OXFORD
UNIVERSITY PRESS

Archives of Clinical Neuropsychology 00 (2021) 1–12

Archives of CLINICAL NEUROPSYCHOLOGY

Comparing Psychometric Properties of the NIH Toolbox Cognition Battery to Gold-Standard Measures in Socioeconomically Diverse Older Adults

R.K. MacAulay*, A. Boeve, A. Halpin

Department of Psychology, University of Maine, Orono, ME 04469, USA

*Corresponding author at: Department of Psychology, University of Maine, 301 Little Hall, Orono, ME 04469, USA. Tel.: 207-581-2044; fax: 207-581-6128. *E-mail address:* Rebecca.macaulay@maine.edu (R.K. MacAulay).

Accepted 18 March 2021

Abstract

Objective: The National Institutes of Health Toolbox-Cognition Battery (NIHTB-CB) is an efficient computerized neuropsychological battery. This study investigated its psychometric properties in terms of sociodemographic characteristics and technology use in adults aged 57–87 (with an average age of 70).

Methods: Community-based participatory research procedures were used to enhance enrollment of adults with lower education and income backgrounds. Study procedures replicated work that compared the NIHTB-CB Crystallized and Fluid composites to analogous gold-standard (GS) measures and extended it by investigation of socioeconomic status and technology use-related differences in performance.

Results: The high correlations among the NIHTB-CB and GS analogous Crystallized and Fluid composites suggested good convergent validity. There was no evidence of significant education- or economic-related group differences in these associations. However, caution is needed as Cronbach's alpha that indicated the NIHTB-CB Fluid composite had questionable internal item consistency. The NIHTB-CB and GS measures demonstrated poor discriminant validity in the high school but not college-educated groups. Regression analyses found that comfort with technology use, income, education, and age predicted better cognitive test performance on the computerized and paper-pencil measures.

Conclusions: There is an urgent need to improve the understanding of socioeconomic disparities influence on test scores and brain health. Lack of discriminant validity in the cognitive tests indicates that these measures could result in diagnostic errors within noncollege-educated older adults. These findings reduce confidence in the use of the NIHTB-CB Fluid composite in older adults and support that there is a significant socioeconomic-related digital divide in comfort with technology use.

Keywords: Diversity; Socioeconomic status; Educational attainment; Mild cognitive impairment; Specificity; Sensitivity

The National Institutes of Health Toolbox-Cognition Battery (NIHTB-CB) is a fully computerized neuropsychological battery with norms across the life span (Heaton et al., 2014; Weintraub et al., 2014). The NIHTB-CB was designed to provide a common set of neuropsychological measures that allows for comparisons across the life span in epidemiologic and clinical research (Gershon et al., 2013). Other benefits of the NIHTB-CB include its use of computer adaptive testing that can significantly reduce testing time and its potential utility to screen for cognitive impairments in medically underserved areas given its easy-to-use computerized format. However, despite these significant strengths, there are concerns regarding its psychometric properties in more culturally and socioeconomically diverse populations.

Growing evidence indicates that the NIHTB's validity of population-based norms for cognitive measures in culturally, ethnically, and socioeconomically diverse populations are not adequately established. Although the NIHTB-CB has provided evidence of good convergent and discriminant validity in its validation studies (see Heaton et al., 2014; Weintraub et al., 2014), these same studies found that the computerized measures varied substantially by age, education, race/ethnicity, sex, and socioeconomic status (SES). In addition to the expected demographic effects of age and education, there were medium to large

doi:10.1093/arclin/acab018

effect size differences (Cohen's $d = 0.62–0.98$) between the low as compared with higher income group on the age-adjusted NIHTB-CB Crystallized, Fluid and Total Cognition Composites (Heaton et al., 2014). Efforts to improve the normative data included fully demographically adjusted scores (age, education, sex, race/ethnicity) for adults and children (Casaletto et al., 2015). However, these norms are based on the original study sample that did not include older adults between the ages of 60–64 years old, which is an important transition point in aging. Sampling procedures also included "some college" (13–15 years of education) in the high-school graduate group. Notably, older adults with some college have been found to demonstrate better neuropsychological performance than those with only a high-school education (Wiederholt et al., 1993). Thus, it is unclear whether the NIHTB-CB categorization procedures potentially influenced study outcomes (e.g., inflating average scores for the "high-school" group).

While the development of the NIHTB-CB included modifications to make the battery more user friendly for older adults, it remains unclear whether computerized cognitive testing introduces construct irrelevance in older adults with less technology exposure. Relevantly, neuropsychological test interpretations can be significantly impacted by sociodemographic related differences in computer familiarity (Bauer et al., 2012), in which there is evidence that greater computer familiarity is associated with better cognitive performance on certain computer administered measures (Iverson, Brooks, Ashton, Johnson, & Gualtieri, 2009). There is also evidence of significant SES-related digital divides in technology use in older adults, in which higher family income and educational attainment is strongly associated with computer/tablet ownership (Anderson & Perrin, 2017). It is also possible that those with less technology exposure may be more susceptible to test anxiety on computerized cognitive measures, which could result in an underestimation of their cognitive abilities. Collectively, these findings suggest that less comfort with technology use could potentially contribute to test bias in lower SES populations.

Accurate assessment of neurocognitive function, both in research and clinical settings, is critical to identifying, understanding, and treating factors related to cognitive decline. Notably, recent research using the NIHTB-CB suggested that almost 50% of adults in the general population would meet psychometric criteria for a diagnosis of a neurocognitive disorder, using the available normative data (Holdnack et al., 2017). This work has suggested that stratifying by education or intellectual ability (using the Crystallized composite) may lead to more accurate classifications.

There is increasing concern regarding the normative data used for neuropsychological assessment in lower SES populations within the literature. Notably, SES is a complex factor comprised of education, income/wealth, occupation, and subjective perceptions of status. Cognitive reserve theories suggest that higher SES-related factors help individuals to maintain cognitive function longer in the presence of brain disease pathology (Stern, 2006). Evidence in support of this work has found that income, wealth, education, and more complex occupation achievements associate with decreased risk for cognitive decline or dementia (e.g., Cadar et al., 2018; Zahodne, Stern, & Manly, 2015; Zhang et al., 2015). There is also evidence to suggest that beyond 9 years of education, the relationship between education and cognitive decline is fully mediated by income in diverse older adults (Zahodne et al., 2015). Furthermore, the effects of low SES on cognitive function are not limited to earlier life, as measures of poverty demonstrate independent associations with measures of processing speed/attention in older adults (Zhang et al., 2015).

Given the future clinical implications of its wide usage in cognitive aging research, to include the large multisite Advancing Reliable Measurement in Alzheimer's Disease and Cognitive Aging (ARMADA) study that describes "paving its way for use in prevention clinical trials" as an objective (Gershon, 2019), it is critical to better understand sociodemographic factors that may contribute to its diagnostic accuracy across a wide age range of older adults.

