Published in final edited form as:
J Int Neuropsychol Soc. 2019 September ; 25(8): 857–867. doi:10.1017/S1355617719000614.

Psychometric Properties of the NIH Toolbox Cognition Battery in Healthy Older Adults: Reliability, Validity, and Agreement with Standard Neuropsychological Tests

Emmi P. Scott[1], **Anne Sorrell**[2], **Andreana Benitez**[1]

[1]Department of Neurology, Medical University of South Carolina, Charleston, South Carolina

[2]Appalachian State University, Boone, North Carolina

Abstract

Objective: Few independent studies have examined the psychometric properties of the NIH Toolbox Cognition Battery (NIHTB-CB) in older adults, despite growing interest its use for clinical purposes. In this paper we report the test-retest reliability and construct validity of the NIHTB-CB, as well as its agreement or concordance with traditional neuropsychological tests of the same construct to determine whether tests could be used interchangeably.

Methods: Sixty-one cognitively healthy adults ages 60-80 completed "gold standard" (GS) neuropsychological tests, NIHTB-CB, and brain MRI. Test-retest reliability, convergent/discriminant validity and agreement statistics were calculated using Pearson's correlations, concordance correlation coefficients (CCC), and root mean square deviations.

Results: Test-retest reliability was acceptable (CCC = .73 Fluid; CCC = .85 Crystallized). The NIHTB-CB Fluid Composite correlated significantly with cerebral volumes (r's = |.35-.41|), and both composites correlated highly with their respective GS composites (r's = .58-.84), although this was more variable for individual tests. Absolute agreement was generally lower (CCC = .55 Fluid; CCC = .70 Crystallized) due to lower precision in Fluid scores and systematic overestimation of Crystallized Composite scores on the NIHTB-CB.

Conclusions: These results support the reliability and validity of the NIHTB-CB in healthy older adults and suggest that the Fluid Composite tests are at least as sensitive as standard neuropsychological tests to medial temporal atrophy and ventricular expansion. However, the NIHTB-CB may generate different estimates of performance and should not be treated as interchangeable with established neuropsychological tests.

Keywords

psychometrics; neuropsychological tests; assessment; Magnetic Resonance Imaging; comparative study; aging

Correspondence concerning this article should be addressed to Emmi P. Scott, Department of Neurology, Medical University of South Carolina, 96 Jonathan Lucas Street MSC 323, Charleston, SC 29425. Contact: scottemm@musc.edu.

Introduction

The NIH Toolbox (NIHTB) for Assessment of Neurological and Behavioral Function was commissioned as part of the NIH Blueprint for Neuroscience Research initiative to provide a common metric among outcome measures in clinical research in the neurosciences (Gershon et al., 2013). The NIHTB Cognition Battery (NIHTB-CB) is a set of computer-based tests of cognitive functions selected based on their relevance to daily functioning and important health outcomes across the lifespan. The computer-based format offers brief yet comprehensive assessment of multiple domains that are typically included in a neuropsychological test battery, relative ease of administration and scoring, testing in English or Spanish, normative data for individuals ages three to 85 years, and increased sensitivity to the full range of functioning and minimization of ceiling and floor effects (Weintraub et al., 2013).

Performance on the individual tests of the NIHTB-CB are summarized into composite scores representing fluid and crystallized cognitive abilities. The Fluid Cognition Composite includes tests of cognitive abilities that are sensitive to aging and neurologic disease (e.g., processing speed, memory, executive functioning), while the Crystallized Cognition Composite is composed of tests that are similar to existing measures of premorbid intellectual functioning (e.g., oral reading recognition), where performance is expected to remain relatively stable across the lifespan (Mungas et al., 2014). Although the NIHTB-CB was not developed as a clinical or diagnostic tool, the co-normed Crystallized composite may provide an individualized estimate of premorbid ability to which one's Fluid abilities may be compared, aiding in the clinical determination of cognitive decline in the presence of neurodegenerative disease or acquired brain injury (Holdnack et al., 2017).

The advantages that the NIHTB-CB offers over traditional neuropsychological test batteries has led to increased interest in its clinical utility and validity in a number of clinical populations including stroke, traumatic brain injury, and spinal cord injury (e.g., Carlozzi et al., 2017; Holdnack et al., 2017; Tulsky et al., 2017; see a special issue of *Rehabilitation Psychology*, 2017, volume 62, issue 4). These studies supported the construct validity of the NIHTB-CB in rehabilitation settings and demonstrated its ability to discriminate acquired brain injury from controls, concluding that the NIHTB-CB may be useful clinically in a screening capacity. Moderate associations with neuropsychological tests of preclinical Alzheimer's disease (Buckley et al., 2017) and hippocampal volumes (O'Shea, Cohen, Porges, Nissim, & Woods, 2016) have been reported and interpreted as preliminary support for its clinical utility with older adults. In a recent investigation of the NIHTB-CB in a memory clinic setting, Hackett and colleagues (2018) found that the NIHTB-CB could differentiate cognitively normal adults from those with Alzheimer's disease dementia, but not from individuals with mild cognitive impairment. Although the authors acknowledge that the NIHTB-CB was not designed to replace comprehensive neuropsychological evaluations, they concluded that it may represent a "key part of the clinical diagnostic evaluation" in memory clinics. Additional studies have established the utility of the NIHTB-CB in neurosurgical evaluations for deep brain stimulation (Loring et al., 2018), cancer treatment (Sinha, Wong, Kallogjeri, & Piccirillo, 2018), and brain tumor resection (Lang et al., 2017), the latter study concluding that the NIHTB-CB represents a "feasible alternative to current neuropsychological batteries," citing length, cost, and variability between test batteries as barriers to traditional neuropsychological evaluations.

Although these studies demonstrate feasibility and criterion validity, more research is needed prior to assuming the equivalence of NIHTB-CB to established measures, and therefore, its clinical utility. That is, if the NIHTB-CB is to be used in clinical practice alongside or interchangeably with traditional neuropsychological tests, it is necessary to examine how closely these measures agree before their scores can be interpreted in the same way. Most validation studies have relied on Pearson's correlations, which are often misinterpreted as indices of agreement or interchangeability, as they only measure the linear relationship between two variables ($Y = aX + b$). By this metric, two measures could be perfectly correlated but not yield a single pair of scores that are equivalent due to systematic differences by either an additive (b) or multiplicative (a) amount.

In contrast, the interchangeability of measures is best assessed through agreement statistics. Agreement can be calculated using a "consistency" definition that accounts for differences in variances but not means (i.e., additivity [$Y = X + a$]), or an "absolute" definition, which takes into account differences in means but also differences in variances and degree of linearity. Absolute agreement describes the degree that two measures on the same scale produce identical scores ($X = Y$). While absolute agreement is indeed a high standard to attain for two measures obtained through different methods (i.e. computerized versus pen-and-paper), within clinical settings the resultant standard scores derived from either method will nonetheless be interpreted similarly. Analyzing absolute discrepancies between individual pairs of scores therefore offers more clinically relevant information than comparing average performances as a whole, including identifying whether one method yields systematically higher or lower scores. No studies to date have examined the degree of absolute agreement between the NIHTB-CB and analogous neuropsychological tests.

