The mini-mental state examination: A comprehensive review. *Journal of the American Geriatrics Society*, *40*, 922–935.
- Tsoi, K. K., Chan, J. Y., Hirai, H. W., Wong, S. Y., & Kwok, T. C. (2015). Cognitive tests to detect dementia: A systematic review and meta-analysis. *JAMA internal medicine*, *175*, 1450–1458.
- Ward, A., Tardiff, S., Dye, C., & Arrighi, H. M. (2013). Rate of conversion from prodromal Alzheimer's disease to alzheimer's dementia: A systematic review of the literature. *Dementia and geriatric cognitive disorders extra*, *3*, 320–332.
- Weintraub, S., Dikmen, S. S., Heaton, R. K., Tulsky, D. S., Zelazo, P. D., Bauer, P. J., & Gershon, R. C. (2013). Cognition assessment using the NIH toolbox. *Neurology*, *80*, S54–S64.
- Weintraub, S., Karpouzian-Rogers, T., Peipert, J. D., Nowinski, C., Slotkin, J., Wortman, K., Ho, E., Rogalski, E., Carlsson, C., Giordani, B., Goldstein, F., Lucas, J., Manly, J. J., Rentz, D., Salmon, D., Snitz, B., Dodge, H. H., Riley, M., Eldes, F., ... Gershon, R. (2022). ARMADA: Assessing reliable measurement in Alzheimer's disease and cognitive aging project methods. *Alzheimer's & Dementia*, *18*(8), 1449–1460. <https://doi.org/10.1002/alz.12497>
- Weintraub, S., Salmon, D., Mercaldo, N., Ferris, S., Graff-Radford, N. R., Chui, H., & Morris, J. C. (2009). The Alzheimer's disease centers' uniform data set (UDS): The neuropsychological test battery. *Alzheimer disease and associated disorders*, *23*, 91–101.
- Wild, K. V., Mattek, N. C., Maxwell, S. A., Dodge, H. H., Jimison, H. B., & Kaye, J. A. (2012). Computer-related self-efficacy and anxiety in older adults with and without mild cognitive impairment. *Alzheimer's & Dementia*, *8*, 544–552.
- Yaffe, K., Falvey, C., Harris, T. B., Newman, A., Satterfield, S., Koster, A., & Simonsick, E. (2013). Effect of socioeconomic disparities on incidence of dementia among biracial older adults: Prospective study. *Bmj*, *347*, f7051–f7051.