Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring 10 (2018) 764-772

Cognitive & Behavioral Assessment

Utility of the NIH Toolbox for assessment of prodromal Alzheimer's disease and dementia

Katherine Hackett[a,*], Robert Krikorian[b], Tania Giovannetti[a], Josefina Melendez-Cabrero[c], Aneela Rahman[f], Emily E. Caesar[d], Jaclyn L. Chen[e], Hollie Hristov[f], Alon Seifan[g], Lisa Mosconi[f], Richard S. Isaacson[f]

[a]*Department of Psychology, Temple University, Philadelphia, PA, USA*
[b]*Department of Psychiatry and Behavioral Neuroscience, University of Cincinnati College of Medicine, Cincinnati, OH, USA*
[c]*Department of Neurology, Weill Cornell Medicine, San Juan, PR, USA*
[d]*Loyola-Stritch School of Medicine, Chicago, IL, USA*
[e]*Stony Brook University School of Medicine, New York, NY, USA*
[f]*Department of Neurology, Weill Cornell Medicine and NewYork-Presbyterian, New York, NY, USA*
[g]*Compass Health Systems, Miami, FL, USA*

Abstract

Introduction: The NIH Toolbox Cognition Battery (NIHTB-CB) is a computer-based protocol not yet validated for clinical assessment.

Methods: We administered the NIHTB-CB and traditional neuropsychological tests to 247 Memory Disorders and Alzheimer's Prevention Clinic patients with subjective cognitive decline, mild cognitive impairment, mild dementia due to Alzheimer's disease, and normal cognition. Principal component analysis, partial correlations, and univariate general linear model tests were performed to assess construct validity. Discriminant function analyses compared classification accuracy.

Results: Principal component analysis identified three conceptually coherent factors: memory (MEM_{NIH}), executive function (EF_{NIH}), and crystallized intelligence (CI_{NIH}). These factors were strongly associated with corresponding traditional tests and differed across diagnostic groups as expected. Both NIHTB and traditional batteries yielded strong overall discriminative ability (>80%).

Discussion: The NIHTB-CB is a valid method to assess neurocognitive domains pertinent to aging and dementia and has utility for applications in a memory clinic setting.

Keywords: Neuropsychology; Clinical assessment; Validation; Dementia; Memory; Technology

1. Introduction

Alzheimer's disease (AD) affects over 26 million people worldwide, a prevalence expected to quadruple by 2050 [1]. The pathophysiological process of AD begins years before clinical symptoms emerge, providing an opportunity for intervention in at-risk individuals [2]. Neuropsychological assessment facilitates diagnosis of AD and improves identification of individuals at risk for dementia [2], but neuropsychological testing can be a lengthy and resource-intensive process, which is not feasible in all clinics. As preventive therapies become available [3–5], efficient and sensitive assessments will be needed for longitudinal evaluations of large samples. Computerized measures offer standardized administration, streamlined scoring and analysis, access to large normative data sets, and automated capture of reaction time and other sensitive response parameters, making them highly effective for detection of subtle cognitive impairment [2,6].

The authors have declared that no conflict of interest exists.

*Corresponding author. Tel.: +914-582-7581; Fax: +215-204-5539.
E-mail address: katherine.hackett@temple.edu

https://doi.org/10.1016/j.dadm.2018.10.002

The NIH Toolbox Cognition Battery (NIHTB-CB) is a computer-administered protocol that assesses key cognitive domains [7]. It demonstrates strong convergent validity against traditional neuropsychological tests in cognitively healthy individuals [8], with normative data available across the life span [9,10]. The NIHTB-CB was not designed for clinical assessment, although investigations are underway to explore its validity in traumatic brain injury, spinal cord injury, and stroke [11–14]. In clinically normal older adults, the NIHTB-CB correlated with traditional tests and was able to identify subtle cognitive impairment [15]. To our knowledge, the NIHTB-CB has not yet been validated in samples with or at risk for AD, a critical gap in investigating its utility in the context of AD staging.

The aim of this study was to investigate the validity and utility of the NIHTB-CB in a clinical setting with patients across the continuum of cognitive decline, including patients seeking risk reduction care. The factor structure of the NIHTB-CB was examined and validated against traditional neuropsychological tests. The discriminative accuracy of the NIHTB-CB for diagnostic grouping was also evaluated. Finally, given the importance of episodic memory assessment for AD, we examined the validity and utility of the NIHTB-CB protocol with and without the delayed recall measure of the Rey Auditory Verbal Learning Test (RAVLT-DR), which is not included in the standard NIHTB-CB protocol.

2. Methods

2.1. Participants

Participants provided informed consent for the IRB-approved Comparative Effectiveness Dementia & Alzheimer's Registry at Weill Cornell Medicine/NewYork-Presbyterian Memory Disorders and Alzheimer's Prevention Clinic. The sample included patients with a range of cognitive conditions from those with family history of AD and no cognitive complaint seeking risk assessment for early intervention to those with advanced cognitive decline. Participants underwent a clinical interview (with informant), medical and neurological examinations, anthropometric and laboratory measures, neuropsychological testing, and structural brain MRI when indicated. Diagnoses were assigned following a consensus conference including a neurologist, family nurse practitioner, and multidisciplinary health care team members who integrated results from (1) clinical history and informant/self-report questionnaires to determine onset and type of cognitive and/or non-cognitive symptoms, ability to perform activities of daily living, and subjective assessment of cognitive function; (2) physical and neurological examination, including brain MRI when available; and (3) the Mini–Mental State Examination [16]. Diagnosis of subjective cognitive decline (SCD) required self-reported decline in memory over the past year according to a standardized questionnaire [17] paired with expressed concern about cognitive difficulties to the clinician, in the absence of objective evidence of impaired cognition [18]. Diagnoses of mild cognitive impairment (MCI) and dementia due to AD were determined using the most recent clinical diagnostic criteria [19,20]. Amnestic and nonamnestic presentations of MCI were classified after consideration of scores on the delayed recall subtest of the Mini–Mental State Examination, along with description and type of cognitive complaint (problems with memory vs. attention, language, or processing speed) [21].