Thus, the purpose of this study was twofold. First, we sought to investigate the psychometric properties of the NIHTB-CB in a sample of cognitively intact older adults. Specifically, we examined test-retest reliability and construct validity. In addition to evaluating convergent and discriminant validity using neuropsychological tests, we also used measures that do not share common-method variance such as demographic (i.e. age, education) and imaging variables (i.e., cerebral volumes) that are pertinent to aging and dementia. With the advent of commercially available metrics of overall and focal cerebral atrophy (Azab, Carone, Ying, & Yousem, 2015), such multi-method validity studies are increasingly possible without the need for neuroimaging analysis expertise. We focused on the NIHTB-CB Fluid and Crystallized Cognition Composites and composites derived from "gold standard" (GS) neuropsychological tests, but also report results from individual tests for comprehensiveness. Second, we examined absolute agreement between age-adjusted scores on the NIHTB-CB and GS tests using methods that are novel to the NIHTB-CB literature. We aimed to evaluate the degree to which the NIHTB-CB and GS tests produce equivalent scores to determine whether their resultant scores can be interpreted similarly or used interchangeably, which has direct implications for the potential clinical use of the NIHTB-CB.

Methods

Participants and Procedures

This sample included older adults ages 60-80 years who were recruited via community advertisements as part of a larger study using advanced MRI for the early detection neurodegenerative disorders. All participants were screened by phone for eligibility criteria, which included speaking English as a primary language, absence of history of significant neurologic disease, serious mental illness, or self-reported cognitive complaints beyond expectations for age. Of the 65 participants enrolled between 2013 and 2015, two were excluded due to meeting actuarial neuropsychological criteria for mild cognitive impairment (Bondi et al., 2014) and two were excluded due to missing NIHTB-CB data. As funding for this study was extended, participants were invited to return to undergo identical procedures to address research questions regarding longitudinal changes in MRI measures. Each participant at follow-up was re-screened to ensure there had been no major changes in medical conditions or incident cognitive decline. Of the remaining 61 participants, 37 returned for follow-up after 15.03 ± 3.11 months. None of these participants met actuarial neuropsychological criteria for MCI.

Participants completed study procedures on the same day, which included the neuropsychological battery of the National Alzheimer's Coordinating Centers' Uniform Data Set (Shirk et al., 2011; Weintraub et al., 2009), followed by the NIHTB-CB, and brain MRI. A licensed clinical neuropsychologist or a trained research assistant under supervision administered all tests and oversaw all scans. Participants were not allowed to take sedating or anxiolytic medications for MRI. All scans were reviewed by the PI, research staff, and MRI technologists, and any incidental findings were sent to a neuroradiologist for review. None had incidental findings that required clinical follow-up. This study was approved by the Institutional Review Board.

Measures

NIH Toolbox Cognition Battery.—Participants completed all seven core measures of the NIHTB-CB on a computer that conformed to hardware and software specifications per NIHTB guidelines. Five measures of attention/inhibitory control, executive functioning (i.e., set shifting) working memory, processing speed, and episodic memory comprise the Fluid Cognition Composite, and two tests of language (i.e., vocabulary comprehension and oral reading decoding) make up the Crystallized Cognition Composite (Gershon et al., 2014). Scores from measures of inhibitory control and set shifting are derived from both accuracy and response speed, while the remaining three Fluid test scores reflect total correct responses. The two language tests are administered using computerized adaptive testing and scored using item response theory. The NIHTB-CB generates three types of scores for each subtest and composite: uncorrected standard scores, age-corrected standard scores, and fully-adjusted scores that account for age, education, gender, and race/ethnicity. Only age-adjusted and unadjusted standard scores, which both have a mean of 100 and standard deviation of 15, were used in the current analyses in the interest of having consistent comparisons with criterion variables. Initial studies indicated that these composites have strong test-retest reliability (i.e., ICCs ranging from .78 to .99) and moderate to strong convergent validity (*r*'s ranging from .48 to .93) with traditional neuropsychological tests (Heaton et al., 2014; Weintraub et al., 2013).

"Gold Standard" neuropsychological test battery.—Table 1 shows each NIHTB-CB measure matched with its corresponding GS measure of the same underlying construct. GS measures largely correspond to those used in the initial NIHTB-CB validation studies (Heaton et al., 2014; Weintraub et al., 2013). We analyzed age-adjusted scores to facilitate comparison with NIHTB-CB scores, as fully-adjusted normative data was not available for all GS tests. Age-adjusted scaled scores and T-scores were derived from the appropriate normative database or test manual (Ivnik, Malec, Smith, Tangalos, & Petersen, 1996; Lucas et al., 2005; Schmidt et al., 1996; Wechsler, 1997) and then converted to standard scores (M=100, SD=15) to be on the same scale as NIHTB-CB scores. The GS Fluid composite score was calculated by averaging the standard scores on five well-established tests of processing speed, working memory, attention, executive functioning, and memory: Digit Symbol Substitution subtest from the Wechsler Adult Intelligence Scale – Third Edition (WAIS-III; Wechsler, 1997), Color-Word trial from the Stroop Color and Word Test (Golden, 1978), Digit Span subtest from the Wechsler Memory Scale-Revised (WMS-R; Wechsler, 1987), Trail Making Test – Part B (Reitan & Wolfson, 1985), and total learning score from the Rey Auditory Verbal Learning Test (RAVLT; Rey, 1964). The American National Adult Reading Test (AmNART), an oral reading test commonly to estimate premorbid intellectual functioning, served as the validation measure for both NIHTB-CB Crystallized Composite subtests. The GS battery did not include a measure of receptive vocabulary to serve as a criterion measure for the NIHTB-CB Picture Vocabulary Test, so correlations provided here are exploratory.

Neuroimaging.—Participants underwent brain MRI using a 3 Tesla Siemens TIM Trio system, including 3D T1-weighted MPRAGE images that were analyzed using the automated segmentation program NeuroQuant 1.0 (Cortechs, San Diego, CA). Combined bilateral cerebral volumes, in cm^3, normalized to percent intracranial volume, for the hippocampi, lateral ventricles, and inferior lateral ventricles were provided in NeuroQuant's Age-related Atrophy Report. These regions are associated with aging, neurodegenerative disease, and cognitive decline (Nestor et al., 2008; Raz et al., 2005). We then computed the hippocampal occupancy score (HOC; hippocampal volume / hippocampal volume + inferior lateral ventricle volume) as an index of medial temporal lobe atrophy, which has been shown to be superior to standard hippocampal volume estimates in predicting conversion from mild cognitive impairment to Alzheimer's disease dementia (Heister, Brewer, Magda, Blennow, & McEvoy, 2011).

Data Analyses

Test-Retest Reliability.—Test-retest reliability and stability were evaluated using Pearson's correlations and concordance correlation coefficients (CCC; Lin, 1989) with 95% confidence intervals, using an absolute definition of agreement. Practice effects were evaluated with paired *t*-tests and effect sizes, computed as scores at follow-up (time 2) minus scores at baseline (time 1), divided by the standard deviation of scores at time 1 (Cohen, 1992).