The primary objective of this study was to investigate the psychometric properties of the NIHTB-CB in comparison to the corresponding gold-standard (GS) measures in terms of sociodemographic characteristics in adults aged 55 and older. The current study replicated testing procedures described by the NIHTB-CB validation series (Heaton et al., 2014; Weintraub et al., 2014). Community-based participatory research (CBPR) procedures were used to enhance the recruitment and enrollment of a socioeconomically diverse group of older adult participants. We aimed to determine whether there were education and/or economic-related group differences in the NIHTB-CB Crystallized and Fluid composites convergent and discriminant validity. We posited that there would be education-related group differences in the NIHTB-CB composites convergent and discriminant validity and that those with a high-school education or less would have significantly lower scores than those with some college or higher. Additionally, multiple hierarchal regression analyses adjusting for relevant variables investigated whether higher levels of income and comfort with technology use associated with better cognitive performance on the computerized NIHTB-CB as compared with paper-and-pencil GS-measures.

Methods

Recruitment Procedures

Participants were recruited as part of the Maine-Aging Behavior Learning Enrichment (M-ABLE study). CBPR procedures were used to enhance recruitment of a socioeconomically diverse sample. Recruitment efforts were conducted throughout

the state with the assistance of directors and other community stake holders at the University of Maine's Center on Aging, Eastern Area of Aging Agency, local health care providers, and low-income independent living community housing residence coordinators. To reduce transportation barriers, there were a total of seven study sites: three university-based locations in different cities/towns and four low-income independent community dwelling residence offices in Penobscot and Kennebec counties in Maine. All study sites were easily accessible with convenient parking. Recruitment and testing for the current study took place from October 2018 to March 2020. Enrolled participants were compensated up to 70 U.S. dollars for completion of the study.

Participants

Study inclusion criteria were purposefully wide to obtain a more representative sample population in order to improve generalizability of the findings to more diverse older adults. Inclusion criteria included: being aged 55–90 years old, willing to undergo neuropsychological assessment, providing information on SES, and having normal or corrected vision and hearing. Exclusion criteria were collected via clinical interview and screening measures. Individuals with a diagnosis of dementia or mild cognitive impairment (MCI), presence of an untreated medical condition (e.g., untreated thyroid disease, B12 deficiency, or psychiatric condition), moderate or severe cognitive impairments (Montreal Cognitive Assessment: MoCA scores <18; Carson, Leach, & Murphy, 2018; Nasreddine et al., 2005), and/or severe depression (Geriatric Depression Scale scores >10; Sheikh & Yesavage, 1986) were excluded. Adults with a known history of an intellectual disability, neurological disease, or other conditions that are known to cause cognitive sequelae (e.g., Parkinson's disease, epilepsy, major stroke, and moderate-to-severe traumatic brain injury [TBI]) were also excluded. Additionally, individuals with a mild TBI within the past 2 years were excluded. Of the 140 older adults screened for enrollment, 121 older adults were eligible for the study. The most common reason for exclusion was not meeting the MoCA cut-score criteria ($n = 9$) followed by no longer being interested or schedule conflicts ($n = 5$). Five participants were unable to complete the entire assessment battery. This study was approved by the University of Maine Institutional Review Board.

Procedures

All study procedures were conducted by well-trained research assistants and supervised by a licensed clinical psychologist. Neuropsychological testing and clinical interviews were administered by advanced clinical psychology doctoral students receiving specialized educational and practicum training in neuropsychology. Given known effects of time of day on neuropsychological test performance in older adults, all appointments were scheduled in the morning with the majority of appointments starting between 8 am and 9 am (depending on the participant's time preference).

Participants were first screened and underwent informed consent followed by the NIHTB-CB and GS measures. Test order was the same used by Heaton and colleagues (NIHTB cognitive team, personal communication via email, May 24, 2018). Examiners presented task instructions and monitored performance throughout the assessment battery. Breaks and snack packs (e.g., bottled water, nuts, dried fruit, cookies, and/or cheese-and-crackers) were provided to maintain motivation and reduce test fatigue. On average, the study procedures took 3.5 hr to complete (including two separate 10-min breaks, screening, and consent procedures). They received 50 U.S. dollars for completing this part of the study. Following the present study's procedures, participants took a lunch break and then returned for physical health and psychological measures for the larger study (approximately an additional hour).

Measures

Additional measures were collected for the larger study, but only the measures relevant to this study are reported here.

Sociodemographic and health characteristics. Information on demographic (age, sex, race, marital status, living/household status, and years of education) characteristics were collected via clinical interview and self-report measures. Income levels were obtained by a confidential survey that asked participants to best describe their approximate family income (including wages, disability payment, retirement income, and welfare). Participants' family income was categorized into nine levels: <$10,000 (2%), $10,000–$19,999 (14%), $20,000–$29,999 (12%), $30,000–$39,999 (7%), $40,000–$49,999 (13%), $50,000–$59,999 (10%), $60,000–$69,999 (12%), $70,000–$100,000 (13%), and >$100,000 (17%). Economic security status was based on the Elder Index (Gerontology Institute, 2012). The Elder Index was established by the Gerontology Institute at the University of Massachusetts to provide more precise estimates of income needed to meet basic needs. Economic insecurity estimates are based on homeowner status, household size (single vs. dual income), geographic location, and general health status. Almost a third of older adults (31.4%) met criteria for being economically insecure.

Table 1. Validation measures by domain

Cognitive Domains		Neuropsychological Measures
CRYSTALIZED	Reading	**Oral Reading Test Theta**
		WRAT-4 Reading Test Total Correct
	Vocabulary	**Picture Vocabulary Test Theta**
		Peabody Picture Vocabulary Test—Fourth Edition Total Score
FLUID	Memory	**Picture Sequence Memory Test Total Correct**
		Rey Auditory Verbal Learning Test Total Sum Trials 1–5
		Brief Visuospatial Memory Test–Revised Total Sum Trials 1–3
	Working Memory	**List Sorting Working Memory Total Correct**
		Paced Auditory Serial Addition Test (PASAT) Total Correct
		WAIS-IV Digit Span Tests Total Correct
	Executive Function/Speed	**Flanker Inhibitory Control and Attention Test Total Score**
		Dimensional Change Card Sort (DCCS) Test Total Score
		Wisconsin Card Sorting Test (WCST) Total Errors[a]
		D-KEFS Color Word Interference Inhibition Total Score[a]
		Pattern Comparison Processing Speed Test Total Correct
		WAIS-IV Coding Total Correct
		WAIS-IV Symbol Search Total Correct

Notes: Bolded tests indicate an NIHTB-CB measure. Theta scores are based on Item Response Theory.
[a]Items were reversed scored; Wide Range Achievement Test-Version 4 (WRAT-4); Wechsler Adult Intelligence Scale—Fourth Edition (WAIS-IV), Delis-Kaplan Executive Function System (D-KEFS).