The current sample included 247 patients with the following diagnoses: cognitively normal (CN; n = 129), SCD (n = 46), amnestic MCI (aMCI; n = 27), nonamnestic MCI (naMCI; n = 19), and mild dementia due to AD (n = 26).

2.2. Cognitive measures

All tests were administered during a single visit (under 75 minutes) by research assistants. Five core and two supplemental NIHTB-CB measures [22] were administered, along with traditional neuropsychological measures [16,23–28] commonly used in dementia evaluations [28–32] (see Table 1).

2.3. Statistical methods

The NIHTB Picture Sequence Memory and List Sorting Working Memory subtests [22] were too challenging for participants with cognitive impairment and were not included due to low completion rates. Of the 247 patients, 52 (21%) had some missing data on the remaining NIHTB-CB protocol. Completion rates did not differ across groups (CN 81%, SCD 89%, MCI 78%) with the exception of AD (53%), where some participants completed abbreviated protocols due to time constraints, inability to complete tests, or because comprehensive testing was not clinically indicated. Missing data for the NIHTB-CB were handled with list-wise deletion.

Principal component analysis (PCA) was performed to examine the factor structure of the NIHTB tests. To test the model's applicability in people without dementia, PCA was repeated without the AD group. To test the model's utility without the additional RAVLT-DR subtest, PCA was repeated without RAVLT-DR. Factor-based scores for the NIHTB and traditional tests were computed using averaged z-scores of tests that strongly loaded onto each factor. Most factor score distributions were not normally distributed according to tests (Shapiro-Wilk, $P < .001$; Kolmogorov-Smirnov, $P < .001$) and associated histograms. Therefore, both nonparametric and parametric tests were performed.

Nonparametric Spearman's partial correlations were performed to assess relations between NIHTB and traditional factor scores within and across corresponding cognitive domains, controlling for age, sex, and education.

Table 1
Neuropsychological measures

Cognitive domain	NIHTB-CB tests	Traditional tests
Learning/Memory	RAVLT 1, 2, 3*	MMSE-DR*
	RAVLT-DR*	Logical Memory immediate recall*
		Logical Memory delayed recall*
		FNAME*
Executive function/Attention/Processing speed	DCCS†	FAS*
	Flanker†	ANT*
	Pattern Comparison*	MMSE-attention*
	ODS*	Trail-Making Test Part B*
Crystallized intelligence	Picture Vocabulary†	
	Oral Reading Recognition†	

Abbreviations: NIHTB-CB, NIH Toolbox Cognition Battery; RAVLT 1, 2, 3, Rey Auditory Verbal Learning Task immediate recall trials 1–3; RAVLT-DR, Rey Auditory Verbal Learning Task delayed recall; DCCS, Dimensional Change Card Sort; Flanker, Flanker Inhibitory Control/Attention; Pattern Comparison, Pattern Comparison Processing Speed; ODS, Oral Digit Symbol; MMSE-DR, Mini–Mental State Examination delayed recall subscore; FNAME, Face Name Associative Memory-cued first letter; FAS, verbal fluency under phonemic constraint to letters F-A-S; ANT, verbal fluency under categorical constraint (animals); MMSE-attention, Mini–Mental State Examination attention subscore.

NOTE. Trail-Making Test Part B score represents time to completion (seconds).

NOTE. Raw and computed scores are unadjusted for demographics.

*Raw score.

†Computed score (provided by the NIH toolbox, used for computer adaptive tests and tests whose score requires combination of accuracy and reaction time vectors).

Univariate general linear model tests with Fisher's least significant difference pairwise comparisons of estimated marginal means were used to test group differences on continuous demographic variables and on NIHTB and traditional factor scores. χ^2 tests were used to compare categorical variables (sex and race) across groups.

Discriminant function analyses examined cognitive factor scores as predictors of diagnostic group reclassification (including CN, aMCI, naMCI, and AD). SCD was not included as we did not expect correspondence with consensus diagnoses when classifying SCD using cognitive scores alone, and including this group would limit overall predictive power. Factor scores from the three unique PCAs were entered as independent predictors in three separate discriminant analyses: (1) scores from PCA 1a (NIHTB-CB including RAVLT-DR); (2) scores from PCA 2a (NIHTB-CB excluding RAVLT-DR); and (3) scores from PCA 3 (traditional tests). Cohen's κ was calculated to account for chance agreement due to differences in diagnostic group sizes. Age, sex, and education were also entered as independent variables in all three analyses. Results were considered significant at $P < .05$.

3. Results

Sample characteristics are shown in Table 2. Participants were 27 to 94 years old, with a mean age of 61 (±15 years). AD, aMCI, and naMCI groups were older than CN and SCD groups (P's < 0.01).