Construct validity.—First, the relationships between unadjusted NIHTB-CB scores with relevant demographic variables (i.e., age and education) and MRI metrics of cerebral volumes were investigated with Pearson's correlations. In testing convergent validity, we expected Fluid Cognition Composite scores to correlate with age and cerebral volumes, and Crystallized Cognition Composite scores to correlate with education. Lower correlations between Crystallized scores and cerebral volumes, and between Fluid scores and educational attainment, were hypothesized to indicate discriminant validity. These analyses were then repeated with GS scores for comparison. Unadjusted standard scores (M=100, SD=15) were used for these analyses, as both neuropsychological and MRI variables are similarly influenced by age.

Second, the relationships between age-adjusted standard scores on NIHTB-CB and GS measures were investigated with Pearson's correlations to test their convergent and discriminant validity. In line with initial NIHTB-CB validation papers (e.g., Weintraub et al., 2014), correlations with criterion measures below .4 were considered poor, .4 to .6 were adequate, and .6 or greater were good evidence of convergent validity. Evidence of discriminant validity consisted of significantly lower correlations between dissimilar measures (e.g., NIHTB-CB Fluid with GS Crystallized), which was confirmed with a Steiger's z (1980) statistic.

Agreement.—We used several methods to evaluate agreement between age-adjusted scores on the NIHTB-CB and corresponding GS measures. First, we calculated CCC's with 95% confidence intervals as indices of absolute agreement (i.e., $X = Y$) to characterize any systematic differences in scores that otherwise would not be captured using a consistency or linear definition. Like Pearson's r, the CCC can range from −1 to +1, but it cannot exceed r, as perfect agreement is only reached when scores on the two measures are equal for each person. The CCC is preferred over Pearson's r for studies that aim to determine whether scores on two different measures agree sufficiently to be used interchangeably. Barnhart et al. (2007) and Carrasco and Jover (2003) demonstrated that the CCC is identical to the intraclass correlation coefficient (ICC; McGraw & Wong, 1996) when assumptions for the latter are met; however, the CCC is more appropriate when comparing two measures that may have different variances. CCCs may be interpreted using the benchmarks proposed by Altman (1991): poor agreement (.20), weak (.21–.40), moderate (.41–.60), good (.61–.80), and very good agreement (.81–1.00).

Root mean squared differences (RMSD) were calculated to measure the average discrepancy between individual pairs of scores (Barchard, 2012). Using an absolute definition of agreement, RMSD values represent the square root of the average squared difference score and are interpreted in the metric of the original scores. Thus, an RMSD of 7.5 would indicate that the average discrepancy is 7.5 standard score points (i.e., 0.5 SD). This value was established as the maximum acceptable disagreement, as larger discrepancies may represent a clinically meaningful difference. CCC and RMSD values were calculated using the Excel spreadsheet provided in the supplementary materials by Barchard (2012).

We also provide base rates for large discrepancies (> 0.5 and 1 SD) and Bland-Altman plots of the NIHTB-CB and GS composite scores, using NCSS version 12 statistical software

(NCSS, Kaysville, UT, USA). Bland-Altman (1983) plots visually depict the differences between measurements by plotting the average of paired scores from each method on the x-axis against the difference of each pair of scores on the y-axis. This illustrates the presence of any constant differences (i.e., differences that are consistent across the entire range of scores), proportional differences (i.e., non-constant differences that are proportional to the obtained score), and degree of imprecision (i.e., random error). Precision can be estimated with the 95% limits of agreement, or limits between which 95% of observations in the population would be expected to lie (i.e., ±1.96 SD of the mean difference).

Results

Characteristics of the sample are shown in Table 2. The sample was predominantly Caucasian (90%; 10% were African American) and had an average of 16 years of education. There were no significant differences on any demographic or cognitive variables between the total sample and the subsample that returned for follow-up (data not shown). MRI data were missing on five participants due to MRI safety concerns or claustrophobia ($n = 4$) and significant motion artifact ($n = 1$). All variables were examined and found to meet the statistical assumptions of the parametric analyses reported here. Descriptive statistics and histograms for age-adjusted NIHTB-CB and GS test scores are provided in Supplementary Table 1 and Figure 1 to facilitate comparisons of the distribution of scores on each battery.

Test-retest reliability

NIHTB-CB Crystallized and Fluid Cognition Composites demonstrated strong test-retest reliability (CCC's = .92 and .73, respectively) over the follow-up interval (Table 3). The reliability of individual measures varied from .46 for List Sorting Working Memory to .87 for Picture Vocabulary. Dimensional Change Card Sort was the only test showing a small but statistically significant practice effect (+3.22 points, ES = 0.33). In comparison, reliability estimates for GS measures were similar or slightly higher for the GS Fluid Cognition Composite (CCC = .85); however, this composite and four of its individual measures showed consistent evidence of practice effects (+3.88 points, ES = 0.38).

Construct Validity

Demographic and Neuroimaging Correlates.—Table 4 shows correlations between the cerebral volumes of interest with NIHTB-CB and GS tests. As expected, unadjusted standard scores on the NIHTB-CB Fluid Cognition Composite were associated with medial temporal lobe volume (HOC $r = .41$, $p = .002$), lateral ventricular volume ($r = -.35$, $p = .008$), and age ($r = -.51$, $p < .001$), but not years of education ($r = .07$, $p = .617$). Only a subset of the individual NIHTB-CB measures (i.e., Dimensional Card Sort, Picture Sequencing Memory, and Pattern Comparison Processing Speed) correlated significantly with cerebral volumes. Conversely, Crystallized Cognition Composite scores significantly correlated with years of education ($r = .42$, $p = .001$), but not age ($r = -.14$, $p = .296$) or cerebral volumes (p's = .840 and .084 for lateral ventricular and medial temporal volumes, respectively), providing evidence of discriminant validity.

In comparison, correlations between GS Fluid Cognition Composite and cerebral volumes were marginally significant and slightly lower in magnitude (see Table 4), although differences between these correlations were not statistically significant for medial temporal (Steiger's $z = 1.09$, $p = .277$) or lateral ventricular volumes ($z = 1.11$, $p = .268$). Of the individual GS tests, Digit Symbol, Stroop Color-Word, and RAVLT demonstrated significant correlations of comparable magnitude to NIHTB-CB tests.

Correlations with GS Measures.—Adequate to strong correlations were found between the two Fluid composites ($r = .58$) and Crystallized composites ($r = .84$), as shown in Table 5 and in the scatterplots on the top row of Figure 1A. Convergent validity coefficients for individual measures were much more variable, ranging from poor ($r = .10$) to very strong ($r = .85$). Correlations between tests of executive functioning were lowest and were not statistically significant. Evidence of discriminant validity was supported by significantly lower correlations between NIHTB-CB Fluid and GS Crystallized Cognition Composite scores ($r = .32$, $p = .018$; Steiger's $z = -2.41$, $p = .027$), and between NIHTB-CB Crystallized and GS Fluid Cognition Composite scores ($r = .43$, $p = .001$; Steiger's $z = -4.71$, $p < .001$). Intercorrelations for all individual NIHTB-CB and GS tests are provided in Supplementary Table 1.