Medical history was collected using the clinician administered National Alzheimer's Coordinating Center Uniform Data Set Subject Health History measure (Form A5, Morris et al., 2006). Variables on this measure are categorized as absent, active, remote/inactive, or unknown. Information includes self-reported clinical history of cardiovascular disease, cerebrovascular disease, presence of neurological conditions (e.g., seizures and/or TBI), psychological history (e.g., depression, anxiety, posttraumatic stress disorder), and other biological indicators of health (hypertension, hypercholesterolemia, diabetes, thyroid disease, B12 deficiency). Enrolled participants self-reported receiving treatments for all medical conditions (e.g., depression, diabetes, hypertension, and hypercholesterolemia).

Neuropsychological measures. Table 1 presents the NIHTB-CB and GS tests organized by their posited cognitive domains. The NIHTB-CB is comprised of seven tests designed to measure aspects of executive function (Flanker Inhibitory Control and Attention Test and Dimensional Change Card Sort [DCCS] Test), crystallized knowledge and language (Picture Vocabulary and Oral Reading Tests), episodic memory (Picture Sequence Memory Test), working memory (List Sorting Working Memory: LSWM Test), and processing speed (Pattern Comparison Processing Speed Test).

The GS measures used in this study were the original measures described in detail in the original series of published NIHTB-CB validation studies (Heaton et al., 2014; Weintraub et al., 2014) with the exception of the Wechsler Adult Intelligence Scale-Version 4 (WAIS-IV) Digit Span Tests (forward, backward, and sequencing) replaced the WAIS-IV Letter-Number Sequencing test, given its more common usage and norm availability for older adults (Wechsler, Coalson, & Raiford, 2008). The selected GS tests collectively evaluate crystallized knowledge [Wide Range Achievement Test—4 (WRAT-4) Reading subtest (Wilkinson & Robertson, 2006)] and the Peabody Picture Vocabulary Test—Fourth Edition (PPVT-4; Dunn & Dunn, 2007)], episodic memory [the Brief Visuospatial Memory Test—Revised (Benedict, 1997) and the Rey Auditory Verbal Learning Test (Schmidt, 1996)], executive function [Delis-Kaplan Executive Function System: D-KEFS Color-Word Interference Inhibition Score (Delis, Kaplan & Kramer 2001) and Wisconsin Card Sorting Test-(WCST; Greve (2001); Heaton et al., 1993)], working memory (WAIS-IV Digit Span Tests and first channel of the Paced Auditory Serial Addition Test, Gronwall, 1977), and processing speed (WAIS-IV Coding and Symbol Search). An additional measure of visuospatial reasoning and construction, the WAIS-IV Block Design test, was administered given evidence of it being one of the most reliable predictors of nonverbal fluid reasoning (Lezak, Howieson, Bigler, & Tranel, 2012).

Technology use. A 10-item survey regarding technology use was developed for this study (available in Supplementary material online). Items were based on common indicators of familiarity and comfort with technology usage (frequency of usage, different domains of usage, technology access/ownership, and presence of ambivalence, comfort or distress with usage) and two experimental items specifically designed to evaluate their overall experience with the iPad administered NIHTB-CB battery ("Using the iPad for the study measures was novel and fun" and "Using the iPad for the study measures was intimidating to me").

Participants rated on a five-point Likert scale the degree to which they agreed with statements designed to evaluate their comfort (e.g., "Most technological devices are too difficult or frustrating to learn how to use" and "I feel comfortable using technology") and familiarity (e.g., frequency in using the internet to pay bills, read the news or use social medias) with technology use. When applicable, items were reversed scored to be on a similar scale. Items were then summed to create a total score with a maximum of 50 points. Higher total scores reflected greater comfort with technology use.

Analyses

Preliminary analyses. Statistical power for the primary study aims was computed a priori using G*Power 3.1 (Faul, Erdfelder, Buchner, & Lang, 2009). The current sample of 121 was determined to be more than sufficient to test the studies primary objectives in order to allow for the expected heterogeneity within the study sample. Effect sizes used for estimation were based on data from Heaton and colleagues (2014) and were large as indicated by Cohen's *d* (Cohen, 1988). Statistical analyses were performed via SPSS (Version 26). Casewise deletion was used for five participants' who were unable to complete the NIHTB-CB measures. All other missing data were managed via SPSS Multiple Imputation (MI). SPSS MI procedures use an iterative Markov chain Monte Carlo method (IBM, 2019). Replaced values followed Rubin's recommendations (1996) and were based on predictions from the observed data. The missing data represented less than 1% of total values.

Descriptive statistics were generated for all variables and data distributions were checked to ensure that the assumptions of normality were met. Group stratification was data driven and based on theory. Analyses examined education-related group differences in the NIHTB-CB composites via one-way ANOVAs to reduce the number of comparisons and followed up significant findings with multiple comparisons to determine appropriate group stratification. The multiple comparisons examined for group differences in those with 10–11 years versus 12 years, 10–12 years versus 13–15 years, 13–15 years versus 16 years, 13–16 years versus 17 years, or greater on the NIHTB-CB measures. Following each comparison, groups were collapsed when there was no evidence of significant differences between groups on the measures. This resulted in education being stratified into the following categories: high school (10–12 years), college (13–16 years), and graduate (17 or higher).

Composite formation. The primary analyses and composite score formation followed procedures described in detail by Weintraub and colleagues (2014) in the NIHTB-CB validation series. Table 1 presents the NIHTB-CB and GS validation measures by domain. The NIHTB-CB and GS Crystallized and Fluid Composite scores were formed by first ranking the unadjusted raw scores from the entire sample, and then transforming these values into normally distributed *z*-scores. These *z*-scores were then converted to scaled scores with a mean of 10 and a standard deviation of 3 for each test. Relevant test scores were then summed and averaged to form the respective composites and again transformed and rescaled to have a mean of 10 and a standard deviation of 3. For the GS measures, Processing Speed (average of Symbol Search and Digit Coding scaled scores) and Episodic Memory (average of RAVLT and BVMT scaled scores) composites scores were formed and rescaled prior to forming the GS Fluid composite score. To allow for score comparisons within education groups, split file analyses procedures repeated the described procedures (ranking, transforming, and scaling scores) to compute scores stratified by education groups. Cronbach's alpha evaluated the reliability of internal consistencies of the composite scores, with alpha coefficients equal to 0.70 indicating the absolute minimum for satisfactory reliability and higher alpha levels ($\alpha > 0.80$) being more favorable (Lance, Butts, & Michels, 2006).