3.1. PCA of NIHTB-CB

PCA was conducted on all 10 NIHTB tests with orthogonal rotation (PCA 1a; Table 3). Three factors were identified, explaining 83% of the total variance: (1) learning/memory (MEM_{NIH}; where $_{NIH}$ = NIHTB); (2) executive function (EF_{NIH}); and (3) crystallized intelligence (CI_{NIH}). The factor structure remained unchanged after excluding participants with AD (PCA 1b; Supplementary Table 1).

PCA conducted on the NIHTB tests excluding RAVLT-DR identified a solution with only two factors explaining

Table 2
Demographic characteristics of participants by diagnostic group

	Total (N = 247)	CN (N = 129)	SCD (N = 46)	naMCI (N = 19)	aMCI (N = 27)	AD (N = 26)	P
Age (mean, SD), years	61.0 (14.7)	52.5 (12.0)	62.9 (11.2)	71.8 (8.1)	74.3 (7.2)	78.4 (8.5)	<.001*
Age range	27–94	27–82	34–80	56–88	58–86	66–94	
Sex (% female)	52%	55%	63%	32%	48%	38%	n.s.
Education (mean, SD)	15.6 (1.2)	15.9 (1.0)	15.6 (1.0)	15.4 (1.3)	15.4 (1.7)	14.7 (1.7)	.001†
Race (% white)	65%	62%	72%	63%	67%	69%	n.s.
Total MMSE	28.2 (3.3)	29.5 (.8)	29.4 (1.4)	28.7 (1.4)	26.3 (2.2)	19.8 (4.4)	<.001*

Abbreviations: CN, cognitively normal; SCD, subjective cognitive decline; naMCI, nonamnestic MCI; aMCI, amnestic MCI; AD, Alzheimer's disease; MMSE, Mini–Mental State Examination; n.s., nonsignificant.

NOTE. MMSE P values reflect General Linear Model test of between-subjects effect after covarying for age, sex, and education.

*$P < .001$.

†$P < .01$.

Table 3
PCAs of NIHTB-CB tests

	PCA 1a (including RAVLT-DR, N = 197)			PCA 2a (excluding RAVLT-DR, N = 198)	
	Component			Component	
	1	2	3	1	2
	Memory (MEM_{NIH})	Executive function (EF_{NIH})	Crystallized intelligence (CI_{NIH})	Executive function/working memory (EF/WM_{NIH})	Crystallized intelligence (CI_{NIH})
RAVLT Trial 2	0.886			0.823	
RAVLT-DR	0.853				
RAVLT Trial 3	0.852			0.838	
RAVLT Trial 1	0.818			0.711	
Flanker		0.897		0.842	
DCCS		0.874		0.84	
Pattern Comparison		0.833		0.867	
ODS		0.741		0.876	
Oral Reading Recognition			0.89		0.891
Picture vocabulary			0.702		0.714
% Explained variance	35%	33%	15%	55%	17%
Total % explained variance		83%		72%	

Abbreviations: NIHTB-CB, NIH Toolbox Cognition Battery; PCA, principal component analysis; RAVLT, Rey Auditory Verbal Learning Test immediate recall; RAVLT-DR, Rey Auditory Verbal Learning Task delayed recall; Flanker, Flanker Inhibitory Control/Attention; DCCS, Dimensional Change Card Sort; Pattern Comparison, Pattern Comparison Processing Speed; ODS, Oral Digit Symbol.

NOTE. PCA was conducted using varimax with Kaiser Normalization. Factors with eigenvalues greater than 1, with a maximum iteration of 25, were extracted—factor loadings shown after orthogonal rotation.

NOTE. PCA 1a rotation converged in 5 iterations. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy = 0.90.

NOTE. PCA 2a rotation converged in 3 iterations. KMO = 0.89.

72% of the variance (PCA 2a; Table 3): (1) EF/working memory (EF/WM_{NIH}) and (2) CI_{NIH}. A final PCA without RAVLT-DR and excluding AD participants yielded a three-factor solution, similar to PCA 1a, explaining 79% of the variance (PCA 2b; Supplementary Table 2).

As we were interested in assessing the utility of the NIHTB-CB in a clinical population, factor-based scores from PCAs including all diagnostic groups were calculated (PCAs 1a and 2a). All tests had adequate unidimensionality (factor loadings at or above 0.70 for a single latent construct) [33] and Kaiser-Meyer-Olkin sampling adequacy [34].

3.2. PCA of traditional tests

PCA of traditional tests identified two factors, explaining 67% of the total variance (PCA 3; Table 4): (1) memory (MEM_T; where $_T$ = traditional tests) and (2) executive function (EF_T). All tests demonstrated adequate unidimensionality and Kaiser Meyer Olkin sampling adequacy [33,34]. Z-scores for each test uniquely loaded onto each component were averaged to calculate factor-based scores for MEM_T and EF_T from PCA 3. Descriptive data for all factor scores are shown in Supplementary Fig. 1.

3.3. Associations between NIHTB and traditional tests

Nonparametric partial correlations controlling for age, sex, and education indicated that NIHTB factor scores were significantly correlated with corresponding traditional factor scores assessing the same cognitive domain (P's < 0.01), demonstrating convergent validity (Table 5). Evidence of discriminant validity included lower correlations with traditional factor scores of different cognitive domains. Similar associations were found when asymptomatic (CN and SCD) versus symptomatic (MCI and AD) groups were analyzed separately (Supplementary Table 3).