Agreement

The NIHTB-CB and GS Crystallized Cognition Composites demonstrated good agreement (CCC = .70), although the average pairwise discrepancy approached one standard deviation (RMSD = 12.15; see Table 5), indicating that individual pairs of scores differed by 12.15 points on average. The Bland-Altman plot in Figure 1B (left panel) shows that the NIHTB-CB (M = 124.33, SD = 14.98) consistently generates higher Crystallized scores than the GS (M = 115.49, SD = 14.19), mean difference = 8.84 (SD =8.41, 95% CI [6.68, 10.99]). A one sample *t*-test confirmed that this differs significantly from zero ($t = 8.21$, $df = 60$, $p < .001$). As shown in the frequency distribution of the differences (Figure 1C, left panel), only 34% of the sample obtained acceptable discrepancy scores within +−7.5 points, while 61% obtained NIHTB-CB scores that were over 7.5 points higher than their GS score. Twenty-eight percent obtained scores that differed by over 15 points (all NIHTB-CB > GS). As described in the Supplementary Material, exploratory analyses indicated that no particular participant subgroup was driving this discrepancy, although the sample sizes are too low to reliably detect demographic differences in this homoegenous sample. Furthermore, while the two oral reading tests (i.e. NIHTB-CB Oral Reading Recognition Test and AmNART) yielded comparable scores, the NIHTB-CB Picture Vocabulary Test yielded scores that were 8-9 standard score points higher than the two reading tests.

The NIHTB-CB and GS Fluid Cognition composites demonstrated moderate agreement (CCC = .55). The proximity of the CCC to Pearson's correlation ($r = .58$) reflects the fact that the two methods have nearly equivalent means, NIHTB-CB = 109.93 (SD = 14.02) and GS = 108.93 (SD = 10.38), mean difference = 1.00 (SD = 11.65, 95% CI [−1.98, 3.99], $t = 0.67$, $df = 60$, $p = .504$). This is confirmed by the Bland-Altman plot and frequency distribution (Figure 1B and C, right panel) showing an approximately equal number of cases above and below the mean difference. However, individual pairs of scores differed by an average of 11.60 points. Fifty-one percent of the sample obtained discrepancy scores greater than 7.5 points, and 23% obtained scores that differed by 15 or more points. The limits of agreement show that 95% of scores on the two Fluid composites are expected to differ between −21.83 and +23.84 points, indicating poor precision but no systematic bias (i.e., over or underestimation). Additionally, the scatter in the Bland-Altman plot appears to trend upward at higher values. As described in the Supplementary Material (Supplementary Figure 2), we explored the presence of proportional bias using non-parametric regression analysis (Passing & Bablok, 1983), which confirmed that the NIHTB-CB overestimates performance for individuals with high scores and underestimates it among individuals with lower scores. Furthermore, the NIHTB-CB appears to overestimate Fluid composite scores in younger participants and underestimate it among African-American participants, although again these results are regarded as preliminary at best given the demographic homogeneity and modest size of the current sample.

Discussion

This study evaluated the psychometric properties of the NIHTB-CB and its agreement with traditional neuropsychological tests in healthy older adults. We found evidence supporting its test-retest reliability over approximately one year and convergent and discriminant validity with criterion variables with which the NIHTB-CB does and does not share method variance. Despite these promising results, our findings nonetheless show that scores from the NIHTB-CB and GS tests are not equivalent, warranting caution should the NIHTB-CB be used for clinical purposes.

First, we found strong test-retest reliability for both Fluid (CCC = .73) and Crystallized (CCC = .92) composite scores. These estimates were very similar to those reported in previous validation studies with much shorter follow-up intervals (*r*'s = .86 and .92 for Fluid and Crystallized; Heaton et al., 2014). Reliability of individual tests comprising the NIHTB-CB Fluid Cognition Composite were generally lower and more variable than their respective GS tests, but displayed minimal evidence of practice effects over the approximate 15-month test-retest interval. This contrasts with the GS Fluid measures, which consistently produced slightly higher scores at follow-up, and suggests that the adaptive format of the NIHTB-CB may be less susceptible to practice effects.

Second, our findings support the construct validity of the Fluid and Crystallized Cognition Composites, based on distinct patterns of associations with age, education, cerebral volumes, and standard neuropsychological tests. Our data suggest that the NIHTB-CB Fluid Cognition Composite tests are at least as sensitive as GS tests in detecting cerebral volume loss that occurs in healthy aging (Dodge et al., 2014) and Alzheimer's disease dementia (Heister et al., 2011). The Crystallized and Fluid Cognition Composites also showed strong convergent and discriminant validity with neuropsychological measures, although convergent validity for individual Fluid tests ranged from poor to adequate. Curiously, although the Dimensional Change Card Sort Test was the strongest correlate of cerebral volume loss, it was poorly correlated with other measures of executive functioning. This likely reflects the multifaceted nature of executive functioning tests and the disparate paradigms that were used to measure cognitive flexibility.

Finally, agreement between NIHTB-CB and GS Crystallized Cognition Composites was good, but the NIHTB-CB systematically overestimated performance by 9 standard score points. In contrast, the two Fluid Cognition Composites yielded roughly equivalent mean scores but had large pairwise differences that reflect poor precision and proportional differences, with increasing discrepancies at the tails of the distributions (NIHTB-CB > GS at higher scores and NIHTB-CB < GS at lower scores). It is possible that this overestimation reflects ceiling effects of the GS tests that are not present in the NIHTB-CB, given that the neuropsychological battery used by UDS was primarily designed for individuals with suspected cognitive impairments and may be less useful for evaluating cognition at the higher end of functioning (Mathews et al., 2014). Follow-up exploratory analyses suggested that investigation of sub-samples in future studies may yield insights into possible demographic differences driving these discrepancies.

Clinical Implications

Our findings are consistent with extant literature supporting the NIHTB-CB Crystallized Cognition Composite as a clinically useful estimate of premorbid intellectual ability that is relatively robust to aging and cerebral volume loss, and may aid in the determination of decline in fluid abilities. The NIHTB-CB Fluid Cognition Composite also appears to be a valid measure of cognitive functioning in older adults, which relates to non-cognitive factors (e.g. cerebral atrophy) that are pertinent to aging and neurodegenerative disease.

However, caution should be taken when interpreting the NIHTB-CB Crystallized Cognition Composite as a measure of premorbid functioning, given the tendency to overestimate performance relative to standard tests. It is possible that a conversion factor could be applied to equate performances on these methods because the discrepancy is constant across the entire range of scores and demographics included in this sample. Exploratory analyses suggest that the Picture Vocabulary test is the primary source of inflation of the NIHTB-CB Crystallized Cognition Composite, yielding scores that were on average 8-9 points higher than the two oral reading tests. Thus, clinicans using the NIHTB-CB may elect to use the oral reading test alone rather than the Crystallized Cognition Composite as a proxy for premorbid IQ.

Caution is also warranted when interpreting the NIHTB-CB Fluid Cognition Composite, insofar as age-adjusted scores are concerned. Several factors appeared to influence the size and direction of the discrepancy, including age, ethnicity, and obtained score, indicating that the NIHTB-CB may overestimate fluid cognition at higher scores and underestimate it at lower scores although lower performance in this sample is still largely within the average range. These findings suggest that additional demographic adjustments may be required when making intra-individual comparisons in clinical settings, particularly with more diverse populations.