Primary analyses. Correlational analyses evaluated the respective concurrent and discriminant validity between the NIHTB-CB Crystallized and Fluid Composite scores with the comparable GS Composite scores. In addition to the original NIHTB-CB validation measures, the associations between the WAIS-IV Block Design scaled scores and the NIHTB-CB composite measures were evaluated. High correlations between the comparable NIHTB-CB and GS measures provided the evidence of convergent validity, while lower correlations between NIHTB-CB and GS measures of a different cognitive construct provided the evidence of discriminant validity. Fisher *r*-to-*z* transformation and Steiger's test were applied to compare the significance of the difference between the relevant correlation coefficients using available online software (Lee & Preacher, 2013).

Sociodemographic-related group differences were evaluated via chi-square tests and MANCOVAs. Pearson and Spearman rank correlations investigated associations among the sociodemographic variables and composite scores. Hierarchal multiple regressions analyses evaluated the effect of the sociodemographic variables and technology use on the Fluid and Crystallized composite scores. Follow-up analyses repeated these procedures to determine whether the respective contributions of these factors were similar to the GS-measures. Adjusted R^2 and standardized beta values are reported for the final regression models as measures of effect size. All tests of significance were two-tailed. The same nomenclature for the effect size of correlation

Table 2. Group differences in participants' characteristics by economic security

Characteristic	Total $N = 121$	Secure $n = 83$	Insecure $n = 38$	$p =$
Age [range 57–87 years]	70.69 (6.47)	70.57 (6.68)	70.97 (6.00)	.758
Education [range 10–20 years]	15.63 (2.69)	16.13 (2.50)	14.40 (2.79)	.001
Race: % Identified White	99.2 ($n = 120$)	99.2%	100%	—
% Female	73.6 ($n = 89$)	66.3%	33.7%	.053
Income range (median dollars)	50,000–59,999	60,000–69,999	10,000–19,999 19$19,999	.001
Montreal Cognitive Assessment	26.21 (2.61)	26.51 (2.49)	25.46 (2.79)	.044
Geriatric Depression Scale	1.08 (1.53)	0.740 (0.923)	1.91 (2.25)	.001
Medical conditions				
Depression	13.2% ($n = 16$)	31.3%	68.8%	.001
Diabetes	6.6% ($n = 8$)	7.0%	5.7%	.410
Hyperlipidemia	43.8% ($n = 53$)	41.9%	48.6%	.739
Hypertension	45.5% ($n = 55$)	40.7%	57.1%	.215

Notes: Chi-square tests (categorical variables) and one-way ANOVAS (continuous variables) investigated group differences. Values indicate Mean (Standard Deviation) unless otherwise noted.

coefficients (<0.3: poor, 0.3–0.6: adequate, and >0.6: good) to determine the evidence of convergent validity was applied (Weintraub et al., 2014).

Post hoc analyses repeated these analyses stratified by an MoCA cut score of 26 to determine if cognitive impairments were responsible for the low internal item consistency. According to the original validation study, a cut score of 26 is deemed to be cognitively normal, while those with scores in the 18–25 range are considered to have MCIs (Nasreddine et al., 2005). The overall pattern of results was highly similar and the overall significance did not change by excluding those with lower MoCA scores; thus, only findings from the entire sample are reported for simplicity.

Results

Descriptive Statistics on Sample Characteristics

Table 2 presents descriptive statistics for the sample by economic security. One-hundred and twenty-one community-dwelling older adults with a broad range of age (range 57–87 years) and years of education (10–20 years) were included within this study. Maine has the highest percentage of non-Hispanic whites of any state (U.S. Census Bureau, 2018), which is reflected in the almost entirely white sample. There was a higher proportion of women to men in the sample. Men and women did not significantly differ in education levels, $F(1, 119) = 1.49$, $p = .225$. There was a trend level difference in age, with men ($M = 72.32$, $SD = 5.77$) being ~2 years older than women ($M = 70.09$, $SD = 6.63$), $F(1, 119) = 2.90$, $p = .091$. Relevant to our analyses, no significant education-related group differences in age $F(2, 118) = 1.87$, $p = .159$ or MoCA scores $F(2, 118) = 1.34$, $p = .266$ were found. Years of education positively associated with income levels ($r = .373$, $p < .001$). Thirty-eight older adults had income levels that fell within the economically insecure range (Elder Index, 2019). Economically insecure individuals had significantly lower years of education and MoCA scores and were more likely to be female and have a history of depression.

Sociodemographic Characteristics Associations with the NIHTB-CB Composite Scores

Similar to the validation series, age was negatively associated with the NIHTB-Fluid ($r = -.445$) and GS-Fluid ($r = -.443$) composites, $ps < .001$. No significant effect on age was found on the NIHTB-Crystallized ($r = -.010$, $p = .911$) and GS-Crystallized ($r = .03$ $p = .746$) composites. MANCOVAs that adjusted for statistical differences in age did not find that men and women significantly differed in their NIHTB-CB Crystallized $F(2, 117) = 0.919$, $p = .340$ or Fluid $F(2, 117) = 2.09$, $p = .151$ composite scores. There were significant group effects of education on the NIHTB-CB Crystallized $F(3, 117) = 5.08$, $p = .002$ and NIHTB-Fluid $F(3, 117) = 22.99$, $p < .001$ composite scores. As hypothesized, multiple comparisons found that those with a high-school education demonstrated lower performance on the NIHTB-CB Crystallized ($p = .039$) and Fluid ($p = .015$) measures than those with "some college"; whereas those with "some college" demonstrated similar performance as compared with those with a college degree on these composites ($ps = .190$ and .886). Additionally, those with a graduate education demonstrated significantly better crystallized knowledge than the other education groups, $p < .001$. These findings supported stratifying education into three levels: High School (10–12 years, $n = 29$), College (13–16 years, $n = 45$), and Graduate (17 years or higher, $n = 47$).

Table 3. The NIHTB-CB fluid composite subdomains' associations with nonverbal intelligence stratified by education groups

NIHTB-CB Fluid Subdomain Tests with PPVT-4	High School	College	Graduate
Executive Function	0.430*	0.021	0.086
Flanker Inhibitory Control and Attention Test			
Dimensional Change Card Sort Test	0.406*	-0.016	-0.031
Episodic Memory: Picture Sequence Memory Test	0.493**	0.031	0.322*
Working Memory (WM): List Sorting WM Test	0.518**	0.305*	0.331*
Processing Speed (PS): Pattern Comparison PS Test	0.285	0.091	0.047

Note: Peabody Picture Vocabulary Test-Version 4 (PPVT-4).
*$p < .05$.
**$p < .001$.