We evaluated the construct validity of CI_{NIH} against education, a common proxy for IQ. Bivariate nonparametric Spearman's correlations showed a significant relation between CI_{NIH} and education ($r_s = 0.468$, $P < .001$), and Fisher's Z tests showed that relation was stronger than the relation between education and all other NIHTB factors ($r_s = 0.312$ with MEM_{NIH}; $r_s = 0.171$ with EF_{NIH}; $r_s = 0.253$ with EF/WM_{NIH}; P's < 0.05). Correlations between CI_{NIH} and education also were observed within the asymptomatic versus symptomatic groups (P's < 0.05; Supplementary Table 3).

3.4. Group differences in factor scores

Independent-samples Kruskal-Wallis tests for nonparametric data demonstrated significant differences on each factor score (NIHTB-CB and traditional) across diagnostic groups (P's < 0.01). Additional parametric analyses (univariate general linear model) were performed to covary for age, sex, and education when examining differences between groups (Supplementary Fig. 1). Overall, analyses showed a significant effect of group for each factor score, and pairwise comparisons demonstrated expected relative performance trends according to type and level of cognitive impairment (CN/SCD < MCI < AD). Detailed differential

performance data on factor scores between diagnostic groups are found in the Supplementary Fig. 1.

3.5. Prediction of group membership

Classification results from discriminant function analyses are presented in Fig. 1.

In the first discriminant function analysis (NIHTB PCA 1a, including RAVLT-DR), the overall χ^2 test was significant (Wilks $\lambda = 0.225$, $\chi^2 = 187.84$, df = 18, $P < .001$). The first function accounted for 94.2% of the variance in diagnostic group membership, with a Canonical correlation of 0.857. Reclassification of cases based on new canonical variables was successful at an overall rate of 84.1% and remained substantial [35] after adjusting for chance (Cohen's $\kappa = 0.66$).

In the second discriminant function analysis (NIHTB PCA 2a, excluding RAVLT-DR), the overall χ^2 test was significant (Wilks $\lambda = 0.258$, $\chi^2 = 172.89$, df = 15, $P < .001$). The first function accounted for 96.7% of the variance in diagnostic group membership, with a Canonical correlation of 0.848. Reclassification of cases based on new canonical variables was successful at an overall rate of 79.7% yet markedly decreased after adjusting for chance (Cohen's $\kappa = 0.58$).

The third discriminant function analysis (traditional tests from PCA 3) demonstrated a significant overall χ^2 test (Wilks $\lambda = 0.214$, $\chi^2 = 173.53$, df = 15, $P < .001$). The first function accounted for 90.7% of the variance in diagnostic group membership, with a Canonical correlation of 0.852. Reclassification of cases based on new canonical variables was successful at an overall rate of 84.7% and remained substantial after adjusting for chance (Cohen's $\kappa = 0.66$).

Regarding specific diagnostic groups, tests from PCA 1a were superior at classifying CN and naMCI, whereas tests from PCA 3 were superior at classifying aMCI and AD. Overall, tests from the NIHTB-CB including RAVLT-DR (PCA 1a) were as good as traditional tests (PCA 3) in classifying clinical groups, and the NIHTB-CB with RAVLT-DR (PCA 1a) performed better than without RAVLT-DR (PCA 2a). Both the NIHTB and traditional batteries demonstrated relative weaknesses in distinguishing MCI. Upon review of the 55%–63% of MCI participants incorrectly classified using the NIHTB-CB, individuals were more likely to be grouped CN (25%–27%) than as the alternate aMCI/naMCI group (13%–20%) or as AD (16%). Among the 52% incorrectly classified using traditional tests, aMCI individuals were equally likely to be grouped as AD or CN (17%), whereas naMCI were equally likely to be grouped as CN or aMCI (31%).

Table 4
PCA of traditional tests

	PCA 3	
	Component	
	1	2
	Memory (MEM_T)	Executive function (EF_T)
Logical Memory immediate recall	0.875	
Logical Memory delayed recall	0.807	
FNAME	0.727	
MMSE-DR	0.626	
FAS		0.83
ANT		0.77
MMSE-attention		0.718
Trails B		−0.648
% Explained variance	34%	33%
Total % explained variance	67%	

Abbreviations: PCA, principal component analysis; FNAME, Face Name Associative Memory-cued first letter; MMSE-DR, Mini–Mental State Examination delayed recall subscore; FAS, verbal fluency under phonemic constraint to letters F-A-S; ANT, verbal fluency under categorical constraint (animals); MMSE-attention, Mini–Mental State Examination attention subscore.

NOTE. PCA was conducted using varimax with Kaiser Normalization. Factors with eigenvalues greater than 1, with a maximum iteration of 25, were extracted—factor loadings shown after orthogonal rotation.

NOTE. PCA 3 rotation converged in 3 iterations. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy = 0.75.

NOTE. Owing to missing data for FNAME and Logical Memory tests added to the protocol at a later date, alternate factor-based scores for the traditional memory domain were calculated leveraging the most available data (i.e., average of MMSE-DR and Logical Memory when FNAME was missing). This factor-based score is noted as MEM_{T1} and is considered in certain analyses by clinical group.

4. Discussion

Our findings support the validity and utility of the NIHTB-CB augmented with a delayed recall subtest for assessment of cognitive status in a clinic setting including younger, middle-aged, and older adults at different stages of cognitive health. We found that (1) the factor structure supported the domains the NIHTB-CB was designed to measure; (2) performance on the NIHTB-CB varied in a manner consistent with performance on traditional neuropsychological tests; (3) the NIHTB-CB explained more variance in cognitive performance and demonstrated a higher agreement rate with consensus diagnoses when RAVLT-DR was included; and (4) the NIHTB-CB with RAVLT-DR demonstrated classification agreement similar to that of the traditional tests. Although the NIHTB-CB was not designed to substitute comprehensive neuropsychological assessment, these data indicate it is a valid tool for broad assessment of cognition in dementia and predementia conditions.