Limitations

The current sample was relatively homogenous, well-educated, and obtained higher scores than those reported in prior validation and normative studies (Casaletto et al., 2015). This may limit the degree to which these results generalize to other normative and clinical populations. Future studies should further investigate whether race, education, or other characteristic disproportionately affects performance on one of the two test batteries. We chose to use age-adjusted normative data on both test batteries to evaluate construct validity and agreement because the full range of demographic adjustments were not available for all GS tests. Additionally, we did not account for order effects on performance as the administration order was fixed per the study protocol. We also did not apply any Type-I error adjustments in our statistical analyses in line with other psychometric papers (e.g., Heaton et al., 2014; Weintraub et al., 2013), as we intended to interpret the overall pattern of associations between the two batteries. Another caveat relates to the selection of GS tests, which do not directly correspond to the administration formats of the NIHTB-CB tests and also have imperfect reliability and validity. Thus, it is important to note that low agreement between NIHTB-CB and GS measures does not indicate the superiority of one measure over the other per se, but rather that the two measures often generate different interpretative results.

Conclusion

This study contributes to the literature by providing clinically relevant information on the comparability of NIHTB-CB and traditional neuropsychological test scores. Agreement or comparability is frequently discussed but rarely systematically examined in the neuropsychological literature. Although the methods used in this study are common in other fields, only a handful of studies in the neuropsychological literature have made use of these methods to assess agreement between scores on a continuous scale (e.g., Berg, Durant, Banks, & Miller, 2016). The current practice is to rely on estimates of linear agreement (e.g., Pearson’s correlations) which fail to capture constant differences, or mean differences (e.g., *t*-tests), which capture constant differences but not random or proportional differences, as large measurement error will result in the incorrect conclusion that the methods are equivalent. Comparability is a question of both random deviation (i.e., error) and systematic deviation (i.e., constant and proportional) and no single statistic can estimate both. Although it would be very unlikely for two different tests to agree perfectly, the *degree* of agreement has critical implications for clinical practice. For example, if a measure of crystallized abilities systematically overestimates premorbid IQ, large discrepancies may be misinterpreted as evidence of decline and ultimately lead to misdiagnoses.

Our findings underlie the need to incorporate both descriptive and statistical analyses of agreement when evaluating new measures against a reference standard to complement classical psychometric methods and enhance the interpretation of validation studies. In conclusion, these findings suggest that the NIHTB-CB appears to be a valid tool for cognitive assessment in aging, but that it may generate different interpretative results from existing measures commonly used in clinical evaluations. Thus, these two methods should not be considered interchangeable.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors thank Brianna Jones, Deborah Nelson, Danae Montgomery, and Clifford Chan for their assistance with participant recruitment and research coordination. They also gratefully acknowledge the study participants for generously offering their time and data for this work. This research was supported by the National Institutes of Health (NCATS UL1TR000062/KL2 TR000060 and NIA K23 AG044434 to A.B.). None of the authors have any conflicts of interest to disclose.

References

Altman DG, & Bland JM (1983). Measurement in Medicine: The Analysis of Method Comparison Studies. Journal of the Royal Statistical Society. Series D (The Statistician), 32(3), 307–317. 10.2307/2987937

Altman Douglas G. (1991). Practical Statistics for Medical Research. London: Chapman and Hall.

Azab M, Carone M, Ying SH, & Yousem DM (2015). Mesial Temporal Sclerosis: Accuracy of NeuroQuant versus Neuroradiologist. AJNR. American Journal of Neuroradiology, 36(8), 1400–1406. 10.3174/ajnr.A4313 [PubMed: 25907519]

Barchard KA (2012). Examining the reliability of interval level data using root mean square differences and concordance correlation coefficients. Psychological Methods, 17(2), 294–308. 10.1037/a0023351 [PubMed: 21574711]

Barnhart HX, Haber MJ, & Lin LI (2007). An Overview on Assessing Agreement with Continuous Measurements. Journal of Biopharmaceutical Statistics, 17(4), 529–569. 10.1080/10543400701376480 [PubMed: 17613641]

Berg J-L, Durant J, Banks SJ, & Miller JB (2016). Estimates of premorbid ability in a neurodegenerative disease clinic population: comparing the Test of Premorbid Functioning and the Wide Range Achievement Test, 4th Edition. The Clinical Neuropsychologist, 30(4), 547–557. 10.1080/13854046.2016.1186224 [PubMed: 27187762]

Bondi MW, Edmonds EC, Jak AJ, Clark LR, Delano-Wood L, McDonald CR, … Salmon DP (2014). Neuropsychological criteria for mild cognitive impairment improves diagnostic precision, biomarker associations, and progression rates. Journal of Alzheimer's Disease: JAD, 42(1), 275–289. 10.3233/JAD-140276 [PubMed: 24844687]

Buckley RF, Sparks KP, Papp KV, Dekhtyar M, Martin C, Burnham S, … Rentz DM (2017). Computerized Cognitive Testing for Use in Clinical Trials: A Comparison of the NIH Toolbox and Cogstate C3 Batteries. The Journal of Prevention of Alzheimer's Disease, 4(1), 3–11. 10.14283/jpad.2017.1

Carlozzi NE, Tulsky DS, Wolf TJ, Goodnight S, Heaton RK, Casaletto KB, … Heinemann AW (2017). Construct validity of the NIH Toolbox Cognition Battery in individuals with stroke. Rehabilitation Psychology, 62(4), 443–454. 10.1037/rep0000195 [PubMed: 29265865]

Carrasco JL, & Jover L (2003). Estimating the generalized concordance correlation coefficient through variance components. Biometrics, 59(4), 849–858. [PubMed: 14969463]

Casaletto KB, Umlauf A, Beaumont J, Gershon R, Slotkin J, Akshoomoff N, & Heaton RK (2015). Demographically Corrected Normative Standards for the English Version of the NIH Toolbox Cognition Battery. Journal of the International Neuropsychological Society: JINS, 21(5), 378–391. 10.1017/S1355617715000351 [PubMed: 26030001]

Cohen J (1992). A power primer. Psychological Bulletin, 112(1), 155–159. [PubMed: 19565683]

Dodge HH, Zhu J, Harvey D, Saito N, Silbert LC, Kaye JA, … Albin RL (2014). Biomarker progressions explain higher variability in stage-specific cognitive decline than baseline values in Alzheimer disease. Alzheimer's & Dementia, 10(6), 690–703. 10.1016/j.jalz.2014.04.513

Gershon RC, Cook KF, Mungas D, Manly JJ, Slotkin J, Beaumont JL, & Weintraub S (2014). Language Measures of the NIH Toolbox Cognition Battery. Journal of the International Neuropsychological Society : JINS, 20(6), 642–651. 10.1017/S1355617714000411 [PubMed: 24960128]

Gershon RC, Wagster MV, Hendrie HC, Fox NA, Cook KF, & Nowinski CJ (2013). NIH Toolbox for Assessment of Neurological and Behavioral Function. Neurology, 80(11 Suppl 3), S2–S6. 10.1212/WNL.0b013e3182872e5f [PubMed: 23479538]

Golden CJ (1978). Stroop color and word test. A manual for clinical and experimental uses. Chicago: Stoelting.