Internal Item Consistency

Cronbach's alphas suggested high internal item consistencies for the NIHTB-CB Crystallized (0.850, $n = 2$), GS Crystallized (0.837, $n = 2$), and GS Fluid (0.825, $n = 8$) composites. Questionable internal item consistency was found for the NIHTB-CB Fluid ($\alpha = 0.615$, $n = 5$) composite. Evaluation of the test items indicated a marginal improvement if the DCCS test was removed from the NIHTB-CB Fluid composite ($\alpha = 0.653$, $n = 4$); this item was thus retained to form the composite given the minor improvement of 0.038. Post hoc analyses investigated whether the lower internal item consistency was primarily driven by those with MoCA scores below 26 (Nasreddine et al., 2005). Investigation into the effect of MoCA scores on reliability found that the scale reliability was substantially higher in those with lower MoCA scores (scores <26) as compared with higher MoCA scores ($\alpha = 0.744$ vs. 0.415). Further examination found adequate interitem reliability in those with a high-school education ($\alpha = 0.764$) but not in those with a college ($\alpha = 0.509$) or graduate education ($\alpha = 0.438$). Examination of the correlations among the measures that make up the NIHTB-CB Fluid Composite indicated weak correlations (defined as ≤0.30), with the exception of the following measures demonstrating moderate associations: episodic memory and working memory (Picture Sequence Memory and LSWM tests, $r = .398$, $p < .001$) and executive function and processing speed (Flanker and PSPC, $r = .511$, $p < .001$).

Convergent and Discriminant Validity

The NIHTB-CB demonstrated good convergent validity with the analogous GS-measures as evidenced by the strong positive associations between the Crystallized ($r = .851$) and Fluid ($r = .799$) composite scores. Inspection for differences in the correlation coefficients by education and economic status groups did not find that the magnitude of effect sizes differed between groups, and these results are thus not reported here for simplicity.

Evidence of discriminant validity was provided by the substantially lower correlations between the NIHTB-CB Crystallized and GS Fluid composites in the college ($r = .273$, $p = .069$) and graduate education ($r = .356$, $p = .013$) groups. Similarly, lower correlations between the NIHTB-CB Fluid and GS Crystallized composites were found in the college ($r = .191$, $p = .208$) and graduate education ($r = .231$, $p = .122$) groups. However, the latter findings must be interpreted with caution given the poor internal item consistency for the NIHTB-CB Fluid composite in the higher education groups. Given this concern, Table 3 presents the associations for NIHTB-CB Fluid measures with the PPVT-4 by education group, as this GS measure was selected to test the discriminant validity of the subdomains in the original validation series (Weintraub et al., 2014). These results indicated adequate discriminant validity for the respective tests that make up the Fluid composite within the college and graduate education groups.

The high correlations among the Crystallized and Fluid composites on the NIHTB-CB and GS-measures indicated poor discriminant validity in the high-school group. Relevantly, this effect was not specific to the NIHTB-CB measures (see Table 4). Results indicated that the magnitude of these correlations significantly differed between those with high-school education and those in the other two education groups. In addition, the WAIS-IV Block Design was significantly associated with the NIHTB-CB Fluid ($r = .691$, $p < .001$) in the high-school group. In contrast, associations among the NIHTB-CB Fluid composite and Block Design were not statistically significant in the college ($r = .276$, $p = .066$) and graduate education ($r = .203$, $p = .171$) groups.

SES and Technology Use

The Technology Use survey's internal item consistency was deemed acceptable (Cronbach's alpha = 0.765). The majority of participants endorsed feeling comfortable with technology ($M = 38.86$ of 50, $SD = 7.52$). Approximately 76% of the participants

Table 4. Correlations among the NIHTB-CB and GS composites in the high-school group

	TB-Crystallized	TB-Fluid	GS-Crystallized
TB-Crystallized	—	—	—
TB-Fluid	0.662**	—	—
GS-Crystallized	0.864**	0.613**	—
GS-Fluid	0.731**	0.691**	0.652**

Note: NIHTB-CB (TB); GS.
**$p < .001$.

reported owning a smartphone or tablet device and over half of the sample reported daily internet usage. The two survey items that evaluated their overall experience with the iPad administered NIHTB-CB battery indicated that 53% of participants agreed that the experience was "novel and fun," while ~35% of the participants endorsed that the iPad administered test was at least somewhat intimidating for them.

In terms of sociodemographic variables, older age ($r = -.237$, $p = .009$) and lower income levels ($r = .304$, $p = .001$) but not education ($r = -.064$, $p = .488$) associated with lower comfort with technology use scores. Economically insecure as compared with secure older adults demonstrated significantly worse NIHTB-CB Fluid performance, $F(1, 118) = 6.25$, $p < .014$, even after adjusting for the significant effect of education. No significant economic-related differences were found in NIHTB-CB Crystalized performance, $F(1, 118) = 0.716$, $p = .399$.

Results also revealed a large effect size for economically insecure as compared with secure older adults reporting being less comfortable with technology use, $F(1, 119) = 15.72$, $p < .001$, Cohen's $d = 74$. Regression analyses adjusting for these relevant variables investigated whether technology use associated with better cognitive performance on the NIHTB-CB and GS-composites. Results indicated that age ($\beta = -0.420$, $p < .001$), education ($\beta = 0.271$, $p = .001$), and technology use ($\beta = 0.206$, $p = .001$) but not income ($\beta = 0.109$, $p = .186$) significantly contributed to the NIHTB-CB Fluid composites in the final model, Adjusted $R^2 = 0.387$, $F(4, 116) = 18.32$, $p < .001$. Technology use ($\beta = 0.173$, $p = .032$) was also a significant predictor of NIHTB-CB Crystallized composite, even after adjusting for the significant effect of education ($\beta = 0.604$, $p < .001$), Adjusted $R^2 = 0.374$, $F(4, 116) = 18.90$, $p < .001$. To better understand these findings, these analyses were repeated with the GS Crystallized and Fluid composite as the dependent variables, in which similar associations with technology use on crystallized knowledge ($\beta = 0.207$, $p = .013$) and fluid cognition ($\beta = 0.262$, $p < .001$) were found in the two respective models (Adjusted R^2 for final models = 0.337 and 0.389).

Discussion

The present study sought to partially replicate previous work on the psychometric properties of the NIHTB-CB and extend it by comparative examinations in terms of sociodemographic characteristics and technology use in a group of socioeconomically diverse older adults. Our data provide further support for the convergent validity of the NIHTB Crystallized and Fluid composites with the analogous GS measures in older adults. However, we found questionable internal item consistency for the NIHTB-CB Fluid composite. There was also evidence of poor discriminant validity for both NIHTB-CB and GS measures in the high-school education group. At last, comfort with technology use was significantly associated with better cognitive test performance; however, as will be discussed, this effect was not specific to the computerized testing medium.