PCA analyses confirmed the convergent validity of the NIHTB tests with strong factor loadings following a pattern consistent with the construct each test was designed to measure [33], similar to that found in the initial validation study in a cognitively unimpaired adult sample [36]. Executive function, learning/memory, and crystallized intelligence tests consistently clustered together across all PCAs both with and without RAVLT-DR and with and without AD participants. When RAVLT-DR was excluded in PCA 2a (including all groups), a two-factor structure emerged, explaining a lower percentage of total variance. These results suggest that when using the NIHTB-CB to measure

Table 5
Nonparametric Spearman's partial correlations NIHTB-CB and traditional factor-based scores

		PCA 1 (NIHTB-CB with RAVLT-DR)			PCA 2a (NIHTB-CB without RAVLT-DR)	
		MEM_{NIH}	EF_{NIH}	CI_{NIH}	EF/WM_{NIH}	CI_{NIH}
PCA 3 (Traditional tests)	MEM_T	0.471*	0.320*	0.196	0.500†	0.196
	MEM_{T1}	0.389†	0.270†	0.314†	0.348†	0.314†
	EF_T	0.423†	0.510†	0.519†	0.547†	0.519†
	EDUCATION	0.312†	0.171*	0.468†	0.253†	0.468†

Abbreviations: NIHTB-CB, NIH Toolbox Cognition Battery; PCA, principal component analysis; MEM_{NIH}, NIHTB-CB memory factor; EF_{NIH}, NIHTB-CB executive function factor; CI_{NIH}, NIHTB-CB crystallized intelligence factor; EF/WM_{NIH}, NIHTB-CB executive function/working memory factor; MEM_T, traditional memory factor; MEM_{T1}, traditional memory factor with most available data; EF_T, traditional executive function factor.

NOTE. Partial correlations controlled for age, sex, and education.

*$P < .05$.

†$P < .001$.

cognition in a heterogeneous sample with memory impairment, the inclusion of a delayed memory test is necessary to explain maximum variance and to fully inform patterns of cognitive performance across multiple domains. This concept is further supported by the higher correspondence with consensus diagnoses resulting from the NIHTB discriminant function model that includes RAVLT-DR, overall and especially for reclassification of dementia due to AD. Although in a nonclinical sample, learning is highly associated with recall performance, this will tend not to be the case in individuals with memory deficits such as those with aMCI and AD. The addition of the RAVLT-DR subtest was also of particular importance in our sample given the fact that many MCI and AD patients were unable to complete the standard NIHTB memory tests (Picture Sequencing and List Sorting), an important consideration for future development of the NIHTB-CB in an AD context.

Our second aim was to compare NIHTB tests to traditional neuropsychological tests. NIHTB factor scores were significantly correlated with traditional factor scores of the same cognitive domains, supporting the convergent validity of the NIHTB-CB. These results held when asymptomatic (CN and SCD) and symptomatic (MCI and AD) groups were analyzed separately. Between-group differences on the NIHTB factors patterned as expected and similarly to traditional tests, which also supported the construct validity of the NIHTB-CB.

As compared with available traditional measures, the NIHTB-CB including RAVLT-DR demonstrated similar classification agreement with clinical groups, even after adjusting for chance. The NIHTB-CB without RAVLT-DR demonstrated the lowest rates of classification agreement. These results suggest that the NIHTB-CB with RAVLT-DR may be just as effective at classifying clinical groups

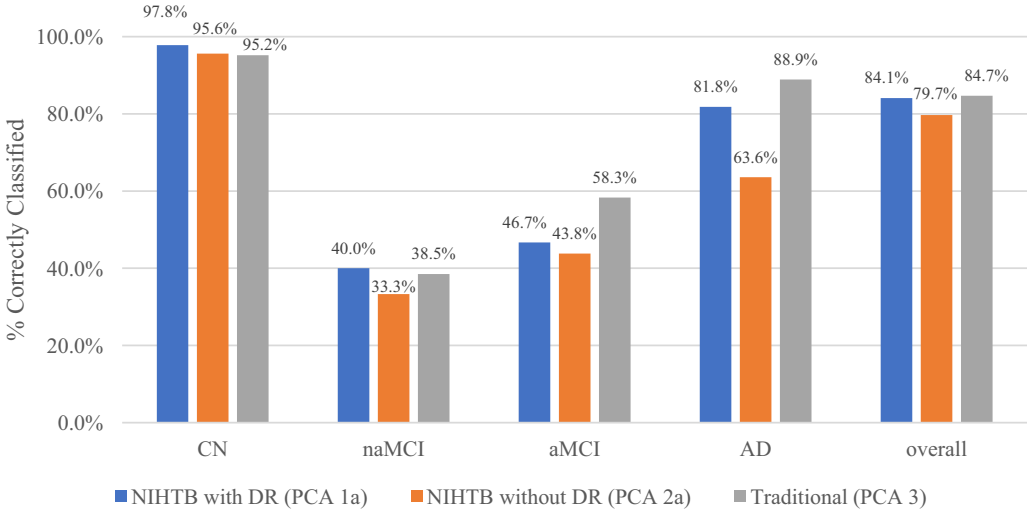

Fig. 1. Discriminant function analyses of factor-based scores from the NIHTB-CB with DR versus NIHTB-CB without DR versus traditional tests. Three discriminant function analyses show the relative rates of accurate diagnostic group reclassification comparing the NIHTB-CB protocol with and without the RAVLT-DR to the traditional test protocol. The y-axis shows the percent of participants correctly reclassified according to their original consensus diagnosis, and the x-axis shows the three cognitive protocols by individual diagnostic group and by the overall sample. Abbreviations: CN, cognitively normal; naMCI, nonamnestic mild cognitive impairment; aMCI, amnestic mild cognitive impairment; AD, dementia due to Alzheimer's disease; NIHTB-CB, NIH Toolbox Cognition Battery; RAVLT-DR, Rey Auditory Verbal Learning delayed recall subtest; PCA, principal component analysis.