Hackett K, Krikorian R, Giovannetti T, Melendez-Cabrero J, Rahman A, Caesar EE, … Isaacson RS (2018). Utility of the NIH Toolbox for assessment of prodromal Alzheimer’s disease and dementia. Alzheimer’s & Dementia : Diagnosis, Assessment & Disease Monitoring, 10, 764–772. 10.1016/j.dadm.2018.10.002

Heaton RK, 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: JINS, 20(6), 588–598. 10.1017/S1355617714000241 [PubMed: 24960398]

Heister D, Brewer JB, Magda S, Blennow K, & McEvoy LK (2011). Predicting MCI outcome with clinically available MRI and CSF biomarkers. Neurology, 77(17), 1619–1628. 10.1212/WNL.0b013e3182343314 [PubMed: 21998317]

Holdnack JA, Tulsky DS, Slotkin J, Tyner CE, Gershon R, Iverson GL, & Heinemann AW (2017). NIH toolbox premorbid ability adjustments: Application in a traumatic brain injury sample. Rehabilitation Psychology, 62(4), 496–508. 10.1037/rep0000198 [PubMed: 29265870]

Ivnik RJ, Malec JF, Smith GE, Tangalos EG, & Petersen RC (1996). Neuropsychological tests’ norms above age 55: COWAT, BNT, MAE token, WRAT-R reading, AMNART, STROOP, TMT, and JLO. The Clinical Neuropsychologist, 10(3), 262–278. 10.1080/13854049608406689

Lang S, Cadeaux M, Opoku-Darko M, Gaxiola-Valdez I, Partlo LA, Goodyear BG, … Kelly J (2017). Assessment of Cognitive, Emotional, and Motor Domains in Patients with Diffuse Gliomas Using the National Institutes of Health Toolbox Battery. World Neurosurgery, 99, 448–456. 10.1016/j.wneu.2016.12.061 [PubMed: 28039096]

Lin LI (1989). A concordance correlation coefficient to evaluate reproducibility. Biometrics, 45(1), 255. doi:10.2307/2532051 [PubMed: 2720055]

Loring DW, Bowden SC, Staikova E, Bishop JA, Drane DL, & Goldstein FC (2019). NIH Toolbox Picture Sequence Memory Test for Assessing Clinical Memory Function: Diagnostic Relationship to the Rey Auditory Verbal Learning Test. Archives of Clinical Neuropsychology, 34(2), 268–276. 10.1093/arclin/acy028 [PubMed: 29608637]

Lucas JA, Ivnik RJ, Willis FB, Ferman TJ, Smith GE, Parfitt FC, … Graff-Radford NR (2005). Mayo’s Older African Americans Normative Studies: normative data for commonly used clinical neuropsychological measures. The Clinical Neuropsychologist, 19(2), 162–183. 10.1080/13854040590945265 [PubMed: 16019702]

Mathews M, Abner E, Kryscio R, Jicha G, Cooper G, Smith C, … Schmitt FA (2014). Diagnostic accuracy and practice effects in the National Alzheimer’s Coordinating Center Uniform Data Set neuropsychological battery. Alzheimer’s & Dementia, 10(6), 675–683. 10.1016/j.jalz.2013.11.007

McGraw KO, & Wong SP (1996). Forming inferences about some intraclass correlation coefficients. Psychological Methods, 1(1), 30–46. 10.1037/1082-989X.1.1.30

Mungas D, Heaton R, Tulsky D, Zelazo PD, Slotkin J, Blitz D, … Gershon R (2014). Factor Structure, Convergent Validity, and Discriminant Validity of the NIH Toolbox Cognitive Health Battery (NIHTB-CHB) in Adults. Journal of the International Neuropsychological Society, 20(06), 579–587. 10.1017/S1355617714000307 [PubMed: 24960474]

Nestor SM, Rupsingh R, Borrie M, Smith M, Accomazzi V, Wells JL, … the Alzheimer’s Disease Neuroimaging Initiative. (2008). Ventricular enlargement as a possible measure of Alzheimer’s disease progression validated using the Alzheimer’s disease neuroimaging initiative database. Brain, 131(9), 2443–2454. 10.1093/brain/awn146 [PubMed: 18669512]

O’Shea A, Cohen R, Porges EC, Nissim NR, & Woods AJ (2016). Cognitive Aging and the Hippocampus in Older Adults. Frontiers in Aging Neuroscience, 8 10.3389/fnagi.2016.00298

Passing H, & Bablok, null. (1983). A new biometrical procedure for testing the equality of measurements from two different analytical methods. Application of linear regression procedures for method comparison studies in clinical chemistry, Part I. Journal of Clinical Chemistry and Clinical Biochemistry. Zeitschrift Fur Klinische Chemie Und Klinische Biochemie, 21(11), 709–720. [PubMed: 6655447]

Raz N, Lindenberger U, Rodrigue KM, Kennedy KM, Head D, Williamson A, … Acker JD (2005). Regional brain changes in aging healthy adults: general trends, individual differences and modifiers. Cerebral Cortex (New York, N.Y.: 1991), 15(11), 1676–1689. 10.1093/cercor/bhi044

Reitan RM, & Wolfson D (1985). The Halstead-Reitan neuropsychological test battery: theory and clinical interpretation. Tucson, Ariz: Neuropsychology Press.

Rey A (1964). L'examen clinique en psychologie (2e éd.). In Le Psychologue (2e éd.). Paris: Presses universitaires de France WorldCat.org.

Schmidt M (1996). Rey Auditory and Verbal Learning Test: A handbook. Los Angeles: Western Psychological Services.

Shirk SD, Mitchell MB, Shaughnessy LW, Sherman JC, Locascio JJ, Weintraub S, & Atri A (2011). A web-based normative calculator for the uniform data set (UDS) neuropsychological test battery. Alzheimer's Research and Therapy, 3(6), 32 10.1186/alzrt94

Sinha P, Wong AWK, Kallogjeri D, & Piccirillo JF (2018). Baseline Cognition Assessment Among Patients With Oropharyngeal Cancer Using PROMIS and NIH Toolbox. JAMA Otolaryngology–Head & Neck Surgery, 144(11), 978 10.1001/jamaoto.2018.0283 [PubMed: 29710116]

Steiger JH (1980). Tests for comparing elements of a correlation matrix. Psychological Bulletin, 87(2), 245–251. 10.1037/0033-2909.87.2.245

Tulsky DS, Carlozzi NE, Holdnack J, Heaton RK, Wong A, Goldsmith A, & Heinemann AW (2017). Using the NIH Toolbox Cognition Battery (NIHTB-CB) in individuals with traumatic brain injury. Rehabilitation Psychology, 62(4), 413–424. 10.1037/rep0000174 [PubMed: 29265862]

Wechsler D (1987). Manual for the Wechsler memory scale-revised. San Antonio, TX: Psychological Corporation.