Psychometric Properties of the NIHTB

The high correlations among the NIHTB-CB and GS analogous crystallized and fluid cognition composites suggested good convergent validity. There was no evidence of statistically significant education- or economic-related group differences in these associations. However, the NIHTB-CB Fluid composite findings must be interpreted with caution given evidence of questionable internal item consistency. When considering the effect of education levels, Cronbach's alpha indicated that the internal consistency for this composite was poor in those who completed some college or higher, while its reliability was deemed to be adequate in those with 10–12 years of education (Lance et al., 2006). It was also noted that the NIHTB-CB Fluid composite was not significantly associated with WAIS-IV Block Design performance in the college and graduate education groups. Further examination indicated that the associations among the NIHTB Fluid measures were generally weak, particularly associations with the DCCS test. Here, it is important to note that these findings do not reflect that the respective tests are not reliable but instead indicates that the NIHTB-CB Fluid composite does not consistently capture the posited construct of Fluid cognition.

In all, these findings align with prior research that has indicated that the NIHTB-CB are not equivalent with GS measures of fluid cognition and that it may overestimate and underestimate fluid cognition at tails of the distribution (Scott, Sorrell, & Benitez, 2019). Collectively, these results suggest that caution should be given in interpreting the NIHTB-CB Fluid composite in older adults. Test reliability is a property of the scores from a given sample population and not an inherent feature of tests. Consequently, the overall patterns of scores that make up this composite should be evaluated for consistency prior to its use. Future work is needed to better understand these differences as its usage could potentially result in diagnostic errors in older adults.

Sociodemographic Effects on Neuropsychological Testing

Consistent with prior research, the negative correlation coefficients between age and the NIHTB-CB and GS Fluid measures were highly comparable and likely reflect the known effect of older age on the Fluid but not Crystallized composite scores. As hypothesized and consistent with previous research, those with only a high-school education demonstrated significantly lower Fluid and Crystallized composite scores than those with "some college"; whereas those with "some college" had similar performance to those with a college degree. These findings add to the evidence that "some college" and "high-school" groups are not comparable in cognitive performance and caution should be taken before collapsing these two groups in cognitive aging research.

When considering the effect of education, both the NIHTB-CB and GS measures demonstrated poor discriminant validity in the high school but not the college-educated groups. Education effects on neuropsychological tests are widely recognized, with evidence indicating that even tests that should be relatively spared from its effects (e.g., digit span and visual memory tests) demonstrate fairly strong associations with education levels (Lezak et al., 2012). Potential reasons underlying the robust relationship between education and cognitive testing are complex and the degree to which lower cognitive scores on normative tests reflects brain-pathology as opposed to environmental and cultural differences in test taking in lower education populations is debatable (Manly, 2005; Romero et al., 2009). Lower years of education is a risk factor for experiencing cognitive decline, and poor discriminant validity paired with significantly lower cognitive scores may be an indication of loss of specificity of brain function (e.g., dedifferentiation theories); however, while this pattern is generally consistent with our findings, in the absence of neuroanatomical evidence, it is not within the scope of this study to determine the reason for the high correlations among measures within the lower education group.

Clinically, it is not uncommon for lower education or non-Westernized individuals to be highly capable of completing fairly complex tasks that rely on higher order cognitive processes (e.g., working memory and executive function) in real-world situations yet be deemed impaired on neuropsychological tests of the same cognitive function. It is possible that the development of novel measures that reflect one's adaptive functions may lead to more ecologically valid measures of cognitive function in less educated and non-Westernized samples.

Technology Use and Testing

The majority of older adults reported being comfortable with technology use and over 76% of the older adult participants reported owning a smartphone or tablet device. Over 53% of participants endorsed that the tablet administered testing experience was "novel and fun." However, there was a large effect size for economically insecure as compared with secure older adults reporting being less comfortable with technology use, and ~35% of the participants endorsed to varying degrees (a little to very much) that "Using the iPad for the study measures was intimidating to me." Older age and lower income levels associated with less comfort with technology use. In terms of cognitive performance, after adjusting for the significant effect of older age and lower income levels, comfort with technology use associated with better crystallized and fluid cognition performance irrespective of the test medium used. Collectively, our findings are consistent with evidence that there is a significant socioeconomic-related digital divide in comfort with technology use. However, we did not find evidence that cognitive performance was significantly worse on the computerized as compared with paper-pencil measures. Taken together, it is possible that better cognitive function contributes to adapting to technology use. In this respect, the degree of technology assimilation may serve as a useful proxy measure of adaptive function rather than it automatically being viewed as a confounding factor.

The use of computerized testing has the capacity to reach underserved populations be it in rural areas or more populated areas with low density of providers. However, the proliferation of computerized cognitive screening measures in medical settings by nonpsychologists has increased concern about levels of training in and who is responsible for the interpretation of computer administered psychological tests (Roebuck-Spencer et al., 2017). Notably, in addition to person-related factors, technical factors to include knowledge of software and updates also impact test administration and reliability (Roebuck-Spencer et al., 2017). Furthermore, computerized testing still poses some significant challenges for its use in older adults, especially those with

socioeconomically and/or culturally diverse backgrounds, as the interactions between diversity and technology remain unclear. Given these concerns, it will be important to continually evaluate the psychometric properties of computerized assessments in relation to ethical, technical (e.g., internet connectivity), and environmental factors guidelines (Bauer et al., 2012).

Study Strengths and Limitations

The novelty of this study includes the geographic location where the study took place and that study visits were conducted at four low-income community-dwelling residences to selectively enhance the sample for more socioeconomically diverse older adults. While our sample of 121 participants is relatively low compared with some studies, it is similar to the 108 older adults that were included in the Heaton and colleagues (2014) study. We also had conducted a priori power analyses to ensure the sample size was adequate for the primary aims. While we did not capture data on the exact geographical location of participants, it is worth noting that Maine is the most rural state in the nation with a particularly low-population density in more northern regions of Maine (e.g., Penobscot county population is 45.3 people per square mile; U.S. Census Bureau, 2010b). This is in contrast to the majority of research that has taken place at large medical centers or university-based sites that are located within major urbanized areas with a high-population density (e.g., the city of Evanston, a primary location for the Heaton study, has a population estimate of 9,575.3 people per square mile; U.S. Census Bureau, 2010a). This study thus provides an entry point to understanding the psychometric properties of the NIHTB-CB in more rural areas of the United States. However, more large-scale studies are needed to clarify associations between cognitive test scores with SES-related factors, including rural as compared with urban areas (Weden, Shih, Kabeto, & Langa, 2017).

As noted above, a strength of this study is the use of CBPR procedures to enhance the representation of noncollege-educated and lower income older adults through working with community stakeholders and targeted recruitment procedures. Using CBPR approaches, 23% of participants had a high-school education or less, as well as almost a third of the older adults in our sample were deemed economically insecure (Gerontology Institute, 2012). Testing was also conducted at multiple community-based sites and weekend testing was offered to reduce barriers to participation. Other strengths include that the time of day for testing was controlled for in this study by scheduling only morning appointments and inclusion of procedures to enhance participant comfort (e.g., provision of snacks and timely breaks).