as traditional tests and has the potential to do so more efficiently. Neither the NIHTB nor traditional batteries were highly effective at classifying aMCI and naMCI, which speaks to the importance of measures that can accurately distinguish prodromal stages of AD. The error trend of the NIHTB tests, which were more likely to classify MCI as CN, may suggest the difficulty level needs to be increased to detect mild impairment seen in MCI. This type of systematic error is preferable to the unsystematic error demonstrated by the traditional tests (equally likely to misclassify aMCI as AD or NC, and naMCI as CN or aMCI), as it can be corrected using data modification techniques.

Considering the influence of demographics on cognitive test performance [37,38], it is worthwhile to note we did not use fully-adjusted scores available through the NIHTB-CB software, which account for age, sex, race, ethnicity, and education [9,10]. Instead, we used nonadjusted raw and computed scores of the NIHTB and traditional tests when calculating factor scores and subsequently covaried for demographic variables concurrently in all other analyses. We chose this method because the full range of demographic adjustments was not available for all traditional tests. Although our methodological decision was important for establishing the convergent validity of the NIHTB-CB, the normative adjustments available through the NIHTB-CB should be applied in clinical settings to account for differences in demographic characteristics such as race and education that may influence cognitive test results.

A future research direction will be validation of the NIHTB-CB in longitudinal assessment and with AD biomarkers, including positron emission tomography, cerebrospinal fluid amyloid, and tau scans [39]. Future studies and accumulation of normative data within a diverse population, including low education groups, will enable identification and dissemination of cut-scores for more efficient classification of different stages. An advantageous feature of the NIHTB-CB is incorporation of a crystallized intelligence composite, which others have noted may lend a unique method to efficiently measure cognitive reserve and increase sensitivity to longitudinal decline [15]. We found that the crystallized intelligence factor (CI_{NIH}) correlated with education across both asymptomatic and symptomatic groups. The utility of adopting criteria for detection of cognitive impairment that considers decrements from estimated premorbid functioning has been proposed [14] and explored in the context of preclinical AD [40]. Further investigation of cognitive profiles seen at preclinical stages of AD [40,41] will be a critical next step toward examining the utility of the NIHTB-CB for earlier detection.

A limitation of this study includes the relatively low agreement rate for classification of MCI. Although the consensus diagnoses represent those that would be expected to be assigned in a neurology clinic setting, it is important to consider that detailed analysis of comprehensive neuropsychological test results has been shown to improve diagnostic accuracy [42]. Although the consensus diagnosis process used here is conventional in many clinical settings, these diagnoses are not considered "gold standard" and may not meet research criteria. However, our intent was to compare typical clinical practice methodology by primary care physicians and general neurologists (among other specialists) with the NIHTB tests to differentiate groups along the AD spectrum. Furthermore, the pattern of scores on both the NIHTB and traditional batteries, which demonstrated increasing impairment across groups (CN < MCI < AD), provides support for overall accuracy of the original diagnoses.

Another limitation resulting from our observational design includes a cohort with diagnostic groups of varying sample sizes. We attempted to address this by examining the consistency of results within asymptomatic and symptomatic groups; however, larger samples and more consistent group sizes will be needed to replicate these results. Strengths include standardized administration of the battery for all clinic patients, and the wide age range and heterogeneous makeup of participants.

Protocols for applying the NIHTB-CB outside of a research context may depend on the setting. Although the NIHTB-CB can be administered by technicians and scored by the program, interpretation should continue to be the domain of clinicians experienced in understanding the application of such tests in combination with clinical history, particularly if diagnoses are being shared or clinical recommendations made. If cut-scores are established in the future, individuals exceeding an asymptomatic threshold or those with borderline scores may proceed to complete comprehensive neuropsychological evaluation, saving resources and avoiding costly evaluations for individuals identified as CN.

In conclusion, we have shown that the NIHTB-CB augmented with a delayed recall subtest is valid and useful for assessing cognitive capability in individuals at risk for, and those diagnosed with, age-related cognitive disorders such as AD. Most participants had little trouble acclimating to the computer-based test procedure, which shows, similar to other studies [15,43,44], that computerized cognitive testing is a feasible method to test younger, middle-aged, and older adults. Data collection was automated, allowing for score reports to be generated and interpreted rapidly. The protocol demonstrated efficiency in the context of a memory clinic and produced a relatively comprehensive evaluation after a brief assessment. Although a traditional neuropsychological evaluation often requires several hours, administration time for our version of the NIHTB-CB ranged from 35 to 40 minutes. These results support the usefulness of the NIHTB-CB as a key part of the clinical diagnostic evaluation and a promising tool for neuroepidemiological studies.

Acknowledgments

The authors would like to thank Chiashin Shih, Jessica Shum, Pedja Stevanovic, and Hemali Patel for assistance with data management and editorial advice.