Wechsler D (1997). Wechsler memory scale : WMS-III (3 ed.). San Antonio: Psychological Corp., Harcourt Brace WorldCat.org.

Weintraub S, Dikmen SS, Heaton RK, Tulsky DS, Zelazo PD, Bauer PJ, … Gershon RC (2013). Cognition assessment using the NIH Toolbox. Neurology, 80(Issue 11, Supplement 3), S54–S64. 10.1212/WNL.0b013e3182872ded [PubMed: 23479546]

Weintraub Sandra, Dikmen SS, Heaton RK, Tulsky DS, Zelazo PD, Slotkin, … Gershon R (2014). The cognition battery of the NIH toolbox for assessment of neurological and behavioral function: validation in an adult sample. Journal of the International Neuropsychological Society: JINS, 20(6), 567–578. 10.1017/S1355617714000320 [PubMed: 24959840]

Weintraub Sandra, Salmon D, Mercaldo N, Ferris S, Graff-Radford NR, Chui H, … Morris JC (2009). The Alzheimer's Disease Centers' Uniform Data Set (UDS): The Neuropsychologic Test Battery. Alzheimer Disease & Associated Disorders, 23(2), 91 10.1097/WAD.0b013e318191c7dd [PubMed: 19474567]

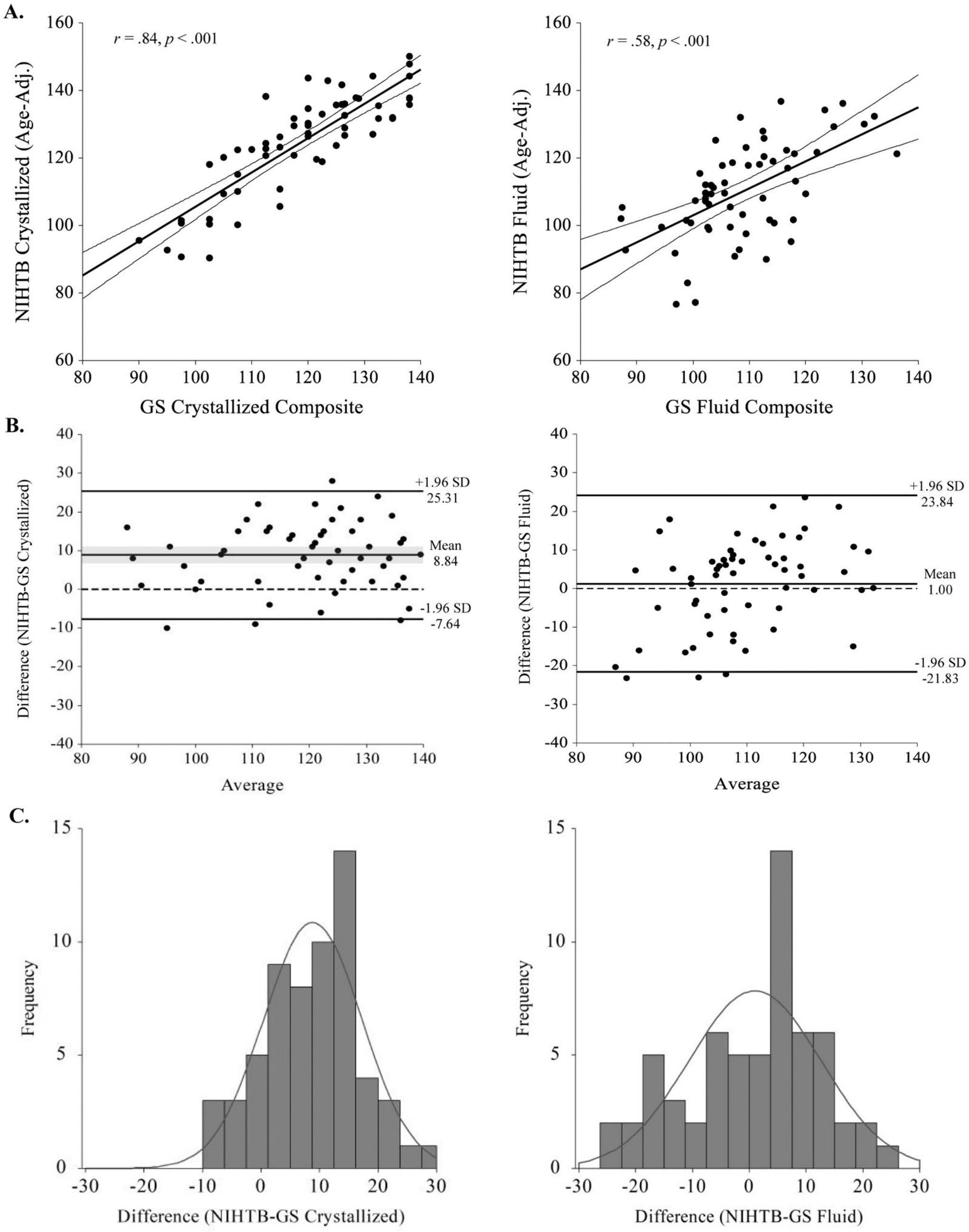

Figure 1.

Agreement between age-adjusted NIHTB-CB and GS composite scores. A.) Scatterplots with regression lines and shading for 95% CIs of age-adjusted NIHTB and GS Crystallized (left) and Fluid (right) Cognition Composite scores. B.) Solid horizontal lines in Bland-Altman plots represent mean differences and upper and lower limits of agreement, illustrating where 95% of differences are expected to fall in the general population. Shading reflects 95% CIs around the mean difference. Dashed lines represent the zero line of perfect agreement. C.) Frequency distributions of discrepancy scores approximate normal distribution.

Table 1.

NIH Toolbox Cognition Battery tests with corresponding gold standard (GS) tests

Cognitive Construct	NIH Toolbox Measure	GS Measure
Fluid Composite		
Executive – Set Shifting	Dimensional Change Card Sort	Trail Making Test – Part B
Executive – Inhibition	Flanker Inhibitory Control & Attention	Stroop Color Word Test (C/W trial)
Episodic Memory	Picture Sequence Memory	RAVLT (total learning trials 1-5)
Working Memory	List Sorting Working Memory	WMS-R Digit Span
Processing Speed	Pattern Comparison Processing Speed	WAIS-III Digit Symbol
Crystallized Composite	Picture Vocabulary + Oral ReadingRecognition	AmNART

RAVLT=Rey Auditory Verbal Learning Test; WMS-R=Wechsler Memory Scale-Revised; WAIS-III=Wechsler Adult Intelligence Scale, Third Edition; AmNART=American National Adult Reading Test.

Table 2.