We also sought to increase the representativeness of our sample by widening our exclusion criteria. Specifically, although we initially set our exclusion criteria to a cut score of 23 based on meta-analytic findings that this score optimizes diagnostic accuracy in older and less educated adults (Carson et al., 2018), it was noted that a disproportionate number of older adults with education of 12 years or less were being excluded from the study based on their MoCA score. Given that lower education populations are underrepresented in research, we revised the criteria to a cut-score of 18 to better understand the relationship between lower education and the neuropsychological measures. Notably, even when the more stringent MoCA cut score of 26 was applied to stratify the sample, our follow-up analyses indicated that the overall pattern and statistical significance of findings were still supported. These findings provide support that the observed low internal item consistency was not being driven by those with greater cognitive impairments as removing participants with lower MoCA scores did not improve the internal item consistency for this composite.

Despite these strengths, the data presented need to be interpreted within the context of the study's limitations. Limitations to the present study include the almost entirely white sample, which is reflective of the 94.7% non-Hispanic white population estimate for the state of Maine (U.S. Census Bureau, 2018). There was also a higher proportion of women to men in the sample, which may affect the generalizability of these findings. As the high-school group ranged from 10 to 12 years of education, it is also possible that those who completed high school may differ in cognitive reserve from those who did not obtain their high-school degrees. Finally, as this study sought to replicate previous work, we did not counterbalance the NIHTB-CB and GS tests. It is thus possible that there is an order effect as the NIHTB-CB was always administered first. It is also worth noting relevant measures, such as delayed memory, were not included in the validation study. Future psychometric studies of the NIHTB-CB that control for order effects and investigate delayed memory function in diverse populations are needed.

Summary

Given the large aging population and anticipated increase in the prevalence of dementia, the need for time efficient, sensitive, and reliable cognitive tests to measure cognitive impairments in diverse older adult populations is unprecedented. Our work supports that there is a significant socioeconomic-related digital divide in comfort with technology use; however, its effect on cognitive testing was not specific to the computerized measures and is worthy of further investigation (e.g., evaluating the role of test anxiety). The current study focused on the NIHTB-CB, a time efficient fully computerized neuropsychological battery. Its growth in popularity and ease of use has created some early calls for its use in clinical trials and as a clinical screening measure

in older adults. This call for its use however may be premature given evidence that its measures may not demonstrate similar psychometric properties as the GS pencil-paper measures in older adults. Relevantly, our results also revealed that both NIHTB-CB and GS measures demonstrated a lack of discriminant validity, potentially suggesting poor specificity, in noncollege-educated older adults.

At the forefront of the field of clinical neuropsychology is concern regarding the adequacy of existing norms for culturally, ethnically, racially, and socioeconomically diverse populations. Comparatively, there is a paucity of cognitive aging research in noncollege-educated, less affluent, older adults, particularly in respect to research that incorporates use of biomarkers and imaging to support the diagnostic conclusions made about low SES populations.

To move the field the forward, it is critical as a field that we commit to obtaining more diverse and clinically representative samples when developing norms and conducting clinical trials to address the problem of specificity. This study provides initial support that widening inclusion/exclusion criteria and increasing testing locations in the community may be a viable way of capturing underrepresented older adults in cognitive aging research. Relevantly, demographic adjustments are only a proxy for education quality, SES, race, and culture; hence, they do not address the unique contributions of these factors on neuropsychological test performance (Manly, 2005). Consequently, while adjusting for years of education, intellectual abilities and other sociodemographic factors can improve diagnostic accuracy in some situations, such adjustments need to be applied with caution as there is the strong possibility of overpathologizing and/or misinforming treatment and services in low SES populations (Romero et al., 2009). Future work is thus urgently needed to disentangle the effect of education from cultural, racial, and socioeconomic disparities effects on brain health.

Supplementary material

Supplementary material is available at *Archives of Clinical Neuropsychology* online.

Acknowledgments

The authors would like to thank community supporters and participants of the M-ABLE study and acknowledge our research assistants (Lisa D'Errico, Michael Fagan, and Savannah Michaud) who assisted in the data collection and scoring.

Funding

This work was supported by a National Academy of Neuropsychology Clinical Research Grant and the University of Maine, Maine Economic Improvement Fund provided to the Primary Investigator (RKM).

Conflicts of Interest

No disclosures to report.

References

Anderson, M., & Perrin, A. (2017). *Technology use among seniors*. Washington, DC: Pew Research Center for Internet & Technology. Retrieved from https://www.silvergroup.asia/wp-content/uploads/2017/07/Technology-use-among-seniors-_-Pew-Research-Center.pdf.

Bauer, R. M., Iverson, G. L., Cernich, A. N., Binder, L. M., Ruff, R. M., & Naugle, R. I. (2012). Computerized neuropsychological assessment devices: Joint position paper of the American Academy of Clinical Neuropsychology and the National Academy of Neuropsychology. *Archives of Clinical Neuropsychology*, *27*(*3*), 362–373. doi: 10.1093/arclin/acs027.

Benedict, R. H. (1997). *Brief visuospatial memory test—revised (BVMT-R)*. Odessa, FL: Psychological Assessment Resources.

Cadar, D., Lassale, C., Davies, H., Llewellyn, D. J., Batty, G. D., & Steptoe, A. (2018). Individual and area-based socioeconomic factors associated with dementia incidence in England: Evidence from a 12-year follow-up in the English longitudinal study of ageing. *JAMA Psychiatry*, *75*(*7*), 723–732.

Carson, N., Leach, L., & Murphy, K. J. (2018). A re-examination of Montreal Cognitive Assessment (MoCA) cutoff scores. *International Journal of Geriatric Psychiatry*, *33*(*2*), 379–388. doi: 10.1002/gps.4756.

Casaletto, K. B., Umlauf, A., Beaumont, J., Gershon, R., Slotkin, J., Akshoomoff, N., et al. (2015). Demographically corrected normative standards for the English version of the NIH toolbox cognition battery. *Journal of the International Neuropsychological Society*, *21*(*5*), 378–391. doi: 10.1017/S1355617715000351.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York, NY: Routledge Academic.

Delis, D. C., Kaplan, E., & Kramer, J. H. (2001). *The Delis-Kaplan Executive Function System*. San Antonio, TX: Psychological Corporation. doi:10.1037/t15082-000.

Dunn, L. M., & Dunn, D. M. (2007). *PPVT-4: Peabody picture vocabulary test*. Minneapolis, MN: Pearson Assessments.

Elder Index. (2019). *The Elder Index™ [public dataset]*. Boston, MA: Gerontology Institute, University of Massachusetts Boston. Retrieved from https://elderindex.org/.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160.

Heaton, R. K., Chelune, G. J., Talley, J. L., Kay, G. G., & Curtiss, G. (1993). *Wisconsin Card Sorting Test manual: revised and expanded*. Odessa, FL: Psychological Assessment Resources.