This study was funded by philanthropic support (Zuckerman Family Foundation, proceeds from the Annual Memories for Mary fundraiser organized by David and Kathy Twardock, the Rimora Foundation, the Washkowitz Family in Memory of Alan Washkowitz, and contributions from grateful patients of the Alzheimer's Prevention Clinic, Weill Cornell Memory Disorders Program), the Women's Alzheimer's Movement, Hilarity for Charity, the Leon Levy Foundation, the Samuel I. Newhouse Foundation, the Weill Cornell Medical College Clinical and Translational Science Center (NIH/NCATS #UL1TR002384), and NIH PO1AG026572. The funders had no role in the study design, data collection, decision to publish, or manuscript preparation.

Supplementary Data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.dadm.2018.10.002.

RESEARCH IN CONTEXT

1. Systematic review: Search engines (PubMed, Web of Science, etc.) were used to identify literature on the development, use, and validity of the NIH Toolbox Cognition Battery (NIHTB-CB).
2. Interpretation: The NIHTB-CB was found to be comparable to traditional neuropsychological measures in characterizing cognitive performance, and the inclusion of a delayed recall subtest substantially improved its validity and utility in this population. The NIHTB-CB is a valid and useful computerized test battery for brief assessment of cognition in a memory clinic and Alzheimer's risk reduction setting.
3. Future directions: Some NIHTB-CB subtests are too difficult for individuals with dementia due to AD; delayed list learning should be added for these patients. Longitudinal evaluation should explore the predictive validity of the NIHTB-CB against AD biomarkers. NIHTB-CB cut-scores for MCI and dementia subgroups should be established. Measures of crystallized intelligence available in the NIHTB-CB may be leveraged to characterize cognitive abilities relative to premorbid intelligence.

References

[1] Brookmeyer R, Johnson E, Ziegler-Graham K, Arrighi HM. Forecasting the global burden of Alzheimer's disease. Alzheimers Dement 2007;3:186–91.

[2] Sperling RA, Aisen PS, Beckett LA, Bennett DA, Craft S, Fagan AM, et al. Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the national institute on aging-Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement 2011;7:280–92.

[3] Sperling RA, Rentz DM, Johnson KA, Karlawish J, Donohue M, Salmon DP, et al. The A4 study: stopping AD before symptoms begin? Sci Transl Med 2014;6:228fs13.

[4] Seifan A, Schelke M, Obeng-Aduasare Y, Isaacson R. Early life epidemiology of Alzheimer's disease-a critical review. Neuroepidemiology 2015;45:237–54.

[5] Schelke MW, Hackett K, Chen JL, Shish C, Shum J, Montgomery ME, et al. Nutritional interventions for Alzheimer's prevention: A clinical precision medicine approach. Ann N Y Acad Sci 2016;1367:50–6.

[6] Racine AM, Clark LR, Berman SE, Koscik RL, Mueller KD, Norton D, et al. Associations between performance on an abbreviated CogState battery, other measures of cognitive function, and biomarkers in people at risk for Alzheimer's disease. J Alzheimers Dis 2016;54:1395–408.

[7] Gershon RC, Wagster MV, Hendrie HC, Fox NA, Cook KF, Nowinski CJ. NIH toolbox for assessment of neurological and behavioral function. Neurology 2013;80:S6.

[8] Weintraub S, Dikmen SS, Heaton RK, Tulsky DS, Zelazo PD, Slotkin J, et al. The cognition battery of the NIH toolbox for assessment of neurological and behavioral function: Validation in an adult sample. J Int Neuropsychol Soc 2014;20:567–78.

[9] Casaletto KB, Umlauf A, Beaumont J, Gershon R, Slotkin J, Akshoomoff N, et al. Demographically corrected normative standards for the English version of the NIH toolbox cognition battery. J Int Neuropsychol Soc 2015;21:378–91.

[10] Casaletto KB, Umlauf A, Marquine M, Beaumont JL, Mungas D, Gershon R, et al. Demographically corrected normative standards for the Spanish language version of the NIH toolbox cognition battery. J Int Neuropsychol Soc 2016;22:364–74.

[11] Tulsky DS, Carlozzi NE, Holdnack J, Heaton RK, Wong A, Goldsmith A, et al. Using the NIH toolbox cognition battery (NIHTB-CB) in individuals with traumatic brain injury. Rehabil Psychol 2017;62:413.

[12] Tulsky DS, Heinemann AW. The clinical utility and construct validity of the NIH toolbox cognition battery (NIHTB-CB) in individuals with disabilities. Rehabil Psychol 2017;62:409.

[13] Carlozzi NE, Tulsky DS, Wolf TJ, Goodnight S, Heaton RK, Casaletto KB, et al. Construct validity of the NIH toolbox cognition battery in individuals with stroke. Rehabil Psychol 2017;62:443.

[14] Holdnack JA, Tulsky DS, Brooks BL, Slotkin J, Gershon R, Heinemann AW, et al. Interpreting patterns of low scores on the NIH toolbox cognition battery. Arch Clin Neuropsychol 2017;32:574–84.

[15] Buckley RF, Sparks KP, Papp KV, Dekhtyar M, Martin C, Burnham S, et al. Computerized cognitive testing for use in clinical trials: A comparison of the NIH toolbox and cogstate C3 batteries. J Prev Alzheimers Dis 2017;4:3–11.

[16] Folstein MF, Folstein SE, McHugh PR. "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 1975;12:189–98.