Sample characteristics at baseline visit

	Total Sample (*N* = 61)	**Sub-sample with follow-up[a] (*n* = 37)**
Basic Demographics		
Age, *M* ± *SD*	67.73 ± 5.26	67.58 ± 4.78
Years of education, *M* ± *SD*	16.38 ± 2.50	16.49 ± 2.52
Caucasian, *n* (%)	55 (90.16)	33 (89.19)
Female, *n* (%)	40 (65.57)	26 (70.27)
Volumetric MRI (bilateral, in cm^3)		
Lateral ventricles	32.01 ± 23.48	28.46 ± 15.96
Inferior lateral ventricles	2.13 ± 0.88	2.06 ± 0.78
Hippocampi	7.58 ± 0.87	7.69 ± 0.89
Medical History[b], *n* (%)		
Hypertension	29 (47.54)	15 (40.54)
Dyslipidemia	19 (31.15)	13 (35.14)
Diabetes	8 (13.11)	5 (13.51)
Thyroid disease	9 (14.75)	5 (13.51)
Heart disease	4 (6.56)	6 (16.22)

[a] No statistically significant differences on demographic variables or cognitive test scores between total sample and subsample returning for follow-up (all $p > .05$).

[b] Medical history includes any lifetime history of diagnoses that participants endorsed on self-report questionnaires.

Table 3.

Test-retest reliability and practice effects (mean interval 15.03 months, SD = 3.11)

Cognitive Test Batteries	*r*	CCC (95% CI)	Time 1 Mean ± SD	Time 2 Mean ± SD	*t* value	*p* value	Cohen’s d
NIHTB-CB Fluid Cognition Composite (*n* =36)	.73***	.73 (.53, .85)	110.57 ± 14.27	110.08 ± 13.75	−0.49	.630	−0.04
Dimensional Card Sort (*n* = 36)	.58***	.54 (.28, .72)	108.05 ± 9.64	111.27 ± 9.57	2.40	.022	**0.33**
Flanker Inhibitory Control	.62***	.54 (.31, .72)	106.62 ± 11.00	102.53 ± 8.17	−2.83	.008	−0.37
Picture Sequencing Memory	.74***	.74 (.55, .85)	114.38 ± 22.07	113.49 ± 20.26	−0.37	.717	−0.04
List Sorting Working Memory	.46**	.46 (.17, .68)	112.84 ± 11.11	112.89 ± 10.21	−0.02	.985	0.00
Pattern Comparison Processing Speed	.61***	.61 (.36, .78)	99.38 ± 16.99	98.48 ± 18.44	−0.34	.734	−0.05
NIHTB-CB Crystallized Cognition Composite	.92***	.92 (.86, .96)	125.76 ± 15.66	126.32 ± 15.89	0.55	.583	0.04
Picture Vocabulary	.88***	.87 (.77, .93)	125.05 ± 15.91	127.07 ± 16.06	1.59	.121	0.13
Oral Reading Recognition	.85***	.85 (.73, .92)	115.35 ± 11.66	114.54 ± 12.08	−0.76	.451	−0.07
GS Fluid Cognition Composite	.92***	.85 (.74, .91)	109.35 ± 10.30	113.23 ± 9.82	5.71	.000	**0.38**
Trail Making Test – Part B	.59***	.54 (.29, .72)	110.22 ± 12.42	115.57 ± 13.40	2.77	.009	**0.43**
Stroop Color Word	.87***	.86 (.74, .92)	106.62 ± 11.00	109.59 ± 15.06	2.23	.032	**0.27**
RAVLT (total learning)	.80***	.78 (.62, .88)	106.73 ± 15.72	111.65 ± 16.41	1.58	.122	**0.31**
WMS-R Digit Span	.67***	.65 (.42, .80)	108.68 ± 18.99	112.46 ± 13.36	1.96	.058	0.20
WAIS-III Digit Symbol	.86***	.77 (.63, .86)	109.03 ± 12.93	116.84 ± 10.20	4.07	.000	**0.60**
GS Crystallized Cognition Composite(AmNART)	.93***	.93 (.86, .96)	116.76 ± 15.79	115.62 ± 15.85	1.17	.251	−0.07

Notes. CCC = concordance correlation coefficient. *p* values reflect significance of two-tailed *t*-tests for differences in paired means. Magnitudes of practice effects are shown with Cohen’s d effect sizes, with values in bold representing significantly higher scores at follow-up (alpha > .05). *n* = 37 unless stated otherwise. All values reflect age-adjusted standard scores (M=100, SD=15).

* $p < .05$;

** $p < .01$;

*** $p < .001$.

Table 4.

Pearson’s correlations between cognitive test scores and cerebral volumes

	LAV		HOC	
Cognitive Tests	***r***	***p***	***r***	***p***
NIHTB-CB Fluid Composite	**−.35**	.008	**.41**	.002
Dimensional Card Sort	**−.36**	.008	**.29**	.037
Flanker Inhibitory Control & Attention	−.18	.198	.13	.344
Picture Sequencing Memory	**−.28**	.043	**.37**	.006
List Sorting Working Memory	−.06	.670	.10	.477
Pattern Comparison Processing Speed	**−.29**	.038	**.29**	.033
NIHTB-CB Crystallized Composite	−.03	.840	.24	.084
Picture Vocabulary	.02	.866	.19	.174
Oral Reading Recognition	−.04	.766	.18	.193
GS Fluid Composite	−.26	.059	.26	.052
Trail Making Test – Part B	.12	.365	−.13	.330
Stroop Color Word	−.24	.072	**.28**	.035
RAVLT	−.23	.091	**.27**	.048
Digit Span	−.01	.970	.16	.240
Digit Symbol	**−.31**	.023	.24	.076
GS Crystallized Composite (AmNART)	.09	.531	.11	.436

Notes. LAV = lateral ventricles; HOC = hippocampal occupancy score. Both LAV and HOC values reflect combined bilateral volumes, in cm^3, normalized to percent intracranial volume. Values in bold are statistically significant at $p < .05$.

Table 5.

Correlations and agreement between age-adjusted standard scores (M=100, SD=15) on NIH Toolbox Cognition Battery and gold standard measures

Cognitive Tests	*r*	CCC (95% CI)	RMSD
Fluid Composites	.58***	.55 (.37, .70)	11.60
Executive-Set Shifting (DCCS-Trails B)	.22	.21 (−.03, .43)	14.74
Executive-Inhibition (FICA-Stroop CW)	.10	.09 (−.15, .32)	16.99
Episodic Memory (PSMT-RAVLT)	.48***	.45 (.24, .62)	20.89
Working Memory (LSWM-Digit Span)	.49***	.45 (.25, .61)	14.15
Processing Speed (PCPS-Digit Symbol)	.43***	.31 (.16, .46)	21.04
Crystallized Composites	.84***	.70 (.58, .79)	12.15
Receptive Vocabulary (PVT-AmNART) [a]	.72***	.62 (.48, .75)	13.65
Reading Decoding (ORRT-AmNART)	.85***	.82 (.73, .89)	7.52

Notes. CCC=concordance correlation coefficient; RMSD=root mean squared difference; DCCS=Dimensional Change Card Sort Test; FICA=Flanker Inhibitory Control and Attention Test; Stroop C/W=Stroop Color-Word trial; PSMT=Picture Sequence Memory Test; RAVLT=Rey Auditory Verbal Learning Test; LSWM=List Sorting Working Memory Test; PCPS=Pattern Comparison Processing Speed; PVT=Picture Vocabulary Test; ORRT=Oral Reading Recognition Test; AmNART=American National Adult Reading Test.

* $p < .05$;

** $p < .01$;

*** $p < .001$.

[a] Values are provided as reference only.