Gerontology Institute, University of Massachusetts Boston (2012). *The National Elder Economic Security Standard Index*. Boston, MA: Gerontology Institute Publications. 75. Retrieved from https://scholarworks.umb.edu/gerontologyinstitute_pubs/75.

Gershon, R. (2019). *ARMADA: Advancing reliable measurement in Alzheimer's disease and cognitive aging*. National Institute of Health-NIH Reporter. Available from: https://projectreporter.nih.gov/project_info_description.cfm?projectnumber=5U2CAG057441-03. Accessed on December 7, 2020.

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*(*Suppl. 3*), S2–S6. doi: 10.1212/WNL.0b013e3182872e5f.

Greve, K. W. (2001). The WCST-64: A standardized short-form of the Wisconsin card sorting test. *The Clinical Neuropsychologist*, *15*(*2*), 228–234. doi: 10.1076/clin.15.2.228.1901.

Gronwall, D. M. A. (1977). Paced auditory serial-addition task: A measure of recovery from concussion. *Perceptual and Motor Skills*, *44*(*2*), 367–373. doi: 10.2466/pms.1977.44.2.367.

Heaton, R. K., Akshoomoff, N., Tulsky, D., Mungas, D., Weintraub, S., Dikmen, S., et al. (2014). Reliability and validity of composite scores from the NIH toolbox cognition battery in adults. *Journal of the International Neuropsychological Society*, *20*(*6*), 588–598.

Holdnack, J. A., Tulsky, D. S., Brooks, B. L., Slotkin, J., Gershon, R., Heinemann, A. W., et al. (2017). Interpreting patterns of low scores on the NIH toolbox cognition battery. *Archives of Clinical Neuropsychology*, *32*(*5*), 574–584.

IBM Corp (2019). *IBM SPSS Missing Values 26*. Armonk, NY: IBM Corp.

Iverson, G. L., Brooks, B. L., Ashton, V. L., Johnson, L. G., & Gualtieri, C. T. (2009). Does familiarity with computers affect computerized neuropsychological test performance? *Journal of Clinical and Experimental Neuropsychology*, *31*(*5*), 594–604. doi: 10.1080/13803390802372125.

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods*, *9*(*2*), 202–220. doi: 10.1177%2F1094428105284919.

Lee, I. A., & Preacher, K. J. (2013). *Calculation for the test of the difference between two dependent correlations with one variable in common [Computer software]*. Retrieved from http://quantpsy.org.

Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel (2012). *Neuropsychological assessment*. New York, New York: Oxford University Press.

Manly, J. J. (2005). Advantages and disadvantages of separate norms for African Americans. *The Clinical Neuropsychologist*, *19*(*2*), 270–275.

Morris, J. C., Weintraub, S., Chui, H. C., Cummings, J., DeCarli, C., Ferris, S., et al. (2006). The uniform data set (UDS): Clinical and cognitive variables and descriptive data from Alzheimer disease centers. *Alzheimer Disease & Associated Disorders*, *20*(*4*), 210–216. doi: 10.1097/01.wad.0000213865.09806.92.

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., et al. (2005). The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, *53*(*4*), 695–699. doi: 10.1111/j.1532-5415.2005.53221.x.

Roebuck-Spencer, T. M., Glen, T., Puente, A. E., Denney, R. L., Ruff, R. M., Hostetter, G., et al. (2017). Cognitive screening tests versus comprehensive neuropsychological test batteries: A national academy of neuropsychology education paper. *Archives of Clinical Neuropsychology*, *32*(*4*), 491–498.

Romero, H. R., Lageman, S. K., Kamath, V., Irani, F., Sim, A., Suarez, P., et al. (2009). Challenges in the neuropsychological assessment of ethnic minorities: Summit proceedings. *The Clinical Neuropsychologist*, *23*(*5*), 761–779.

Rubin, D. B. (1996). Multiple imputation after 18 years. *Journal of the American Statistical Association*, *91*, 473–489.

Schmidt, M. (1996). *Rey auditory verbal learning test: A handbook*. Los Angeles, CA: Western Psychological Services.

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*(*8*), 857–867.

Sheikh, J. I., & Yesavage, J. A. (1986). Geriatric depression scale (GDS): Recent evidence and development of a shorter version. *Clinical Gerontologist: The Journal of Aging and Mental Health*, *5*(*1–2*), 165–173. doi: 10.1300/J018v05n01_09.

Stern, Y. (2006). Cognitive reserve and Alzheimer disease. *Alzheimer Disease & Associated Disorders*, *20*(*2*), 112–117.

U.S. Census Bureau (2010a). *QuickFacts Evanston city*, Illinois. Retrieved March 11, 2021 from https://www.census.gov/quickfacts/evanstoncityillinois.

U.S. Census Bureau (2010b). *QuickFacts Penobscot County*, Maine. Retrieved March 11, 2021 from https://www.census.gov/quickfacts/fact/table/penobscotcountymaine/POP060210#POP060210.

U.S. Census Bureau (2018). *QuickFacts Maine*. Retrieved March 11, 2021 from https://www.census.gov/quickfacts/ME.

Wechsler, D., Coalson, D. L., & Raiford, S. E. (2008). *WAIS-IV. Wechsler adult intelligence scale: Fourth edition. Technical and interpretative manual*. San Antonio, TX: NCS Pearson.

Weden, M. M., Shih, R. A., Kabeto, M. U., & Langa, K. M. (2017). Secular trends in dementia and cognitive impairment of US rural and urban older adults. *American Journal of Preventive Medicine*.

Weintraub, S., Dikmen, S. S., Heaton, R. K., Tulsky, D. S., Zelazo, P. D., Slotkin, J., et al. (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. doi: 10.1017/S1355617714000320.

Wiederholt, W. C., Cahn, D., Butters, N. M., Salmon, D. P., Kritz-Silverstein, D., & Barrett-Connor, E. (1993). Effects of age, gender and education on selected neuropsychological tests in an elderly community cohort. *Journal of the American Geriatrics Society*, *41*(*6*), 639–647. doi: 10.1111/j.1532-5415.1993.tb06738.x.

Wilkinson, G. S., & Robertson, G. J. (2006). *Wide range achievement test 4 (WRAT4)*. Lutz, FL: Psychological Assessment Resources.

Zahodne, L. B., Stern, Y., & Manly, J. J. (2015). Differing effects of education on cognitive decline in diverse elders with low versus high educational attainment. *Neuropsychology*, *29*(*4*), 649–657.

Zhang, M., Gale, S. D., Erickson, L. D., Brown, B. L., Woody, P., & Hedges, D. W. (2015). Cognitive function in older adults according to current socioeconomic status. *Aging, Neuropsychology, and Cognition*, *22*(*5*), 534–543.