[17] Jorm AF, Christensen H, Korten AE, Henderson AS, Jacomb PA, Mackinnon A. Do cognitive complaints either predict future cognitive decline or reflect past cognitive decline? A longitudinal study of an elderly community sample. Psychol Med 1997;27:91–8.

[18] Molinuevo JL, Rabin LA, Amariglio R, Buckley R, Dubois B, Ellis KA, et al. Implementation of subjective cognitive decline criteria in research studies. Alzheimers Dement 2017;13:296–311.

[19] Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, et al. The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the national institute on aging-Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement 2011;7:270–9.

[20] McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR, Kawas CH, et al. The diagnosis of dementia due to Alzheimer's disease: recommendations from the national institute on aging-Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement 2011;7:263–9.
[21] Petersen RC. Mild cognitive impairment as a diagnostic entity. J Intern Med 2004;256:183–94.
[22] Slotkin J, Kallen M, Griffith J, Magasi S, Salsman H, Nowinski C, et al. NIH toolbox technical manual. Bethesda, MD: National Institutes of Health; 2012.
[23] Wechsler D. WMS-R: Wechsler memory scale-revised. New York: Psychological Corporation; 1987.
[24] Rentz DM, Amariglio RE, Becker JA, Frey M, Olson LE, Frishe K, et al. Face-name associative memory performance is related to amyloid burden in normal elderly. Neuropsychologia 2011;49:2776–83.
[25] Benton AL, Hamsher KD, Varney NR. Spreen 0. Contributions to neuropsychological assessment. New York: Oxford UP; 1983.
[26] Parker DM, Crawford JR. Assessment of frontal lobe dysfunction. In: Crawford JR, Parker DM, McKinnley WW, eds. Erlbaum: A handbook of neuropsychological assessment. East Sussex, England; 1992. 267–91.
[27] Reitan RM. Manual for administration of neuropsychological test batteries for adults and children. Tucson, Arizona: Reitan Neuropsychological Laboratory; 1979.
[28] Spreen O, Strauss E. A compendium of neuropsychological tests: Administration, norms, and commentary. New York, NY: Oxford; 1991.
[29] Monsch AU, Bondi MW, Butters N, Salmon DP, Katzman R, Thal LJ. Comparisons of verbal fluency tasks in the detection of dementia of the Alzheimer type. Arch Neurol 1992;49:1253.
[30] Tombaugh TN, Kozak J, Rees L. Normative data stratified by age and education for two measures of verbal fluency: FAS and animal naming. Arch Clin Neuropsychol 1999;14:167–77.
[31] Ala TA, Hughes LF, Kyrouac GA, Ghobrial MW, Elble RJ. The Mini Mental State Exam may help in the differentiation of dementia with lewy bodies and Alzheimer's disease. Int J Geriatr Psychiatry 2002; 17:503–9.
[32] Razani J, Wong JT, Dafaeeboini N, Edwards-Lee T, Lu P, Alessi C, et al. Predicting everyday functional abilities of dementia patients with the mini-mental state examination. J Geriatr Psychiatry Neurol 2009;22:62–70.
[33] Mentzer JT, Flint DJ. Validity in logistics research. J Bus Logist 1997; 18:199.
[34] Hutcheson CD, Sofroniou N. The multivariate social scientist: Introductory statistics using generalized linear models. London: Sage; 1999.
[35] Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20:37–46.
[36] Mungas D, Heaton R, Tulsky D, Zelazo PD, Slotkin J, Blitz D, et al. Factor structure, convergent validity, and discriminant validity of the NIH toolbox cognitive health battery (NIHTB-CHB) in adults. J Int Neuropsychol Soc 2014;20:579–87.
[37] Manly JJ, Jacobs DM, Sano M, Bell K, Merchant CA, Small S, et al. Cognitive test performance among nondemented elderly African Americans and whites. Neurology 1998;50:1238–45.
[38] Manly JJ, Jacobs DM, Touradji P, Small SA, Stern Y. Reading level attenuates differences in neuropsychological test performance between African American and White elders. J Int Neuropsychol Soc 2002; 8:341–8.
[39] Karlawish J, Jack CR Jr, Rocca WA, Snyder HM, Carrillo MC. Alzheimer's disease: The next frontier—Special Report 2017. Alzheimers Dement 2017;13:374–80.
[40] Seifan A, Isaacson R. The Alzheimer's prevention clinic at Weill Cornell Medical College/New York-Presbyterian Hospital: risk stratification and personalized early intervention. J Prev Alzheimers Dis 2015; 2:254.
[41] Grober E, Hall CB, Lipton RB, Zonderman AB, Resnick SM, Kawas C. Memory impairment, executive dysfunction, and intellectual decline in preclinical Alzheimer's disease. J Int Neuropsychol Soc 2008;14:266–78.
[42] Bondi MW, Edmonds EC, Jak AJ, Clark LR, Delano-Wood L, McDonald CR, et al. Neuropsychological criteria for mild cognitive impairment improves diagnostic precision, biomarker associations, and progression rates. J Alzheimers Dis 2014;42:275–89.
[43] Wild K, Howieson D, Webbe F, Seelye A, Kaye J. Status of computerized cognitive testing in aging: A systematic review. Alzheimers Dement 2008;4:428–37.
[44] Fredrickson J, Maruff P, Woodward M, Moore L, Fredrickson A, Sach J, et al. Evaluation of the usability of a brief computerized cognitive screening test in older people for epidemiological studies. Neuroepidemiology 2010;34:65–75.